Natural or Computer Generated: A Computational Study Towards Digital Image Forensics and Understanding Algorithmic Fairness

A thesis submitted in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

IN

COMPUTER SCIENCE

In the Faculty of Science

By

MANJARY P. GANGAN

Under the Guidance of

Dr. LAJISH V. L.

DEPARTMENT OF COMPUTER SCIENCE
UNIVERSITY OF CALICUT
KERALA, INDIA

April 2024

UNIVERSITY OF CALICUT
DEPARTMENT OF COMPUTER SCIENCE

Dr. Lajish. V. L
Associate Professor & Head

Calicut University P. O.
Kerala - 673635, India

Certificate

This is to certify that the Thesis entitled **"Natural or Computer Generated: A Computational Study Towards Digital Image Forensics and Understanding Algorithmic Fairness"**, submitted by **Ms. Manjary P. Gangan**, to the University of Calicut, for the partial fulfilment of the requirements for the award of the degree of **Doctor of Philosophy (Ph.D.)** in Computer Science, is a bonafide research work done by Ms. Manjary P. Gangan under my supervision and guidance in the Department of Computer Science, University of Calicut, Kerala. The content embodied in this Thesis, in full or in parts, have not been submitted to any other University or Institute for the award of any degree.

Dr. LAJISH V. L.
Associate Professor & Head
Department of Computer Science
& Director, Calicut University Computer Center
University of Calicut, Kerala, India

University of Calicut
April 04, 2024

Declaration

I, **Manjary P. Gangan**, declare that this Thesis entitled, **"Natural or Computer Generated: A Computational Study Towards Digital Image Forensics and Understanding Algorithmic Fairness"** is based on the original work done by me under the supervision and guidance of Dr. Lajish V. L. in the Department of Computer Science, University of Calicut. I confirm that:

- The work presented in this Thesis has not been submitted previously for the award of any degree either to this University or to any other University or Institution.
- I have followed the guiding principles given by the University in organizing the Thesis.
- Whenever I have used materials (theoretical analysis, data, figures, and text) from other sources, I have given due credit to them by citing them in the Thesis and giving their particulars in the references.

MANJARY P. GANGAN

University of Calicut
April 04, 2024

⚠ *The examples provided in this Thesis, specifically those related to the work of identifying algorithmic bias in visual transformer based models, may be offensive in nature and may hurt one's moral beliefs. All such instances are not the perspectives of the author, but are purely the outputs of various algorithms.*

Abstract

The path-breaking advancements in technology involving high-resolution imaging devices and inexpensive storage make an all-easy acquisition, storing, and sharing of high-quality digital images. But the constant growth of image manipulation harms the success of digital imaging. Nowadays, there exists much software that allows even relatively inexperienced users to edit digital images or create fake images with such perfection without leaving any trace of tampering, deceiving the human eyes that it is hard to distinguish these fake images from the original ones. And hence, we are no longer in a world where seeing is believing. This brings into account the significance of assessing the trustworthiness of an image through digital image forensic techniques, by applying scientific methods to investigate the identity of a digital image. This Thesis is a computational study in the direction of digital image forensics to distinguish natural images taken by a camera from the computer generated images such as computer graphics and Generative Adversarial Network (GAN) images, and also to understand the perspectives of algorithmic fairness of the forensic systems classifying natural and computer generated images.

The initial contribution of the Thesis is a digital image forensics algorithm, *MC-EffNet*, that distinguishes natural images from computer generated images, including both computer graphics and GAN images. The algorithm employs a parallel fusion of three fine-tuned EfficientNet networks that operate in different colorspaces chosen after studying the efficacy of a variety of colorspace transformations specifically towards this image forensics problem. The experimental results of this study in the Thesis shows that the proposed model could obtain high performance accuracies and outperform the state-of-the-art baselines. The study compares the performance of the proposed algorithms with a manual classification performance and points out the necessity of computational algorithms for the task of distinguishing natural images from computer generated images. The study also analyzes the behavior of the proposed model by visualizing image regions responsible for the model's decisions and compares these model explanations with manual explanations provided by human participants.

Despite the forensic task of distinguishing natural and computer generated images achieving high accuracies with the support and advancements in deep neural networks and transformer based architectures, these forensic models are seen to fail over post-processed images. Post-processing operations such as JPEG compression,

gaussian noising, etc., are usually performed over the images to trick the forensic algorithm. Hence the second contribution of the Thesis proposes an approach towards distinguishing natural and computer generated images including both computer graphics and GAN generated images, that produces high classification accuracies as well as is highly robust against the post-processing operations. The proposed model uses a fusion of two vision transformers where each of the transformer networks operates in different color spaces. The experimental results of this study shows that the proposed approach achieves higher performance, robustness, and generalizability when compared to the state-of-the-art baselines. Also, the features of the proposed model are observed to attain higher inter-class separability than the baseline features. Visualizing the attention maps of the networks of the fused model shows that the proposed methodology can capture more image information relevant to the forensic task of classifying natural and generated images.

As like in any other machine learning based algorithms, biased forensic algorithms can cause serious societal harm and security concerns. Hence, besides developing forensic algorithms it is also essential to identify any bias in such forensics systems. Accordingly, the third contribution of this Thesis is in the direction of exploring algorithmic fairness in forensic systems particularly built using the large visual transformers, which are the most commonly employed recent deep learning architectures due to their capability to produce high classification performances. The study tries to identify gender, racial, affective, and intersectional biases in forensics systems classifying natural and computer generated images using a bias evaluation corpora and a vast set of bias evaluation measures. The study uses a two phase evaluation setting to examine whether the most common post-processing operation of image compression in any way influences the model biases, and observes that image compression impacts model biases.

Keywords: Digital Image Forensics, Computer-generated images, GAN images, Algorithmic Fairness

Acknowledgements

The journey of my research work would not have come true without the great support and encouragement of my research supervisor, collaborators, and colleagues. First and foremost, I would like to express my deepest gratitude to my research supervisor *Dr. Lajish V. L.*, Associate Professor and Head, Department of Computer Science, University of Calicut, India, for the great opportunity to pursue my Ph.D. research under his guidance. I am greatly thankful to him for his guidance and support throughout the course and the edges of life.

I'm highly thankful to my research colleague, *Dr. Anoop Kadan*, Postdoctoral Researcher, Queen's University Belfast, United Kingdom, who helped contribute to the work and provided noteworthy technical inputs. The discussions with him throughout the research work were very insightful and helpful.

I would like to thank *Dr. Esakkirajan S.*, Professor, Instrumentation and Control Engineering, PSG College of Technology, Coimbatore, India, and *Dr. Badri Narayan Subudhi*, Associate Professor, Department of Electrical Engineering, Indian Institute of Technology Jammu, India, for their valuable time and fruitful interactions that were helpful for my research work.

I whole-heartedly thank *Dr. Deepak P.*, Associate Professor, School of Electronics, Electrical Engineering and Computer Science, Queen's University Belfast, United Kingdom, the intergroup discussions with him have helped and motivated me a lot to understand various perspectives of research in computing. My sincere thanks to *Dr. Deepak Mishra*, Professor, Avionics Department, Indian Institute of Space Science and Technology (IIST), Trivandrum, India, *Dr. S. Balasubramanian*, Associate Professor, Department of Mathematics & Computer Science, Sri Sathya Sai Institute of Higher Learning (SSSIHL), Andhra Pradesh, India, *Dr. Darshan Gera*, Associate Professor, Mathematics and Computer Science, SSSIHL, Bengaluru, India, *Dr. Kumar Rajamani*, Senior Manager, KLA Tencor, Bengaluru, India, and *Mr. Sumanth Reddy Kaliki*, Lead AI Expert, Ice Edge Business Solutions, Canada, whose insightful lectures and discussions provided valuable insights into the theoretical and practical frameworks of deep learning, enabling me to establish a solid foundation in the field.

I sincerely thank the *Department of Science and Technology (DST)* of the Government of India, for granting financial support to my research project related to identifying computer generated images under the Women Scientist Scheme-A (WOS-A) for Research in Basic/Applied Science. A special thanks to *Dr. Vandana Singh*, Scientist-F, Department of Science & Technology, Ministry of Science & Technology, New Delhi, India, for her support throughout the conduct of the project. I also sincerely thank the

financial support provided by the *Government of Kerala* through the E-Grantz scholarship scheme.

I especially thank the Master's students (CS2019-21) and staff at the Department of Computer Science, University of Calicut for their involvement and cooperation in conducting the psychophysics experiments in my research work. Also, my sincere and heartfelt thanks to the love and support of my fellow researchers, faculty members, non-teaching staff, and the students of the Department of Computer Science, and the University of Calicut.

Manjary P. Gangan

Contents

List of Figures

List of Tables

List of Abbreviations

CG	**C**omputer **G**enerated
GAN	**G**enerative **A**dversarial **N**etworks
DIF	**D**igital **I**mage **F**orensics
LDA	**L**inear **D**iscriminant **A**nalysis
SVM	**S**upport **V**ector **M**achine
CFA	**C**olor **F**ilter **A**rray
PRNU	**P**hoto **R**esponse **N**on-**U**niformity
PRCG	**P**hoto-**R**ealistic **C**omputer **G**raphics
RBF	**R**adial **B**asis **F**unction
CNN	**C**onvolutional **N**eural **N**etwork
RNN	**R**ecurrent **N**eural **N**etwork
SC-EffNet	**S**ingle **C**olorspace **Eff**icient**Net**
MC-EffNet	**M**ulti-**C**olorspace fused **Eff**icient**Net**
MBConv	**M**obile inverted **B**ottleneck **Conv**olution
PG^2	**P**ose **G**uided **P**erson **G**eneration
LDRD	**L**evel-**D**esign **R**eference **D**atabase
DFDC	**D**eep **F**ake **D**etection **C**hallenge
t-SNE	**t**-Distributed **S**tochastic **N**eighbor **E**mbedding
Grad-CAM	**Grad**ient-weighted **C**lass **A**ctivation **M**apping
MCE-ViT	**M**ulti-**C**olorspace fused and **E**nriched **Vi**sion **T**ransformer
ViT	**Vi**sion **T**ransformer
CvT	**C**onvolutional **V**ision **T**ransformer
Swin	**S**hifted **win**dow
FFHQ	**F**lickr-**F**aces-**HQ**
FPR	**F**alse **P**ositive **R**ate
FNR	**F**alse **N**egative **R**ate
ACS	**A**verage **C**onfidence **S**core
DP	**D**emographic **P**arity
EO	Equal **O**pportunity

To

my husband
and my family, teachers, and friends

Chapter 1

Introduction

Abstract: This chapter provides an introduction to the Thesis, with the research objectives that the Thesis accomplishes, the motivation of the research study in this Thesis, and the major Thesis contributions. An overview of the entire Thesis is also presented in this chapter.

1.1 Digital Image Forensics

The advent of modern imaging devices and plenty of social media platforms available for rapid information exchange led to an exponential growth in the number of digital images produced and circulated through different media sources. These digital images constitute photographs taken by a camera (natural images or real images), or their edited or manipulated versions, or even the computer generated images (CG or fake images). It is essential to understand the authenticity of a digital image because the manipulated or artificially generated images used along with fake news can support the news story convincingly very well and thereby leading to the spread of lies and propaganda quickly.

The art of making fake images is not new; such cases of reports are seen even from the 19th century. One of the earlier examples is the fake image montage 'Ulysses S. Grant at City Point'[1] composed of several images to claim the presence of General Ulysses S. Grant with his troops at City Point during the American Civil War; but actually, the General was not present at the City Point (figure 1.1). From then till now, there are many such fake image examples, used to persuade peoples' minds and bring their belief over different fake stories. One such recent example is a fake X-ray image presented with fake content stating a snake exists inside a woman's pelvic area and circulated on social media. The Snopes fact-checker website helped to reveal that the image originally appeared on a website meant for creating strange and fake digital artwork[2].

[1] https://en.wikipedia.org/wiki/File:Ulysses_S._Grant_at_City_Point.jpg, *accessed: 17-12-2023*

[2] https://www.snopes.com/fact-check/woman-snake-xray/, *accessed: 17-12-2023*

FIGURE 1.1: Images used for creating tampered image "Ulysses S. Grant at City Point"

Apart from such image manipulations, fully automated fake image generation is also possible using advanced Computer Graphics and rendering software and the recent Generative Adversarial Networks (GAN) algorithms. These techniques are exceptionally efficient to artificially (in the sense that they do not originate from an imaging device or camera) generate images convincingly, to deceive human eyes, and hence are referred to as Photo-Realistic Computer Generated images. The process of computer generation of images tends to be such realistic nowadays that it is impossible to differentiate natural images from computer generated images. Figure 1.2 shows some sample images where one can observe the extent of photorealism attained in

GAN generated (left image[3]) and computer graphics generated (middle image[4]) images making it hard to classify them as computer generated images and for the natural image (right image[5]) at first glance, one might have an impression that it is a computer graphics image. Such artificial photo-realistic computer generated images are ubiquitously available on the internet. Hence, no longer can we believe the images we see. The digital images that reach us always have a question of authenticity, that is, whether the image is a projection of a real-world event taken by a camera or is it manipulated or computer-generated content. Here comes the significance of Digital Image Forensics (DIF).

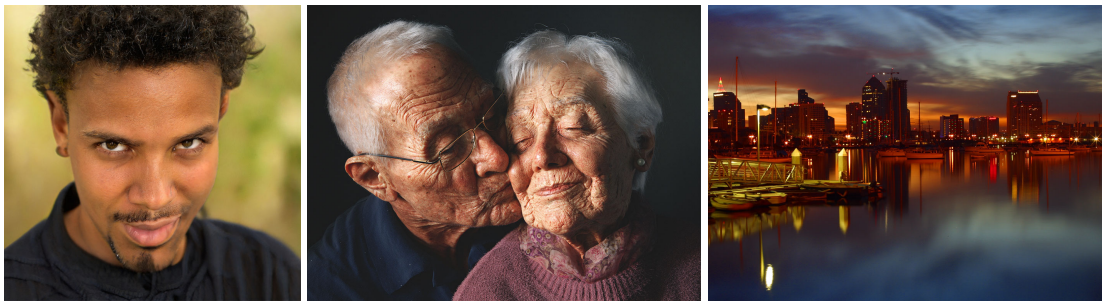


FIGURE 1.2: A sample of GAN generated (left), computer graphics (middle) and natural (right) image

Digital Image Forensics is one of the recent active research areas investigating the trustworthiness of a digital image by utilizing scientific methods that can furnish shreds of evidence of image authenticity for accident, eyewitness, civil or criminal cases. Digital Image Forensics deals with computational algorithms for the automated detection of computer-generated, recaptured, or manipulated images. This Thesis is a computational study in the research area of digital image forensics, focusing on the natural and fully generated computer generated images.

1.2 Research Objectives

This Thesis is a computational study in the area of Digital Image Forensics, particularly focusing on natural images taken from a camera, and fully generated photo-realistic computer graphics and GAN images. The Thesis initially attempts to develop a computational algorithm, *MC-EffNet* for classifying natural and computer generated images. Secondly, the Thesis attempts to develop an approach to classify natural and

[3]Source: https://github.com/NVlabs/stylegan2, *accessed: 01-05-2022*
[4]Source: https://cgsociety.org/c/featured/1f9s/the-forever, *accessed: 01-05-2022*
[5]Source: Computer Graphics versus Photographs dataset [1]

computer generated images, *MCE-ViT*, that is robust towards post-processing operations. Finally, the Thesis attempts to explore fairness in digital image forensic systems classifying natural and computer generated images. Three of these works attempted in this Thesis are discussed below.

Multi-Colorspace fused EfficientNet (*MC-EffNet*) There is an exponential growth in the number of digital images being produced and circulated through different media sources day by day. This includes natural images of the real-world scenes taken by a camera and computer generated images. The computer generated images include images that are generated by different graphics and rendering software and also those generated using the latest deep learning algorithms called Generative Adversarial Networks. Hence images that reach us always have a question of authenticity, that is, whether the image is a projection of real-world events taken by a camera or is it computer generated content? Hence, this Thesis attempts to develop a computational algorithm, *MC-EffNet*, to distinguish natural images from computer generated images (in chapter 3). Unlike, the previous works that either address *natural images versus computer graphics* or *natural images versus GAN images* at a time, this Thesis approaches the problem of distinguishing natural images from computer generated images by considering both categories of computer generated images, i.e., computer graphics and GAN images. Such an approach would be more suitable for a real-world image forensic scenario since it considers all categories of image generation, because in most cases image generation is unknown.

Multi-Colorspace fused and Enriched Vision Transformer (*MCE-ViT*) In the real world, the images transferred through social media are generally post-processed, especially will have mostly undergone a JPEG compression operation [2]. Moreover, to deceive the forensic algorithms, image forgeries are followed by some post-processing operations such as JPEG compression, addition of Gaussian noise, etc., or even multiple post-processing operations [3]. Even though advanced deep learning algorithms have paved the way for high-accuracy forensics algorithms classifying natural and computer generated images, in the real world most of these algorithms are seen to drastically fail over the post-processed images. Hence, this Thesis attempts to develop a robust computational approach, *MCE-ViT*, to distinguish natural images and computer generated images including both computer graphics and GAN images (in chapter 4).

Exploring Fairness in Natural and GAN Generated Image Detection Systems Fairness studies are becoming increasingly popular in most of the computational research involving data-driven machine learning based algorithms. Fairness studies are also highly relevant in the area of digital image forensics. Unfair forensic algorithms classifying natural and computer generated images can lead the images of certain social groups of people to be more likely to be predicted as natural/real images even if they are actually fake images. Similarly, this can also lead the images of certain social groups of people to be more likely to be predicted as fake images even if they are actually real images. Such situations cause serious security concerns, making it crucial to assess the fairness of image forensics systems. Hence, this Thesis attempts to explore fairness in the forensic algorithms distinguishing natural and computer generated images (in chapter 5).

1.3 Research Motivation

Easy availability of image acquiring devices, massive publicly accessible image datasets, rapid progress and a wide variety of generative algorithms, and user-friendly easily available apps producing high quality super realistic images have drastically increased the production of fake images all around. Even though computer generated images are mostly seen produced for creative art, entertainment, advertisement, joke, or satirical purposes they have a high potential to easily propagate through social media causing misinformation, particularly when presented with fake stories or fake news [4]. They also have much darker sides like the earlier incidents of claiming pornographic images of children as computer generated graphics images to escape from legal actions[6] to the recent incidents of creating fake accounts in social media platforms using GAN generated faces[7], creating nude photographs of people from their original photographs through GAN algorithms[8], fake images used as evidence for supporting fake news, defamation, false light portrayals [4, 5, 6], etc. The deficiencies in human perception to distinguish natural and computer generated images without the assistance of any additional tools [7] highly demands and points out the necessity of computational algorithms in digital image forensics to investigate images since the authenticity of an image legally depends on whether it is a natural image or

[6] www.sciencedaily.com/releases/2016/02/160218144928.htm, *accessed: 17-12-2023*

[7] https://www.wired.com/story/facebook-removes-accounts-ai-generated-photos/, *accessed: 17-12-2023*

[8] www.technologyreview.com/2020/10/20/1010789/ai-deepfake-bot-undresses-women-and-underage-girls/, *accessed: 17-12-2023*

computer generated. This shows the importance of digital image forensics and motivates the research direction of distinguishing natural (or real) images from computer generated images (or fake), one of the fundamental and most actively researched problems in digital image forensics.

1.4 Thesis Contributions

The contributions of this Thesis include a deep learning based digital image forensics algorithm to detect natural and computer generated images, a visual transformer based image forensics algorithm to detect natural and computer generated images that is robust against post processing operations, and an exploration of fairness in digital image forensics systems classifying natural and computer generated images. Each of these three contributions are elaborated below.

Multi-Colorspace fused EfficientNet The contribution *MC-EffNet* of the Thesis (explained in chapter 3) proposes a novel deep learning based approach for distinguishing natural and computer generated images including both computer graphics and GAN generated images. The study proposes the Multi-Colorspace fused EfficientNet model by parallelly fusing three EfficientNet networks that utilize transfer learning methodology. Each of the three networks in the fused model works in a distinct colorspace. The colorspaces of each of the EfficientNet network is selected by analyzing the effectiveness of various colorspace transformations specifically towards the forensic task of distinguishing natural, computer graphics and GAN images. The performance of the proposed model is observed to be higher than the baselines in terms of accuracy, robustness towards post-processing, and generalizability towards other datasets. The chapter conducts psychophysics experiments to understand how accurately humans can distinguish natural, computer graphics, and GAN images. The study investigates the behavior of the proposed model through visual explanations to understand salient regions that contribute to the model's decision making. The study also compares these model explanations with manual explanations provided by human participants in the form of region markings to understand any similarities between the model and manual explanations.

Multi-Colorspace fused and Enriched Vision Transformer The forensic classification models distinguishing natural and computer generated images are seen to fail

over the images that have undergone some post-processing operations usually performed to mislead the forensic algorithms, such as JPEG compression, gaussian noising, etc. The contribution work *MCE-ViT* of the Thesis (explained in chapter 4) proposes a robust approach towards distinguishing natural and computer generated images including both, computer graphics and GAN generated images. The proposed model utilizes a combination of two vision transformers where each of the vision transformer networks operates in a distinct color space. The proposed approach attains higher performance improvement, higher robustness, and generalizability when compared to a set of baselines. The feature visualizations of the proposed model are seen to better separate the classes than the baseline features. The attention map visualizations of the networks of the fused model show that *MCE-ViT* is able to capture more image information for classifying natural and computer generated images.

Exploring Fairness in Natural and GAN Generated Image Detection Systems The contribution of the Thesis (explained in chapter 5) attempts to investigate bias in the image forensic algorithms that classify natural and GAN generated images. To investigate bias in gender, racial, affective, and intersectional domains, this study procures a bias evaluation corpora and employs a vast set of individual and pairwise bias evaluation measures. This study also examines the role of image compression on model bias, since robustness of the algorithms against image compression is crucial for forensic tasks. Hence the study conducts a two phase evaluation setting. That is the bias evaluation experiments are conducted in the uncompressed evaluation setting and also in the compressed evaluation setting, to analyze the influence of image compression on model bias.

1.5 Publications and Research Grants

Publications based on this Thesis

Journals

1. **Manjary P. Gangan**, Anoop K., Lajish V. L., "Distinguishing natural and computer generated images using Multi-Colorspace fused EfficientNet", *Journal of Information Security and Applications*, Elsevier, Vol. 68, August 2022, pp. 103261, ISSN: 2214-2126, DOI: https://doi.org/10.1016/j.jisa.2022.103261 (***SCIE Indexed, Impact Factor: 5.6***) – *Chapter 3*

2. **Manjary P. Gangan**, Anoop K., Lajish V. L., "Image Tampering Detection using Deep Learning", *Calicut University Research Journal (CURJ), Special Edition: Responsible AI for Social Good* (*Accepted for Publication*) – *Chapter 1 and 2*

3. **Manjary P. Gangan**, Anoop K., Lajish V. L., "A Robust Approach Towards Distinguishing Natural and Computer Generated Images using Multi-Colorspace fused and Enriched Vision Transformer", *arXiv preprint*, DOI: https://doi.org/10.48550/arXiv.2308.07279 (*In communication*) – *Chapter 4*

4. **Manjary P. Gangan**, Anoop K., Lajish V. L., "Exploring Fairness in Pre-trained Visual Transformer based Natural and GAN Generated Image Detection Systems and Understanding the Impact of Image Compression in Fairness", *arXiv preprint*, DOI: https://doi.org/10.48550/arXiv.2310.12076 (*In communication*) – *Chapter 5*

Research Grants

- **Women Scientist Scheme-A** (WOS-A) for Research in Basic/Applied Science
  Funding Agency: Department of Science and Technology (DST), Government of India
  Grant No.: SR/WOS-A/PM-62/2018
  Principal Investigator: Manjary P. Gangan, and **Mentor:** Dr. Lajish V. L.

1.6 Overview of the Thesis

This Thesis has three parts as illustrated in figure 1.3. The introduction of the Thesis, and a background and review of related works are presented in the initial chapters, chapter 1 and chapter 2. The research contributions are presented in the chapter 3, chapter 4 and chapter 5. The conclusion and future directions of research are presented in chapter 6.

Chapter 2 presents the background and literature review on digital image forensics, particularly focusing classification of natural and computer generated images. Section 2.1 presents an introduction to this chapter followed by section 2.2 presenting the earlier attempts and section 2.3 presenting the origin of forensic perspective. Section 2.4 and 2.5 provides review of literature on computational methods detecting computer graphics and GAN images, respectively. The manual approaches to distinguish natural and computer generated images are presented in section 2.6. The state-of-the-art works that analyze the robustness of the methods towards post-processing operations are discussed in section 2.7. Section 2.8 presents a review of works that

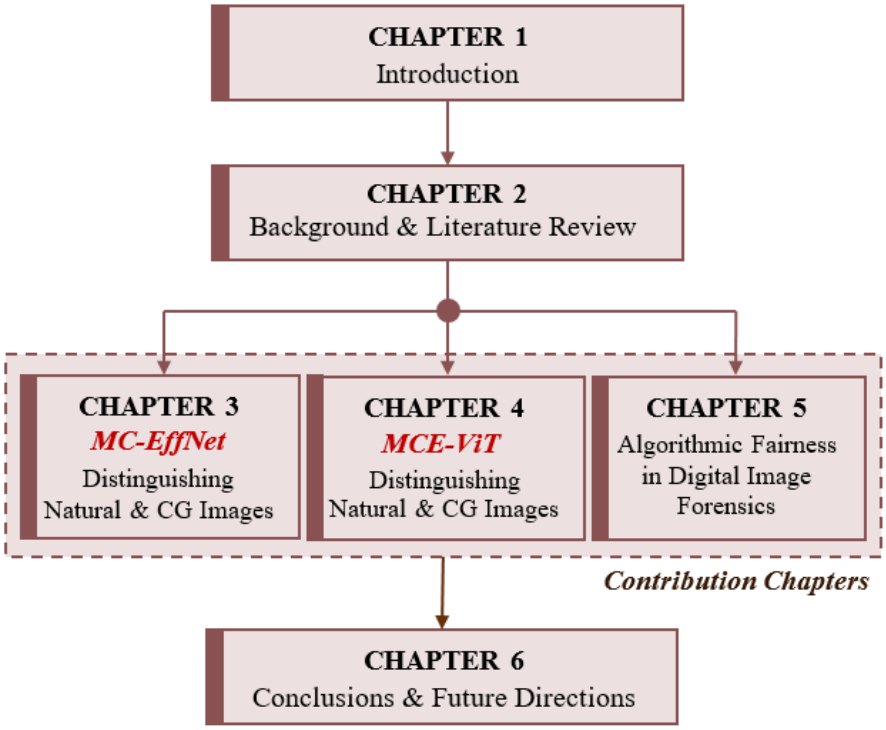


FIGURE 1.3: Overview of the Thesis

attempt explainability in this area of digital image forensics and section 2.9 presents a review of the works that explore bias in image forensics systems. Section 2.10 finally summarizes this chapter.

Chapter 3 presents the novel contribution of the Thesis, *MC-EffNet*, a deep learning based digital image forensic model for distinguishing natural and computer generated images. Section 3.1 presents an introduction of the chapter. Section 3.2 describes in detail the dataset used in this study. Section 3.3 discusses in detail the methodology of the proposed work. Section 3.4 gives the details of experimental settings and baselines used for comparing against the proposed work. Section 3.5 illustrates results and discussion including statistical significance, robustness, generalizability, feature visualization, psychophysics experiments and model behavior analyses using activation maps. Finally section 3.6 summarizes the chapter.

Chapter 4 presents the novel contribution of the Thesis, *MCE-ViT*, a visual transformer based digital image forensic model for distinguishing natural and computer generated images that is highly robust against post processing operations. Section 4.1 presents an introduction of the chapter. Section 4.2 gives details of the dataset used in this study. Section 4.3 discusses the methodology of the proposed work along with the

motivation and detailed description of the proposed model. Section 4.4 presents the experimental settings and details of the baseline models used for comparing the proposed model. Section 4.5 presents the results and discussion including the results of the experiments for robustness, generalizability, feature visualization, and the analysis of attention maps. Section 4.6 finally summarizes the chapter.

Chapter 5 presents the novel contribution of the Thesis, an exploration of fairness in the image forensic systems distinguishing natural and GAN generated images. Section 5.1 presents an introduction of the chapter. The rest of the chapter is organized as section 5.2 discusses in detail the construction of transformer based models for the task of classifying natural and GAN images. Section 5.3 explains in detail the evaluation domains, evaluation corpora, and evaluation measures used for bias analysis experiments. Section 5.4 presents the results and discussions of both the uncompressed and compressed evaluation settings and finally section 5.5 presents the summary of the chapter.

Chapter 6 presents the conclusions of the Thesis, with Thesis contributions emphasized in section 6.1, and section 6.2 discussing the future scope and research directions.

This brings the conclusion of this first chapter which provides an introduction of the Thesis with details of the research objectives, research motivation, Thesis contributions, research publications and grants received based on this Thesis and finally presenting an overview of the entire Thesis.

Chapter 2

Background and Literature Review

Abstract: This chapter illustrates a broad background of Digital Image Forensics and the taxonomy with respect to various digital image forgeries, which is required for a sound understandability of the subsequent evolutions in this research area, specifically focusing on computer generated images including both the computer graphics and GAN categories.

2.1 Introduction

Digital Image Forensic methods are categorized either as Active or Passive approaches, Blind or Non-blind approaches, or Signal-based, Scene-based, or Metadata based approaches [8]. Passive approaches cannot interfere with the process of image generation. These approaches analyze the image using the intrinsic characteristics of the camera or the processing artifacts, without taking into account any image semantics. Each camera has its inherent variations due to diverse manufacturers or technological imperfections, which enables understanding the origin of a given image. Instead, in Active approaches, image generation is purposefully altered by embedding some additional information to leave some identifying traces. e.g., embedding a digital watermark into the image [9].

Blind and Non-blind approaches is another categorization of DIF approaches [10]. In contrast to Non-blind, the Blind image forensic approaches don't know about the original scene, image generation, or any post-processing employed during the time of analysis. The Active and Non-blind image forensic approaches are unviable in practical settings because they preclude the examination of any random images of unknown origin. Passive blind forensic techniques examine traces at different levels viz. Signal-based, Scene-based, and Metadata-based analysis. The Signal-based analysis methods ignore the semantics of the images and consider them as sequences of discrete symbols. The Scene-based analysis takes scene properties into account, further divided into Physics-based and Semantics-based analysis. Physics-based relates to real-world projections, whereas Semantics-based refers to the meaning of the image content. Metadata-based analysis exploits auxiliary digital data accompanying

the photograph. The overview of digital image forensic approaches and the areas of their applications are illustrated in figure 2.1.

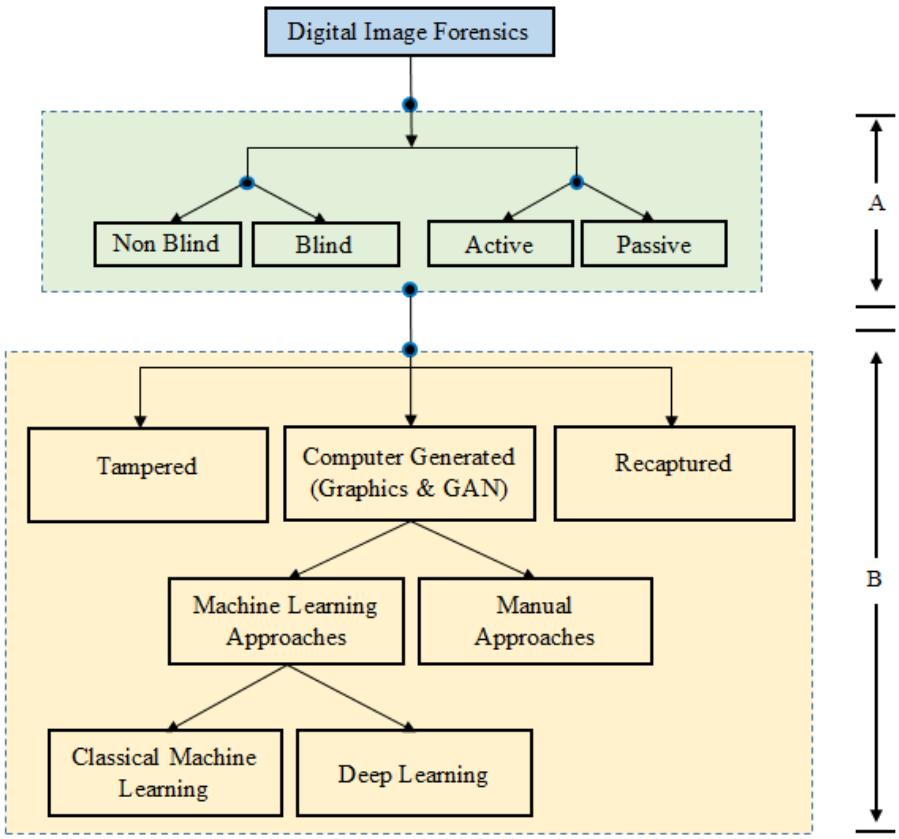


FIGURE 2.1: A general anatomy Digital Image Forensic approaches (A: Approaches, B: Applications)

2.1.1 Organization of the Chapter

The rest of this chapter is organized as section 2.2 presents a background on the earlier attempts and section 2.3 gives the origin of the forensic perspective in the category of works classifying natural and computer generated images. Section 2.4 and 2.5 provides the review of literature on various computational methods involving both handcrafted features and the deep neural networks used for detecting computer graphics and GAN images respectively. The manual approaches to distinguish natural and computer generated images are presented in section 2.6. Background and literature of the state-of-the-art works that analyze the robustness of the methods towards post-processing operations are discussed in section 2.7. Section 2.8 throws light on the attempts of explainability in this area of digital image forensics classifying natural and

computer generated images. Section 2.9 presents the background and literature of the works that explore bias in image forensics systems. Section 2.10 finally summarizes this chapter.

2.2 The Early Picture of Distinguishing Natural and Computer Generated Images

The works distinguishing natural and computer generated images have been reported since the 1970s. Most of these works came into picture as a part of constructing some image or video search engines. The work done at Columbia University in 1996 by Smith et al. [11], originally meant to build an image and video search engine prototype called WebSEEk[9] for cataloging and indexing images and videos on the World-Wide-Web is one among the earlier approaches that considers categorizing computer graphics, rather than considering it as a work to distinguish computer graphics and natural photographs. The image content based feature used in this work is the color histogram. This work has been referred to by Athitsos et al. [12] to build an automated image search engine system again, called Web Seer[10] to classify web images as computer graphics or natural photographs based on their color based features, image dimensions and the associated text. The features offered in the work are based on some simple and obvious statistical observations over the image content of computer graphics or natural photographs. Number of colors, prevalent color, farthest neighbor, color saturation, color histogram, farthest neighbor histogram are the color related image content features, and smallest dimension and dimension ratio are the image dimension related features used in the work. A few other similar works can be seen during this period [13, 14] utilizing the color based image content features for categorizing computer graphics images. All these works suit only for categorizing very simple computer graphics images or graphical drawings that are not troublesome for humans to decide whether they are computer graphics images or real photographs.

During the start of the 3rd millennium an improved attempt was made by Hartmann et al. [15] to classify images into various categories namely, graphical, realistic looking computer graphics, and real photographs, using color, region, edge and noise based features. In the graphical category they consider graphical drawings, buttons, navigation elements etc. In the realistic looking computer graphics images category they consider the raytracing images and images from different graphic tools and

[9] https://www.ee.columbia.edu/ln/dvmm/researchProjects/MultimediaIndexing/WebSEEK/WebSEEK.htm, *accessed: 17-12-2023*

[10] http://hint.fm/seer/, *accessed: 17-12-2023*

games. For the real photograph category they consider the natural scenes and family photos. Apart from the previous works they try to bring a clear distinction between simple graphical drawings which are easily recognizable by humans from the realistic looking computer graphics images.

2.3 The Origin of Forensic Perspective

The entertaining frame of mind through which one viewed a computer graphics image till then, changed to a digital image forensic perspective with Farid et al. [16] in 2003 pointing out the necessity of distinguishing computer graphics images from photographs by explaining the difficulty of law enforcement agencies in prosecuting the child pornographers claiming their images are computer generated. They come up with a higher order wavelet based statistical model towards various applications of digital image forensics. Implementation of this idea can be seen in one of the the state-of-the-art works in digital image forensics by Lyu et al. [17] showing that the differences of the first and higher order wavelet statistics for the photo-realistic computer graphics and real photographs are significant enough to classify them. They use the separable quadrature mirror filters to decompose the image color channels with respect to different scales and orientations, after which the mean, variance, skewness and kurtosis are extracted from each of the sub bands to obtain the first and higher order statistics. The process is also replicated over the error coefficients produced from the linear predictor of the image magnitude. Both these statistical features together form the feature vector to discriminate photo-realistic computer graphics from real photographs. The dataset used in this work is an unbalanced set of 40000 photographs and 6000 graphics images of outdoor and indoor scenes. The computer graphics images collected are created using graphics and rendering software. Different dimensions of images are accommodated in the dataset by central cropping the images to a size of 256×256. For classification they utilize the Linear Discriminant Analysis (LDA) classifier, and non-linear Support Vector Machine (SVM) classifier.

2.4 Computational Approaches to Distinguish Natural and Computer Graphics Images

Different categories of approaches are seen to be adopted for computationally distinguishing natural and computer graphics images. This includes the handcrafted

feature based classical machine learning approaches and the deep learning based approaches. Both these categories of works are reviewed separately in the following subsections.

2.4.1 Handcrafted Feature based Classical Machine Learning Approaches

Differences in the generation of natural images and computer graphics, camera or device properties, etc., are considered for the handcrafted traditional feature based classification works [18, 19, 20, 21, 22, 23, 24]. Dirik et al. [20] proposes chromatic aberration and demosaicking based features to distinguish photographs and computer generated images. Khanna et al. [21] proposes digital image forensics techniques to distinguish photographs, computer graphics and images generated using scanners, based on the residual pattern noise in images, independent of the image content. Gao et al. [22] proposes a hybrid feature of white balance, Color Filter Array (CFA) and Photo Response Non-Uniformity (PRNU) noise artifacts that captures the intrinsic properties of optical imaging to distinguish photographs and computer graphics images. Peng et al. [23] proposes a method to distinguish natural images and computer graphics, using CFA interpolation on PRNU, and Long et al. [24] proposes a method to identify natural images and computer graphics using PRNU properties.

Many other feature based works using color, texture, and shape based statistical features followed this category of image forensics study [25, 17, 26, 27, 28, 29, 1, 30]. The work by Ng et al. [25] of Columbia University in 2004, utilizes Natural Image Statistics from the power spectrum, local image patches, and wavelet transform to classify 800 Photo-Realistic Computer Graphics (PRCG) images collected from 3D graphics and artists' websites and 800 real photographs taken from camera. In 2005 the dataset used in this paper was released under the name 'Columbia Photographic Images and Photorealistic Computer Graphics Dataset' [31], which was a major milestone in this area providing a benchmark dataset to conduct experiments in many research works [32, 26, 33, 34, 35, 36, 28, 29, 29].

Tian-Tsong et al. [26] formulates a passive-blind online system for classifying computer-generated images and photographs, based on the state of the art algorithms extracting the geometrical, wavelet, and cartoon features. The study also focuses on the variation of classification accuracy with image size reduction. In order to improve performance, they perform subclass-based bagging techniques. The dataset in this study consists of 800 PRCG images belonging to categories architecture, game, nature, objects, etc., 800 personal photograph collections, mainly travel images like indoor, outdoor, people, objects, etc., 800 images from Google Image Search following

keywords like architecture, people, scenery, forest, etc., and 800 Non PRCG images including cartoons, drawings, 2D graphics, logos, etc. The authors club the Personal and Google image categories into a class, and the PRCG and Non PRCG form the opposite class, though they are very different types of images. The classification is supervised by Radial Basis Function (RBF) kernel based SVM and five-fold cross-validation.

Another work to classify computer graphics images and photographs during this period is by Dehnie et al. [27] where they exploit the similarity in the pattern noise of photographs taken using different digital cameras, the property which is absent for computer generated images. Patchara et al. [28] proposes features derived from the statistical moments of the photo-realistic computer graphics images and photographs in the YCbCr color system. They perform a correlation analysis over the components and only choose the Y and Cb color components for feature generation due to the high correlation between Cb and Cr to minimize the computational complexities. To reduce the influence of the image content they consider the prediction errors, and to improve the performance of the model with reduced dimensions they use the Discrete AdaBoost technique. The dataset contains 15,200 photographs and 7,492 computer graphics images.

Rong et al. [29] addresses the problem of distinguishing photo-realistic computer graphics images and photographs by analyzing the statistical property of local edge patches in the images. They generate a Voronoi cell based visual vocabulary and represent images as a binned histogram of visual words. The dataset used in this work consists of 1000 photographs and 900 graphic images. The photographs are chosen by them from different cameras, since the Columbia dataset [31] may contain images that are post processed or compressed. In the graphics category, 800 are from the Columbia dataset, and 100 are from a graphics website with a high degree of visual realism. They use non-linear SVM for classification. Their results imply that the structure of local edges is capable enough to capture differences between photographs and graphics.

Another work of significance in the category of handcrafted feature based classification of photo-realistic computer graphics and real photographs is the one contributed by Tokuda et al. [1]. In this work, the authors present an extensive study of 13 previous representative works in the literature and devise feature-classifier fusion approaches for improving the classifier accuracies. They also contribute a large dataset comprising 4850 images both in photo-realistic computer graphics images and real photographs.

Peng et al. [30] conduct a classification work based on the motivation that photographs and graphics are initiated from different pipelines of image acquisition. They

consider the Columbia dataset [31], Dresden image dataset [37], and collections from graphic sites to obtain their dataset with 3000 photographs and 3000 graphics. They find the residuals of the images drawn using linear regressions to study the texture difference with the multifractal spectrum and histogram. Their feature selection is simple with fewer number of features which reduces the time complexity considerably.

Pan et al. [38] extract a set of features from wavelet based Hidden Markov Tree to classify natural images and computer graphics images using SVM on HSV and RGB color space. The dataset consists of a diverse collection of 3000 natural images taken from the Washington image database and personal collections and 3000 computer graphics images from the graphics websites. Leveraging such transform domain features from various image transformation domains like wavelet [16], contourlet [39], quaternion wavelet [40], etc., is another approach that can be seen in the traditional feature based category of works. A snapshot of related handcrafted feature based works in the literature classifying natural and computer graphics images is given in the table 2.1.

TABLE 2.1: Related handcrafted feature based works classifying natural and computer graphics images

Work	Feature	Feature category
Farid et al. [16]	Wavelet based	Statistical features
Ng et al. [25]	Wavelet transform	Statistical features
Dehnie et al. [27]	Pattern noise	Device properties
Tian-Tsong et al. [26]	Geometrical, wavelet, cartoon based	Statistical features
Dirik et al. [20]	Chromatic aberration, demosaicking	Device properties
Khanna et al. [21]	Residual pattern noise	Device properties
Patchara et al. [28]	Color moments	Statistical features
Pan et al. [38]	Wavelet based	Statistical features
Rong et al. [29]	Local edge patches	Statistical features
Ozparlak et al. [39]	Contourlet based	Statistical features
Tokuda et al. [1]	Texture, Edge, Pattern noise	Statistical features and Device properties
Gao et al. [22]	White balance, CFA, PRNU	Device properties
Peng et al. [23]	PRNU properties	Device properties
Wang et al. [40]	Quaternion wavelet	Statistical features
Long et al. [24]	PRNU properties	Device properties

2.4.2 Deep Learning based Approaches

The process of manually designing and extricating discriminative features majorly depends on the prior knowledge humans have in the corresponding area. But, many times these features are not found to attain the presupposed performances, notably when the datasets are huge or complex. The revolutionary progress in the area of machine learning with the formulation of deep neural networks like Convolutional Neural Networks (CNN) simplified the contributions avoiding the burdensome process of handcrafted feature extraction, feature selection, dimensionality reduction, etc. These networks have unified end-to-end optimizations and strong learning capacities that enable them to attain excellent performances when compared to the classical approaches. From 2017 onwards a lot of research works can be seen based on deep learning towards the task of classifying photo-realistic computer graphics images from photographs and exhibit comparatively higher performance than the traditional feature based classification approaches [41, 42, 43, 44, 45, 46, 47].

Yu et al. [41] propose two classification models, the object recognition VGG network and a CNN without pooling to identify photorealistic computer graphics on random input image patches of size 32×32. They design the system to detect the synthesized regions of the images too. Their dataset contains 1000 photographs taken using cell phones and digital cameras and the graphics images from the Columbia dataset [31] and photorealism competitions. Based on the number of patches marked as graphics, the image is classified as computer graphics. One of the drawbacks of this model is that the detection of the synthesized region is possible only in cases where the patches marked as graphics are uniformly distributed in the image.

Rahmouni et al. [42] proposes CNN with custom pooling to extract the statistical features like lower order moments, for the task of classifying photo-realistic computer graphics images from photographs. Since the available benchmark datasets were created a few years back and are not comparable with the computer graphics images during the time of this work, the authors collected their own dataset with 1800 computer graphics images after judging that they were photo-realistic enough. The high resolution photographs are collected from the RAISE dataset [48]. The classifiers are built using the randomly extracted image patches of size 100×100. Their method is also able to visualize the modified regions in an image.

Another CNN based classification work for real-time forensic tasks is suggested by Cui et al. [43]. To include the PRNU information they subject the input images first to high pass filtering and then supply it to the ResNet-50 model with short cut connection. Even though they choose a deep learning architecture their work is implemented

and evaluated on the Columbia Dataset [31] with considerably less amount of images.

Yao et al. [44] proposes a five layer CNN and three high pass filters based model for photograph and computer generated graphics classification. The high pass filters remove the low frequency image content and provide sensor pattern noise instigated by the imaging device. The images are clipped into patches of size 650×650 and then fed into the model. Hence this model is not widely applicable for the discrimination of images in any other dataset that does not meet the dimension requirement of width and height greater than 650 pixels.

Unlike the previous works, He et al. [49] conducts a set of preprocessing steps. Initially, the texture information of the images transformed to the YCbCr color domain is enhanced using Schmid filtering. They then employ a dual path CNN combined with directed acyclic graph Recurrent Neural Networks (RNN) to classify photographs and graphics. Their experiments are conducted over 6800 graphics images collected from graphics websites and 6800 photographs of outdoor and indoor scenes using different cameras.

Nguyen et al. [50] proposes a modular technique, where the pre-trained VGG-19 network without fine tuning is chosen for feature extraction, statistical CNN is chosen for feature transformation and the best machine learning model among the state-of-the-art works is chosen for classification. The dataset used in the work is from [42] which is not so small but the diversity of images is less, which is the reason for choosing the pre-trained VGG-19. Similarly, since statistical features are proven better in the literature, the extracted features are converted to statistical features using the statistical CNN transformation module. Both the photographs and graphics are high resolution images. Compared to certain previous works the authors do not crop these images because the cropped images are also of high quality, in contrast to the real world scenarios where the images are of low quality. So to diversify the image quality before dividing it into patches they choose to resize the images to 360 pixels using interpolation. The testing procedure in their work shows that the use of high resolution images only for training purposes is not enough to address real world forensic attacks.

Since most of the approaches to distinguish photographs and computer graphics in the literature require uniform processing of the entire image pixels, they are expensive in terms of memory and computation time. Therefore, Tariang et al. [46] presents a Stochastic Convolutional Recurrent Attention Network with comparatively a fewer number of parameters that can be broken up into Glimpse Network, RNN, Classifier, Emission Network. To effectively manage the computation overhead due to a large number of parameters, their model makes use of a glimpse network. Their attention based technique permits to process selected image region sequences locally.

The dataset they use contains 800 graphics from Columbia Dataset [31] and 1000 photographs from RAISE dataset [48].

Zhang et al. [47] in their work extracts channel correlation using a self coding module and pixel correlation using a CNN. Their work is based on the motivation that graphics images show weak pixel and channel correlation since they are produced by rendering instead of interpolation as in photographs. They utilize the dataset in [49] for their work and to assess the generalizability towards unseen data they use the Columbia dataset [31]. They observe the generalizability of their self coding component that, it can be integrated directly with prevailing CNN architectures to enhance their performance further.

A few deep learning based works that employ transfer learning methodology by using pre-trained off-the-shelf networks [51, 52, 53] can also be seen in the literature. Apart from considering full-sized images, some works crop images into fixed size patches and derive results of full-sized images from these image patches [42, 54, 49].

2.5 Computational Approaches to Distinguish Natural and GAN Generated Images

Works in this category classifying natural and GAN images are comparatively much more recent than the works classifying natural and computer graphics images, ensuing the advent of the new and powerful class of deep learning algorithms called Generative Adversarial Networks. This category also consists of works employing both the traditional handcrafted feature based classification and deep learning based classification. Both these categories are reviewed separately in the following subsections.

2.5.1 Handcrafted Feature based Classical Machine Learning Approaches

Unlike the previous category of works classifying natural and computer graphics images, very few works employ the traditional handcrafted feature based approach for classifying natural and GAN images [55, 56, 57]. McCloskey and Albright [55] analyze color and saturation statistics based cues for distinguishing GAN generated images from natural images. Their analyses are centered on the generation process of only one category of GAN algorithm [58]. They observe that the saturation based statistics features that measure the frequency of over exposed and under exposed image pixels along with a linear SVM for classification is a better feature for detecting GAN images than the color statistics based feature.

Li et al. [56] proposes a compact set of co-occurrence based features for classifying natural images and images generated from deep neural networks, based on the motivation that these two class of images have differences in chrominance components, which is much more observable in the residual domain. Their experiments show that the differences are much more evident in the residual domain chrominance components of the HSV and YCbCr color spaces than the RGB color space, and extract co-occurrence based features from these components.

Marra et al. [57] conducts a study of various handcrafted features as well as state-of-the-art models and off-the-shelf deep learning models on detecting image-to-image translation category of GAN images both in uncompressed scenario and compressed scenario. They test the features and models over the dataset consisting of 36000 images collected from the CycleGAN generative algorithm. The features considered in this study are the steganalysis features originally proposed in [59] that are utilized in [60] for image forgery detection. The feature is extracted from the co-occurrence matrices computed from the high pass filtered images. Their work shows that even though most of the features and models perform well in the ideal uncompressed scenarios, only deep learning based models are able to be a little more robust towards compressed scenarios. A similar observation can be seen in the work by Rössler et al. [61], that this steganalysis based feature is not able to endure compressed scenarios. A snapshot of related handcrafted feature based works in the literature classifying natural and computer graphics images is given in the table 2.2.

TABLE 2.2: Related handcrafted feature based works classifying natural and GAN images

Work	Feature	Feature category
McCloskey and Albright [55]	Color and saturation	Statistical features
Li et al. [56]	Co-occurrence based	Statistical features
Marra et al. [57]	Co-occurrence based	Steganalysis features

2.5.2 Deep Learning based Approaches

A majority of works in the literature for classifying natural and GAN images, approach this problem using deep learning algorithms; and there are a very large number of studies reported using deep learning algorithms for the task. Many studies are seen to utilize convolutional neural network architectures for classifying natural and GAN generated images [42, 62, 57, 63, 64]. Some of the off-the-shelf convolutional neural network architectures are also seen to be utilized directly or even as backbone

networks for classifying natural and GAN generated images, such as, VGG network [65] in the work [66], InceptionNet [67] in [57], and XceptionNet [68] in many studies like [57, 61, 69, 70, 71]. Other learning strategies such as pairwise learning [72], incremental learning [73], and one-shot learning [74] are also seen to be experimented in this category of works. Besides the works targeting to detect images generated from a single GAN algorithm [57, 75, 76], there is also another line of work for the attribution of known GANs that are used to generate fake images [77, 78, 79, 71]. Many works in this category are seen to specifically work over GAN generated human faces rather than considering heterogeneous image content [80, 69, 81]. Few works are also seen to utilize transformer based architectures to detect the GAN class, but most of these works try to specifically detect the DeepFake videos [82, 83, 84, 85, 86].

2.6 Manual Approaches to Distinguish Natural and Computer Generated Images

Manual approaches to assess computer graphics images can be seen in some of the works [87, 88, 89, 90] which are meant to enhance the visual realism of computer graphics to photo-realistic computer graphics. Manually determining the factors of image realism is also important in the field of digital image forensics which might help to improve the computational models that classify photographs from photo-realistic graphics. The awareness of these factors would even be helpful in uncovering the fundamental cognitive and perceptual mechanisms of viewers to discern visual realism. But unfortunately, this has been a challenging task due to the lack of exploration of the factors that influence visual realism.

Besides the objective studies in the aforesaid categories of computational works, there are only a very few numbers of subjective studies that involve humans to distinguish natural images and computer graphics images using certain psychophysics experiments [91, 92, 93, 94, 95]. An initial record on manual classification of photographs and graphics can be seen in the work by Farid and Bravo [91]. For this purpose, they collect 180 graphics images of high quality from the past six years, 2000 to 2006, with image contents including man made, natural or human. For each of the graphics images, a photograph that closely matches the image content of the graphics image was gathered. They then provide these 360 images randomly to 10 participants with unlimited time for classification. The authors also inspected the time taken by the participants for classification where they saw a trade-off between speed and accuracy, longer inspection time gave greater performances. They observe that in the

graphics image set the human image category obtains highest performance, but a decrease in the performance can be seen in the images produced in the previous year, i.e., 2006. This work during 2007 suggests that even though the graphics technologies are improving greatly, the visual system of humans still achieves good accuracies for photograph versus graphics classification.

Building over this, another work [92] by the same set of authors in 2012 explores the potential of manual classification of 30 photographs and 30 photo-realistic graphics of human faces rendered during 2007 to 2010 without any background using psychophysical experiments. The study involves two sets of observers (one set from their laboratory and the other from the Amazon Mechanical Turk) without any pre-training to reliably direct that a given human face image is a photograph (mainly to evaluate the child pornography cases, stating that it's illegal if the images are photographs). The authors measure the probability of an image being a photograph when it is manually classified as a photograph. The authors note that most of the time when the observer judges an image to be a photograph, in fact, it is a photograph.

In [93] Fan et al. does an image decomposition approach combined with signal detection theory to find out the image realism factors including the cognitive factors that effects human classification of face photographs and graphics. For each of the background removed 10 pairs of photographs (collected from face recognition datasets) and photo-realistic computer graphics images (collected from the graphics websites) they considered the color, grayscale and reflectance versions of the images. The image pairs are provided for classification to graphic experts and lay persons with unconstrained viewing distance, time and terminals and they were asked to indicate what gave the clue to their classification i.e., is it the skin, eye, nose, lip, expression, color, illumination or any others. They find that the shading factor is important than color and the graphics experts perform better in their classification than laypersons. When the authors utilized the dataset mentioned in the work of [92] they find that ethnicity is also a factor for photograph and graphics classification.

As an extended study of [91, 92, 93], the work proposed by Holmes et al. [94] conducts a study to understand whether the advancements brought about in the scope of computer graphics from the time of their past studies till the time of this study have impacted the peoples' capability to distinguish natural and generated images. They observe that due to increased photorealism, people find it much more difficult to identify the generated images than in their previous studies, and more often people report the images as natural images than computer graphics. Also, they observe that the difficulty in classification is being able to be improved by providing some training on these classes of images. Pursuing this work, a sequence of experiments are performed

in [95] to enhance the classification capability of people to detect generated images, which includes providing feedback and incentives.

The works addressing manual classification of natural images and GAN images are also very few [96, 97]. Nightingale et al. [96] conducts manual classification of StyleGAN2 [98] generated fake faces and matching real faces that are used in training StyleGAN2 algorithm, through three sets of experiments. In the first experiment, each face is manually classified into real or GAN generated fake faces, the second set of experiments involves training and feedback, and the third set of experiments involves a trustworthiness rating of the faces.

The study of Lago et al. [97] conducts the manual classification of 150 GAN images selected from three different algorithms PGGAN [58], StyleGAN [99] and StyleGAN2 [98], and 150 real images from FFHQ dataset [99]. The participants of the study were provided with 30 random samples of images. Their study observes that images of earlier GAN algorithms are easy to be detected as fake images whereas those generated using the recent algorithms are mostly classified to be real, especially the StyleGAN2 faces are detected as real images more frequently than the originally real faces.

Another work by Dang et al. [70] studies the human capabilty to distinguish 110 random natural and fake images from their dataset. They observe that manual classification is difficult, and manual classification is mostly targeted on the semantic notions like artifacts, lighting, quality of image etc., and mostly do not concentrate on the fine details in an image.

2.7 Robustness Towards Post-processing Operations

One of the earlier works in the category of natural versus graphics image classification, discussing the robustness against post processing operations would be that of Farid et al. in [92]. Their work is a psychophysical experiment based manual classification of human face photographs and photo-realistic graphics considering the effect of image compression quality, resolution, color adjustments, and orientations on the classification performance. The study mainly aims to deal with child pornography cases and utilizes two types of observers to check a given human face image is a photograph. Most of the times the observers judge an image to be a photograph, it is a photograph itself. Interestingly more than the full resolution images, better results are obtained at half the image resolutions. For a wide quality range of image compression, the manual classification performance remains almost nearly the same, except for very low compression values. But the loss of color information by conversion to the grayscale domain greatly degrades the manual classification accuracies.

Rong et al. in [29] does a feature based work to distinguish photo-realistic computer graphics images and photographs using Voronoi cell based visual vocabulary. They find the performance of their method towards the JPEG compression attacks. For three different values of compression quality factors, they observe their method to be stable for the graphics and photographs in two out of the three cases.

The residual texture based regression work of Peng et al. [30] looks over the robustness of their system on the post processing operations including JPEG compression and resizing, noise addition, and rotation. They analyze the performance across four values of quality factor for JPEG compression, scaling factor for scaling, angles for rotation, and signal to noise ratio of gaussian white noise for additive noise. The performance of the system is affected slightly by JPEG compression, noise addition, and rotation. Even though noise addition can impact the texture information the low pass gaussian filtering used in their system could resist the gaussian white noise. They observe a decrease in the system performance while the images are scaled up to larger sizes which they understand is due to the alterations in the texture information caused by the interpolation of pixels.

Yu et al. [41] proposes two classification models one using VGG and the other using a CNN for image patches but which seems not to be robust against post processing operations such as JPEG compression and resizing. Another deep learning work that discusses the robustness in post processing is the one by He et al. [49] utilizing a CNN combined along with an RNN network. They evaluate their model after adding Gaussian noise and JPEG compression which does not produce traces that are visually appreciable. They notice a slight degradation occurring in the detection performance when their model is applied over the graphic images in the test set which are artifacted by weak post processing. They also observe that stronger post processing operations leads to a worse detection performance, especially for JPEG compression, maybe because these operations are capable enough to destruct the inherent properties of graphics images. Hence their method could only withstand robustness against some degree of postprocessing.

Zhang et al. in [47] uses deep learning to model pixel and channel correlation information in photographs and graphics. They inspect the effect of seven different JPEG compressions on photographs and graphics by compressing them in such a way without a visual difference happening. The system is not trained on the compressed images and they observe that their system is mostly stable towards JPEG compressions.

Robustness against post processing operation of JPEG compression is also considered in some of the works classifying natural and GAN generated images. The work

proposed by Wang et al. [100] claims a classifier that is trained on a particular GAN generative algorithm is generalizable to other generative algorithms after performing a thorough process of data augmentation because there are potential chances of these generated algorithms to share typical properties different from natural images. As a part of their experiments, they conduct generalizability tests over JPEG compressed data and find that the model is robust since it is trained with data augmentation by including the post processed data. Many more other studies in the category classifying natural and GAN generated images conduct robustness experiments towards JPEG compression [78, 63, 101, 71, 69].

2.8 The Role of Explainability in Digital Image Forensics

There are many studies progressing in the category of deep learning based approaches to distinguish natural and computer generated images due to their high capability to produce remarkable classification accuracies. But these machine learning models are blackboxes that do not exemplify the predictions in a manner that humans can comprehend. Hence explainability of the deep learning models is investigated in most of the different tasks that utilize these deep learning algorithms, conveyed through different ways such as visualizing attention maps, gradient activation maps etc. These explanations would answer why an input produces a certain output; hence paving ways to understand the blackboxes. Such model behavior analyses can be especially beneficial to analyze and develop computational algorithms employing sensitive data within the areas of forensics, criminal justice, law, health care, etc. However, even with innumerable studies evolving in the category of deep learning based approaches for image forensics, there are only a very few studies analyzing, at least at a primitive level, the behavior of the models distinguishing natural and computer generated images.

Among the works distinguishing natural and graphic images, a work that attempts to analyze the model behavior is the one proposed by Quan et al. [54]. Their work follows a patch based approach, where an entire image is divided into patches to perform CNN classification and later the classification decision of the entire image is derived from the decisions of the patches. Their work analyzes what has been learned by their CNN model by utilizing visualization tools such as layer-wise relevance propagation [102] and deep visualization toolbox [103]. They could observe that lighting and color transition are the important factors their proposed CNN considers for classification. Their analysis also comes up with observations that natural images contain

more amount of variabilities making them more complex than graphics images which are simpler with large color primitives.

Among the works distinguishing natural and GAN generated images, a work that attempts to analyze the model behavior is the one proposed by Guo et al. [104]. Their work tries to detect GAN generated face images by analyzing both eyes in a face for any corneal specular highlight based artifacts. For this purpose, they initially carry out localization of eye region, and later utilize a residual attention network for detecting GAN images by locating the inconsistencies. Attention map visualizations of their qualitative analysis also reveal that there are dissimilarities in the regions captured by the attention between eye regions of the GAN and natural images.

2.9 Exploring Fairness in Image Forensics systems

Many works are seen to be reported in the literature studying fairness in image based research problems, such as in the areas of face recognition [105], image classification [106], medical image processing [107], etc. But comparably only a very few studies explore bias in forensics systems, and amongst those studies, most of them work on videos, i.e., Deep Fake videos.

Trinh and Liu [108] explore bias in three deep fake detection models Xception [68], MesoInception-4 [109] and Face X-Ray [110], using gender and race balanced Deep Fake face datasets. Their study observes high racial bias in the predictions of these Deep Fake detection models. They could also observe that one of the most popularly used datasets for training the models for Deep Fake detection, FaceForensics++ [61], is also highly biased towards female Caucasian social groups.

Hazirbas et al. [111] proposes a video dataset to analyze the robustness of top-winning five models of DFDC dataset [112] for the domains gender, skin type, age, and lighting. They could observe that all the models are biased against dark skin people and hence find that these five models are not generalizable to all groups of people.

Pu et al. [113] explores gender bias in one of the Deep Fake detection models MesoInception-4, in the presence of certain make-up anomalies, using the FaceForensics dataset. Their study is centered on analyzing these models at various prominence levels of the anomaly in the female and male social groups. Their observations are that the model is biased towards both genders, but mostly towards the female group.

Xu et al. [114] explores bias in three Deep Fake detection models EfficientNet-B0 [115], Xception [68], and Capsule-Forensics-v2 [116], by conducting evaluations

on five Deep Fake datasets which are annotated with 47 attributes including non-demographic and demographic attributes. Their observations state that these models are highly unfair towards many of these attributes.

2.10 Summary

This chapter illustrated a brief background of Digital Image Forensics, the literature review of computational and manual approaches to identify computer generated images including both computer graphics and GAN images, and the the literature review of works exploring bias in digital image forensic systems. Besides the review of related works presented in this chapter, the proposed contributions of this Thesis are delineated in the context of the state-of-the-art works, in the corresponding chapters itself (in sections 3.1.2, 4.1.2 and 5.1.2), for a better comprehension of the research gap that each of the chapters focuses on.

Chapter 3

MC-EffNet: Distinguishing Natural and Computer Generated Images using Multi-Colorspace fused EfficientNet

Abstract: This chapter proposes a novel approach for distinguishing natural and computer generated images, attempting the problem as a three-class classification task classifying natural, computer graphics, and GAN images. For the task, the chapter proposes a Multi-Colorspace fused EfficientNet model by parallelly fusing three EfficientNet networks that follow transfer learning methodology. Each of the three networks in the fused model operates in a different colorspace, one in RGB, the other in LCH, and the third in HSV, which are chosen after analyzing the efficacy of various colorspace transformations specifically towards this image forensics problem. The proposed model outperforms the baselines in terms of accuracy, robustness towards post-processing, and generalizability towards other datasets. The study conducts psychophysics experiments to understand how accurately humans can distinguish natural, computer graphics, and GAN images and could observe that humans find difficulty in classifying these images, particularly the computer generated images. This indicates the necessity of computational algorithms for the task. The study also analyzes the behavior of the proposed model through visual explanations to understand salient regions that contribute to the model's decision making and compare with manual explanations provided by human participants in the form of region markings. The experiments show similarities in both the model and manual explanations, indicating the powerful nature of the proposed model to make decisions meaningfully.

This work was supported by the Women Scientist Scheme-A (WOS-A) for Research in Basic/Applied Science from the Department of Science and Technology (DST) of the Government of India

3.1 Introduction

The problem of distinguishing natural images from photo-realistic computer generated ones either addresses *natural images versus computer graphics* or *natural images versus GAN images* at a time. For the *natural image versus computer graphics*

problem, when an image is not computer graphics it shall fall into the natural image category, but it may sometimes actually belong to the GAN category of computer generated images which is not considered for the task, a similar issue may also occur for the *natural images versus GAN images* problem. Therefore in a real-world scenario, to provide a complete forensic solution to distinguish natural images from computer generated images, since the image generation is unknown, it is highly essential to consider all categories of image generation including natural images taken by a camera, computer graphics and GAN images. This work for the first time, to the best knowledge, attempt to address this gap of a generalized algorithm in digital image forensics to distinguish natural images from photo-realistic computer generated images including both computer graphics and GAN images, as a three-class classification task by proposing a Multi-Colorspace fused EfficientNet model.

3.1.1 Research Question

This chapter addresses the following research questions.

***RQ1*:** Can a generalized algorithm be proposed for the forensic task of distinguishing natural and computer generated images including both computer graphics and GAN images, rather than the usual task of natural images versus either of the computer generated images?

***RQ2*:** Does the combination of multiple colorspace transformations help towards increasing the performance of distinguishing natural and computer generated images including computer graphics and GAN images?

3.1.2 Delineating the Proposed Work in the Context of Literature

In the literature, works either address only *natural images versus computer graphics* problem or *natural images versus GAN images* problem at a time. However, such a closed set will not suit the real-world scenario that requires a single forensic system to authenticate an image by investigating multiple types of image generations where, in most cases the image generation is unknown. Therefore unlike previous image forensic works that had always been dealt with as a two-class classification problem, the proposed work, for the first time, to the best knowledge, attempt the image forensic task of distinguishing natural images from computer generated images as a three-class classification problem classifying natural images, computer graphics and GAN images.

The proposed work performs a deep neural network based classification with transfer learning methodology that avoids the burdensome processes such as feature extraction and feature selection present in the case of conventional feature extraction based approaches. Different from the transfer learning based work to distinguish natural images from computer graphics proposed by Rezende et al. [51] using a ResNet architecture with 25.6M parameters, the choice of network in this proposed work is an EfficientNet which is 4.9 times smaller with just 5.3M parameters. When compared to the deep learning based works proposed by Cui et al. [43], and Quan et al. [54] that utilize an earlier dataset, the Columbia dataset (Columbia Photographic Images and Photorealistic Computer Graphics dataset [31]), with considerably less amount of images for a deep learning task (800 images per class), the choice of dataset in this proposed work is more challenging in the real world forensic scenario by maintaining heterogeneity in each of the three classes so that to build a generalized robust model that is unbiassed towards any particular image category, origin or generating algorithm without compromising the number of images in the dataset (4000 images per class). Also, this work avoids patch based implementation in the deep learning approach because firstly such patch based approaches are computationally very expensive than taking full-sized images for checking whether an image is fully computer generated or taken by a camera (e.g., Quan et al. [54] extracts 200 patches from a single image) and moreover, such patch based implementations might be more suitable for image forgery problems to detect manipulated image regions.

Apart from [117, 118] that choose certain colorspace transformations in their work to distinguish natural images from computer generated images, this work examines in detail which colorspaces provide high classification accuracies for the task of distinguishing natural images from computer generated images including computer graphics and GAN images and also the chances of improvement in accuracy by fusing the networks operating in different colorspaces. Among various works in literature to distinguish natural images from computer generated images, not many works are seen to discuss the interpretability or behavior of the model, a work in this regard would be [54] that tries to understand what the model learns to differentiate natural and computer graphics images. Whereas this proposed three-class classification work for natural, graphics and GAN images, besides visualizing explanations of correct and wrong predictions for model behavior analysis, also compares visual explanations of the proposed model with human explanations labeled as region markings during psychophysics experiments. This helps to look for any similarities between the model and human explanations and to understand whether the proposed model is predicting the decisions meaningfully.

3.1.3 Contributions

The major contributions of this chapter include:-

- This chapter introduces a deep learning based image forensic solution to distinguish natural images from photo-realistic computer generated images including both computer graphics and GAN images
- This chapter propose a Multi-Colorspace fused EfficientNet model built by parallelly fusing three EfficientNet networks that follow transfer learning methodology where each of the three networks operates in a different colorspace, one in RGB, the other in LCH, and the third in HSV that are chosen after analyzing the efficacy of colorspace transformations in this forensic problem
- The proposed Multi-Colorspace fused EfficientNet model obtains good forensic performance outperforming baselines in terms of accuracy, robustness towards post-processing, and generalizability towards other datasets
- The study also conducts psychophysics experiments to assess the capability of humans to classify natural images from photo-realistic computer generated images including computer graphics and GAN images
- The study analyzes behavior of the proposed model through visual explanations to understand salient regions that contribute to the model's decision making and compare it with manual explanations provided by human participants in the form of region markings procured during the psychophysics experiments

3.1.4 Chapter Organization

The rest of the chapter is organized as follows. Section 3.2 describes in detail the dataset used in this study. Section 3.3 discusses in detail the methodology of the proposed work. Section 3.4 gives the details of experimental settings and baselines used for comparing against the proposed work. Section 3.5 illustrates results and discussion including statistical significance, robustness, generalizability, feature visualization, psychophysics experiments and model behavior analyses using activation maps. Finally section 3.6 summarizes the chapter.

3.2 Dataset

The proposed study utilizes a total of 12000 images where GAN, Graphics, and Real classes contain 4000 images each. For Graphics and Real classes, images are collected

from the Computer Graphics versus Photographs dataset [1], a challenging dataset with diversity in image category, origin, quality, and content. The class Graphics of the dataset include photorealistic images that are not easily predictable manually as computer graphics images and excludes graphical icons. Similarly to incorporate heterogeneity and to avoid bias in class GAN, the study collects images from four different GAN algorithms ProgressiveGAN [58], StyleGAN [99], StyleGAN2 [98] and StyleGAN2-ADA [119] considering their excellent performances to generate high-quality realistic images. From the ProgressiveGAN generated images, almost 75 images are collected from each of the 31 image categories such as airplane, bedroom, bicycle, bird, boat, bottle, bridge, bus, church outdoor, classroom, conference room, etc. Almost 100 images are collected from each of the four image categories, bedroom, car, cat, and face of StyleGAN generated images, five image categories, car, cat, church, face, and horse of StyleGAN2 generated images, and the rest of the images from six categories, cat, dog, wild, brecahad, face, and metface of StyleGAN2-ADA generated images. Thus, the entire dataset maintains heterogeneity in every class with several different categories like outdoor and indoor scenes, objects, animals, characters, landscapes, architectures, etc. Table 3.1 shows a sample set of GAN, Graphics, and Real

TABLE 3.1: A sample set of images from the dataset used in the proposed study

GAN

Graphics

Real


images from the dataset. The entire dataset is split into the ratio of 60:20:20 to form the train, validation, and test sets, where the total number of images belonging to various categories are split proportionally across each set.

3.3 Methodology

This study formulates the image forensic task of distinguishing natural images from computer generated images as a three-class classification task with the classes being Real, GAN, and Graphics where class Real indicates natural images and the classes GAN and Graphics indicate computer generated images. Even though both GAN and Graphics images are computer generated, they are maintained as separate classes since they follow entirely different processes of image generation. Accordingly, this study puts up an amendment to the depiction of the general framework followed for *natural images versus computer generated images* problem as outlined by Quan et al. [54] by incorporating the three-class classification approach, as shown in figure 3.1.

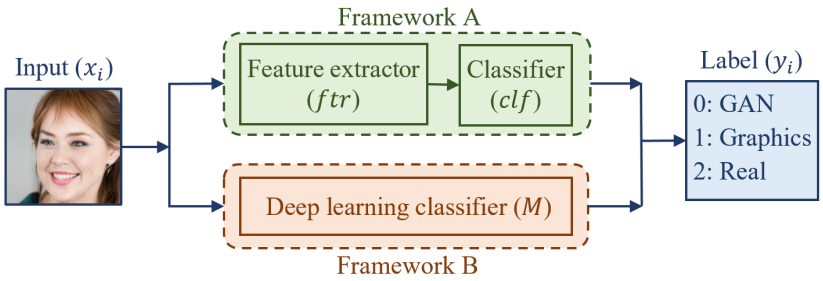


FIGURE 3.1: The general machine learning frameworks for digital image forensic problem

Framework A indicates conventional feature based classification that finds a mapping $y = clf(ftr(x))$ between the training data x and corresponding label y using a good choice of feature set (ftr) and classifier (clf) combination. Whereas framework B indicates deep neural network based classification that avoids the tiresome process of hand-crafted feature extraction and feature selection. This work follows framework B of deep neural network based classification with an aim to find the best-fit mapping function $M : y = M(x)$ for the training data $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ where x_i indicates i^{th} image in training set and y_i indicates corresponding image label denoted as 0 for the class GAN, 1 for the class Graphics and 2 for Real. Deep neural networks also allow the option of transfer learning where a network pre-trained over huge datasets for some n-class classification task is utilized for another m-class classification task even with less number of training data. Such a knowledge transfer from the source

network helps to obtain high accuracy in the target network of different tasks. This work incorporates transfer learning methodology that helps to transfer information from an object classification network trained on the huge ImageNet [120] dataset with 1000 classes, to the proposed three-class forensic task.

3.3.1 Motivation

Some state-of-the-art works for distinguishing natural images from computer generated images discuss the future scope of fusing deep learning models to create ensembled architectures that can improve classification accuracies [51, 54]. The concept of network fusion for distinguishing natural from computer generated images is motivated by ColorNet [121] where authors demonstrate that colorspace transformations can significantly affect classification accuracies and also observe that there is no hundred percentage correlation between different colorspace transformed images.

The choice of base network for network fusion is centered on the motivation that it should have less number of parameters so that to reduce network complexity without compromising on classification accuracy and further make chances of network fusion easier by not much shooting up the complexity of the fused network. Hence among the wide range of deep neural network architectures to hand, the proposed work chooses one of the latest networks EfficientNet [115], as the base network for the study that shows high performance in ImageNet recognition challenge [120]. Classification based on transfer learning methodology using a pre-trained EfficientNet-B0 model helps to reduce training complexity by keeping the number of trainable parameters of a single EfficientNet-B0 network to a very short number of only 3843.

3.3.2 Single Colorspace EfficientNet Network (*SC-EffNet*)

The family of EfficientNet networks viz., EfficientNet-B0 to EfficientNet-B7 emanates from the baseline network (called EfficientNet-B0) by systematically experimenting with scaling of network dimensions using a user-specified compound co-efficient [115]. The idea of uniform scaling carefully balances network depth, width, and resolution by empirically quantifying the relationship among these dimensions, unlike the usual scaling techniques that arbitrarily scale only one of these dimensions. The scaling co-efficient ϕ controls resources obtainable for model scaling and employs constants α, β and γ estimated by a grid search to allocate these extra resources to network depth d,

width w and resolution r respectively.

$$\begin{aligned} \text{i.e., depth: } & d = \alpha^{\phi} \\ \text{width: } & w = \beta^{\phi} \\ \text{resolution: } & r = \gamma^{\phi} \\ \text{s.t. } & \alpha.\beta^2.\gamma^2 \approx 2 \\ & \alpha \geq 1, \beta \geq 1, \gamma \geq 1 \end{aligned} \tag{3.1}$$

Such a principled procedure of compound scaling helps the network achieve state-of-the-art classification performance. The baseline network EfficientNet-B0 is constructed inspired by Mnasnet architecture [122] by utilizing multi-objective neural architecture search to optimize accuracy and FLOPS (Floating Point Operations Per Second) using a trade-off parameter. Initially, the compound co-efficient ϕ is fixed to 1 and grid search estimates optimal values for EfficientNet-B0 as $\alpha = 1.2, \beta = 1.1, \gamma = 1.15$ under the constraint of $\alpha.\beta^2.\gamma^2 \approx 2$. Later these values of α, β and γ are kept constant with different values of ϕ in equation 3.1 for scaling up the model to obtain EfficientNet-B1 to B7.

Among the EfficientNet family of networks, this study utilizes EfficientNet-B0 for the task, since it has the advantage of good classification performance as well as less number of parameters and FLOPS that can reduce training complexity, especially when trying to implement network fusion. The major building block of EfficientNet-B0 is the mobile inverted bottleneck convolution (MBConv) that utilize in-depth separable convolutions to reduce complexity of calculations [123] along with squeeze-and-excitation optimization [124], batch normalization and swish activation [125] that helps to further improve model performance. Figure 3.2 shows architecture of the EfficientNet-B0 network.

This study performs classification based on transfer learning methodology using a pre-trained EfficientNet-B0 model by removing its top dense layer with 1000 neurons and instead, fitting a fully connected dense layer with 3 neurons and *softmax* activation for this three-class classification task. All other layers in the EfficientNet-B0 network are kept frozen while training and validating the task. The initial phase of classification is performed on the dataset (described in section 3.2) by considering input images without any color conversion i.e., in the RGB colorspace itself (named as *SC-EffNet*$_{RGB}$). EfficientNet-B0 network can intake input images within the data range 0-255, since data normalization is included as a part of its architecture. Hence while implementing an EfficientNet-B0 model for RGB images, the input images are not rescaled to the range 0-1 as like the normal procedure of multiplication with 1./255,

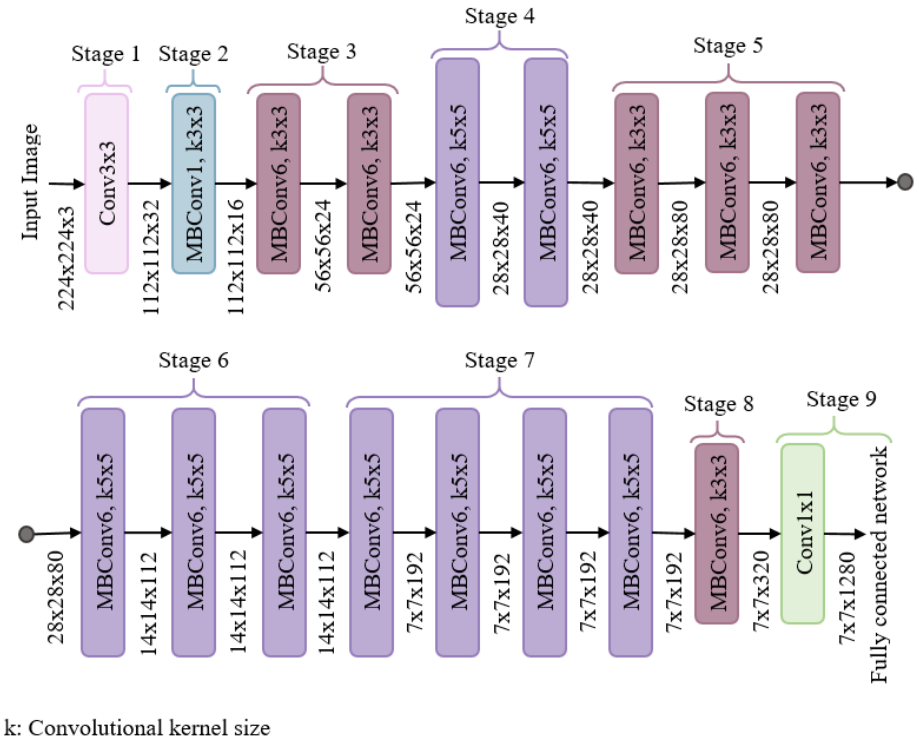


FIGURE 3.2: The architecture of EfficientNet-B0 network
(k: Convolutional kernel size)

which is most commonly performed while implementing many off-the-shelf deep neural network architectures.

In this image forensic task of classifying GAN, Graphics, and Real images, there was curiosity to know the colorspaces that significantly affect classification accuracies. Hence, the study performs classifications using the EfficientNet-B0 model over colorspace transformed images. Colorspaces chosen for the set of experiments include HLS, HSV, LAB, LCH, XYZ, YCbCr, YDbDr, YIQ, YPbPr, and YUV, which are the most commonly known and used color spaces. Except HLS and YDbDr, all the other colorspaces chosen in this study are also experimented in ColorNet [121] for object classification task, because of their easiness in the transformation from RGB. For compiling the model, the study utilizes categorical cross-entropy as the loss function, *Adam* optimizer with a learning rate of 0.001, batch size of 256 and 100 epochs.

In the case of colorspaces other than RGB, an additional intensity rescaling procedure is performed over the transformed images as their intensities do not follow the range 0-255 for being admitted to the EfficientNet-B0 model. The image intensity rescaling procedure that is followed in the proposed work is given in algorithm 3.1.

Algorithm 3.1: Rescale image intensities to the data range 0-255

Input: Colorspace transformed image T_{img} with color channels $[ch_1, ch_2, ch_3]$
Output: Intensity rescaled image R_{img}

Initialize R_{img} to be empty
for $i \leftarrow 1$ ***to*** 3 **do**
  $min(i) = min(T_{img}[ch_i])$
  $max(i) = max(T_{img}[ch_i])$
  $R_{img}[ch_i] = round\left(\frac{T_{img}[ch_i] - min(i)}{max(i) - min(i)} \times 255\right)$
end
return R_{img}

The test accuracies of classification for RGB images and for all the colorspace transformed images without rescaling and after rescaling are shown in table 3.2. There was curiosity on whether rescaling colorspace transformed images would in any case reduce the classification accuracies. But, it is found that the classification accuracies instead improved on rescaling the colorspace transformed images to the range 0-255, for every colorspace transformation, particularly for LAB and LCH colorspaces (as observed in table 3.2). Also, the classification in HLS colorspace was able to be implemented only after rescaling.

TABLE 3.2: Accuracy of *SC-EffNet* for different colorspaces in percentage (The highest accuracy is given in boldface)

Colorspace	Without Rescaling	After Rescaling
RGB	**82.13**	-
HLS	-	77.79
HSV	77.96	80.38
LAB	40.29	77.42
LCH	36.33	80.52
XYZ	80.00	80.26
YCbCr	74.66	75.75
YDbDr	75.13	75.58
YIQ	74.83	76.54
YPbPr	74.17	75.92
YUV	74.38	75.79

The three-class forensic classification task of distinguishing GAN, Graphics, and Real images obtains the highest accuracy of 82.13 percent when images are in RGB colorspace itself, as against ColorNet [121] performed over the object classification task that obtains its highest accuracy in LAB colorspace. Also, we observe that three

other colorspaces, HSV, LCH, and XYZ, have their accuracies near to RGB after rescaling, unlike ColorNet where except LAB every other colorspace have similar values of accuracies. A much more detailed view of classification results showing accuracies of each class separately is shown in table 3.3.

TABLE 3.3: Class accuracy and total accuracy of *SC-EffNet* for different colorspaces in percentage (The highest two accuracies in each class and total accuracy are given in boldface)

Colorspace	GAN	Graphics	Real	Total Accuracy
RGB	88.75	**79.88**	**77.75**	**82.13**
HLS	90.13	70.75	72.50	77.79
HSV	**93.88**	**75.63**	71.63	80.38
LAB	88.13	70.38	73.75	77.42
LCH	**92.87**	74.88	73.82	**80.52**
XYZ	90.75	74.61	**75.42**	80.26
YCbCr	84.38	68.88	74.00	75.75
YDbDr	87.38	66.25	73.13	75.58
YIQ	86.38	69.88	73.37	76.54
YPbPr	84.13	70.88	72.75	75.92
YUV	81.00	71.00	75.38	75.79

It can be observed that the accuracy of each class varies highly for different colorspaces. The accuracy of class GAN is comparatively higher than the other two classes for all the colorspaces. The highest accuracy for class GAN is observed in HSV colorspace and for classes Graphics and Real is observed in RGB colorspace. LCH and XYZ are the other colorspaces that show the nearest higher accuracies for these classes. Since the highest accuracies for each of the classes when viewed individually are obtained in different colorspaces, there is scope for increasing the total accuracy of the task by combining these colorspaces. Therefore, the study tries to combine the *SC-EffNet* networks of colorspaces that obtain the highest accuracy for each class when treated individually and also the colorspaces that obtains the highest overall accuracy i.e., the combinations of RGB, HSV, LCH, and XYZ to form a Multi-Colorspace fused EfficientNet.

3.3.3 Multi-Colorspace fused EfficientNet Network (*MC-EffNet*)

For combining the networks operating in different colorspaces, each colorspace except RGB is rescaled and passed through a separate EfficientNet-B0 model pre-trained over the ImageNet dataset. The top dense layer of each EfficientNet-B0 with 1000 neurons

is removed and all other layers are kept frozen for the training phase similar to the *SC-EffNet* based classification. The EfficientNet-B0 networks without the top dense layer now returns a feature vector of size 1280, for each colorspace network. A parallelly fused model is constructed, where the outputs of all the colorspace networks used for fusion are concatenated and provided to a dense layer with three neurons, suitable for the proposed classification task. Test accuracies of the different fused models constructed from RGB, HSV, LCH, and XYZ colorspace networks are shown in table 3.4. The fusion technique shows an increase in overall accuracy, especially the combination of the three colorspace networks RGB, LCH and HSV that produce a high accuracy of 87.96 percent, an increase of 5.83 percentage points from *SC-EffNet*$_{RGB}$ model. But the addition of XYZ colorspace network again to this fused model is seen to slightly degrade the accuracy. Hence the study adheres to the three colorspaces, RGB, LCH and HSV, to build the Multi-Colorspace fused EfficientNet model, *MC-EffNet-1*.

TABLE 3.4: Accuracy of *MC-EffNet* for colorspace network combinations in percentage (The highest accuracy is given in boldface)

Colorspace network combination	Accuracy
RGB + HSV	86.04
RGB + LCH	86.63
RGB + XYZ	82.63
RGB + LCH + HSV	**87.96**
RGB + LCH + HSV + XYZ	86.83

Since image forensic classification models, apart from providing high accuracies should also show a good amount of robustness towards post-processed images, *MC-EffNet-1* is tested over JPEG compressed images. It could be observed that even though *MC-EffNet-1* gives high classification accuracy for original images without any post-processing, the model accuracy decreases highly for JPEG compressed images, even for a quality factor of 90 (shown in table 3.5). Interestingly, the decrease in test accuracy for JPEG compressed images is comparatively higher for class GAN than Graphics and Real when observed class-wise. The accuracies of the base *SC-EffNet* networks over JPEG compressed images is also provided in table 3.5. For *SC-EffNet*$_{RGB}$, it can be observed that with an increase in compression (or decrease in quality factor), there exists a decrease in classification accuracy, but the rate of decrease is not as high as for *MC-EffNet-1*. But while checking *SC-EffNet*$_{LCH}$ and *SC-EffNet*$_{HSV}$ a very quick decay in their accuracies for compressed images can be observed. This helps to finalize that even though LCH and HSV colorspace transformations can highly increase the classification accuracies of images without any post-processing, they do not behave well

with JPEG compressed images.

TABLE 3.5: Model accuracy over original images and JPEG compressed images in percentage for different quality factors (qf)

Model	Original images	JPEG compressed images				
		qf=90	qf=80	qf=70	qf=60	qf=50
MC-EffNet-1	87.96	74.79	68.20	64.75	63.37	62.16
SC-EffNet$_{RGB}$	82.13	80.20	77.04	77.63	78.00	78.13
SC-EffNet$_{LCH}$	80.79	62.96	59.29	56.00	54.17	54.46
SC-EffNet$_{HSV}$	80.38	67.58	62.08	59.63	58.54	57.25

On further investigation, blocking artifacts is observed in the compressed images, particularly in the compressed GAN images, which on colorspace transformations becomes very much visible as blocks with uniform intensity values, replacing original intensities and region or shape information in that area of compressed images. This creates differences in image contents between the same image moving through the RGB pipeline and the LCH or HSV pipeline of *MC-EffNet-1* for compressed images, which might be the major cause for such degradation in model accuracy for compressed images. This difference in image content between compressed images passed through the RGB pipeline and the LCH or HSV pipeline increases as the quality factor decreases.

To maintain the advantage of high model accuracy provided by LCH and HSV colorspace transformations and to eliminate the negative effect of blocking artifacts when dealing with compressed images, an additional pre-processing block is attached to the LCH and HSV pipelines that employ a laplacian of gaussian filter over the images and add the residuals to corresponding images to avoid loss of information. The advantage of passing an image through different filters before computations are well explored in many fields of image forensics [62, 60, 59]. Such pre-processing operations can very well study hidden data representations to understand natural image statistics or any deviations from these statistics. Inspired by such image forensics works, the usage of laplacian of gaussian filter to pre-process colorspace transformed images refines *MC-EffNet-1* model to a more robust *MC-EffNet-2* model that achieves improved robustness towards JPEG compressions. Also, with the addition of laplacian of gaussian pre-processing block in *MC-EffNet-2*, the overall test accuracy improves to 89.38 percent an increase of 1.42 percentage points from *MC-EffNet-1*. The overall architecture of the proposed Multi-Colorspace fused EfficientNet model *MC-EffNet-2* is given in figure 3.3.

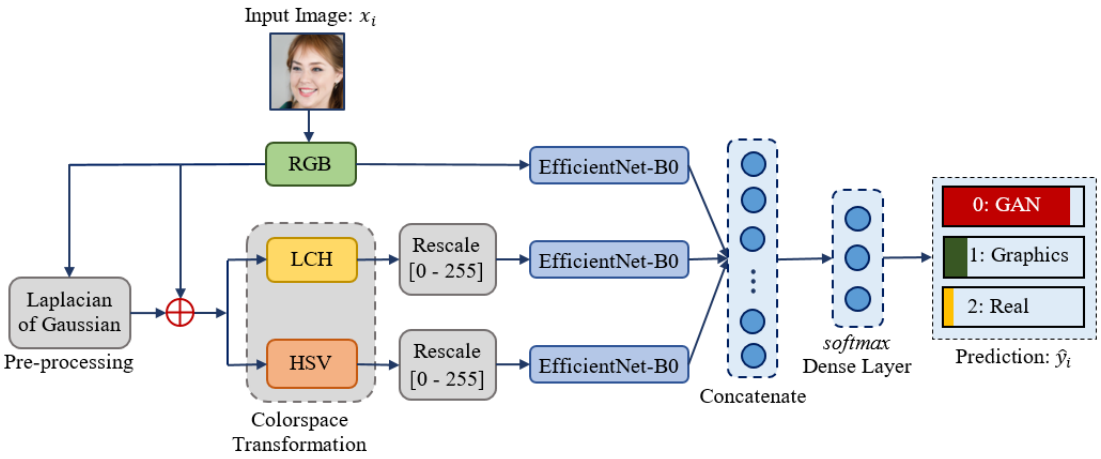


FIGURE 3.3: The overall architecture of Multi-Colorspace fused EfficientNet model *MC-EffNet-2*

3.4 Experimental Settings

In the proposed model *MC-EffNet-2*, with the use of transfer learning methodology, concatenation of feature outputs from the three colorspace network pipelines (RGB, LCH and HSV) produces a feature vector of length 3840 which is then provided to a dense layer with 3 neurons and *softmax* activation making the total trainable parameters of the model to be 11523. The model is compiled with categorical cross-entropy as loss function, *Adam* optimizer with learning rate 0.001, batch size of 256 and 100 epochs. Performance of the proposed model *MC-EffNet-2* is compared with a set of baselines discussed below.

3.4.1 Baselines

This work is the first of its kind considering the task of distinguishing natural images from photo-realistic computer generated images including both computer graphics and GAN images, as a three-class classification task. Therefore, baseline comparison is performed for the *MC-EffNet-2* model by fine-tuning state-of-the-art works belonging to the categories *natural images versus computer graphics*, *natural images versus GAN images*, and one another off-the-shelf deep neural network architecture, as three-class classification tasks with the dataset used in this study.

- Quan et al. [54] (*natural images versus computer graphics*): A CNN based work that proposes a local-to-global strategy for predicting classification results of local patches after which the global classification results of the full-sized images are

derived by majority voting. They compare their work with another patch-based CNN approach [42] and four other state-of-the-art feature based works [126, 32, 30, 29] where their work obtains higher accuracies and robustness. Hence their work is chosen as a baseline to compare the proposed work by replacing the final dense layer of two neurons in their CNN model with three neurons to suit this three-class classification task.

- Rezende et al. [51] (*natural images versus computer graphics*): A work that uses ResNet-50 for classifying natural images and computer generated images. They perform a result comparison between their 7 deep learning experimental settings and 17 approaches implemented in [1]. Among all the 24 results, their experimental setting of transfer learning combined with a shallow classifier SVM with RBF kernel obtains the highest accuracy when compared to their other deep learning settings and feature based approaches. Hence their high accuracy experimental setting is chosen as one of the baselines for comparing the proposed work by replacing the top layer to suit this three-class classification task.

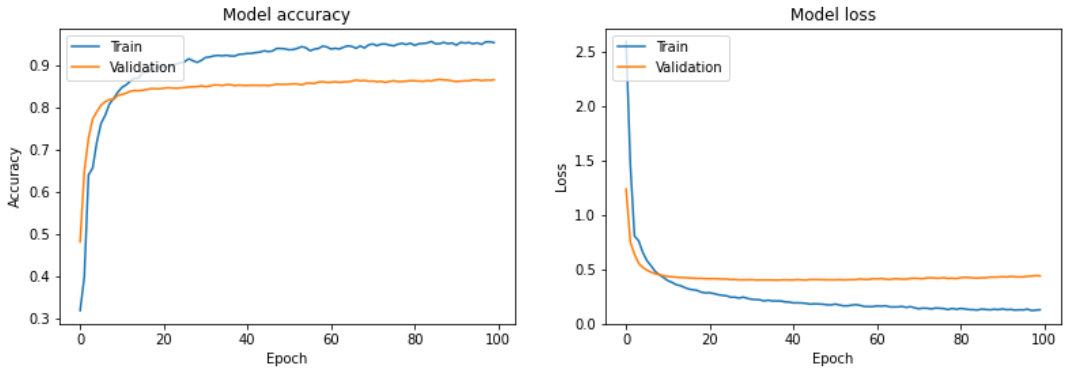
- Nataraj et al. [63] (*natural images versus GAN images*): A CNN based work to detect GAN images by using co-occurrence matrix of the RGB channels. Their work obtains higher accuracy when compared to three state-of-the-art works, a work based on steganalysis features [59, 60], a deep learning based work extracting residual features [62], and another work that fine-tunes generic deep learning architecture of XceptionNet [68] pre-trained on ImageNet. Hence their work [63] is chosen as a baseline to compare the proposed work by replacing their final sigmoid layer with a dense layer of 3 neurons and *softmax* activation.

- InceptionResNet [67] (Off-the-shelf deep neural network): A model that shows high classification accuracy for ImageNet classification task with almost 55.8M parameters. Transfer learning methodology on InceptionResNetV2 pre-trained over ImageNet dataset is attempted by freezing all its layers during the training phase and replacing the final prediction layer of 1000 neurons with three neurons. Other hyperparameters include batch size 256, *Adam* optimizer with a learning rate of 0.01 and 100 epochs.

3.5 Results and Discussions

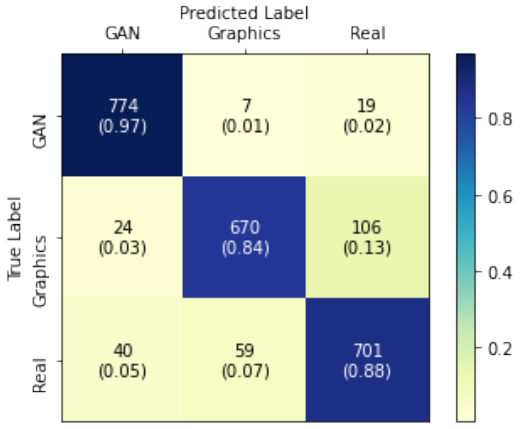
The proposed model *MC-EffNet-2* achieves a test accuracy of 89.38 percent, a gain of 1.42 percentage points when compared to *MC-EffNet-1*, and a gain of 7.25 percentage points when compared to *SC-EffNet*$_{RGB}$ that achieves highest accuracy among single

colorspace models. Figure 3.4 shows the plots of train-validation accuracy and loss, and also the confusion matrix of the test result for the proposed model *MC-EffNet-2*. Class GAN obtains a higher accuracy than the other two classes, Graphics and Real. For class GAN, *MC-EffNet-2* obtains an accuracy of 96.75 percent i.e., a gain of 2.87 percentage points when compared to *SC-EffNet*$_{HSV}$ that obtains the highest accuracy of 93.88 percent for class GAN among single colorspace models. Similarly, class Graphics achieves an accuracy of 83.75 percent i.e., a gain of 3.87 percentage points when compared to the highest accuracy of 79.88 percent for class Graphics obtained for *SC-EffNet*$_{RGB}$ among single colorspace models. Class Real seems to be most advantaged of the fusion technique since it achieves 87.63 percentage accuracy, a high gain of 9.88 percentage points when compared to the highest accuracy of 77.75 percent obtained for class Real of *SC-EffNet*$_{RGB}$ among the single colorspace models.



(A) Train-validation accuracy

(B) Train-validation loss



(C) Confusion matrix of test results

FIGURE 3.4: Train-validation accuracy, loss and confusion matrix of the proposed model *MC-EffNet-2*

Table 3.6 presents a comparison of model performances in terms of test accuracies of individual classes and total accuracy, for the proposed model *MC-EffNet-2* against the chosen baselines. Results indicate that the proposed model obtains the highest overall accuracy, and highest accuracies even for individual classes. Among the baselines, the next higher accuracy shown by InceptionResnet is less than the proposed model by 9.92 percentage points. All the baselines including those proposed for *natural images versus computer graphics* problem, except the baseline model proposed by Quan et al. [54], shows a similar trend of higher accuracy for class GAN followed by class Real and then Graphics. Whereas the model proposed by Quan et al. [54], shows higher accuracy for the class Graphics for which that model was originally proposed for, but not higher than the accuracy for class Graphics achieved by the proposed model *MC-EffNet-2*.

TABLE 3.6: Comparison of model performance accuracies in percentage (The highest accuracy is given in boldface)

Model	GAN	Graphics	Real	Total Accuracy
Quan et al. [54]	56.08	72.88	69.79	66.25
Rezende et al. [51]	59.40	52.83	57.40	56.54
Nataraj et al. [63]	76.00	48.25	57.13	60.46
InceptionResNet [67]	85.25	73.50	79.63	79.46
MC-EffNet-2 (Proposed model)	**96.75**	**83.75**	**87.63**	**89.38**

3.5.1 Statistical significance

In addition to the significant gains achieved by the proposed model *MC-EffNet-2* in terms of accuracy over various baselines, statistical significance tests is conducted between the proposed model and the baselines. The study performs the Stuart-Maxwell[11] test with conventional significance level i.e., a p-value of 0.05. A p-value of 0.00187 is obtained between the proposed model *MC-EffNet-2* and the model that obtained the highest accuracy among baselines (i.e., InceptionResNet). Also, a p-value of 0.01036 is obtained between the proposed model *MC-EffNet-2* and the model that obtained highest accuracy among the Single Colorspace EfficientNet Networks (i.e., *SC-EffNet*$_{RGB}$). This provides evidence to conclude that the results of the proposed model *MC-EffNet-2* are statistically significant over the best performing baselines.

[11] http://www.john-uebersax.com/stat/mcnemar.htm#stuart, *accessed: 17-12-2023*

3.5.2 Robustness Against Post-processing

Post-processing operations are quite common when uploading images to the web or social media. Therefore apart from producing good accuracies on original images in the dataset, an effective algorithm for image forensics should also be robust over post-processing operations. Hence this study evaluates robustness of the proposed model and the baselines towards typical post-processing operation of JPEG compression, where the models trained on original data are tested over ten different JPEG compression quality factors within the range 100 to 10, in steps of 10. Results of the robustness test is shown in figure 3.5. It can be observed that even though accuracy of *MC-EffNet-2* drops with a decrease of quality factor, it always achieves better performance than the baselines for all the quality factors. Also, it can be observed that *MC-EffNet-2* attains highly improved robustness than *MC-EffNet-1* with the inclusion of pre-processing block to colorspace transformations. Similar to the classification results of original images without compression, here also InceptionResnet is the baseline that shows the next higher results for different quality factors. The entire results over different compression quality factors indicate that the proposed model *MC-EffNet-2* achieves better robustness towards post-processing based on JPEG compression than the baselines.

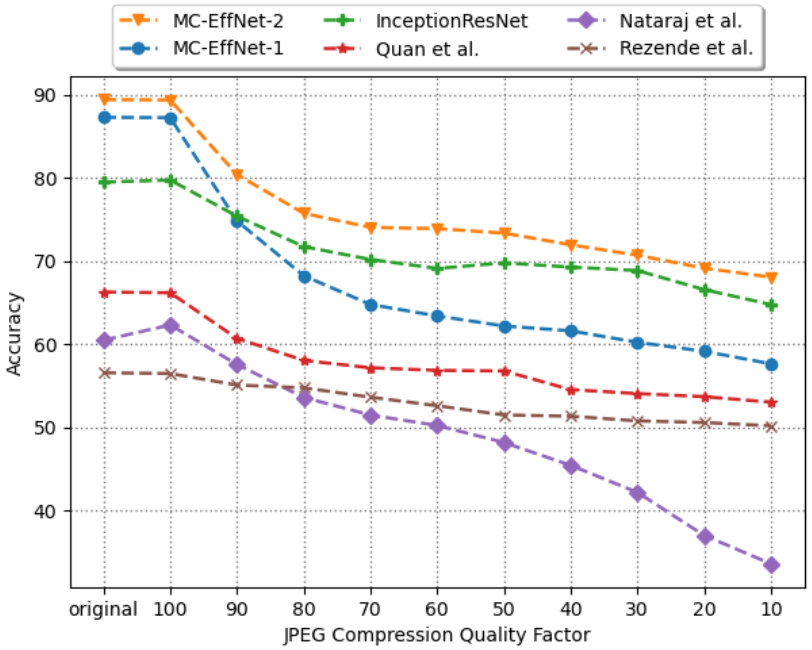


FIGURE 3.5: Classification accuracies of the models for various JPEG compression quality factors

3.5.3 Generalizability

Generalizability of the proposed model *MC-EffNet-2* and the baselines is analyzed by testing over three dataset combinations that are unseen during the training phase. In the first dataset, images for Real and Graphics classes are collected from PIM-Google (photographic images from Google Image Search) and PRCG (photo-realistic computer graphics images) sets of Columbia dataset [31] respectively. The images for class GAN is collected from the generated images of PG^2 (Pose Guided Person Generation [127]) GAN algorithm that produces high quality and realistic person images. The Columbia dataset [31] contains large diversity in image content. Many previous state-of-the-art works utilize this dataset for the two-class *natural images versus computer graphics* classification problem [32, 54, 43, 128]. From this dataset with 800 images per class, only twenty percent of the data is collected i.e., 160 images per class randomly (without considering the five images which were associated with incorrect labels in PIM-Google class as per findings in [54]), since it is only for testing model generalizability. To build a balanced generalizability test dataset, the same number of 160 random images is collected from PG^2 for class GAN. Apart from PIM-Google and PRCG, the Columbia dataset [31] also consists of another set, PIM-Personal (photographic images from the authors' personal collections) with 800 natural images, that constitutes the second dataset for the generalizability test by replacing class Real of the first dataset with 160 images randomly collected from PIM-Personal class.

Another dataset combination mostly used in many previous state-of-the-art works [42, 54, 128] for the *natural images versus computer graphics* problem is the RAISE [48] versus LDRD (Level-Design Reference Database) [129]. From the RAISE dataset that consists of high-resolution raw and uncompressed images specifically for image forensic research investigations, 160 images are collected randomly and these images are converted directly to JPEG format to form the class Real of the third dataset. LDRD consisting of screenshots from various video games can be seen utilized in [42] by selecting only those screenshots which seem to be photo-realistic, followed by cropping them to remove game information like dialogues, text bars, etc. From these images provided by [42], 160 images are collected to form the class Graphics of the third dataset. To form the class GAN, images generated by CycleGAN [130] that generates high-quality realistic images utilized in many state-of-the-art works to detect GAN images [63, 73, 71] are collected. From CycleGAN images, instead of choosing a single category of images amongst various unpaired image-to-image translation categories of objects and scenes, 160 images are collected randomly from the horses-to-zebra, zebra-to-horse, apple-to-orange and orange-to-apple categories.

Table 3.7 shows the results of generalizability tests for the proposed model and the baselines when tested over the three datasets, where one can observe that the proposed model outperforms all the baselines. Higher accuracies obtained for the proposed model *MC-EffNet-2* when tested over the datasets on which the model was not originally trained, indicates the promising nature of the proposed approach for tackling future challenges in computer generated images. Also, it can be observed that even though the proposed model is trained on the category of GAN algorithms that generate whole new images, such as StyleGAN, it could perform well on CycleGAN which belongs to the attribute transfer category of GAN algorithms.

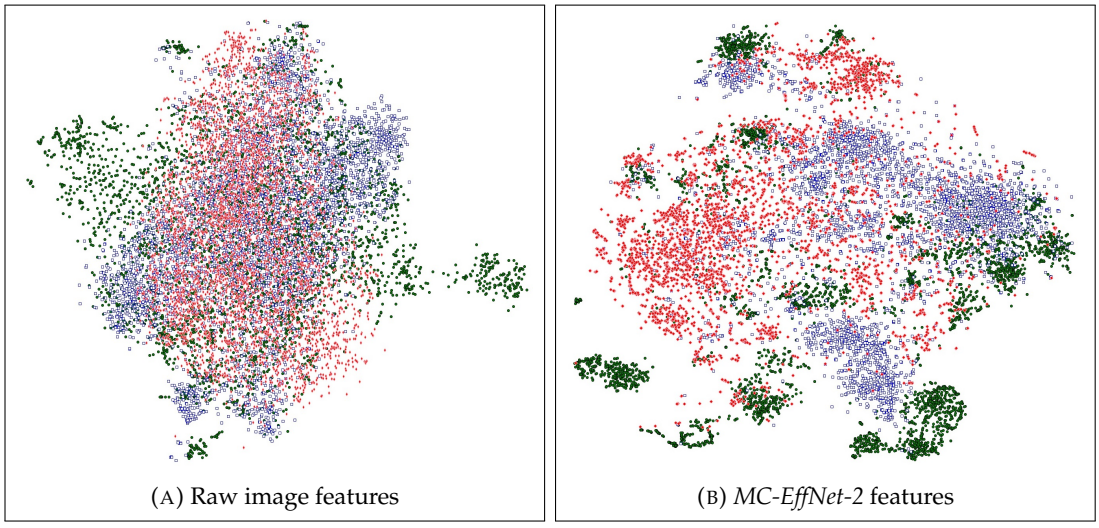
TABLE 3.7: Model generalizability over different datasets in percentage (The highest accuracy is given in boldface)

Model	PG^2 × PRCG × PIM-Google	PG^2 × PRCG × PIM-Personal	Cycle GAN × Raise × LDRD
Quan et al. [54]	54.22	56.27	60.01
Rezende et al. [51]	51.25	51.21	50.92
Nataraj et al. [63]	49.17	53.63	48.75
InceptionResNet [67]	62.08	67.16	71.74
MC-EffNet-2 (Proposed model)	**81.04**	**85.21**	**84.79**

The study also checks the applicability of the proposed model in distinguishing GAN generated videos, computer graphics videos and real videos by classifying the corresponding video frames into GAN, graphics and real categories. For the experiments, 30 short length videos are utilized i.e., 10 GAN generated and 10 real videos from the Deep Fake Detection Challenge (DFDC) dataset [131] and 10 graphics videos downloaded from CGSociety [132]. Each video in DFDC dataset contains almost 300 frames and hence from each graphic video 300 frames are selected randomly after removing the frames with title or other text descriptions. The frames corresponding to each video are passed through the proposed model *MC-EffNet-2* for testing. A majority voting scheme over the prediction results for entire frames of a video classifies the corresponding video into GAN generated, graphics or real video. A classification accuracy of 61.83 percent is obtained, indicating the applicability of the proposed model even to classify GAN, graphics and real videos.

3.5.4 Feature Visualization

The proposed model *MC-EffNet-2* projects raw pixels of the input images with dimension 224×224×3 or the feature vector of size 150528 to a lower dimension feature vector of size 3840, with an intention to provide a good amount of separability between the three classes, so that the top classifier layer attains a high classification accuracy. To understand the separability of features projected from the proposed model, a technique for dimensionality reduction called t-Distributed Stochastic Neighbor Embedding (t-SNE) [133] is implemented. t-SNE can visualize high dimensional features into a two-dimensional plane. Both the raw image features and output features from *MC-EffNet-2* are projected into two-dimensional plots with three different colors indicating three different classes. The plots are given in figure 3.6, where green circles represent class GAN, blue squares represent class Graphics and pink diamonds represent class Real. As can be seen from t-SNE visualizations, raw image features are more clustered particularly towards the center of the plot (figure 3.6a), whereas output features from the proposed model *MC-EffNet-2* are seen to be more separated (figure 3.6b). Thus, t-SNE visualizations prove that the proposed model *MC-EffNet-2* suits the forensic task of classifying natural images from computer generated images including both computer graphics and GAN images, by projecting raw image pixels to much better and separable feature space.



(A) Raw image features (B) *MC-EffNet-2* features

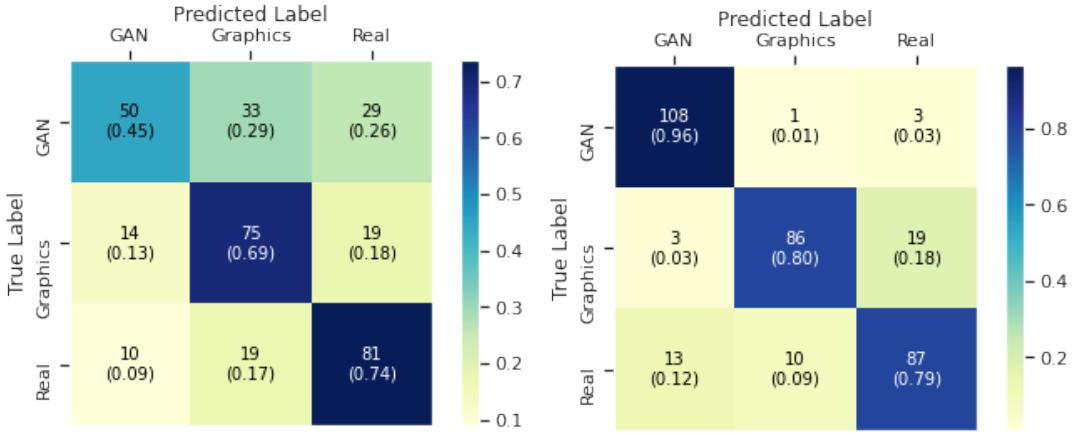
FIGURE 3.6: t-SNE visualization of the feature vectors (● indicates GAN, □ indicates Graphics and, ◆ indicates Real images)

3.5.5 Psychophysics Experiments

Based on an inquisitive interest to know how accurately humans can distinguish natural images from photo-realistic computer generated images including computer graphics and GAN images, this study performs a manual classification test for GAN, Graphics, and Real classes of images. The study gathers information from eleven human participants on a set of 330 images randomly selected from test data used in this study. The participants belong to the age group 22 to 40 years with normal or corrected-to-normal visual acuity and normal color vision. An annotation tool VIA [134] is utilized that helps participants to label each image as GAN, Graphics or Real and also mark parts/regions in the image that explains their decisions. This way of asking human participants to provide evidence/explanation in the form of region markings allows to omit the chances of lucky guesses and moreover provides insight into what the participants perceive as a suspect and/or evidence in the images. In VIA participants can zoom in and zoom out images for better analysis and no other constraints like viewing distance, time for observation, etc., are imposed. Each participant is asked to label thirty images randomly chosen and assigned to them and each image is annotated only once. For a participant, the entire experiment on thirty images including marking their explanations in the images takes nearly 45 minutes.

For comparison, accuracy of the proposed model *MC-EffNet-2* is also computed over the same set of 330 images selected for psychophysics experiments. Figure 3.7 shows the confusion matrices of manual classification performed by human participants (figure 3.7a) and that of the proposed model *MC-EffNet-2* (figure 3.7b), over the 330 images. For manual classification, a total accuracy of 62.42 percent is obtained, whereas for the same set of images the proposed model *MC-EffNet-2* obtains a higher accuracy of 85.15 percent i.e., a very high gain of 22.73 percentage points. It can be observed that the ability of humans to classify Real images is almost near to the proposed model *MC-EffNet-2*, with a decrease of 5 percentage points for manual classification, but for Graphics images *MC-EffNet2* highly outperforms manual classification. Similarly, in the case of GAN images manual classification accuracy is almost half of *MC-EffNet-2*. The overall results indicate that the ability of humans to identify photo-realistic computer generated images are very low and hence there is a high necessity for image forensic algorithms that can computationally aid to distinguish natural images and photo-realistic computer generated images. The proposed model *MC-EffNet-2* with a high classification accuracy, especially for photo-realistic computer generated images, is thus a better solution for the forensic task of distinguishing natural images

from photo-realistic computer generated images. The Stuart-Maxwell statistical significance test is also computed between the proposed model and manual classification where the proposed model obtains a p-value of $2.481E-06$ indicating the significance of *MC-EffNet-2* model over manual classification.



(A) Manual classification (B) *MC-EffNet-2*

FIGURE 3.7: Confusion matrices of classification performed by human participants and the proposed model *MC-EffNet-2*

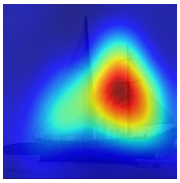
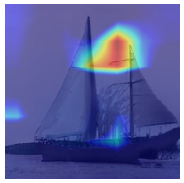
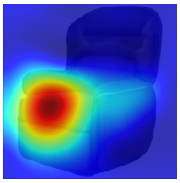
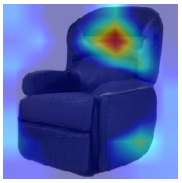
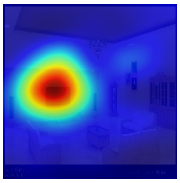
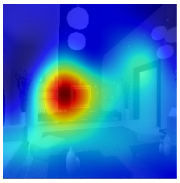
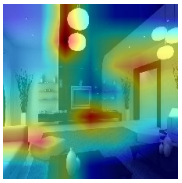
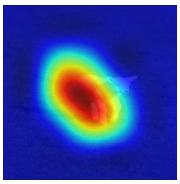
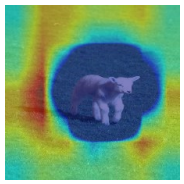
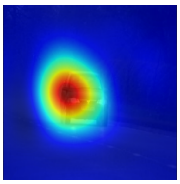
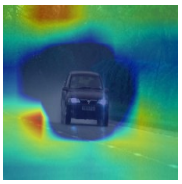
3.5.6 Understanding the Explanations

Apart from analyzing the model performance, behavior of the proposed model *MC-EffNet-2* is investigated so that one can trust the predictions of the proposed deep learning model. Hence, this study utilizes a Gradient-weighted Class Activation Mapping (Grad-CAM) [135] that makes use of class-specific gradient information to make the deep learning model more transparent through visual explanations. To obtain visual explanations Grad-CAM is employed at the penultimate layer of the fully trained and saved *MC-EffNet-2* model that constructs coarse localization maps of salient regions in input images that are significantly important for the predictions.

Since this task is formulated as an image classification problem build using transfer learning methodology over the source networks of pre-trained EfficientNets originally meant for the object classification task, the study initially tries to identify whether the proposed fused model is still looking for objects while making the decisions, or is it looking for regions significant for classifying natural images and computer generated ones in a forensic perspective. To address this question Grad-CAM explanations of a set of images from the dataset used in this study, are taken for the proposed fused model *MC-EffNet-2*, and also for the base network EfficientNet-B0 pre-trained on the

ImageNet dataset. Grad-CAM explanations from both the models for GAN, Graphics and Real images are shown in table 3.8. The EfficientNet-B0 network which is pre-trained on the ImageNet dataset as obvious mainly highlights the objects present in the images. But it can be observed that even though the proposed fused model *MC-EffNet-2* is built over the base network of EfficientNet which was originally designed for the object classification task, after applying transfer learning to this forensic task, does not primarily give importance to objects as in source task, rather highlights regions that are significant in classifying images as GAN, Graphics or Real in a forensic perspective. This indicates the fitness of the proposed model *MC-EffNet-2* as a forensic solution to classify GAN, Graphics and Real images.

TABLE 3.8: Grad-CAM explanations from the base network EfficientNet-B0 and the proposed model *MC-EffNet-2* for GAN, Graphics and Real images

Original image	Grad-CAM explanations		Original image	Grad-CAM explanations	
	EfficientNet	*MC-EffNet-2*		EfficientNet	*MC-EffNet-2*
GAN					
					
Graphics					
					
Real					
					

Accordingly, the study tries to understand what makes the proposed *MC-EffNet-2* model label an image as GAN, Graphics or Real in the context of image forensics.

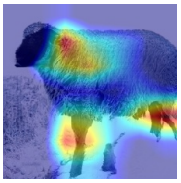
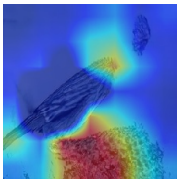
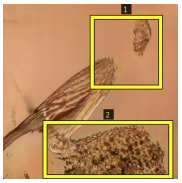
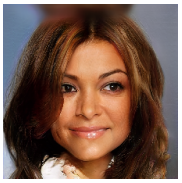
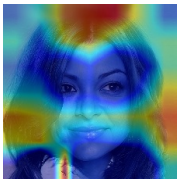
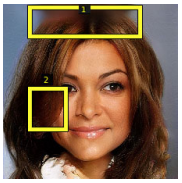
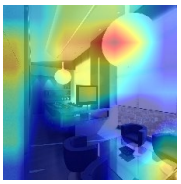
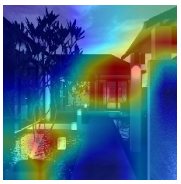
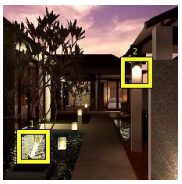
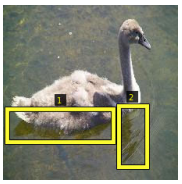
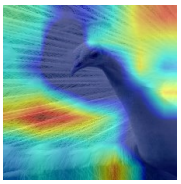
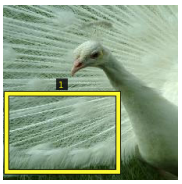
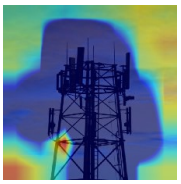
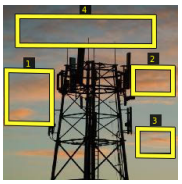
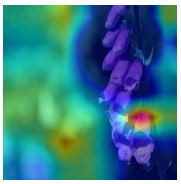
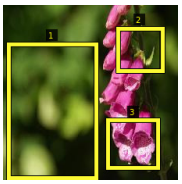
For this study, manual explanations in the form of region markings (yellow bounding boxes[12]) from human participants of the psychophysics experiments are also utilized, to compare whether Grad-CAM explanations given by the proposed model have any similarities with the manual explanations. First, the case of images for which the model and human participants provide correct predictions are considered and the Grad-CAM and manual explanations of these images are analyzed. Table 3.9 shows the examples of this case, where each set of two rows indicates GAN, Graphics and Real classes, respectively, and for each class, four sample images with their corresponding Grad-CAM and manual explanations are provided.

In the set of GAN generated images in the table, for the first image of a sheep it can be observed that the image is not completely formed and there are missing regions like the legs of the sheep. While looking at the Grad-CAM explanations provided by the *MC-EffNet-2* model for this image, it can be seen that the model captures image regions of legs and also slight differences in fur color and texture near the neck region. For the same image, the human participant has marked the leg region as an explanation for classifying that image as a GAN image. Similarly, in the second GAN image of a bird where the image is not completely formed at the head region and the texture of the bird is misplaced towards the bottom of the image, Grad-CAM captures both these regions along with the tail region and also some parts of the surroundings. The human participant has manually marked the head region and misplaced texture at the bottom of the image. In the next GAN image of a human face, where no misformations or misplacements can be seen evidently, Grad-CAM mainly highlights the texture of hair, some regions of the face and also some parts of the surroundings. For this image, the human participant has highlighted the hair region as an explanation. In the GAN image of the tower also, both the Grad-CAM explanation of the model and the manual annotation, highlights almost the similar region of the image. Hence in the GAN image examples, similarities can be observed between Grad-CAM explanations provided by the proposed model and the manual explanations.

Next, the Grad-CAM and manual explanations of Graphics images in the second set of rows are analyzed. For all correct predictions of Graphics images, Grad-CAM explanations are most commonly seen to highlight uneven illuminations or illuminated regions in images. Human participants are also seen to mark such regions of uneven illuminations. In the third set of Real images, Grad-CAM explanations are commonly seen to be centered on the surroundings focusing on complex variabilities in the background regions. In the first Real image, along with background regions, a

[12]Boundaries of the bounding boxes are thickened in this paper for better visibility

TABLE 3.9: Explanations of images for which the proposed model *MC-EffNet-2* and human participants both produce correct predictions

Original image	Grad-CAM explanation	Manual explanation	Original image	Grad-CAM explanation	Manual explanation
GAN					
					
					
Graphics					
					
					
Real					
					
					

major significance can be seen given to the shadow of the swan in the water by Grad-CAM as well as the human participant. Similarly, the feather regions in the second

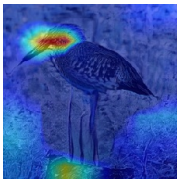
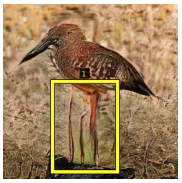
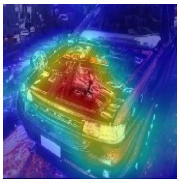
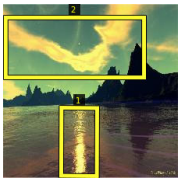
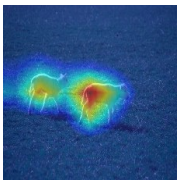
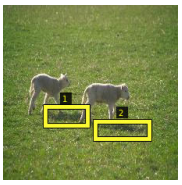
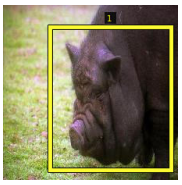
image, clouds in the third image, and the blurred natural background in the fourth image can be seen highlighted in both Grad-CAM and manual explanations. From the overall results, it can be summed up that explanations of the *MC-EffNet-2* model are mostly similar to manual explanations, and also the proposed model is able to identify more number of salient regions than human participants to distinguish between natural and computer generated images. Thus, the explanations demonstrate the powerful nature of the proposed model *MC-EffNet-2* to take decisions meaningfully.

This study also tries to get insights on the wrong predictions by examining images for which human participants provide correct predictions but the proposed model *MC-EffNet-2* produces wrong predictions. A few examples of this case are given in table 3.10. The first image (a) is a GAN image that is manually predicted as GAN itself but *MC-EffNet-2* misclassifies it as Real. Grad-CAM explanation from *MC-EffNet-2* shows that the decision is produced from the head region, rock at bottom of the image and surroundings, but not mainly from the misformed regions like the leg region. For Graphics image (c), which is manually predicted as a Graphics itself, *MC-EffNet-2* misclassifies it as GAN. Grad-CAM explanation shows that it produces decision from only the object region of one of the photographs in the image and not from regions of illuminations as usually seen in the case of Graphics images. Similarly, the Graphics image (d) is misclassified as Real taking into account the regions of mountains and some parts of the background, rather than considering regions of illuminations in the image as usually seen in the case of graphics images. Similarly, in the Real image (e) object regions of the sheeps are highlighted by *MC-EffNet-2* than the meaningful explanations like surroundings or shadows as usually seen for Real images which might be the reason for its misclassification as GAN. Whereas in the images (b) and (f), *MC-EffNet-2* highlights uneven illuminations which might be the reason for its misclassification as Graphics.

3.6 Summary

This work proposed a deep learning based Multi-Colorspace fused EfficientNet model to classify natural images and photo-realistic computer generated images including both computer graphics and GAN images, as against the state-of-the-art works that have always discussed either *natural images versus computer graphics* or *natural images versus GAN images* problem at a time. The study compared the proposed model with state-of-the-art methods, where the proposed model outperforms all the baselines in terms of performance accuracy, robustness against typical post-processing operation of JPEG compression and generalizability towards other datasets which demonstrates

TABLE 3.10: Explanations of images for which human participants provide correct predictions but *MC-EffNet-2* produces wrong predictions

Original image	Grad-CAM explanation	Manual explanation	Original image	Grad-CAM explanation	Manual explanation
Ground-truth: GAN					
					
(a) GAN misclassified as Real			(b) GAN misclassified as Graphics		
Ground-truth: Graphics					
					
(c) Graphics misclassified as GAN			(d) Graphics misclassified as Real		
Ground-truth: Real					
					
(e) Real misclassified as GAN			(f) Real misclassified as Graphics		

the utility of the proposed model in real-world forensic applications. The study also conducted psychophysics experiments to realize how capable humans are in classifying natural images and photo-realistic computer generated images, where the results of manual classification accuracies were lower than the proposed model accuracies, particularly in classifying the photo-realistic computer generated images, indicating the necessity and usefulness of the proposed computational model for the task. The chapter also analyzed the behavior of the proposed model by visualizing salient regions in images that are responsible for classification decisions. These visual explanations of the proposed model were compared with explanations manually labeled by human participants for their correct predictions. Similarities were observed between explanations of the proposed model and manual explanations indicating that

the proposed model takes decisions meaningfully. To the best knowledge, such a comparison of visual explanations to understand whether the model behaves alike human explanations to produce decisions meaningfully is a new attempt that might even be useful in other digital image forensics or multimedia security tasks. To aid future research, these manual classifications along with the manually labeled visual explanations and other relevant materials including datasets and source codes are made publicly available at https://github.com/manjaryp/GANvsGraphicsvsReal and https://dcs.uoc.ac.in/cida/projects/dif/mceffnet.html.

Chapter 4

MCE-ViT: A Robust Approach Towards Distinguishing Natural and Computer Generated Images using Multi-Colorspace fused and Enriched Vision Transformer

Abstract: Even though the forensic classification task of distinguishing natural and computer generated images gets the support of the new convolutional neural networks and transformer based architectures that can give remarkable classification accuracies, they are seen to fail over the images that have undergone some post-processing operations usually performed to deceive the forensic algorithms, such as JPEG compression, gaussian noising, etc. This work proposes a robust approach towards distinguishing natural and computer generated images including both, computer graphics and GAN generated images using a fusion of two vision transformers where each of the transformer networks operates in different color spaces, one in RGB and the other in YCbCr color space. The proposed approach achieves high performance gain when compared to a set of baselines, and also achieves higher robustness and generalizability than the baselines. The features of the proposed model when visualized are seen to obtain higher separability for the classes than the input image features and the baseline features. This work also studies the attention map visualizations of the networks of the fused model and observes that the proposed methodology can capture more image information relevant to the forensic task of classifying natural and generated images.

This work was supported by the Women Scientist Scheme-A (WOS-A) for Research in Basic/Applied Science from the Department of Science and Technology (DST) of the Government of India

4.1 Introduction

With the advent of convolutional neural network architectures (CNN) and transformer based architectures, the image classification systems are achieving high classification accuracies. Utilizing these techniques the forensic algorithms or tools to distinguish natural and computer generated images while moving to attain some success have been mostly vulnerable to different variations in images caused due to

image quality, resolution, compression quality, or color which are capable enough to dramatically modify or restructure the underlying image properties[13]. Image forgeries are almost always followed by these operations like adding some noise or applying JPEG compression, to obscure any traces of image generation or forgeries and thereby deceiving image forensic algorithms and tools. Therefore, besides producing a high performance system for classifying photographs and computer generated images of both categories (computer graphics and GAN), it is equally essential to review the robustness of these forensic systems towards various post processing operations. Hence, unlike the works in the literature that deals the forensic task of distinguishing natural and computer generated images as a binary-class classification task either by considering only computer graphics or only GAN images, this work proposes a generalized and robust forensic algorithm for classifying real, computer graphics and GAN images. The proposed methodology focuses on building a classification model such that it achieves high accuracy and is highly robust against various post-processing operations.

4.1.1 Research Question

This chapter addresses the following research questions.

***RQ1*:** Do the image categories such as natural images, computer graphics, and GAN generated images, behave similarly towards the post-processed images?

***RQ2*:** If not, can these differences be exploited to build a forensic system that distinguishes natural images, computer graphics, and GAN images with high performance as well as with high robustness against post-processing operations?

4.1.2 Delineating the Proposed Work in the Context of Literature

To the best knowledge, the work of *MC-EffNet* proposed in the previous chapter 3 is the only work in the literature that distinguishes natural versus computer generated images by considering both computer graphics and GAN images in the category of computer generated images. The work proposed in the previous chapter utilizes an EfficientNet CNN architecture for the three-class classification task. But, to the best

[13] https://www.nytimes.com/interactive/2023/06/28/technology/ai-detection-midjourney-stable-diffusion-dalle.html?auth=register-google&utm_source=pocket-newtab-intl-en, *accessed: 17-12-2023*

knowledge, no works are seen to be reported using the high performance recent deep learning architectures viz., transformer based architectures, for classifying natural images versus both categories of computer generated images, i.e., Graphics and GAN generated images. This work utilizes vision transformers and proposes a fusion based approach to distinguish natural versus computer generated images including both Graphics and GAN image categories.

Despite other works in the literature that utilize certain color space transformations for the task of distinguishing natural images from computer-generated images of either of the category [117, 118], the proposed work utilizes two color spaces in the methodology, based on rigorous set of experiments, one for improving the overall accuracy of the task of distinguishing natural images from computer-generated images of both categories, and the other color space that deals with improving the robustness of the proposed model against post-processing operations such as JPEG compression, gaussian noising, etc. The proposed methodology follows a transfer learning approach which helps to improve the accuracy of the task without involving the burdensome process of pre-training and thereby not increasing the training complexity. The proposed work is centered on an overlooked facet of leveraging the variability in vulnerabilities of different classes of images i.e., GAN, graphics, and real images towards post-processing operations for the task of distinguishing these classes of images with high accuracy and robustness.

4.1.3 Contributions

The major contributions of this chapter are:

- This chapter proposes a transformer based approach for the forensic task of distinguishing natural images and computer generated images including both computer graphics and GAN images.
- This work majorly focuses on building a classification model that is highly robust against post-processing operations such as JPEG compression, addition of gaussian noise, etc., and therefore a fusion of two color space transformations is employed in this approach.
- The chapter compares the performance of the proposed model with a set of baselines and could observe that the performance of the model outperforms the baselines, is highly robust against post processing operations, and is also generalizable to unseen test data.

- The chapter visualizes feature representations of the proposed model and compares with the input image features and feature representation of a baseline method, where it is observed that the proposed model provides better separability between the three different classes, helping towards improved classification performance.
- The chapter also visualizes the attention maps of the networks of the proposed fused model, to study the capability of the proposed methodology for the forensic task of classifying natural and computer generated images.

4.1.4 Chapter Organization

The rest of this chapter is organized as follows. Section 4.2 gives details of the dataset used in this study. Section 4.3 discusses the methodology of the proposed work along with the motivation and detailed description of the proposed model. Section 4.4 presents the experimental settings and details of the baseline models used for comparing the proposed model. Section 4.5 presents the results and discussion including the results of the experiments for robustness, generalizability, feature visualization, and the analysis of attention maps. Section 4.6 finally summarizes the chapter.

4.2 Dataset

The work proposed in this chapter also utilizes the same GAN, Graphics, Real image dataset used in the previous chapter 3 (explained in detail in section 3.2). In brief, the dataset comprises 12000 images in total, where each of the classes contains 4000 images. The class GAN consists of images collected from a variety of generative algorithms such as ProgressiveGAN [58], StyleGAN [99], StyleGAN2 [98], and StyleGAN2-ADA [119]. Whereas, the images for the classes Graphics and Real are obtained from the Computer Graphics versus Photographs dataset [1]. The entire dataset is challenging and contains a diverse category of images in terms of image content (indoor and outdoor scenes, animals, objects, etc.), origin (different generative/graphics algorithms or cameras, etc), and quality. Also, the computer generated images i.e., both GAN and graphics generated images are photorealistic, i.e., they are not manually easy to predict as computer generated ones. For the training, validation, and testing the dataset is proportionally split in the ratio 60:20:20 respectively.

4.3 Methodology

In this work, the forensic task of distinguishing natural and computer generated images is formulated as a three-class classification task, where the three classes are GAN, Graphics, and Real. GAN and graphics images, even though both being computer generated, are maintained as separate classes as they have different generation processes. For this three-class classification problem, transformer based deep neural networks is utilized to find the best-fit mapping function $M : y = M(x)$ for the train data $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ where for each image x_i in the training set, $y_i \in (0, 1, 2)$, indicating either of the classes GAN, Graphics or Real, respectively.

4.3.1 Motivation

The images transmitted through social media and online platforms are subject to image compression, mostly JPEG compression, where different online platforms follow different JPEG compression standards [2, 136, 137]. Also in the cases of intentionally propagating fake images to support some piece of fake news to convince the audience, these images are seen to be mostly subject to multiple image compressions [3, 4]. Hence, any model proposed for distinguishing natural and computer generated images, apart from obtaining high classification accuracy must also be highly robust towards compression. But most of the computational models in the literature proposed for the forensic task of distinguishing natural and computer generated images, are less robust to post processing operations, particularly the JPEG compression [138]. That is, as the compression factor of the images increases (or the quality factor of the images decreases), the accuracy of the computational models to distinguish natural and computer generated images decreases. The classification accuracies of different computational models to distinguish GAN, graphics, and real images are analyzed at varying levels of JPEG compression, including the state-of-the-art vision transformer based architectures ViT-Base and ViT-Large [139], state-of-the-art CNN based architecture InceptionResNet [67] and a few other baselines that classify natural and computer generated images [140, 54, 51, 63]. Figure 4.1 shows the results of classification accuracies of these models for various compression factors. As can be seen from the figure, the accuracy of all the models decreases as the compression increases.

To propose a robust model for the task, hence, the class accuracies of these models are further analyzed. Figure 4.2 shows the individual class accuracies of the models for varying levels of JPEG compression i.e., accuracies of the class GAN generated images at varying levels of compression factors in figure 4.2a, graphics generated images in figure 4.2b, and real images in figure 4.2c. It can be observed that, for uncompressed

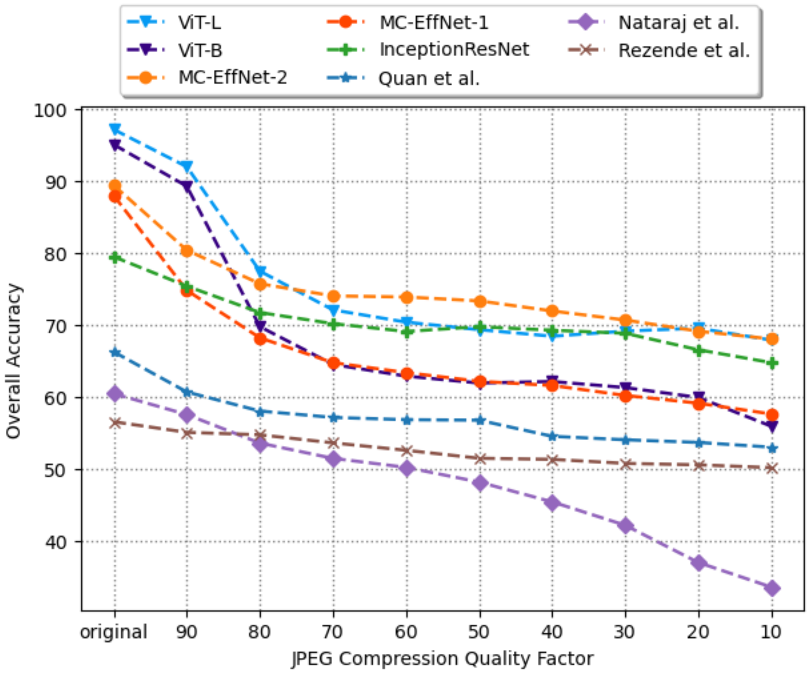
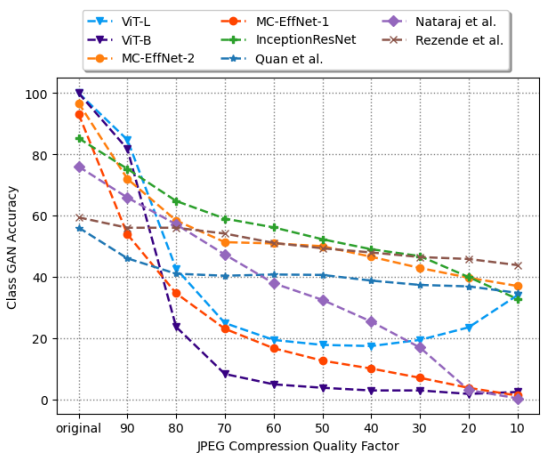
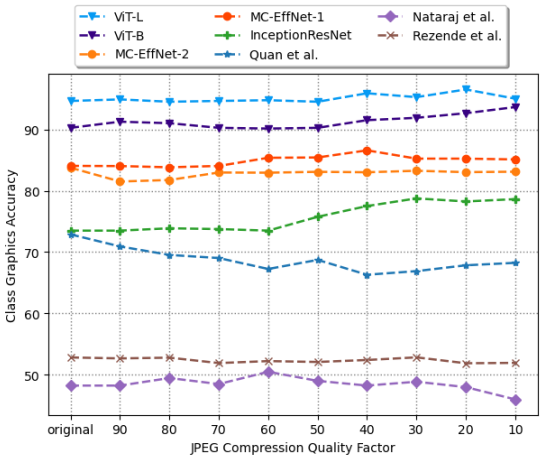


FIGURE 4.1: Classification accuracies of the models for various JPEG compression quality factors

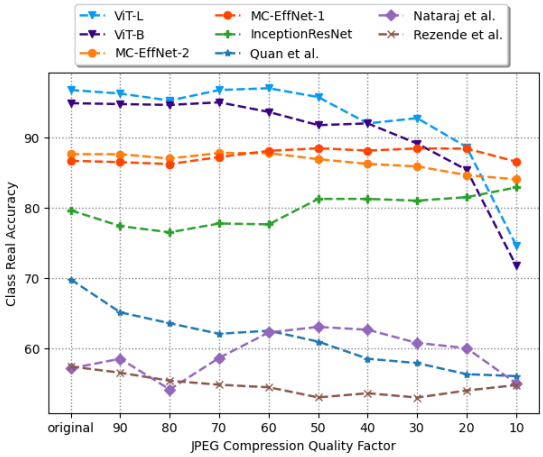
images (original images with zero compression factor), class accuracies of GAN generated images are always higher than the accuracies of graphics and real images, for all the models. But at the same time, as the compression increases, class GAN is the most affected (as shown in figure 4.2a). That is, there is a very high drop/decay in the class GAN accuracies of the models with even small increments in JPEG compression. Whereas, for the class Graphics in figure 4.2b, there is negligible drop in class accuracies of the models, it has almost stable accuracies as the compression increases. Also, for the class Real in figure 4.2c, the drop in class accuracies of the models as the compression increases are very less when compared to the class GAN; but there are certain slight drop or increase in accuracies for the class Real more than the rates of class Graphics. Altogether, analyzing the class accuracies shows that the rate of decrease in class accuracies with the increase in compression differs for different classes. And hence, how can this vulnerability of the decrease in accuracy with an increase in compression, which differs for each class be exploited is the motivation for the proposed approach of a robust classification model for the forensic task of distinguishing natural and computer generated images.



(A) GAN class accuracies



(B) Graphics class accuracies



(C) Real class accuracies

FIGURE 4.2: Class accuracies of the models for various JPEG compression quality factors

4.3.2 Network Architecture

The methodology of this study is devised in such a way that the proposed model should obtain high classification accuracy and should also be highly robust. A Multi-Colorspace fused and Enriched Vision Transformer (*MCE-ViT*) model is proposed by parallely combining two transformer based networks that operates in two different color spaces, one for obtaining high classification accuracy and the other network dedicated to improve the robustness of classification. The entire architecture of the proposed model is shown in figure 4.3.

For obtaining high classification accuracies the study chooses as the first network of the fused model, one of the recent state-of-the-art transformer based models, the

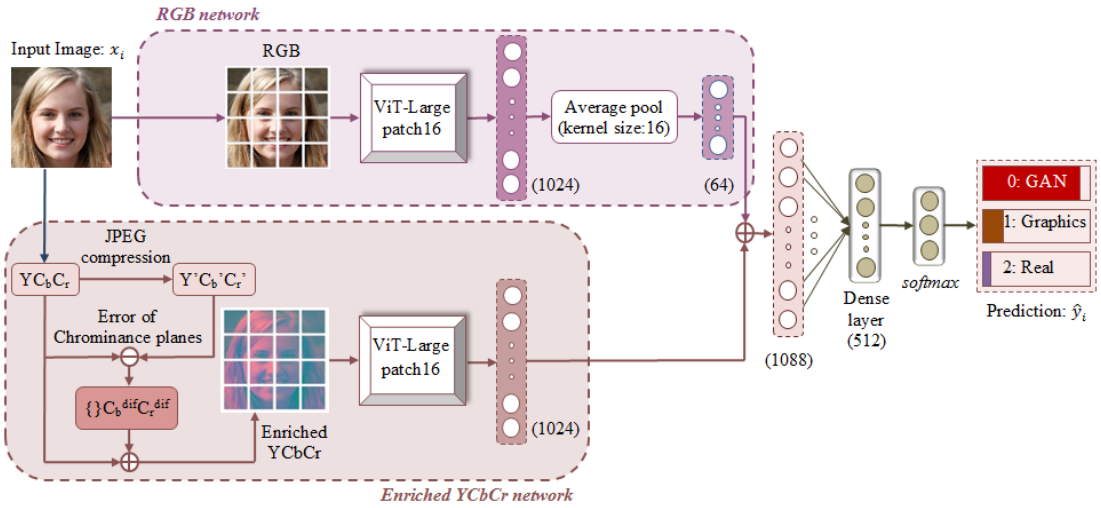


FIGURE 4.3: Overall architecture of the Multi-Colorspace fused and Enriched Vision Transformer (*MCE-ViT*)

Vision Transformer (ViT) [139]. The ViT network divides each of the images into non overlapping patches of fixed size which are linearly embedded, along with a position embedding for each of the image patches and a learnable token for classification.

The first ViT network used for the study takes as input the RGB images, and is named as *RGB ViT network* in this study. The study follows a transfer learning strategy where the ViT network pre-trained on the Imagenet 21-k dataset [141] is chosen, and fine-tuned using the task specific GAN, Graphics and Real images dataset (detailed in section 4.2). Even though this first *RGB ViT network* is found to obtain high classification accuracies for uncompressed images, it also has a performance decay when the images are JPEG compressed, as already seen in figure 4.1. Since the rate of performance decay is different for different classes (as detailed in section 4.3.1), allowing the model to also learn about these differences in performance decay among the classes, would help to better identify these classes in compressed scenarios. Therefore, a strategy is designed for the second network to form a fused model in such a way as to make the model learn these differences and thereby improve the robustness.

In the second network of the fused model, the input RGB images from the dataset are initially converted to YCbCr color space, the color space commonly used in the process of JPEG compression [142]. The process of JPEG compression mainly downsamples the chrominance planes Cb and Cr because the changes brought up in these planes are less visually discernible [142, 143]. On this basis, the second *Enriched YCbCr ViT network* enriches the YCbCr images with the information on the error of chrominance planes between the original images and their corresponding JPEG compressed

versions as follows.

$$\begin{aligned} &\mathrm{YCbCr} \xrightarrow{\text{JPEG compression}} \mathrm{Y'Cb'Cr'} \\ &\mathrm{YCbCr} - \mathrm{Y'Cb'Cr'} \xrightarrow{\text{Error of chrominance planes}} \{\}\mathrm{Cb}^{\mathrm{dif}}\mathrm{Cr}^{\mathrm{dif}} \\ &\mathrm{YCbCr} + \{\}\mathrm{Cb}^{\mathrm{dif}}\mathrm{Cr}^{\mathrm{dif}} \xrightarrow{\text{Enrichment}} \mathrm{YCbCr}_{\mathrm{enriched}} \end{aligned} \tag{4.1}$$

This strategy is motivated by the fact that, since the rate of differences between original and compressed versions of images are different for different classes, enriching the images with this information on the rate of differences would be highly beneficial to distinguish natural and generated images with high classification accuracies, even for images in post-processed scenarios thereby improving robustness of the model. Accordingly, the second pre-trained ViT network is fine-tuned using these enriched YCbCr images.

Both the first *RGB ViT network* and the second *Enriched YCbCr ViT network* produce an output feature vector. Even though the feature vector from the first RGB ViT network is very powerful to classify GAN, graphics, and real images with high accuracy in uncompressed scenarios, in order to decrease its effect towards performance decay in compressed scenarios, an average pooling is performed on this feature vector to obtain only the representative feature of the RGB ViT network. Thus, pooling the feature vector from the RGB ViT network helps to not affect the robustness of the model in compressed scenarios, as well as to maintain high model accuracies. Later, to build the fused model, the feature vectors from both the ViT networks are concatenated to form a single feature vector which is passed to a fully connected neural network with a dense layer and a 3-class output layer that determines the class label of the images.

4.4 Experimental Settings

The proposed study utilizes the vision transformer 'vit-large' with patch size 16 and input image size 224 x 224, for both *RGB* and the *Enriched YCbCr* networks of the fused model. The *Enriched YCbCr ViT network* utilizes JPEG compression with a quality factor of 90. The study follows transfer learning strategy where both the 'vit-large' networks are pre-trained on ImageNet-21k [141] and fine-tuned on the task specific dataset used in the study. Each ViT network produces an output feature vector of size 1024. After manually analysing the results of average pooling of the feature vector from the first *RGB ViT network* with various kernel sizes and strides, a kernel size of 16

without any stride is selected as a representative setting, and the resultant feature vector of size 64 is concatenated with the feature vector from the second *Enriched YCbCr ViT network*. The entire concatenated feature vector of size 1088 is fed to a dense layer of 512 neurons with *ReLU* activation function, followed by an output dense layer of 3 neurons with *softmax* activation function, making 559107 number of trainable parameters. The other hyperparameters are batch size 16, *categorical crossentropy* loss function, *Adam* optimizer, and 50 epochs.

The experiments are conducted on the deep learning workstation equipped with Intel Xeon Silver 4208 CPU at 2.10 GHz, 256 GB RAM, and two GPUs of NVIDIA Quadro RTX 5000 (16GB each), using the libraries Tensorflow (version 2.8.0), Keras (version 2.8.0), Torch (version 1.13.1+cu116), PyTorch Lightning (version 1.9.0), Transformer (version 4.17.0), and Albumentations (version 1.3.1).

4.4.1 Baselines

Performance of the proposed model is compared with the performance of the models proposed in the previous chapter 3, i.e., the *MC-EffNet-1* and *MC-EffNet-2* models designed to classify GAN, Graphics, and Real images. The proposed model is also compared with the other baselines detailed in the previous chapter 3. That is, an off-the-shelf deep neural network architecture of InceptionResNet [67], the works of Quan et al. [54] and Rezende et al. [51] that classifies natural and computer graphics images, and the work of Nataraj et al. [63] that classifies natural and GAN generated images, by fine-tuning these baseline works to the three-class classification task using the GAN, graphics, real dataset used in this study. The experimental settings of all these baselines are followed same as mentioned in the previous chapter (in section 3.4.1).

4.5 Results and Discussions

The proposed model achieves a test accuracy of 94.25 percent. The confusion matrix of the test result of the proposed model is shown in figure 4.4. Table 4.1 presents a comparison of the test results of the proposed model and the baselines in terms of total accuracy and class-wise accuracy. The test results indicate that the proposed model *MCE-ViT* outperforms the baselines in terms of overall test accuracy and even in the case of individual class-wise accuracy. *MCE-ViT* obtains a gain of 4.87 percentage points in terms of overall accuracy when compared to *MC-EffNet-2*, the best performing model among the baselines which is the work proposed in the previous chapter 3. Among the three classes, GAN obtains the highest individual class accuracy of

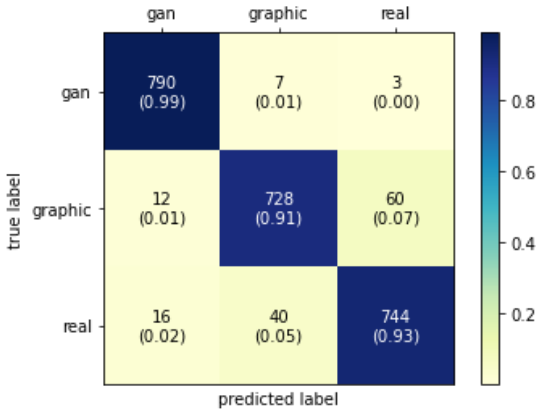


FIGURE 4.4: Confusion matrix of *MCE-ViT*

98.75 percent, achieving a gain of 2 percentage points when compared to the highest class accuracy of 96.75 percent obtained by *MC-EffNet-2* amongst the baselines. Class Graphics achieves an accuracy of 91.00 percent, a very high gain of 7.25 percentage points when compared to the highest class accuracy of 83.75 percent for the *MC-EffNet-2* model amongst the baselines. Class Real achieves 93.00 percentage accuracy, i.e., a gain of 5.37 percentage points when compared to the highest class accuracy of 87.63 percent obtained by *MC-EffNet-2* amongst the baselines.

TABLE 4.1: Comparison of model performance accuracies in percentage
(The highest accuracy is given in boldface)

Model	GAN	Graphics	Real	Total accuracy
MCE-ViT (Proposed model)	**98.75**	**91.00**	**93.00**	**94.25**
MC-EffNet-2 (Chapter 3)	96.75	83.75	87.63	89.38
MC-EffNet-1 (Chapter 3)	96.25	81.75	85.88	87.96
InceptionResNet	85.25	73.50	79.63	79.46
Quan et al. [54]	56.08	72.88	69.79	66.25
Nataraj et al. [63]	76.00	48.25	57.13	60.46
Rezende et al. [51]	59.40	52.83	57.40	56.54

4.5.1 Robustness Against Post-processing

Besides achieving high classification accuracies on original images or images that are not being post-processed, an efficient image forensic algorithm should also provide robust classification output over post-processed images. Hence, the robustness of the proposed model *MCE-ViT* is evaluated towards the typical post-processing operation of JPEG compression and is compared against the baselines. Apart from the baseline

models discussed in section 4.4.1, the robustness of the proposed model is also compared with the off-the-shelf pre-trained vision transformer networks, ViT-B/16 and ViT-L/16 that are fine-tuned using the dataset used in this study. Here, the ViT-L/16 results will serve as an ablation study for the proposed methodology that incorporates color space fusion and enrichment of YCbCr with the error of chrominance planes information using ViT-L/16.

Every model trained on uncompressed data (or original train data in the dataset) is tested separately over ten different test data created using various JPEG compression quality factors within the range 100 to 10, in steps of 10. Figure 4.5 shows the robustness test results, where it can be observed that compared to the baselines, the proposed model achieves improved robustness towards post-processing based on JPEG compression, for all the compression quality factors. Here also, similar to the classification results of original uncompressed images (table 4.1), *MC-EffNet-2* (the work proposed in the previous chapter) is the baseline model that shows the next improved robustness among all the baselines for different compression quality factors.

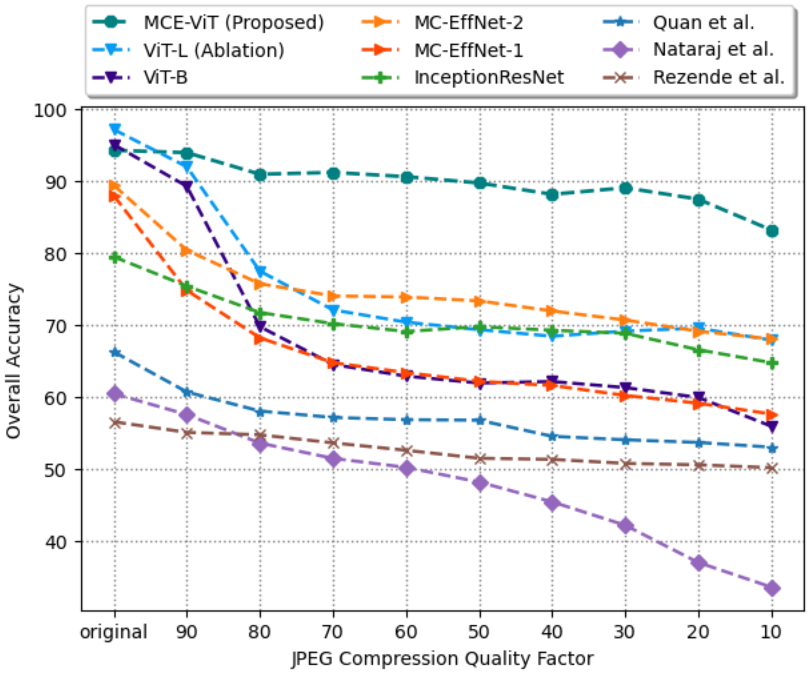


FIGURE 4.5: Classification accuracies of the proposed model and the baselines for various JPEG compression quality factors

Besides analyzing robustness against JPEG compression in terms of overall accuracy, the robustness of the proposed model is analyzed specifically for the class GAN,

since it is the class that is most affected by post-processing operations and also the major contributing factor towards the decrease in overall accuracy, as already discussed in section 4.3.1. Figure 4.6 shows the robustness test results of the proposed model and the baselines, specifically for the class GAN. It can be observed that the proposed model achieves very high class accuracy for all the compression factors. The class accuracy obtained for the proposed model while classifying original images (i.e., without compression) is 98.75 percent. At the same time, for the images compressed at a compression factor of 10 (i.e., at very high compression), the class accuracy of the proposed model is 85.13 percent. That is, for a very high compression factor of 10, the proposed model obtains a very high gain of 41.32 percentage points when compared to the next higher accuracy of 43.81 percent at compression factor 10 for the baseline work of Rezende et al. [51].

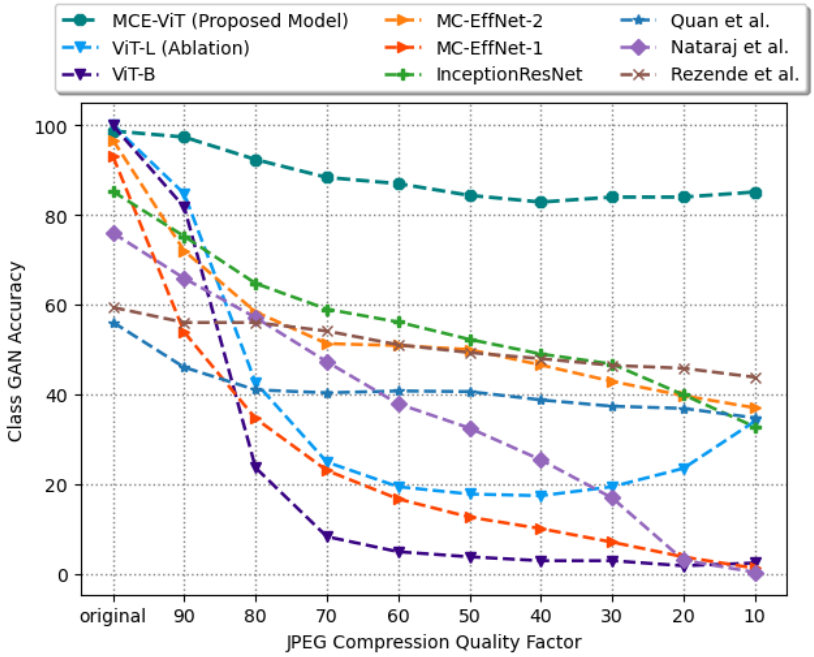


FIGURE 4.6: GAN class accuracies of the proposed model and the baselines for various JPEG compression quality factors

The class accuracy of the baseline ViT-L network (ablation) at a compression factor 10 is only 34.00 percent. That is, the proposed model even being built using the ViT-L network, but the methodology of color space fusion and the enrichment of YCbCr with the error of chrominance planes information between original and compressed images has proven to be highly advantageous in improving the robustness of the proposed model thereby achieving a class accuracy of 85.13 percent even at a compression factor

of 10, i.e., a very huge gain of 51.13 percentage points when compared to the ViT-L network (ablation).

The study also finds the robustness of the proposed model against other post processing operations and compares it with the baselines. The accuracy of the proposed model and the baselines at various standard deviation (σ) values of gaussian noise is shown in figure 4.7. It can be observed that compared to the baselines the proposed model achieves better robustness against Gaussian noising also.

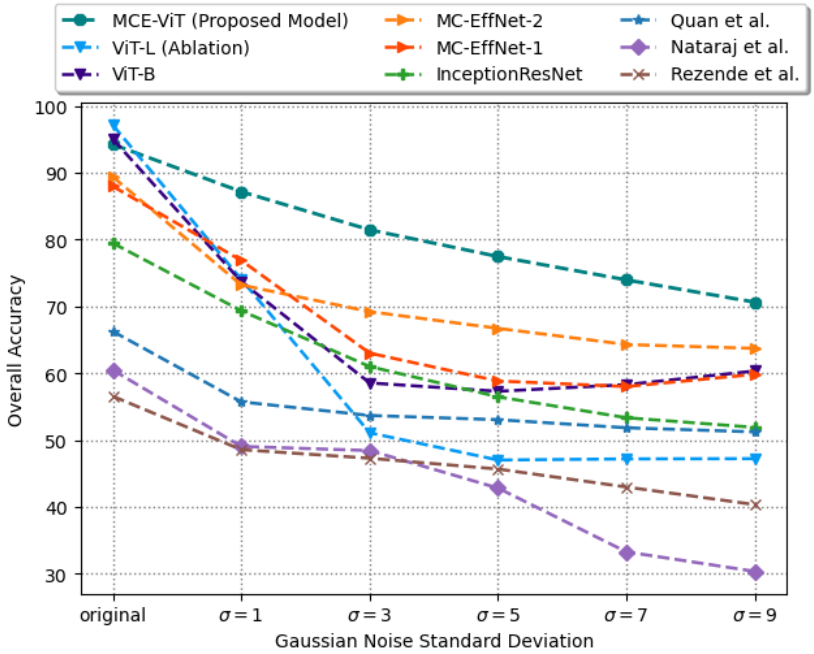


FIGURE 4.7: Classification accuracies of the proposed model and the baselines for various Gaussian noise standard deviations (σ)

4.5.2 Generalizability

This study also analyzes the generalizability of the proposed model *MCE-ViT* by testing over unseen data. The proposed model *MCE-ViT* which is fine-tuned on the dataset used in this study is tested over three other combinations of unseen GAN, Graphics, Real images, i.e., PG^2 versus PRCG versus PIM-Google [127, 31], PG^2 versus PRCG versus PIM-Personal [127, 31], and Cycle GAN versus Raise versus LDRD [130, 48, 129] datasets, with 160 images in each of the class of the three datasets, as experimented in the previous chapter 3. The generalizability test results are shown in table 4.2, where it can be observed that compared to the baselines the proposed model is able to obtain higher test accuracies. The results hence prove that the proposed

model *MCE-ViT* has better generalizability, which can even help towards the future challenges in generated image categories.

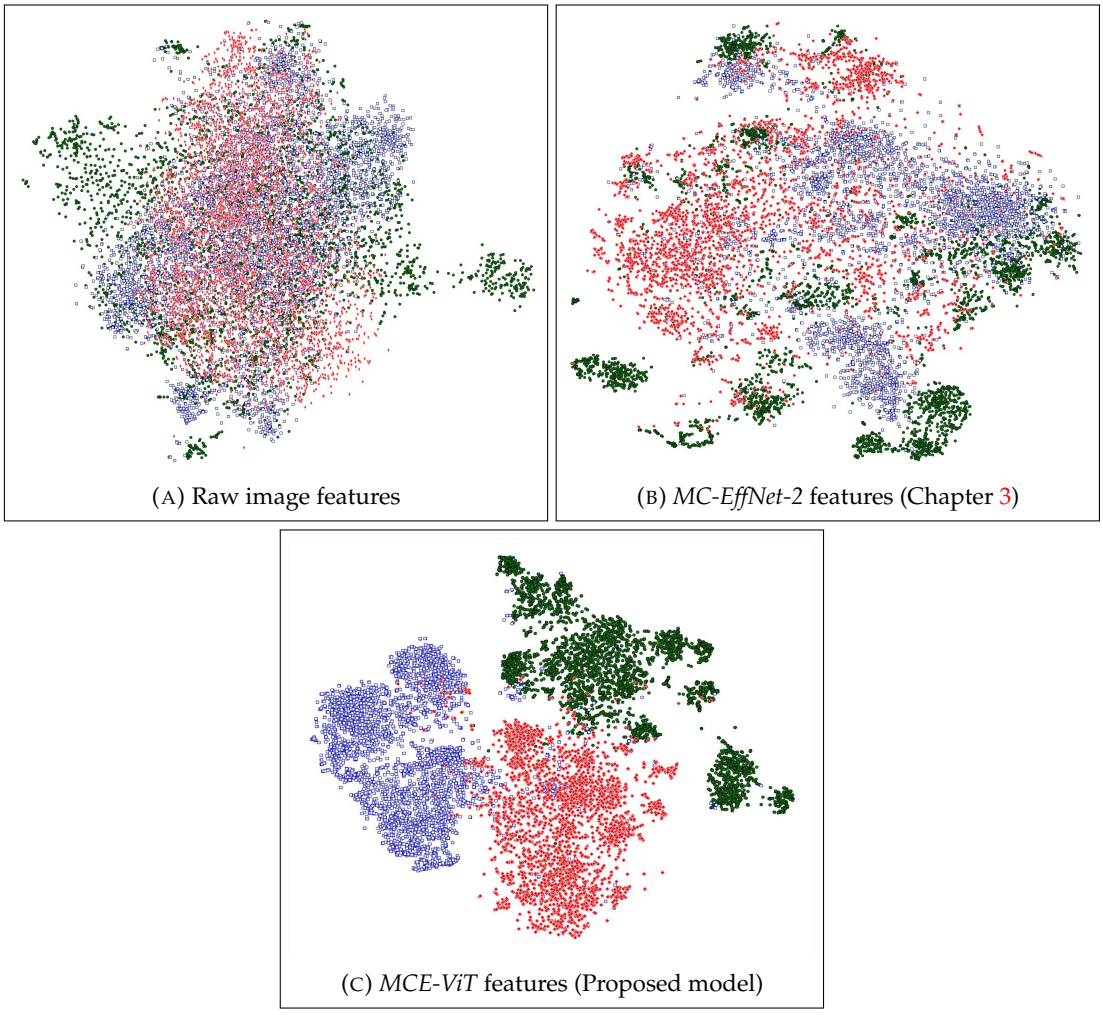
TABLE 4.2: Generalizability of the models over three datasets in percentage (The highest accuracies are given in boldface)

Model	PG^2 × PRCG × PIM-Google	PG^2 × PRCG × PIM-Personal	Cycle GAN × Raise × LDRD
MCE-ViT (Proposed model)	**87.84**	**90.31**	**91.75**
MC-EffNet-2 (Chapter 3)	81.04	85.21	84.79
InceptionResNet	62.08	67.16	71.74
Quan et al. [54]	54.22	56.27	60.01
Rezende et al. [51]	51.25	51.21	50.92
Nataraj et al. [63]	49.17	53.63	48.75

4.5.3 Feature Visualization

The dimension of images that are input to the proposed *MCE-ViT* model is 224 × 224 × 3, i.e., the length of the raw image features is 150528. *MCE-ViT* projects these raw image features into a smaller dimension of length 1088, with an aim to provide better separability between the three classes. The separability potential of the feature vector of the proposed model *MCE-ViT* is analyzed by comparing against the feature vectors of raw input image pixels and the best baseline model *MC-EffNet-2* (the work proposed in previous chapter 3) that obtains the next higher classification accuracy among all the baselines. The t-SNE [133] dimensionality reduction technique can visualize high dimensional features into a two-dimensional plane and thereby helps to easily compare the separability potential of the feature vectors. Using the t-SNE technique the raw image features, feature vector output from the baseline model *MC-EffNet-2* and feature vector output from the proposed model *MCE-ViT* are projected into two-dimensional plots, shown in figure 4.8. In each of the plots, three different colors are used to indicate three different classes, green circles are used to represent GAN image class, blue squares are used to represent Graphics images class, and pink diamonds are used to represent the class of Real images. From the t-SNE visualizations, it is clearly understandable that the raw image features do not have the potential to separate the classes, all the classes are clustered together, more particularly towards the center of the plot (figure 4.8a). The output features projected from the baseline model *MC-EffNet-2* (figure 4.8b) seems to produce separability between the classes better than the raw image pixels. Whereas, the output features projected from

the proposed model *MCE-ViT* (figure 4.8c) seems to produce much better separability between the three classes when compared to both raw pixels and the best baseline *MC-EffNet-2*. This demonstrates that the proposed model *MCE-ViT* has better capability to transform the raw image pixels to a much better separable feature space for the forensic task of distinguishing natural images from computer-generated images including both computer graphics and GAN images.



(A) Raw image features

(B) *MC-EffNet-2* features (Chapter 3)

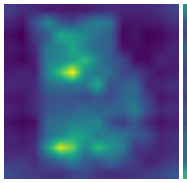
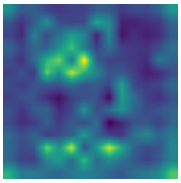
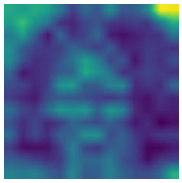
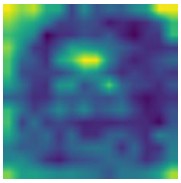
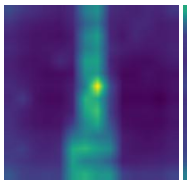
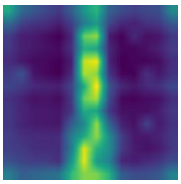
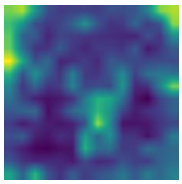
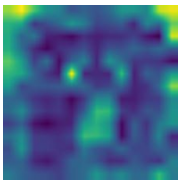
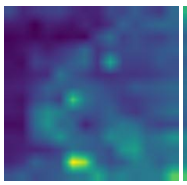
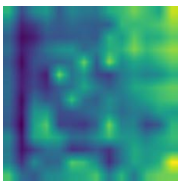
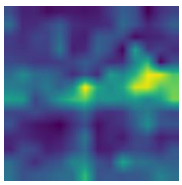
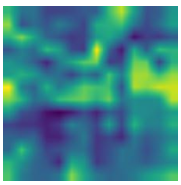
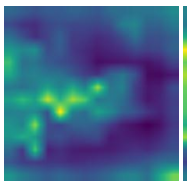
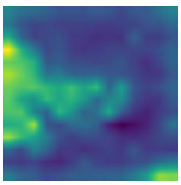
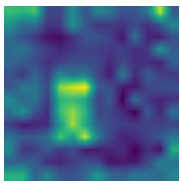
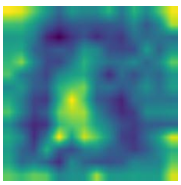
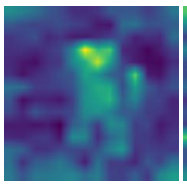
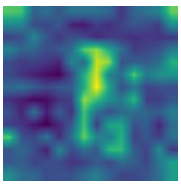
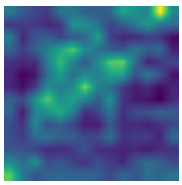
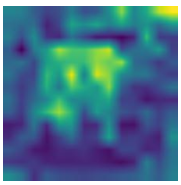
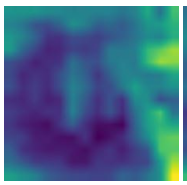
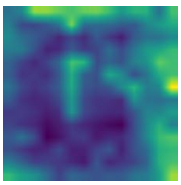
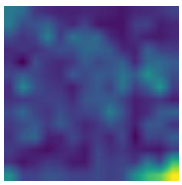
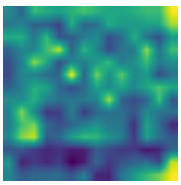
(C) *MCE-ViT* features (Proposed model)

FIGURE 4.8: t-SNE visualizations of the feature vectors (● indicate GAN images, □ indicate Graphics images and, ◆ indicates Real images)

4.5.4 Attention Visualization

The attention map visualizations of *RGB network* and *Enriched YCbCr network* for a set of GAN, Graphics and Real images is shown in table 4.3. While comparing the

TABLE 4.3: Attention map visualizations of the *RGB network* and the *Enriched YCbCr network*

Original image	Attention map visualization		Original image	Attention map visualization	
	RGB network	*Enriched YCbCr network*		*RGB network*	*Enriched YCbCr network*
GAN					
					
					
Graphics					
					
					
Real					
					
					

attention maps of both the networks of the proposed fused model i.e., the *RGB network* and the *Enriched YCbCr network*, it can be observed that compared to the *RGB network*, the *Enriched YCbCr network* could potentially identify more regions or information in an image that are helpful for classifying images into the classes GAN, Graphics or Real. For example in the first generated image of a cat in the set of GAN images, the RGB network mainly captures the edges of the cat, whereas the *Enriched YCbCr network* captures some of the information captured by the *RGB network* and even more including from the bacground, irrespective of the regions of the object in the image. Similarly, in the second generated face image, the forehead, right eye, and the image background are given more attention by the *Enriched YCbCr network* than the *RGB network*.

Similarly in the set of graphics images also, it can be observed that the *Enriched YCbCr network* is able to capture more relevant image information in detail than the *RGB network*, such as the regions of uneven illuminations in the image. Also, mostly, the attention given to the captured image regions by the *Enriched YCbCr network* are higher than the attention given by the *RGB network*.

In the real class of images also it can be observed that the *Enriched YCbCr network* could capture more image information than the *RGB network*. For example in the first image of the dog and the second image of the cow, the attention given to the image regions are higher for *the Enriched YCbCr network* than the *RGB network*. Also, more image regions relevant for classification are seen to be captured by the attention maps of the *Enriched YCbCr network*.

Thus for all the classes, it can be observed that the image information captured by the *Enriched YCbCr network* for this forensic task is comparatively much more relevant than the *RGB network*. This, in fact, shows the powerful nature of the proposed methodology incorporating YCbCr color space transformation and enriching the color space with the information on rate of differences between original and corresponding compressed versions of images in different categories, to capture image information that is highly relevant to the forensic task of classifying natural and generated images.

4.6 Summary

In this work a robust approach towards distinguishing natural and computer generated images including both, computer graphics and GAN generated images is proposed, unlike the works in the literature to distinguish natural and computer generated images that consider only either of the generated images category. The proposed work utilized a fusion of two vision transformers one of them operating in

RGB color space and the other in YCbCr color space. Transformer based architectures can provide good classification performances, while the proposed methodology focuses on improving the robustness of the proposed model against post processing operations such as JPEG compression, by maintaining the high model accuracy, because usually the forensic algorithms and tools even with high performance accuracies are deceived using these post processed images. Experiments are conducted to analyze the performance of the model and are compared against a set of baselines. The proposed model achieved higher accuracy than the baselines and is found to be highly robust and generalizable. Visualizing the features showed better separability capability of the proposed model than the baseline. The work also studied the attention map visualizations of the networks of the fused model and observed that the proposed methodology could capture more image information relevant to the forensic task of classifying natural and generated images. To aid the future research, the relevant materials of this study including the source code are made publicly available at https://github.com/manjaryp/MCE-ViT and https://dcs.uoc.ac.in/cida/projects/dif/mcevit.html along with the publication.

Chapter 5

Exploring Fairness in Pre-trained Visual Transformer based Natural and GAN Generated Image Detection Systems and Understanding the Impact of Image Compression in Fairness

Abstract: It is not only sufficient to construct computational models that can accurately classify or detect fake images from real images taken from a camera, but it is also important to ensure whether these computational models are fair enough or produce biased outcomes that can eventually harm certain social groups or cause serious security threats. Exploring fairness in forensic algorithms is an initial step towards correcting these biases. Since visual transformers are recently being widely used in most image classification based tasks due to their capability to produce high accuracies, this study tries to explore bias in the transformer based image forensic algorithms that classify natural and GAN generated images. By procuring a bias evaluation corpora, this study analyzes bias in gender, racial, affective, and intersectional domains using a wide set of individual and pairwise bias evaluation measures. As the robustness of the algorithms against image compression is an important factor to be considered in forensic tasks, this study also analyzes the role of image compression on model bias. Hence to study the impact of image compression on model bias, a two phase evaluation setting is followed, where a set of experiments is carried out in the uncompressed evaluation setting and the other in the compressed evaluation setting.

5.1 Introduction

A lot of studies are reported proposing various methods for distinguishing fake images from real images taken from a camera, which can help to understand or even serve as an evidence to prove image authenticity [109, 110, 140, 144]. Although there is a lot of research in this area of distinguishing fake and real images, there are only a very few studies that explore algorithmic bias in such image forensics systems [111, 113]. Exploring bias in the image forensics systems is very significant because unfair forensic systems can lead images of certain social groups to be more likely to be predicted as fake images even if they are actually real images. Unfair models may

also lead images of certain social groups to be more likely to be predicted as real images even if they are actually fake images creating security concerns. Therefore it is essential to test the fairness of image forensics systems.

Motivated by the transformer based networks that were initially designed dedicatedly for natural language based tasks, vision transformers were developed that handle images as sequences of patches [139]. Recently, visual transformer based models have drawn considerable attention due to their impressive performances for a variety of downstream tasks, for example, transformer based models for image classification (e.g. ViT [139]), object detection (e.g. DEtection TRansformer (DETR) [145], RT-DETR [146], ViT-YOLO [147]), segmentation (e.g. SegFormer [148], Segmenter [149]), image generation (e.g. Transgan [150]), etc., [151, 152]. The area of image forensics also reports many works in the literature, utilizing these visual transformers [82, 83, 84, 153, 144]. Due to the recent widespread use of visual transformers in image forensics, this study tries to explore bias, if any, in the visual transformers for the forensics task of distinguishing natural and GAN generated images.

Images shared through social media websites, unlike other post-processing operations, almost always go through compression knowingly or unknowingly [2]. Also, to deceive the forensic models detecting fake images and to spread fake news, the fake images are usually compressed and propagated through social media [3, 4]. Therefore, in the image forensic task of detecting natural and GAN generated images, the robustness of the forensic algorithms towards post-processing operations, particularly image compression, is a very important factor to be considered. Hence, studies in the literature that build high performance fake image detector systems also analyze the robustness of those models [144].

Most studies report a high accuracy drop for the models in compressed scenarios [140, 144]. In this regard, one of the interests of this study, apart from identifying bias in visual transformers based classification of natural and GAN images, is to explore whether image compression impacts model bias. To study these objectives, this study conducts bias analysis experiments in two evaluation settings, one in the original uncompressed evaluation setting and the other in the compressed setting, using the same set of evaluation measures. This helps to understand and identify any bias in the transformer based models and also to analyze whether the model bias is impacted by image compression. Figure 5.1 shows the overall workflow of the proposed work, with an example set of input images and prediction scenario to better understand the workflow and how this study conducts the bias exploration. This example only depicts the case of analyzing bias in GAN images[14], but the study considers analysis

[14]The GAN images in this example is collected from the StyleGAN2 [98] generated images

over both the real and GAN class of images.

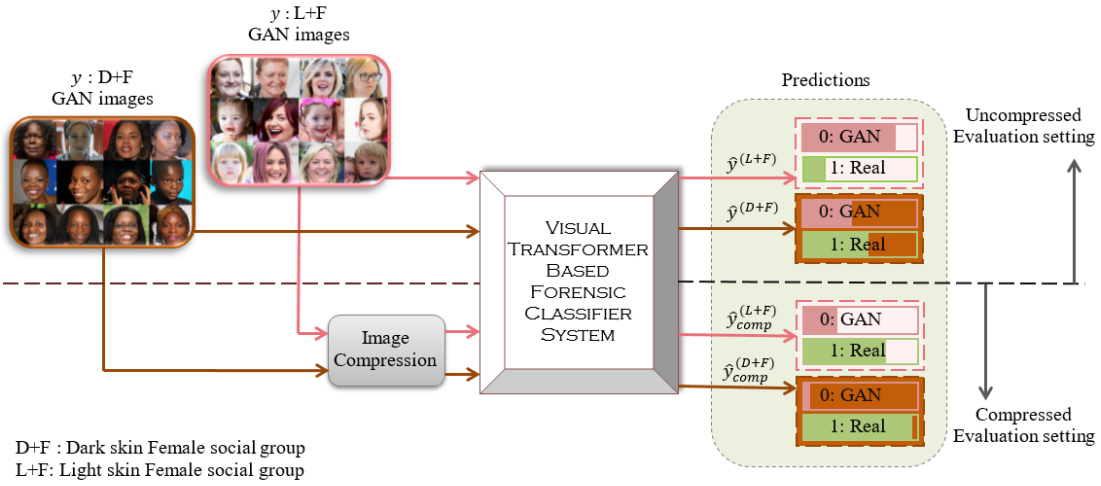


FIGURE 5.1: Overall workflow of the proposed work (the visual transformer based forensic classifier system is illustrated in figure 5.2)

5.1.1 Research Question

This chapter addresses the following research questions.

***RQ1*:** Do visual transformers produce biased outcomes for the task of distinguishing natural and GAN generated images?

***RQ2*:** Does image compression impact or amplify bias in these models classifying natural and GAN generated images?

5.1.2 Delineating the Proposed Work in the Context of Literature

In the context of the previous works in the literature that analyze bias in image forensic algorithms classifying natural and GAN generatead images [108, 111, 113, 114], the proposed work is the first work, to the best knowledge, that explores bias in transformer based image forensic models classifying natural and GAN generated images. Also, the proposed work is the first work, to the best knowledge, to study the role/impact of image compression in model biases. The work tries to unveil any existence of bias in gender, racial, affective, and even intersectional domains using a vast set of individual and pairwise evaluation measures, and sets aside the mitigation of these biases outside the scope of this work, for future studies.

5.1.3 Contributions

The major contributions of the proposed work are:

- This chapter explores bias in the transformer based forensic systems that classify natural and GAN generated images
- The chapter tries to understand the impact of image compression on model bias by analyzing and comparing the model performances across uncompressed and compressed evaluation settings
- The work procures a bias evaluation corpora to analyze bias in gender, racial, affective, and intersectional domains
- The chapter conducts extensive bias evaluations in each of the domains using individual and pairwise evaluation measures

5.1.4 Chapter Organization

The rest of the chapter is organized as section 5.2 discusses in detail the construction of transformer based models for the task of classifying natural and GAN images. Section 5.3 explains in detail the evaluation domains, evaluation corpora, and evaluation measures used for bias analysis experiments. Section 5.4 presents the results and discussions of both the uncompressed and compressed evaluation settings and finally section 5.5 presents the summary of the chapter.

5.2 Classification of Natural and GAN Generated Images

This section discusses the visual transformer based deep learning models that are investigated for fairness in this study, the dataset used to fine-tune these transformer based models, and the construction of classifier models using these visual transformers for the task of classifying natural and GAN generated images.

5.2.1 Transformer based Deep Learning Models

This work tries to identify bias in three popular transformer based deep learning models, viz. Vision Transformer (ViT) [139], Convolutional Vision Transformer (CvT) [154] and Swin transformer [155]. The ViT architecture divides the images into fixed size patches in order. These non overlapping patches are then linearly embedded. These embeddings along with the position embeddings of the patches and a learnable classification token are supplied to the transformer encoder block for classification task

[139]. CvT architecture utilizes convolutions within the ViT architecture with an aim to improve the performance of ViT. The major difference includes using a set of transformers with convolutional token embedding, convolutional projection and convolutional transformer block [154]. Swin transformer follows hierarchical architecture based on Shifted WINdow approach [155]. All these transformer based architectures are recently very popular and widely used in many of the image based tasks due to their capability to produce high classification accuracies [151].

5.2.2 Fine-tuning Corpora

To build transformer based forensic classifier systems that classify GAN and Real images, each of the pre-trained transformer based models are fine-tuned using a GAN versus Real image dataset that consists of a total of 10,000 images; each class containing 5000 images. The GAN images are collected from the StyleGAN2 image generative algorithm [98] and the Real class of images are collected from the Flickr-Faces-HQ (FFHQ) dataset [99]. The total fine-tuning corpora is split in the ratio 60:20:20 for training, validation and testing respectively.

5.2.3 Natural Image versus GAN Image Classifier Model

Natural image versus GAN image classification is formulated as a two class classification task that can classify images under evaluation into either of the two classes GAN or Real. The classifiers are fed with the training data $x_1, x_2, \ldots, x_N$ (x_i indicates i$^{\text{th}}$ image in train data) and associated ground truth classes $y_1, y_2, \ldots, y_N$ ($y \in \{\text{GAN, Real}\}$) such that to find a best fitting model $M : y = M(x)$. To build the classifier models, the three pre-trained transformer networks are fine-tuned using the task specific GAN versus Real image dataset. To build the ViT based classifier the ViT-Large network that employs a patch size of 16 and pre-trained on the ImageNet-21K [141] image dataset, is used. To build the CvT based classifier the CvT-21 network pre-trained on ImageNet-1k [120] image dataset is used. And, to build the Swin transformer based classifier the Swin-Large network that employs a patch size of 4 and window size of 7 and pre-trained on ImageNet-21K dataset is used. The size of the input image for all the three networks is 224×224. A diagrammatic representation of the visual transformer based natural image versus GAN image classifier model is shown in figure 5.2.

The fine-tuning experiments of transformers are conducted on on the deep learning workstation equipped with Intel Xeon Silver 4208 CPU at 2.10 GHz, 256 GB RAM, and two GPUs of NVIDIA Quadro RTX 5000 (16GB each), using the libraries Torch (version 1.13.1+cu116), PyTorch Lightning (version 1.9.0), Transformer (version 4.17.0),

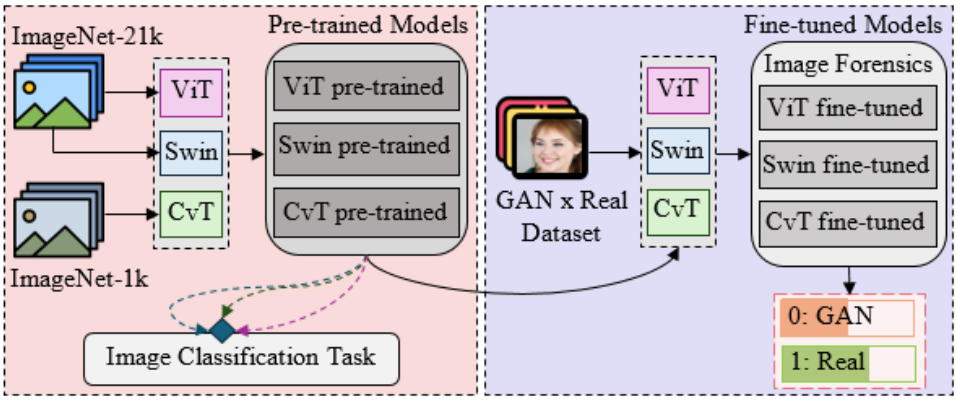


FIGURE 5.2: Visual transformer based forensic classifier system

Tensorflow (version 2.8.0), and Keras (version 2.8.0). Table 5.1 shows the model parameters. Table 5.2 shows the test accuracy of the fine-tuned transformer based models in classifying natural and GAN images.

TABLE 5.1: Model parameters

Parameter	ViT	CvT	Swin
Network	ViT-Large	CvT-21	Swin-Large
Patch size	16	-	4
Window size	-	-	7
Pre-training dataset	ImageNet-21K [141]	ImageNet-1k [120]	ImageNet-21K [141]
Input image	224×224	224×224	224×224
Learning rate	$2e-5$	$2e-5$	$2e-5$
Batch size	4	4	4
Optimizer	Adam	Adam	Adam
Epoch	25	25	25
Trainable parameters	303 M	31.2 M	194 M

TABLE 5.2: Fine-tuned model accuracies

Model	Total test accuracy	GAN accuracy	Real accuracy
ViT	91.75	94.9	88.6
CvT	99.60	99.6	99.6
Swin	99.70	99.9	99.5

5.3 Fairness Analysis in Image Forensic Classifier Systems

This study tries to identify bias (if any), in the transformer based *Natural image versus GAN generated image* classifier systems. Fairness analysis is conducted in the gender, racial, affective, and also in the intersectional domains. Gender domain based bias analysis considers the female and male social groups, the racial domain considers the dark skin people and light skin people social groups, the affective domain considers the smiling face and non smiling face groups and intersectional bias analysis considers two domains simultaneously, such as dark skin female, light skin male, etc. Apart from analyzing bias by comparing performances of each social group against the other using individual evaluation measures, this study also performs pairwise analysis of social groups. Bias analysis in this forensic task of classifying natural and GAN generated images using transformer based models is conducted using two categories of evaluation corpora, one consisting of the original uncompressed GAN and Real evaluation corpora and the other is the JPEG compressed version of the same evaluation corpora. That is, in the first phase of bias analysis, the transformer based models are evaluated over the uncompressed evaluation corpora using a set of evaluation measures, and in the second phase of analysis the same evaluation corpora is JPEG compressed with a quality factor of 90 and analyzed using the same evaluation measures. The evaluation corpora and evaluation measures are detailed below.

5.3.1 Fairness Evaluation Domains and Corpora

This work procures an evaluation corpora for bias analysis with respect to gender, racial, and affective domains. To procure the evaluation corpora the study utilizes Natural images from the FFHQ dataset [99] and GAN images from the StyleGAN2 [98] generated images. From both Natural and GAN generated images 1000 female face images and 1000 male face images each for the gender bias analysis, 1000 dark skin and 1000 light skin face images for racial bias analysis, and 1000 smiling and 1000 non smiling face images for affective bias analysis are collected. This also gives chances for intersectional bias analysis with 500 images each in the category of dark skin female, dark skin male, light skin female, and light skin male faces. A sample set of GAN images from the evaluation corpora used in this study is provided in figure 5.3 (even though the real class of images in the evaluation corpora are collected from the publicly available FFHQ dataset which is properly cited as [99], the chapter avoids portraying the images of real people for showing the examples of each social groups, and only use the sample images from the class GAN).

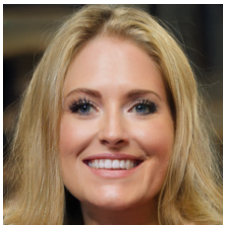
(A) Dark skin female

(B) Light skin female

(C) Non smiling face

(D) Dark skin male

(E) Light skin male



(F) Smiling face

FIGURE 5.3: A sample of GAN face images from the evaluation corpora used in this study

5.3.2 Fairness Evaluation Measures

Bias analysis in this study focuses on comparing the classification performance of the transformer based models over different social groups (or groups) within the same domain using certain evaluation measures. These analyses are performed to compare social groups within a single domain (e.g. Male vs. Female in gender domain) as well as to compare social groups within intersectional domains (e.g. Dark skin Male vs. Light skin Male). Apart from the measures that evaluate individual social groups, this study also utilizes pairwise evaluation measures to quantify bias associated with a pair of social groups in single domain or intersectional groups in a domain. The measures considering individual social groups in a domain and pairwise measures considering two social groups simultaneously are detailed below.

Individual Measures

These measures are defined by the probability of correct and incorrect classifications in a social group within a domain. Social groups over which the individual measures are evaluated include, Female (F) and Male (M) social groups in the gender domain, Dark skin (D) and Light skin (L) social groups in the racial domain, Non-smiling (Ns) and Smiling (S) groups in the affective domain, and Dark skin Female (DF), Dark skin Male (DM), Light skin Female (LF) and Light skin Male (LM) groups in the intersectional domain.

- Total Accuracy [105, 156]: This popular classification measure computes the total classification accuracy of a model over a social group in a domain. Total accuracy gives the percentage of images in a social group that is correctly classified into the natural image category and the GAN generated image category.

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{5.1}$$

  where, TP and TN are the number of true positives and true negatives, and FP and FN denote false positives and false negatives, respectively.

- GAN Class Accuracy: This measure gives the accuracy of the class GAN images, i.e., the number of GAN images correctly classified as GAN images. This measure gives the True Positive Rate (TPR) [156] of the model.

$$Acc_{gan} = \frac{TP}{TP + FN} \tag{5.2}$$

- Real Class Accuracy: This measure gives the accuracy of the class of natural images, i.e., the number of natural images correctly classified as natural images. This measure is the True Negative Rate (TNR) [156] of the model.

$$Acc_{real} = \frac{TN}{TN + FP} \tag{5.3}$$

- False Positive Rate (FPR) [105, 156]: For this classification task, FPR gives the ratio of Real images misclassified as GAN images, among the total number of Real images.

$$FPR = \frac{FP}{FP + TN} \tag{5.4}$$

- False Negative Rate (FNR) [156]: FNR gives the ratio of GAN images misclassified as Real images, among the total number of GAN images.

$$FNR = \frac{FN}{TP + FN} \tag{5.5}$$

During evaluation, the results obtained for each of these individual measures across the social groups within a domain are correspondingly compared, rather than looking for ideal high classification results.

Pairwise Measures

The pairwise evaluations are computed on a pair of social groups $g^{(a)}$ and $g^{(b)}$ within a domain. $y(g_i^{(a)})$ indicates the ground truth class of i$^{\text{th}}$ image in the social group $g^{(a)}$ (for $i \in A$), and $y(g_j^{(b)})$ indicates the ground truth class of j$^{\text{th}}$ image in the group $g^{(b)}$ (for $j \in B$), where A and B indicates total number of instances in the social groups $g^{(a)}$ and $g^{(b)}$, respectively. Also, $y_{class}(g_i^{(a)})$ and $y_{class}(g_j^{(b)})$ indicate the corresponding prediction classes, and $y_{score}(g_i^{(a)})$ and $y_{score}(g_j^{(b)})$ indicate prediction intensities (confidence scores of prediction), of $g^{(a)}$ and $g^{(b)}$, respectively. Pairwise measures are evaluated over the pairs, Female vs. Male (F $\times$ M) in gender domain, Dark skin vs. Light skin (D $\times$ L) in racial domain, Non-smiling vs. Smiling (Ns $\times$ S) in affective domain, and Dark Female vs. Dark Male (D+F $\times$ D+M), Light Female vs. Light Male (L+F $\times$ L+M), Dark Female vs. Light Female (D+F $\times$ L+F), Dark Male vs. Light Male (D+M $\times$ L+M), Dark Female vs. Light Male (D+F $\times$ L+M) and Light Female vs. Dark Male (L+F $\times$ D+M) in the intersectional domain.

- Average Confidence Score (ACS) [157]: This measure is computed using the ratio between average prediction intensities of the two social groups under evaluation.

$$ACS = 1 - \frac{\frac{1}{A}\left(\sum_{i=1}^{A} y_{score}(g_i^{(a)})\right)}{\frac{1}{B}\left(\sum_{i=1}^{B} y_{score}(g_j^{(b)})\right)} \tag{5.6}$$

  An ideal unbiased scenario gives ACS = 0 for a pair. Positive values of ACS show that the prediction intensities of the social group $g^{(a)}$ are lower than $g^{(b)}$, whereas negative ACS indicates that the prediction intensities of the social group $g^{(a)}$ are higher than $g^{(b)}$.

- Demographic Parity (DP) [157, 158]: This is one of the popular measures to quantify bias in a classification model, by analyzing similarity (or dissimilarity) in the classifications of the model for two social groups in a domain.

$$DP = \frac{P\left(y_{class}(g_i^{(a)}) = c \mid z = g^{(a)}\right)}{P\left(y_{class}(g_j^{(b)}) = c \mid z = g^{(b)}\right)} \tag{5.7}$$

  where, $P\left(y_{class}(g_i^{(a)}) = c \mid z = g^{(a)}\right)$ and $P\left(y_{class}(g_j^{(b)}) = c \mid z = g^{(b)}\right)$ are the probabilities of the groups $z \in \{g^{(a)}, g^{(b)}\}$, respectively, for being classified into a class $c \in$ (GAN, Real) where, in the $g^{(a)} \times g^{(b)}$ pair, $g^{(a)}$ is the group with higher

probability. That is, the measure DP recommends that the probability of predicting a class c needs to be similar for both the social groups $g^{(a)}$ and $g^{(b)}$ within a domain. Hence, an ideal unbiased case is indicated by DP = 1 for a pair, and lower values of DP indicate higher bias. A threshold of 0.80 is commonly used for identifying lower DP values, indicating high model bias [159].

- Equal Opportunity (EO) [158, 159]: This measure is also similar to DP, but EO considers the ground truth in addition to the predicted classes.

$$EO = \frac{P\left(y_{class}(g_i^{(a)}) = c \;\&\&\; y(g^{(a)}) = c \;\mid z = g^{(a)}\right)}{P\left(y_{class}(g_j^{(b)}) = c \;\&\&\; y(g^{(b)}) = c \mid z = g^{(b)}\right)} \tag{5.8}$$

where, $y(g^{(a)} = c)$ and $y(g^{(b)} = c)$ indicates the ground truth class c of group $y(g^{(a)})$ and $y(g^{(b)})$. Similar to DP, an ideal unbiased case is indicated by EO = 1 for a pair, and lower values of EO indicate higher bias.

5.4 Results and Analysis

Since this study follows a two-phase evaluation setting, where the bias evaluation experiments are carried out in both the uncompressed and compressed settings, this section initially analyses the results of bias evaluation of each of the transformer based models, ViT, CvT, and Swin over the original uncompressed evaluation corpora and later the results of bias evaluation of the models over the compressed evaluation corpora.

5.4.1 Fairness Analysis in the Uncompressed Evaluation setting

Vision Transformer

Bias evaluation results of the ViT based model in the uncompressed setting is shown in table 5.3. The top portion of the table presents the results of individual measures of bias analysis of ViT and the bottom portion presents the results of pairwise measures of bias analysis, within gender, race, affective, and intersectional domains.

While looking into the results of individual measures (in the top portion of table 5.3), in the gender domain, the total model accuracy (Acc) of ViT over the female group (F) is less than the male group (M) by 4.45 percentage points, indicating biased prediction. This bias is observed to be very high for class Real (Acc_{real}), i.e. accuracy of the female group is less than male by 9.3 percentage points, indicating high gender bias against the female social group. Whereas, for class GAN (Acc_{gan}), the accuracy of

TABLE 5.3: Evaluation results of **ViT** in **uncompressed** setting

Individual measures based analysis										
Metric	Gender		Race		Affective		Intersection			
	F	M	D	L	Ns	S	D+F	D+M	L+F	L+M
Acc	88.20	92.65	92.50	88.35	92.70	89.75	90.70	94.30	85.70	91.00
Acc_{gan}	93.10	92.70	91.50	94.30	93.90	95.10	89.80	93.20	96.40	92.20
Acc_{real}	83.30	92.60	93.50	82.40	91.50	84.40	91.60	95.40	75.00	89.80
FPR	0.167	0.074	0.065	0.176	0.085	0.156	0.084	0.046	0.250	0.102
FNR	0.069	0.073	0.085	0.057	0.061	0.049	0.102	0.068	0.036	0.078

Pairwise measures based analysis									
	Gender	Race	Affect	Intersection					
	F × M	D × L	Ns × S	D+F × D+M	L+F × L+M	D+F × L+F	D+M × L+M	D+F × L+M	L+F × D+M
GAN									
ACS	-0.0036	+0.0173	+0.0049	+0.0173	-0.0232	+0.0368	-0.0029	+0.0145	-0.0202
DP	0.9117	0.8758	0.9250	0.9959	0.8435	**0.7989**	0.9551	0.9590	**0.7956**
EO	0.9957	0.9703	0.9874	0.9635	0.9564	0.9315	0.9893	0.9740	0.9668
Real									
ACS	+0.0252	-0.0212	-0.0200	+0.0139	+0.0398	-0.0368	-0.0095	+0.0045	+0.0489
DP	0.9029	0.8637	0.9150	0.9961	0.8053	**0.7721**	0.9550	0.9588	**0.7691**
EO	0.8996	0.8813	0.9224	0.9602	0.8352	0.8188	0.9413	0.9804	0.7862

the male group is less than female only by a very small value of 0.4 percentage points, a negligible difference to indicate any bias. Also, in the gender domain, the measure FPR is higher for the female group than male. This indicates Real images of females are more likely to be misclassified as GAN generated images than those of males (an observation similar to the one reported in [108]). Whereas, the very low difference in FNR values between male and female groups indicates negligible chances that GAN images of males get misclassified as Real images.

In the racial domain, the total accuracy of ViT over the light skin (L) group is less than dark skin (D) by 4.15 percentage points, indicating biased prediction against light skin people, which is much more evident in the case of class Real with a difference of 11.1 percentage points, indicating high racial bias against light skin people. Whereas, in the case of class GAN, the accuracy of the dark skin group is less than light skin by 2.8 percentage points, indicating bias against dark skin. The measure FPR shows a

higher value for light skin group than dark skin, which indicates Real images of light skin people are more likely to be misclassified as GAN images, and FNR indicates slight chances for GAN images of dark skin people being misclassified as Real images.

In the affective domain, the total accuracy of ViT over the group of smiling faces (S) is less than non-smiling faces (Ns) by 2.95 percentage points. A similar pattern is shown in class Real, with a difference of 7.1 percentage points, indicating affective bias against the group with smiling faces. Whereas in the case of class GAN, the accuracy of smiling faces is higher than non-smiling faces by 1.2 percentage points. FPR shows high value for smiling faces, which indicates Real images of smiling people are more likely to be misclassified as GAN images. The slightly higher values of FNR for non-smiling faces indicate slight chances for GAN images of non-smiling faces being misclassified as Real images.

In the intersectional domain, it can be observed that the total accuracy varies across different intersectional groups. The highest total accuracy is observed for dark skin male group (D+M), and lowest for light skin female (L+F), with a difference of 8.6 percentage points, which indicates bias against light skin female group. Whereas for class GAN, an accuracy of 96.4 percent is obtained for light skin female group, which is the highest accuracy obtained across various groups among both classes and even compared to the total accuracy. The lowest accuracy in class GAN is for the dark skin female group (D+F), a difference of 6.6 percentage points compared to the highest accuracy group, indicating biased prediction. In class Real, the highest accuracy is obtained for dark skin male group and the lowest accuracy of 75.0 percent is obtained for light skin female group, which is the lowest accuracy obtained across various groups among both classes and even compared to the total accuracy. That is, both these groups have a very high difference of 20.4 percentage points, indicating very large intersectional bias against light skin female. FPR stands highest for the light skin female group indicating *Real images of light skin females have a very high probability of being misclassified as GAN images*. FNR is highest for the dark skin female group indicating *GAN images of dark skin females have a very high probability of being misclassified as Real images*.

The bottom portion of the same table 5.3 presents the results of pairwise measures of bias analysis of ViT for both GAN and Real classes. In the gender domain, for class GAN, the negative value of the measure ACS for the Female vs. Male pair (F$\times$M) shows that the prediction intensities of the female group are higher than males. The measure DP has a low value, but since it is not less than the threshold of 0.80 this measure does not report bias in the Female vs. Male pair. The measure EO has a high value and does not report gender bias in class GAN predictions. For class Real,

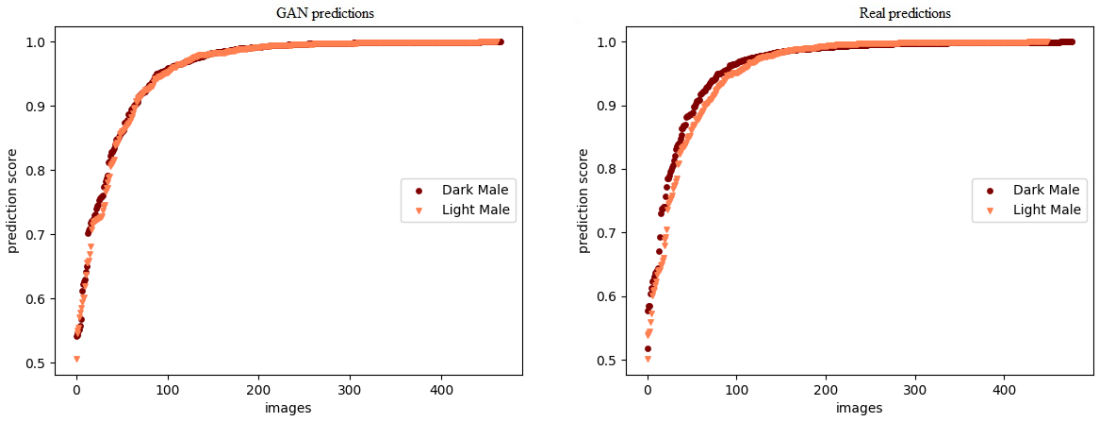
positive ACS for Female vs. Male pair shows that the prediction intensities of the male group are higher than females. The measures DP and EO have low values. But since DP is not lower than the threshold 0.80, it does not report gender bias in class Real predictions.

In the racial domain, for class GAN, the positive ACS value for Dark skin vs. Light skin pair (D×L) shows that the prediction intensities of the light skin group are higher than dark skin. The measure DP has a low value, but since it is not less than the threshold of 0.80 this measure does not report racial bias. EO has a high value and does not report racial bias in the class GAN predictions. For the class Real, negative ACS for the pair shows that the prediction intensities of the dark skin group are higher than light skin. The measures DP and EO have low values, where DP is not lower than the threshold of 0.80 and hence do not report racial bias in the class Real predictions.

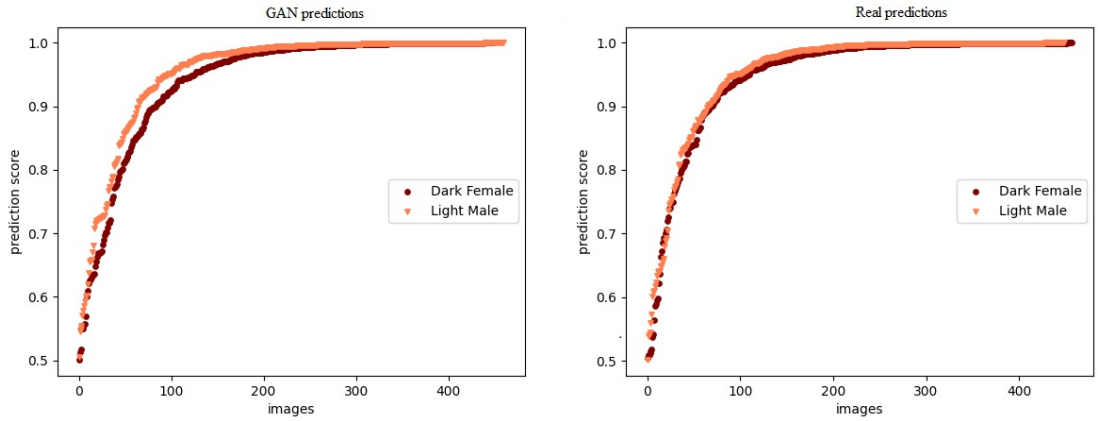
In the affective domain, for class GAN, the positive ACS value for the Non-smiling vs. Smiling pair (Ns×S) shows that the prediction intensities of the smiling group are higher than the non-smiling group. The measure DP has a low value, but since it is not less than the threshold of 0.80 this measure does not report bias in this pair. EO has a high value and does not report bias in the class GAN predictions of this pair. For the class Real, negative ACS for the pair shows that the prediction intensities of the non-smiling group are higher than the smiling group. The measures DP and EO have low values. But since DP is not lower than the threshold of 0.80, this measure do not report bias in Real predictions of this pair.

In the intersectional domain, for the class GAN, the measure DP is very low for the pairs involving light skin female group, i.e., {Light skin Female vs. Light skin Male} (L+F × L+M), {Dark skin Female vs. Light skin Female} (D+F × L+F) and {Light skin Female vs. Dark skin Male} (L+F × D+M). Particularly for the pairs {Dark skin Female vs. Light skin Female} and {Light skin Female vs. Dark skin Male}, the measure DP is less than the threshold of 0.80, which indicates the existence of intersectional bias. Similarly in class Real also, pairs involving the light skin female group show bias with very low values for DP and even EO. That is, pairwise bias analysis measures could unveil existence of bias in the {Light skin Female vs. Light skin Male}, {Dark skin Female vs. Light skin Female} and {Light skin Female vs. Dark skin Male} intersectional pairs. The prediction intensity (confidence) plots of the images of intersectional pairs in the bias evaluation corpora that shows unbiased and biased results are shown in figs. 5.4 and 5.5. The horizontal axes of the plots indicate images in the bias evaluation corpora (that contains a total of 500 GAN/Real images) and the vertical axis indicates the prediction intensity score of each of these images. The figs. 5.4a and 5.4b are the intensity predictions of the unbiassed intersectional pairs {Dark skin Male vs. Light

skin Male} and {Light skin Female vs. Light skin Male}, and it can be observed that there is not much difference in prediction intensities within these pairs, for both the classes, GAN and Real. Whereas, in figs. 5.5a and 5.5b, the comparatively much more difference in prediction intensities within the pairs for both GAN and Real classes, subsidize the quantitative results in table 5.3 that indicates the existence of bias in the intersectional pair {Dark skin Female vs. Light skin Female} and {Light skin Female vs. Dark skin Male}

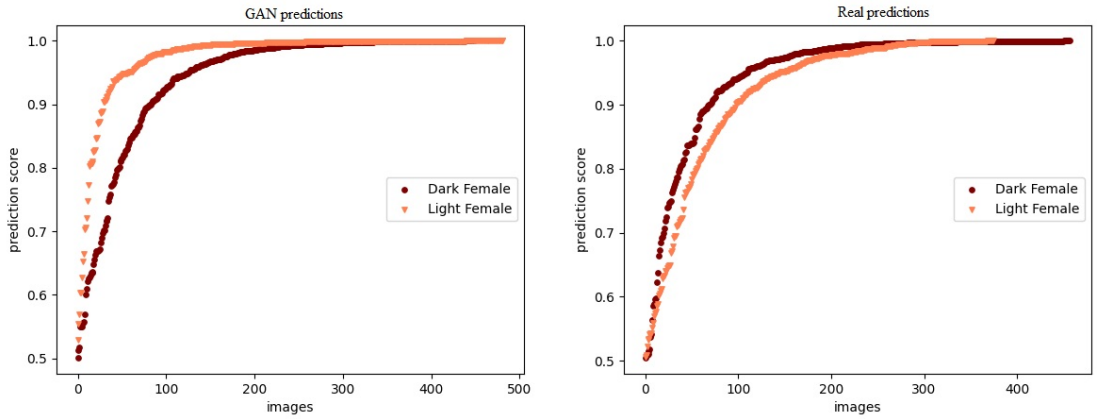


(A) Dark skin Male vs. Light skin Male (D+M × L+M)

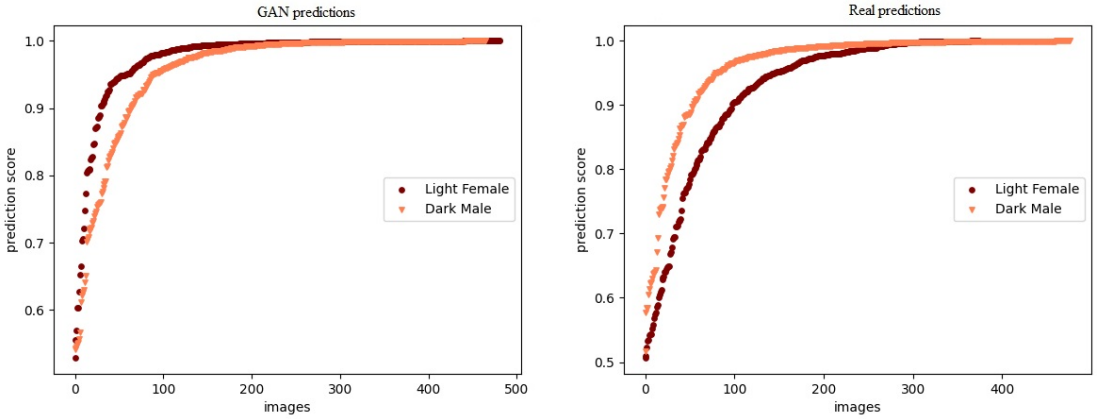


(B) Dark skin Female vs. Light skin Male (D+F × L+M)

FIGURE 5.4: ViT prediction intensity plots of a sample set of **unbiased** intersectional pairs in the bias evaluation corpora



(A) Dark skin Female vs. Light skin Female (D+F × L+F)



(B) Light skin Female vs. Dark skin Male (L+F × D+M)

FIGURE 5.5: ViT prediction intensity plots of a sample set of **biased** intersectional pairs in the bias evaluation corpora

Convolutional Vision Transformer

The bias evaluation results of the transformer based model CvT for various domains is shown in table 5.4. From the top portion of the table showing the results of individual measures, it can be observed that the model shows high and similar accuracies for all categories of social groups within each of the domains. The FPR and FNR values are also very low and similar across the social groups within each domain. The bottom portion of the table presents the results of pairwise analysis of CvT for various domains. The measures DP and EO also report very high values, nearly similar to an ideal unbiased scenario. Altogether, the individual and pairwise measures do not show the existence of significant bias in the CvT based transformer model.

TABLE 5.4: Evaluation results of **CvT** in **uncompressed** setting

Individual measures based analysis										
Metric	Gender		Race		Affective		Intersection			
	F	M	D	L	Ns	S	D+F	D+M	L+F	L+M
Acc	99.05	99.15	98.80	99.40	99.45	99.40	98.80	98.80	99.30	99.50
Acc_{gan}	98.40	98.60	98.10	98.90	99.10	99.20	98.00	98.20	98.80	99.00
Acc_{real}	99.70	99.70	99.50	99.90	99.80	99.60	99.60	99.40	99.80	100.0
FPR	0.003	0.003	0.005	0.001	0.002	0.004	0.004	0.006	0.002	0.000
FNR	0.016	0.014	0.019	0.011	0.009	0.008	0.020	0.018	0.012	0.010

Pairwise measures based analysis									
	Gender	Race	Affect	Intersection					
	F × M	D × L	Ns × S	D+F × D+M	L+F × L+M	D+F × L+F	D+M × L+M	D+F × L+M	L+F × D+M
GAN									
ACS	+0.0050	+0.0151	-0.0010	+0.0064	+0.0036	+0.0165	+0.0138	+0.0200	-0.0103
DP	0.9980	0.9960	0.9970	0.9960	1.0000	0.9939	0.9980	0.9939	0.9980
EO	0.9980	0.9919	0.9990	0.9980	0.9980	0.9919	0.9919	0.9899	0.9939
Real									
ACS	-0.0005	+0.0063	+0.0012	-0.0004	-0.0007	+0.0064	+0.0062	+0.0058	-0.0068
DP	0.9980	0.9961	0.9970	0.9961	1.0000	0.9941	0.9980	0.9941	0.9980
EO	1.0000	0.9960	0.9980	0.9980	0.9980	0.9980	0.9940	0.9960	0.9960

Swin Transformer

The bias evaluation results of the Swin transformer based model for various domains is shown in table 5.5. Similar to the previous model, it can be observed from the top portion of the table with individual measures that, the Swin transformer shows high and similar accuracies for all categories of social groups. The FPR and FNR values are also very low and similar across the social groups within each domain. Bottom portion of the table presents the results of pairwise analysis of the Swin transformer for various domains. The measures DP and EO also shows very high values, nearly similar to an ideal unbiased scenario. Altogether, the individual and pairwise measures do not show any existence of significant bias in the Swin transformer model.

TABLE 5.5: Evaluation results of **Swin** transformer in **uncompressed** setting

Individual measures based analysis										
Metric	Gender		Race		Affective		Intersection			
	F	M	D	L	Ns	S	D+F	D+M	L+F	L+M
Acc	99.30	99.95	99.70	99.55	99.80	99.40	99.40	100.0	99.20	99.90
Acc_{gan}	99.40	100.0	99.60	99.80	99.80	99.60	99.20	100.0	99.60	100.0
Acc_{real}	99.20	99.90	99.80	99.30	99.80	99.20	99.60	100.0	98.80	99.80
FPR	0.008	0.001	0.002	0.007	0.002	0.008	0.004	0.000	0.012	0.002
FNR	0.006	0.000	0.004	0.002	0.002	0.004	0.008	0.000	0.004	0.000

Pairwise measures based analysis									
	Gender	Race	Affect	Intersection					
	F × M	D × L	Ns × S	D+F × D+M	L+F × L+M	D+F × L+F	D+M × L+M	D+F × L+M	L+F × D+M
GAN									
ACS	+0.0026	+0.0034	-0.0002	+0.0059	-0.0007	+0.0067	+0.0002	+0.0060	-0.0009
DP	0.9990	0.9930	0.9960	0.9960	0.9940	0.9881	0.9980	0.9940	0.9921
EO	0.9940	0.9980	0.9980	0.9920	0.9960	0.9960	1.0000	0.9920	0.9960
Real									
ACS	-0.0014	+0.0010	+0.0008	-0.0006	-0.0022	+0.0018	+0.0002	-0.0004	-0.0024
DP	0.9990	0.9930	0.9960	0.9960	0.9940	0.9880	0.9980	0.9940	0.9920
EO	0.9930	0.9950	0.9940	0.9960	0.9900	0.9920	0.9980	0.9980	0.9980

5.4.2 Fairness Analysis in the Compressed Evaluation Setting

Vision Transformer

The results of individual and pairwise measures of bias analysis of the transformer based model ViT for various domains on the JPEG compressed evaluation corpora is shown in table 5.6. Similar to the observations discussed in the study [144], it can be observed that for compressed data the accuracies of the models decreases, and this decrease in accuracy is much higher for the class GAN than the class Real. In this compressed evaluation setting of ViT, it can also be observed that the difference in GAN accuracies (Acc_{gan}) between the groups within each of the domains has increased than its previous uncompressed evaluation setting. For example, in the previous uncompressed evaluation setting of ViT, the difference in GAN accuracies between female

TABLE 5.6: Evaluation results of **ViT** in **compressed** setting

Individual measures based analysis										
Metric	Gender		Race		Affective		Intersection			
	F	M	D	L	Ns	S	D+F	D+M	L+F	L+M
Acc	85.90	89.10	89.15	85.85	89.40	88.40	87.30	91.00	84.50	87.20
Acc_{gan}	86.10	84.60	82.80	87.90	86.00	89.20	80.40	85.20	91.80	84.00
Acc_{real}	85.70	93.60	95.50	83.80	92.80	87.60	94.20	96.80	77.20	90.40
FPR	0.143	0.064	0.045	0.162	0.072	0.124	0.058	0.032	0.228	0.096
FNR	0.139	0.154	0.172	0.121	0.140	0.108	0.196	0.148	0.082	0.160

Pairwise measures based analysis									
	Gender	Race	Affect	Intersection					
	F × M	D × L	Ns × S	D+F × D+M	L+F × L+M	D+F × L+F	D+M × L+M	D+F × L+M	L+F × D+M
GAN									
ACS	-0.0068	+0.0157	+0.0066	+0.0212	-0.0323	+0.0411	-0.0113	+0.0101	-0.0208
DP	0.9064	0.8386	0.9173	0.9751	0.8168	**0.7522**	0.9444	0.9209	**0.7714**
EO	0.9826	0.9420	0.9641	0.9437	0.9150	0.8758	0.9859	0.9571	0.9281
Real									
ACS	+0.0165	-0.0121	-0.0168	+0.0129	+0.0216	-0.0174	-0.0085	+0.0045	0.0298
DP	0.9138	0.8509	0.9214	0.9807	**0.8026**	**0.7504**	0.9534	0.9350	**0.7652**
EO	0.9156	0.8775	0.944	0.9731	0.8540	0.8195	0.9339	0.9597	0.7975

and male groups within the gender domain is 0.4 percentage points (in table 5.3); But, in this compressed setting, this difference has increased to 1.5 percentage points. Similarly, 2.8 percentage points of difference in racial domain between dark skin and light skin groups in the previous uncompressed evaluation setting have increased to 5.1 percentage points in this compressed setting, and 1.2 percentage points of difference between non-smiling and smiling groups of affective domain have increased to 3.2 percentage points. Similar to the previous uncompressed setting, here also in the class GAN of intersectional domain, light skin females obtains the highest accuracy and dark skin female group obtains the lowest accuracy, but the difference in accuracies (Acc_{gan}) between these groups increases to 11.4 percentage points compared to the difference of 6.6 percentage points in the previous uncompressed setting. Altogether, the individual measures show that in the class GAN, the gender bias against

the male group (lower accuracy for male than female), racial bias against dark skin, affective bias against non-smiling group, and intersectional biases, particularly against dark skin female, has increased in the compressed evaluation setting the the previous uncompressed evaluation setting. Also, similar to the uncompressed setting, this compressed setting have higher values of FPR for female group in gender domain, light skin in racial domain, smiling face group in affective doman and light skin females in the intersectional domain indicating that Real images of these groups are highly likely to be misclassified as GAN images.

Bottom portion of the table 5.6 presents the results of pairwise measures of bias analysis of the transformer based model ViT for various domains on the JPEG compressed evaluation corpora. The tabulated results of pairwise analysis show that, in this compressed setting, for class GAN there is a decrease in DP and EO values when compared to its previous uncompressed evaluation setting. For example, the DP of {Dark skin vs. Light skin} for class GAN has decreased from 0.8758 (in previous uncompressed evaluation setting, table 5.3) to 0.8386 (in current compressed evaluation setting, table 5.6), DP of {Dark skin Female vs Light skin Female} has decreased from 0.7989 to 0.7522, etc. Thus, the pairwise evaluations on ViT also shows that, for class GAN, the biases increases over the compressed images than the uncompressed images. That is, this indicates biases in the class GAN gets amplified with compression.

Convolutional Vision Transformer

The results of individual and pairwise measures of bias analysis of the transformer based model CvT for various domains on the JPEG compressed evaluation corpora is shown in table 5.7. The top portion of the table shows results of individual measures. Similar to ViT, compression decreases the accuracies of the CvT model, particularly the class GAN accuracy (Acc_{gan}), whereas class Real (Acc_{real}) maintains its high accuracy. But compared to the model ViT, the drop in the accuracies for class GAN of the CvT model is massively very high. Also, this accuracy decay in CvT is not similar across different social groups within a domain, indicating high bias.

Bottom portion of the table presents the results of pairwise analysis of CvT for various domains on the JPEG compressed evaluation corpora. From the table, it can be understood that, for the class GAN of the CvT model, the ideal unbiased scenario which was seen in the previous uncompressed evaluation setting of CvT (in table 5.4) has been completely overturned to a very largely biased scenario due to compression. This is because the drop in GAN accuracies are not similar for various groups within a domain (except for the dark skin female vs. light skin male pairs). Whereas, it can

TABLE 5.7: Evaluation results of **CvT** in **compressed** setting

Individual measures based analysis										
Metric	Gender		Race		Affective		Intersection			
	F	M	D	L	Ns	S	D+F	D+M	L+F	L+M
Acc	51.65	51.15	51.05	51.75	51.70	52.05	50.80	51.30	52.50	51.00
Acc_{gan}	34.00	23.00	22.00	35.00	34.00	42.00	1.80	2.60	5.00	2.00
Acc_{real}	99.90	100.0	99.90	100.0	100.0	99.90	99.80	100.0	100.0	100.0
FPR	0.001	0.000	0.001	0.000	0.000	0.001	0.002	0.000	0.000	0.000
FNR	0.966	0.977	0.978	0.965	0.966	0.958	0.982	0.974	0.950	0.980

Pairwise measures based analysis									
	Gender	Race	Affect	Intersection					
	F × M	D × L	Ns × S	D+F × D+M	L+F × L+M	D+F × L+F	D+M × L+M	D+F × L+M	L+F × D+M
GAN									
ACS	-0.0267	+0.0591	-0.0331	-0.0886	+0.0432	-0.0012	+0.1200	+0.0420	-0.0872
DP	**0.6571**	**0.6571**	**0.7907**	**0.7692**	**0.3999**	**0.3999**	**0.7692**	1.0000	**0.5199**
EO	0.6765	0.6280	0.8095	0.6923	0.3999	0.3599	0.7692	0.8999	0.5199
Real									
ACS	+0.0027	-0.0021	-0.0027	+0.0024	+0.0029	-0.0024	-0.0019	+0.0005	0.0048
DP	0.9939	0.9939	0.9954	0.9970	0.9849	0.9849	0.9970	1.0000	0.9878
EO	0.9990	0.9990	0.9990	0.9980	1.0000	0.9980	1.0000	0.9980	1.0000

be observed that the class Real of the CvT still maintains the ideal unbiased scenario as in the previous uncompressed evaluation setting.

Swin Transformer

The results of individual and pairwise measures of bias analysis of the Swin transformer based model on the JPEG compressed evaluation corpora is shown in table 5.8. The top portion of the table shows the results of individual measures. In this model also the GAN accuracy (Acc_{gan}) decreases due to compression, thereby decreasing the total model accuracy. Contrary to the previous uncompressed setting of Swin transformer where similar and high accuracies are obtained for all the social groups within each domain, this compressed evaluation setting has eventually brought up differences in GAN accuracies across social groups within each of the domains. That is, the GAN accuracy of the male group is less than the female group by 2.5 percentage

points in the gender domain, the dark skin group is less than the light skin group by 4.9 percentage points in the racial domain, and the non-smiling group is less than smiling group by 3.2 percentage points in the affective domain. In the intersectional domain, the highest GAN accuracy is obtained for the light skin female group and lowest for the dark skin female group, a very high accuracy difference of 16.4 percentage points is observed between these two intersectional groups for class GAN. Thus these accuracy differences, indicate high bias in the compressed setting for the class GAN of Swin transformer.

TABLE 5.8: Evaluation results of **Swin** transformer in **compressed** setting

Individual measures based analysis										
Metric	Gender		Race		Affective		Intersection			
	F	M	D	L	Ns	S	D+F	D+M	L+F	L+M
Acc	83.10	82.45	81.95	83.60	84.39	85.70	79.60	84.30	86.60	80.60
Acc_{gan}	68.20	65.70	64.50	69.40	69.60	72.80	60.00	69.00	76.40	62.40
Acc_{real}	98.00	99.20	99.40	97.80	99.20	98.60	99.20	99.60	96.80	98.80
FPR	0.020	0.008	0.006	0.022	0.008	0.014	0.008	0.004	0.032	0.012
FNR	0.318	0.343	0.355	0.306	0.304	0.272	0.40	0.310	0.236	0.376

Pairwise measures based analysis									
	Gender	Race	Affect	Intersection					
	F × M	D × L	Ns × S	D+F × D+M	L+F × L+M	D+F × L+F	D+M × L+M	D+F × L+M	L+F × D+M
GAN									
ACS	-0.0008	-0.0028	-0.0020	+0.0242	-0.0254	+0.0213	-0.0284	-0.0035	0.0029
DP	0.9473	0.9092	0.9488	0.8761	**0.7989**	**0.7638**	0.9164	0.9559	0.8718
EO	0.9633	0.9294	0.9560	0.8695	0.8167	0.7853	0.9043	0.9615	0.9031
Real									
ACS	+0.0025	-0.0057	-0.0016	-0.0006	+0.0057	-0.0089	-0.0026	-0.0032	0.0083
DP	0.9723	0.9518	0.9707	0.9382	0.8826	0.8649	0.9574	0.9798	0.9218
EO	0.9879	0.9839	0.9940	0.9959	0.9797	0.9758	0.9919	0.9959	0.9718

Bottom portion of the table presents the results of pairwise analysis of the Swin transformer on the JPEG compressed evaluation corpora. Compared to the previous uncompressed setting of Swin transformer (in table 5.5) that reports nearly an ideal unbiased scenario, in this compressed setting the DP and EO measures decrease highly

for the class GAN indicating an increase in bias in the class GAN. A very high bias for class GAN can be observed particularly in the pairs, light skin female vs. light skin male and dark skin female vs. light skin female.

5.4.3 Discussion

The bias evaluation results shows that in uncompressed evaluation settings, the evaluation corpora and measures could identify bias in the ViT based forensic classifier model, such as, bias in pairs involving light skin female groups e.g., bias in light skin female vs. dark skin male, dark skin female vs. light skin female, etc. Also, bias analysis in ViT based model shows other interesting inferences such as, Real images of light skin females have a very high probability of being misclassified as GAN images, and GAN images of dark skin females have a very high probability of being misclassified as Real images. However, the uncompressed setting could not identify any bias in the CvT and the Swin transformer based models.

The compressed evaluation setting, on the other hand, identifies high bias in all three transformer based models, particularly in the class GAN predictions. In the compressed evaluation setting, for the class GAN predictions of all the models, gender bias against the male group, racial bias against dark skin group, affective bias against non-smiling group, and intersectional biases, particularly against dark skin female group, has increased when compared to the uncompressed evaluation setting. Bias is identified in all the domains for the CvT based model, in the compressed evaluation setting. That is, the study could observe that model bias is impacted by image compression. Moreover, the model bias identified in the uncompressed setting is observed to be amplified in the compressed setting, particularly for the class GAN predictions. Also, given that the results indicate Real images of certain social groups such as light skin females are more likely to be misclassified as GAN images, and GAN images of dark skin females are more likely to be misclassified as Real images, etc., the images of these social groups when compressed can even more increase the risk of security threats. Therefore, image forensics works that utilize visual transformers for the task of distinguishing natural and GAN generated images, besides assessing the robustness of the algorithms towards image compression, should also study the existence of bias in these algorithms, even in the compressed setting.

ViT and Swin transformer based models chosen for this study are pre-trained on the ImageNet-21K dataset [141]. As already stated above, more than the uncompressed evaluation settings, these models show a higher bias in their corresponding compressed evaluation settings. On the other hand, the model CvT is pre-trained on

the ImageNet-1k dataset [120]. But unlike ViT and Swin transformer, CvT has comparatively a very high transition from an ideal unbiased scenario in the uncompressed evaluation setting to a very largely biased model in the compressed evaluation setting. Hence, pre-training corpora of the visual transformers might be one of the factors inducing bias in these models.

5.5 Summary

This study explored bias in the visual transformer based image forensic algorithms that classify natural and GAN generated images. The study utilized three visual transformers viz., ViT, CvT and Swin, for constructing image forensic algorithms to classify natural and GAN images. The pre-trained visual transformers are fine-tuned using the task-specific natural image versus GAN image dataset, and are examined for any existence of bias in the gender, racial, affective, and even intersectional domains. Hence, a bias evaluation corpora consisting of social groups belonging to the evaluation domains are procured for the study. Individual and pairwise bias evaluation measures are used for identifying any existence of bias in these transformer based forensic models. Since, robustness towards image compression is significant for the forensic algorithms, this study also examines the role of image compression on model bias. To the best knowledge, this is the first work to study the impact of image compression on model bias, particularly focusing on the task of classifying natural and GAN images. Hence, this work conducts the bias evaluation experiments in two separate settings; one set of experiments on the original uncompressed evaluation corpora and the other on the compressed version of the same evaluation corpora, where both these experiments rely on same evaluation measures.

This study helped to identify the existence of bias in the transformer based models for the task of distinguishing natural and GAN generated images. The two-phase bias evaluation strategy helped to identity bias in the uncompressed and compressed scenarios and also to study the impact of image compression on the model bias. The study observed that image compression impacts model biases, and particularly compression amplifies the biases of the class GAN predictions. To help towards the future research, all relevant materials of this study including the source codes will be made publicly available at `https://github.com/manjaryp/ImageForgeryFairness` and `https://dcs.uoc.ac.in/cida/projects/dif/Imageforgeryfairness.html` along with the publication.

Chapter 6

Conclusion

Abstract: This chapter provides a conclusion to the Thesis. The chapter summarizes the Thesis contributions and also discusses the future scope and research directions.

6.1 Summary of the Thesis

This Thesis presented a computational study in the direction of digital image forensics, i.e., towards distinguishing natural and computer generated images, and also explored the fairness in digital image forensics systems classifying natural and GAN generated images. The major contributions of this Thesis are summarized below.

Multi-Colorspace fused EfficientNet The contribution of the Thesis, *MC-EffNet*, proposed in chapter 3 is a deep learning based model, Multi-Colorspace fused EfficientNet model, to classify natural images and photo-realistic computer generated images, including both computer graphics and GAN images, unlike the state-of-the-art works that usually deal with either *natural images versus computer graphics* or *natural images versus GAN images* problem at a time. The results demonstrated that the proposed model outperformed the state-of-the-art baselines. Psychophysics experiments conducted to learn how capable humans are in classifying natural images and photo-realistic computer generated images showed that manual classification accuracies were lower than the proposed model accuracies, particularly in classifying the photo-realistic computer generated images. This indicated the necessity and usefulness of the proposed computational model for the task. The behavior of the proposed model was analyzed by visualizing salient regions in the images that are responsible for classification decisions. Similarities were observed when the explanations of the proposed model were compared with manual explanations labeled by human participants, indicating that the proposed model takes decisions meaningfully.

Multi-Colorspace fused and Enriched Vision Transformer The contribution of the Thesis, *MCE-ViT*, proposed in chapter 4 is a robust approach towards distinguishing natural and computer generated images. The proposed work employed a combination of two vision transformers, where each of the transformer based networks utilized a different color space transformation. The proposed methodology focused on increasing the classification performance, as well as improving the robustness against post-processing operations such as JPEG compression, because the forensic algorithms are usually fooled using these post-processed images. The performance of the proposed model when compared against a set of baselines achieved higher accuracy, outperforming the baselines and is found to be highly robust and generalizable. Better separability is observed than the baseline when the features of the proposed model are visualized. The attention map visualizations of the networks of the fused model when analyzed, it was observed that the proposed methodology could capture more image information relevant to the forensic task of classifying natural and generated images.

Exploring Fairness in Natural and GAN Generated Image Detection Systems The contribution of the Thesis proposed in chapter 5 investigated any existence of gender, racial, affective, and even intersectional biases in image forensic algorithms that classify natural and GAN generated images using visual transformers. The study procured an evaluation corpora consisting of social groups belonging to different domains, and bias evaluations are performed using individual and pairwise evaluation measures. Two sets of evaluation experiments are conducted to study the role of image compression on model bias; One set of experiments on the original uncompressed evaluation corpora and the other set on the compressed version of the same evaluation corpora. The study unveils the existence of bias in the transformer based models classifying natural and GAN generated images. The results of the study also show that image compression influences model biases.

To help towards future research, all relevant materials of each contribution in this Thesis, including the source codes, datasets used in each of the study, etc., are offered publicly at:

- Multi-Colorspace fused EfficientNet (*MC-EffNet*): https://github.com/manjaryp/GANvsGraphicsvsReal and https://dcs.uoc.ac.in/cida/projects/dif/mceffnet.html
- Multi-Colorspace fused and Enriched Vision Transformer (*MCE-ViT*): https://github.com/manjaryp/MCE-ViT and https://dcs.uoc.ac.in/cida/projects/dif/mcevit.html
- Exploring Fairness in Natural and GAN Generated Image Detection Systems: https://github.com/manjaryp/ImageForgeryFairness and https://dcs.uoc.ac.in/cida/projects/dif/Imageforgeryfairness.html

6.2 Future Research Directions

In the future, the contributions of this Thesis towards distinguishing natural and computer generated images can be expanded for identifying other forensic attacks like the recaptured images. Apart from images generated by GANs, a lot of diffusion models are also recently gaining popularity for image synthesis. Studies in literature have reported that the images generated by GAN algorithms and diffusion models have differences in their characteristics [160]. Therefore, in the future, the work can be extended by also including images generated by diffusion models. Future studies can also consider improving the applicability of these models in classifying natural and computer generated videos, by incorporating multi-model approaches.

Due to the lack of computer graphics evaluation corpora, the contribution of this Thesis has only considered exploration of bias in GAN generated category of computer generated images. In the future, the curation of a computer graphics evaluation corpora can help towards identifying bias in forensic systems classifying natural versus computer generated images including both computer graphics and GAN images. This work can be expanded to analyze the fairness of forensic models that can also detect images generated by diffusion models. There are plans to extend this work to analyze various factors that cause or originate these biases. Although it is cumbersome to procure datasets with balanced groups within a wide variety of domains, this could, in the future, help in determining various sources of biases, especially in identifying the existence of any pre-train and fine-tune data biases. The evaluation corpora can also be expanded and annotated to explore bias in many other domains

such as age, occupation, religion, etc. Another important direction is to identify bias from video forensics systems detecting computer generated videos and, recaptured and tampered image/video forensics systems. Also, there is a large scope for the mitigation of these biases from the models to develop fair forensic systems that one can trust when deployed in the real world.

Thus concluding this Thesis, expecting the Thesis contributions will be worthwhile to the digital image forensics research community.

Bibliography

[1] Eric Tokuda, Helio Pedrini, and Anderson Rocha. "Computer generated images vs. digital photographs: A synergetic feature and classifier combination approach". In: *Journal of Visual Communication and Image Representation* 24.8 (2013), pages 1276–1292. DOI: https://doi.org/10.1016/j.jvcir.2013.08.009.

[2] Tatsuya Chuman, Kenta Iida, and Hitoshi Kiya. "Image manipulation on social media for encryption-then-compression systems". In: *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. 2017, pages 858–863. DOI: 10.1109/APSIPA.2017.8282153. URL: https://ieeexplore.ieee.org/document/8282153.

[3] Jianyuan Wu and Wei Sun. "Towards multi-operation image anti-forensics with generative adversarial networks". In: *Computers & Security* 100 (2021), page 102083. ISSN: 0167-4048. DOI: https://doi.org/10.1016/j.cose.2020.102083. URL: https://www.sciencedirect.com/science/article/pii/S0167404820303564.

[4] K Anoop, P Gangan Manjary, P Deepak, and VL Lajish. "Leveraging heterogeneous data for fake news detection". In: *Linking and Mining Heterogeneous and Multi-view Data*. Cham: Springer International Publishing, 2019, pages 229–264. ISBN: 978-3-030-01872-6. DOI: 10.1007/978-3-030-01872-6_10.

[5] Stamatis Karnouskos. "Artificial Intelligence in Digital Media: The Era of Deepfakes". In: *IEEE Transactions on Technology and Society* 1.3 (2020), pages 138–147. ISSN: 2637-6415. DOI: 10.1109/TTS.2020.3001312.

[6] Jan Kietzmann, Linda W. Lee, Ian P. McCarthy, and Tim C. Kietzmann. "Deepfakes: Trick or treat?" In: *Business Horizons* 63.2 (2020). ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING, pages 135–146. ISSN: 0007-6813. DOI: https://doi.org/10.1016/j.bushor.2019.11.006. URL: https://www.sciencedirect.com/science/article/pii/S0007681319301600.

[7] Victor Schetinger, Manuel M Oliveira, Roberto da Silva, and Tiago J Carvalho. "Humans are easily fooled by digital images". In: *Computers & Graphics* 68 (2017), pages 142–151. DOI: https://doi.org/10.1016/j.cag.2017.08.010.

[8] Weiqi Luo, Zhenhua Qu, Feng Pan, and Jiwu Huang. "A survey of passive technology for digital image forensics". In: *Frontiers of Computer Science in China* 1.2 (2007), pages 166–179. DOI: 10.1007/s11704-007-0017-0.

[9] Guojuan Zhou and Dianji Lv. "An Overview of Digital Watermarking in Image Forensics". In: *2011 Fourth International Joint Conference on Computational Sciences and Optimization*. Kunming and Lijiang City, China: IEEE, 2011, pages 332–335. DOI: 10.1109/CSO.2011.85.

[10] Tanzeela Qazi, Khizar Hayat, Samee U. Khan, Sajjad A. Madani, Imran A. Khan, Joanna Kołodziej, Hongxiang Li, Weiyao Lin, Kin Choong Yow, and Cheng-Zhong Xu. "Survey on blind image forgery detection". In: *IET Image Processing* 7.7 (2013), pages 660–670. DOI: https://doi.org/10.1049/iet-ipr.2012.0388. eprint: https://ietresearch.onlinelibrary.wiley.com/doi/pdf/10.1049/iet-ipr.2012.0388. URL: https://ietresearch.onlinelibrary.wiley.com/doi/abs/10.1049/iet-ipr.2012.0388.

[11] John R Smith and Shih-Fu Chang. "Searching for images and videos on the world-wide web". In: *IEEE multimedia magazine* (1996). URL: https://www.ee.columbia.edu/ln/dvmm/publications/96/smith96e.pdf.

[12] Vassilis Athitsos, Michael J Swain, and Charles Frankel. "Distinguishing photographs and graphics on the world wide web". In: *1997 Proceedings IEEE Workshop on Content-Based Access of Image and Video Libraries*. IEEE. 1997, pages 10–17. DOI: 10.1109/IVL.1997.629715.

[13] John R Smith and Shih-Fu Chang. "Multi-stage classification of images from features and related text". In: *4th Europe EDLOS Workshop, San Miniato, Italy, Aug*. 1997. URL: https://www.ercim.eu/publication/ws-proceedings/DELOS4/smith.pdf.

[14] Yuanhao Chen, Zhiwei Li, Mingjing Li, and Wei-Ying Ma. "Automatic classification of photographs and graphics". In: *2006 IEEE International Conference on Multimedia and Expo*. IEEE. 2006, pages 973–976. DOI: 10.1109/ICME.2006.262695.

[15] Alexander Hartmann and Rainer W Lienhart. "Automatic classification of images on the web". In: *Storage and Retrieval for Media Databases 2002*. Volume 4676. International Society for Optics and Photonics. SPIE, 2001, pages 31–40. DOI: 10.1117/12.451108. URL: https://doi.org/10.1117/12.451108.

[16] Hany Farid and Siwei Lyu. "Higher-order wavelet statistics and their application to digital forensics". In: *2003 Conference on Computer Vision and Pattern Recognition Workshop*. Volume 8. IEEE. 2003, pages 94–94. DOI: 10.1109/CVPRW.2003.10093.

[17] Siwei Lyu and Hany Farid. "How realistic is photorealistic?" In: *IEEE Transactions on Signal Processing* 53.2 (2005), pages 845–850. ISSN: 1941-0476. DOI: 10.1109/TSP.2004.839896.

[18] A.C. Popescu and H. Farid. "Exposing digital forgeries in color filter array interpolated images". In: *IEEE Transactions on Signal Processing* 53.10 (2005), pages 3948–3959. ISSN: 1941-0476. DOI: 10.1109/TSP.2005.855406.

[19] Micah K. Johnson and Hany Farid. "Exposing Digital Forgeries through Chromatic Aberration". In: MM&Sec '06. Geneva, Switzerland: Association for Computing Machinery, 2006, 48–55. ISBN: 1595934936. DOI: 10.1145/1161366.1161376. URL: https://doi.org/10.1145/1161366.1161376.

[20] A. E. Dirik, S. Bayram, H. T. Sencar, and N. Memon. "New Features to Identify Computer Generated Images". In: *2007 IEEE International Conference on Image Processing*. Volume 4. IEEE. 2007, pages IV –433–IV –436. DOI: 10.1109/ICIP.2007.4380047.

[21] Nitin Khanna, George T. C. Chiu, Jan P. Allebach, and Edward J. Delp. "Forensic techniques for classifying scanner, computer generated and digital camera images". In: *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE. 2008, pages 1653–1656. DOI: 10.1109/ICASSP.2008.4517944.

[22] Shang Gao, Cong Zhang, Chan-Le Wu, Gang Ye, and Lei Huang. "A Hybrid Feature Based Method for Distinguishing Computer Graphics and Photo-Graphic Image". In: *Digital-Forensics and Watermarking*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2014, pages 303–313. ISBN: 978-3-662-43886-2. DOI: https://doi.org/10.1007/978-3-662-43886-2_22.

[23] Fei Peng and Die lan Zhou. "Discriminating natural images and computer generated graphics based on the impact of CFA interpolation on the correlation of PRNU". In: *Digital Investigation* 11.2 (2014), pages 111–119. ISSN: 1742-2876. DOI: https://doi.org/10.1016/j.diin.2014.04.002. URL: https://www.sciencedirect.com/science/article/pii/S1742287614000425.

[24] Min Long, Fei Peng, and Yin Zhu. "Identifying natural images and computer generated graphics based on binary similarity measures of PRNU". In: *Multimedia Tools and Applications* 78 (2019), pages 489–506. DOI: https://doi.org/10.1007/s11042-017-5101-3.

[25] Tian-Tsong Ng and Shih-Fu Chang. "Classifying photographic and photorealistic computer graphic images using natural image statistics". In: *ADVENT* (2004). URL: https://www.ee.columbia.edu/ln/dvmm/publications/04/ng-tech-report-nis-04.pdf.

[26] Tian-Tsong Ng and Shih-Fu Chang. "An online system for classifying computer graphics images from natural photographs". In: *Security, Steganography, and Watermarking of Multimedia Contents VIII*. Volume 6072. International Society for Optics and Photonics. SPIE, 2006, page 607211. DOI: 10.1117/12.650162. URL: https://doi.org/10.1117/12.650162.

[27] Sintayehu Dehnie, Taha Sencar, and Nasir Memon. "Digital image forensics for identifying computer generated and digital camera images". In: *2006 International Conference on Image Processing*. IEEE. 2006, pages 2313–2316. DOI: 10.1109/ICIP.2006.312849.

[28] Patchara Sutthiwan, Jingyu Ye, and Yun Q Shi. "An enhanced statistical approach to identifying photorealistic images". In: *Digital Watermarking*. Springer. Berlin, Heidelberg, 2009, pages 323–335. ISBN: 978-3-642-03688-0. DOI: https://doi.org/10.1007/978-3-642-03688-0_28.

[29] Rong Zhang, Rang-Ding Wang, and Tian-Tsong Ng. "Distinguishing photographic images and photorealistic computer graphics using visual vocabulary on local image edges". In: *Digital Forensics and Watermarking*. Berlin, Heidelberg: Springer, 2011, pages 292–305. ISBN: 978-3-642-32205-1. DOI: https://doi.org/10.1007/978-3-642-32205-1_24.

[30] Fei Peng, Die-lan Zhou, Min Long, and Xing-ming Sun. "Discrimination of natural images and computer generated graphics based on multi-fractal and regression analysis". In: *AEU-International Journal of Electronics and Communications* 71 (2017), pages 72–81. ISSN: 1434-8411. DOI: https://doi.org/10.1016/j.aeue.2016.11.009.

[31] Tian-Tsong Ng, Shih-Fu Chang, Jessie Hsu, and Martin Pepeljugoski. "Columbia photographic images and photorealistic computer graphics dataset". In: *Columbia University, ADVENT Technical Report* (2005), pages 205–2004. URL: https://www.ee.columbia.edu/ln/dvmm/downloads/PIM_PRCG_dataset/.

[32] Tian-Tsong Ng, Shih-Fu Chang, Jessie Hsu, Lexing Xie, and Mao-Pei Tsui. "Physics-Motivated Features for Distinguishing Photographic Images and Computer Graphics". In: *Proceedings of the 13th Annual ACM International Conference on Multimedia*. MULTIMEDIA '05. Hilton, Singapore: Association for Computing Machinery, 2005, 239–248. ISBN: 1595930442. DOI: 10.1145/1101149.1101192.

[33] Ashwin Swaminathan, Min Wu, and KJ Ray Liu. "Digital image forensics via intrinsic fingerprints". In: *IEEE transactions on information forensics and security* 3.1 (2008), pages 101–117. DOI: 10.1109/TIFS.2007.916010.

[34] Andrew C Gallagher and Tsuhan Chen. "Image authentication by detecting traces of demosaicing". In: *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*. IEEE. 2008, pages 1–8. DOI: 10.1109/CVPRW.2008.4562984.

[35] Christine McKay, Ashwin Swaminathan, Hongmei Gou, and Min Wu. "Image acquisition forensics: Forensic analysis to identify imaging source". In: *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE. 2008, pages 1657–1660. DOI: 10.1109/ICASSP.2008.4517945.

[36] Gopinath Sankar, Vicky Zhao, and Yee-Hong Yang. "Feature based classification of computer graphics and real images". In: *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE. 2009, pages 1513–1516. DOI: 10.1109/ICASSP.2009.4959883.

[37] Thomas Gloe and Rainer Böhme. "The'Dresden Image Database'for benchmarking digital image forensics". In: *Proceedings of the 2010 ACM Symposium on Applied Computing*. SAC '10. Sierre, Switzerland: Association for Computing Machinery, 2010, pages 1584–1590. ISBN: 9781605586397. DOI: 10.1145/1774088.1774427. URL: https://doi.org/10.1145/1774088.1774427.

[38] Feng Pan and Jiwu Huang. "Discriminating Computer Graphics Images and Natural Images Using Hidden Markov Tree Model". In: *Digital Watermarking*. Berlin, Heidelberg: Springer, 2011, pages 23–28. ISBN: 978-3-642-18405-5. DOI: https://doi.org/10.1007/978-3-642-18405-5_3.

[39] Levent Ozparlak and Ismail Avcibas. "Differentiating between images using wavelet-based transforms: a comparative study". In: *IEEE Transactions on Information Forensics and Security* 6.4 (2011), pages 1418–1431. DOI: 10.1109/TIFS.2011.2162830.

[40] Jinwei Wang, Ting Li, Yun-Qing Shi, Shiguo Lian, and Jingyu Ye. "Forensics feature analysis in quaternion wavelet domain for distinguishing photographic images and computer graphics". In: *Multimedia tools and Applications* 76.22 (2017), pages 23721–23737. DOI: https://doi.org/10.1007/s11042-016-4153-0.

[41] In-Jae Yu, Do-Guk Kim, Jin-Seok Park, Jong-Uk Hou, Sunghee Choi, and Heung-Kyu Lee. "Identifying photorealistic computer graphics using convolutional neural networks". In: *2017 IEEE International Conference on Image Processing (ICIP)*. IEEE. 2017, pages 4093–4097. DOI: 10.1109/ICIP.2017.8297052.

[42] Nicolas Rahmouni, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. "Distinguishing computer graphics from natural images using convolution neural networks". In: *2017 IEEE Workshop on Information Forensics and Security (WIFS)*. IEEE. 2017, pages 1–6. DOI: 10.1109/WIFS.2017.8267647.

[43] Qi Cui, Suzanne McIntosh, and Huiyu Sun. "Identifying materials of photographic images and photorealistic computer generated graphics based on deep CNNs". In: *Comput. Mater. Continua* 55.2 (2018), pages 229–241. DOI: 10.3970/cmc.2018.01693.

[44] Ye Yao, Weitong Hu, Wei Zhang, Ting Wu, and Yun-Qing Shi. "Distinguishing computer-generated graphics from natural images based on sensor pattern noise and deep learning". In: *Sensors* 18.4 (2018), page 1296. ISSN: 1424-8220. DOI: 10.3390/s18041296. URL: https://www.mdpi.com/1424-8220/18/4/1296.

[45] Huy H Nguyen, Junichi Yamagishi, and Isao Echizen. "Capsule-forensics: Using capsule networks to detect forged images and videos". In: *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2019, pages 2307–2311. DOI: 10.1109/ICASSP.2019.8682602.

[46] Diangarti Bhalang Tariang, Prithviraj Senguptab, Aniket Roy, Rajat Subhra Chakraborty, and Ruchira Naskar. "Classification of Computer Generated and Natural Images based on Efficient Deep Convolutional Recurrent Attention Model." In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. 2019, pages 146–152. URL: https://openaccess.thecvf.com/content_CVPRW_2019/html/Media_Forensics/Tarianga_Classification_of_Computer_Generated_and_Natural_Images_based_on_Efficient_CVPRW_2019_paper.html.

[47] Rui-Song Zhang, Wei-Ze Quan, Lu-Bin Fan, Li-Ming Hu, and Dong-Ming Yan. "Distinguishing Computer-Generated Images from Natural Images Using Channel and Pixel Correlation". In: *Journal of Computer Science and Technology* 35 (2020), pages 592–602. DOI: https://doi.org/10.1007/s11390-020-0216-9.

[48] Duc-Tien Dang-Nguyen, Cecilia Pasquini, Valentina Conotter, and Giulia Boato. "Raise: A raw images dataset for digital image forensics". In: *Proceedings of the 6th ACM multimedia systems conference*. Association for Computing Machinery, 2015, pages 219–224. ISBN: 9781450333511. DOI: 10.1145/2713168.2713194.

[49] Peisong He, Xinghao Jiang, Tanfeng Sun, and Haoliang Li. "Computer graphics identification combining convolutional and recurrent neural networks". In: *IEEE Signal Processing Letters* 25.9 (2018), pages 1369–1373. DOI: 10.1109/LSP.2018.2855566.

[50] Huy H Nguyen, T Ngoc-Dung Tieu, Hoang-Quoc Nguyen-Son, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. "Modular convolutional neural network for discriminating between computer-generated images and photographic images". In: *Proceedings of the 13th International Conference on Availability, Reliability and Security*. ARES '18. Germany: Association for Computing Machinery, 2018. ISBN: 9781450364485. DOI: 10.1145/3230833.3230863.

[51] Edmar RS De Rezende, Guilherme CS Ruppert, Antonio Theophilo, Eric K Tokuda, and Tiago Carvalho. "Exposing computer generated images by using deep convolutional neural networks". In: *Signal Processing: Image Communication* 66 (2018), pages 113–126. ISSN: 0923-5965. DOI: https://doi.org/10.1016/j.image.2018.04.006.

[52] Min Long, Sai Long, Fei Peng, and Xiao-hua Hu. "Identifying natural images and computer-generated graphics based on convolutional neural network". In: *International Journal of Autonomous and Adaptive Communications Systems* 14.1-2 (2021), pages 151–162. DOI: 10.1504/IJAACS.2021.114295.

[53] Kunj Bihari Meena and Vipin Tyagi. "Distinguishing computer-generated images from photographic images using two-stream convolutional neural network". In: *Applied Soft Computing* 100 (2021), page 107025. ISSN: 1568-4946. DOI: https://doi.org/10.1016/j.asoc.2020.107025.

[54] Weize Quan, Kai Wang, Dong-Ming Yan, and Xiaopeng Zhang. "Distinguishing between natural and computer-generated images using convolutional neural networks". In: *IEEE Transactions on Information Forensics and Security* 13.11 (2018), pages 2772–2787. DOI: 10.1109/TIFS.2018.2834147.

[55] Scott McCloskey and Michael Albright. "Detecting gan-generated imagery using color cues". In: *arXiv preprint arXiv:1812.08247* (2018). DOI: https://doi.org/10.48550/arXiv.1812.08247.

[56] Haodong Li, Bin Li, Shunquan Tan, and Jiwu Huang. "Identification of deep network generated images using disparities in color components". In: *Signal Processing* 174 (2020), page 107616. DOI: 10.1016/j.sigpro.2020.107616.

[57] Francesco Marra, Diego Gragnaniello, Davide Cozzolino, and Luisa Verdoliva. "Detection of gan-generated fake images over social networks". In: *2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*. IEEE. 2018, pages 384–389. DOI: 10.1109/MIPR.2018.00084.

[58] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. "Progressive growing of GANs for Improved Quality, Stability, and Variation". In: *International Conference on Learning Representations (ICLR)*. 2018. URL: https://openreview.net/forum?id=Hk99zCeAb.

[59] Jessica Fridrich and Jan Kodovsky. "Rich Models for Steganalysis of Digital Images". In: *IEEE Transactions on Information Forensics and Security* 7.3 (2012), pages 868–882. DOI: 10.1109/TIFS.2012.2190402.

[60] Davide Cozzolino, Diego Gragnaniello, and Luisa Verdoliva. "Image forgery detection through residual-based local descriptors and block-matching". In: *2014 IEEE International Conference on Image Processing (ICIP)*. IEEE. 2014, pages 5297–5301. DOI: 10.1109/ICIP.2014.7026072.

[61] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Niessner. "FaceForensics++: Learning to Detect Manipulated Facial Images". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2019, pages 1–11. URL: https://openaccess.thecvf.com/content_ICCV_2019/html/Rossler_FaceForensics_Learning_to_Detect_Manipulated_Facial_Images_ICCV_2019_paper.html.

[62] Davide Cozzolino, Giovanni Poggi, and Luisa Verdoliva. "Recasting residual-based local descriptors as convolutional neural networks: an application to image forgery detection". In: *Proceedings of the 5th ACM Workshop on Information Hiding and Multimedia Security*. Association for Computing Machinery, 2017, pages 159–164. ISBN: 9781450350617. DOI: 10.1145/3082031.3083247.

[63] Lakshmanan Nataraj, Tajuddin Manhar Mohammed, BS Manjunath, Shivkumar Chandrasekaran, Arjuna Flenner, Jawadul H Bappy, and Amit K Roy-Chowdhury. "Detecting GAN generated fake images using co-occurrence matrices". In: *Electronic Imaging* 2019.5 (2019), pages 532–1. DOI: 10.2352/ISSN.2470-1173.2019.5.MWSF-532.

[64] Xinsheng Xuan, Bo Peng, Wei Wang, and Jing Dong. "On the Generalization of GAN Image Forensics". In: *Biometric Recognition*. Cham: Springer International Publishing, 2019, pages 134–141. ISBN: 978-3-030-31456-9. DOI: https://doi.org/10.1007/978-3-030-31456-9_15.

[65] Karen Simonyan and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition". In: (2015). URL: https://www.robots.ox.ac.uk/~vgg/publications/2015/Simonyan15/simonyan15.pdf.

[66] Nhu-Tai Do, In-Seop Na, and Soo-Hyung Kim. "Forensics face detection from GANs using convolutional neural network". In: *ISITC* 2018 (2018), pages 376–379. URL: https://www.researchgate.net/profile/Nhu-Tai-Do/publication/327905310_Forensics_Face_Detection_From_GANs_Using_Convolutional_Neural_Network/links/5bac84e7a6fdccd3cb768b1c/Forensics-Face-Detection-From-GANs-Using-Convolutional-Neural-Network.pdf.

[67] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A. Alemi. "Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning". In: *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*. AAAI'17. San Francisco, California, USA: AAAI Press, 2017, 4278–4284. URL: https://dl.acm.org/doi/10.5555/3298023.3298188.

[68] François Chollet. "Xception: Deep Learning with Depthwise Separable Convolutions". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017, pages 1800–1807. DOI: 10.1109/CVPR.2017.195.

[69] Nils Hulzebosch, Sarah Ibrahimi, and Marcel Worring. "Detecting CNN-Generated Facial Images in Real-World Scenarios". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 2020, pages 2729–2738. DOI: 10.1109/CVPRW50498.2020.00329.

[70] Hao Dang, Feng Liu, Joel Stehouwer, Xiaoming Liu, and Anil K. Jain. "On the Detection of Digital Face Manipulation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020. URL: https://openaccess.thecvf.com/content_CVPR_2020/html/Dang_On_the_Detection_of_Digital_Face_Manipulation_CVPR_2020_paper.html.

[71] Michael Goebel, Lakshmanan Nataraj, Tejaswi Nanjundaswamy, Tajuddin Manhar Mohammed, Shivkumar Chandrasekaran, and BS Manjunath. "Detection, attribution and localization of gan generated images". In: *Electronic Imaging* (2021). DOI: doi:10.2352/ISSN.2470-1173.2021.4.MWSF-276.

[72] Chih-Chung Hsu, Yi-Xiu Zhuang, and Chia-Yen Lee. "Deep fake image detection based on pairwise learning". In: *Applied Sciences* 10.1 (2020), page 370. ISSN: 2076-3417. DOI: 10.3390/app10010370.

[73] Francesco Marra, Cristiano Saltori, Giulia Boato, and Luisa Verdoliva. "Incremental learning for the detection and classification of gan-generated images". In: *2019 IEEE International Workshop on Information Forensics and Security (WIFS)*. IEEE. 2019, pages 1–6. DOI: 10.1109/WIFS47025.2019.9035099.

[74] Hadi Mansourifar and Weidong Shi. "One-shot gan generated fake face detection". In: *arXiv preprint arXiv:2003.12244* (2020). DOI: https://doi.org/10.48550/arXiv.2003.12244.

[75] Huaxiao Mo, Bolin Chen, and Weiqi Luo. "Fake faces identification via convolutional neural network". In: *Proceedings of the 6th ACM Workshop on Information Hiding and Multimedia Security*. 2018, pages 43–47. ISBN: 9781450356251. DOI: 10.1145/3206004.3206009.

[76] Shahroz Tariq, Sangyup Lee, Hoyoung Kim, Youjin Shin, and Simon S Woo. "Detecting both machine and human created fake face images in the wild". In: *Proceedings of the 2nd international workshop on multimedia privacy and security*. 2018, pages 81–87. ISBN: 9781450359887. DOI: 10.1145/3267357.3267367.

[77] Francesco Marra, Diego Gragnaniello, Luisa Verdoliva, and Giovanni Poggi. "Do GANs Leave Artificial Fingerprints?" In: *2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*. IEEE, 2019, pages 506–511. ISBN: 978-1-7281-1199-5. DOI: 10.1109/MIPR.2019.00103.

[78] Ning Yu, Larry S Davis, and Mario Fritz. "Attributing fake images to gans: Learning and analyzing gan fingerprints". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019, pages 7556–7566. DOI: 10.1109/ICCV.2019.00765.

[79] Matthew Joslin and Shuang Hao. "Attributing and Detecting Fake Images Generated by Known GANs". In: *2020 IEEE Security and Privacy Workshops (SPW)*. IEEE. 2020, pages 8–14. DOI: 10.1109/SPW50608.2020.00019.

[80] Mauro Barni, Kassem Kallas, Ehsan Nowroozi, and Benedetta Tondi. "CNN detection of GAN-generated face images based on cross-band co-occurrences analysis". In: *2020 IEEE International Workshop on Information Forensics and Security (WIFS)*. IEEE. 2020, pages 1–6. DOI: 10.1109/WIFS49906.2020.9360905.

[81] Shu Hu, Yuezun Li, and Siwei Lyu. "Exposing GAN-generated Faces Using Inconsistent Corneal Specular Highlights". In: *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2021, pages 2500–2504. DOI: 10.1109/ICASSP39728.2021.9414582.

[82] Deressa Wodajo and Solomon Atnafu. "Deepfake video detection using convolutional vision transformer". In: *arXiv preprint arXiv:2102.11126* (2021). DOI: https://doi.org/10.48550/arXiv.2102.11126.

[83] Davide Alessandro Coccomini, Nicola Messina, Claudio Gennaro, and Fabrizio Falchi. "Combining EfficientNet and Vision Transformers for Video Deepfake Detection". In: *Image Analysis and Processing – ICIAP 2022*. Cham: Springer International Publishing, 2022, pages 219–229. ISBN: 978-3-031-06433-3. URL: https://doi.org/10.1007/978-3-031-06433-3_19.

[84] Junke Wang, Zuxuan Wu, Wenhao Ouyang, Xintong Han, Jingjing Chen, Yu-Gang Jiang, and Ser-Nam Li. "M2TR: Multi-Modal Multi-Scale Transformers for Deepfake Detection". In: *Proceedings of the 2022 International Conference on Multimedia Retrieval*. ICMR '22. Newark, NJ, USA: Association for Computing Machinery, 2022, 615–623. ISBN: 9781450392389. DOI: 10.1145/3512527.3531415. URL: https://doi.org/10.1145/3512527.3531415.

[85] Young-Jin Heo, Woon-Ha Yeo, and Byung-Gyu Kim. "Deepfake detection algorithm based on improved vision transformer". In: *Applied Intelligence* 53.7 (2023), pages 7512–7527. DOI: https://doi.org/10.1007/s10489-022-03867-9.

[86] Sudarshana Kerenalli, Vamsidhar Yendapalli, and C. Mylarareddy. "Fake Face Image Classification by Blending the Scalable Convolution Network and Hierarchical Vision Transformer". In: *Proceedings of Fourth International Conference on Computer and Communication Technologies*. Singapore: Springer Nature Singapore, 2023, pages 117–126. ISBN: 978-981-19-8563-8. DOI: https://doi.org/10.1007/978-981-19-8563-8_12. URL: https://link.springer.com/chapter/10.1007/978-981-19-8563-8_12.

[87] Gary W Meyer, Holly E Rushmeier, Michael F Cohen, Donald P Greenberg, and Kenneth E Torrance. "An experimental evaluation of computer graphics imagery". In: *ACM Transactions on Graphics (TOG)* 5.1 (1986), pages 30–50. DOI: https://doi.org/10.1145/7529.7920.

[88] Ann M McNamara. "Exploring perceptual equivalence between real and simulated imagery". In: *Proceedings of the 2nd symposium on Applied Perception in Graphics and Visualization*. 2005, pages 123–128. DOI: https://doi.org/10.1145/1080402.1080425.

[89] Rachel McDonnell, Martin Breidt, and Heinrich H Bülthoff. "Render me real? Investigating the effect of render style on the perception of animated virtual humans". In: *ACM Transactions on Graphics (TOG)* 31.4 (2012), pages 1–11. DOI: https://doi.org/10.1145/2185520.2185587.

[90] Shaojing Fan, Tian-Tsong Ng, Bryan Lee Koenig, Jonathan Samuel Herberg, Ming Jiang, Zhiqi Shen, and Qi Zhao. "Image visual realism: From human perception to machine computation". In: *IEEE transactions on pattern analysis and machine intelligence* 40.9 (2018), pages 2180–2193. DOI: 10.1109/TPAMI.2017.2747150.

[91] Hany Farid and Mary Bravo. "Photorealistic rendering: How realistic is it?" In: *Journal of Vision* 7.9 (2007), pages 766–766. DOI: https://doi.org/10.1167/7.9.766.

[92] Hany Farid and Mary J Bravo. "Perceptual discrimination of computer generated and photographic faces". In: *Digital Investigation* 8.3-4 (2012), pages 226–235. DOI: https://doi.org/10.1016/j.diin.2011.06.003.

[93] Shaojing Fan, Tian-Tsong Ng, Jonathan S. Herberg, Bryan L. Koenig, and Shiqing Xin. "Real or Fake? Human Judgments about Photographs and Computer-Generated Images of Faces". In: *SIGGRAPH Asia 2012 Technical Briefs*. SA '12. Singapore, Singapore: Association for Computing Machinery, 2012. ISBN: 9781450319157. DOI: 10.1145/2407746.2407763.

[94] Olivia Holmes, Martin S. Banks, and Hany Farid. "Assessing and Improving the Identification of Computer-Generated Portraits". In: 13.2 (2016). ISSN: 1544-3558. DOI: 10.1145/2871714. URL: https://doi.org/10.1145/2871714.

[95] Brandon Mader, Martin S Banks, and Hany Farid. "Identifying computer-generated portraits: The importance of training and incentives". In: *Perception* 46.9 (2017), pages 1062–1076. DOI: 10.1177/0301006617713633. URL: https://doi.org/10.1177/0301006617713633.

[96] Sophie J. Nightingale and Hany Farid. "AI-synthesized faces are indistinguishable from real faces and more trustworthy". In: *Proceedings of the National Academy of Sciences* 119.8 (2022), e2120481119. DOI: 10.1073/pnas.2120481119. eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.2120481119. URL: https://www.pnas.org/doi/abs/10.1073/pnas.2120481119.

[97] Federica Lago, Cecilia Pasquini, Rainer Böhme, Hélène Dumont, Valérie Goffaux, and Giulia Boato. "More Real Than Real: A Study on Human Visual Perception of Synthetic Faces [Applications Corner]". In: *IEEE Signal Processing Magazine* 39.1 (2022), pages 109–116. ISSN: 1558-0792. DOI: 10.1109/MSP.2021.3120982.

[98] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. "Analyzing and Improving the Image Quality of StyleGAN". In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020, pages 8107–8116. DOI: 10.1109/CVPR42600.2020.00813.

[99] Tero Karras, Samuli Laine, and Timo Aila. "A Style-Based Generator Architecture for Generative Adversarial Networks". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43.12 (2021), pages 4217–4228. DOI: 10.1109/TPAMI.2020.2970919.

[100] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A. Efros. "CNN-Generated Images Are Surprisingly Easy to Spot... for Now". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020. URL: https://openaccess.thecvf.com/content_CVPR_2020/html/Wang_CNN-Generated_Images_Are_Surprisingly_Easy_to_Spot..._for_Now_CVPR_2020_paper.html.

[101] Xu Zhang, Svebor Karaman, and Shih-Fu Chang. "Detecting and Simulating Artifacts in GAN Fake Images". In: *2019 IEEE International Workshop on Information Forensics and Security (WIFS)*. IEEE, 2019, pages 1–6. DOI: 10.1109/WIFS47025.2019.9035107.

[102] Sebastian Lapuschkin, Alexander Binder, Grégoire Montavon, Klaus-Robert Müller, and Wojciech Samek. "The LRP Toolbox for Artificial Neural Networks". In: *Journal of Machine Learning Research* 17.114 (2016), pages 1–5. URL: http://jmlr.org/papers/v17/15-618.html.

[103] Jason Yosinski, Jeff Clune, Anh Nguyen, Thomas Fuchs, and Hod Lipson. "Understanding neural networks through deep visualization". In: *arXiv preprint arXiv:1506.06579* (2015). DOI: https://doi.org/10.48550/arXiv.1506.06579.

[104] Hui Guo, Shu Hu, Xin Wang, Ming-Ching Chang, and Siwei Lyu. "Robust Attentive Deep Neural Network for Detecting GAN-Generated Faces". In: *IEEE Access* 10 (2022), pages 32574–32583. ISSN: 2169-3536. DOI: 10.1109/ACCESS.2022.3157297.

[105] Joy Buolamwini and Timnit Gebru. "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification". In: *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*. Volume 81. Proceedings of Machine Learning Research. PMLR, 2018, pages 77–91. URL: https://proceedings.mlr.press/v81/buolamwini18a.html.

[106] Nina Schaaf, Omar de Mitri, Hang Beom Kim, Alexander Windberger, and Marco F. Huber. "Towards Measuring Bias in Image Classification". In: *Artificial Neural Networks and Machine Learning – ICANN 2021*. Cham: Springer International Publishing, 2021, pages 433–445. ISBN: 978-3-030-86365-4. DOI: https://doi.org/10.1007/978-3-030-86365-4_35.

[107] John S. H. Baxter and Pierre Jannin. "Bias in machine learning for computer-assisted surgery and medical image processing". In: *Computer Assisted Surgery* 27.1 (2022), pages 1–3. DOI: 10.1080/24699322.2021.2013619. eprint: https://doi.org/10.1080/24699322.2021.2013619. URL: https://doi.org/10.1080/24699322.2021.2013619.

[108] Loc Trinh and Yan Liu. "An Examination of Fairness of AI Models for Deepfake Detection". In: *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*. Edited by Zhi-Hua Zhou. Main Track. International Joint Conferences on Artificial Intelligence Organization, Aug. 2021, pages 567–574. DOI: 10.24963/ijcai.2021/79. URL: https://doi.org/10.24963/ijcai.2021/79.

[109] Darius Afchar, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. "MesoNet: a Compact Facial Video Forgery Detection Network". In: *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*. IEEE, 2018, pages 1–7. DOI: 10.1109/WIFS.2018.8630761.

[110] Lingzhi Li, Jianmin Bao, Ting Zhang, Hao Yang, Dong Chen, Fang Wen, and Baining Guo. "Face X-Ray for More General Face Forgery Detection". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020, pages 5000–5009. DOI: 10.1109/CVPR42600.2020.00505. URL: https://ieeexplore.ieee.org/document/9157215.

[111] Caner Hazirbas, Joanna Bitton, Brian Dolhansky, Jacqueline Pan, Albert Gordo, and Cristian Canton Ferrer. "Towards Measuring Fairness in AI: The Casual Conversations Dataset". In: *IEEE Transactions on Biometrics, Behavior, and Identity Science* 4.3 (2022), pages 324–332. DOI: 10.1109/TBIOM.2021.3132237.

[112] Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer. "The deepfake detection challenge (dfdc) dataset". In: *arXiv preprint:2006.07397* (2020). DOI: https://doi.org/10.48550/arXiv.2006.07397.

[113] Muxin Pu, Meng Yi Kuan, Nyee Thoang Lim, Chun Yong Chong, and Mei Kuan Lim. "Fairness Evaluation in Deepfake Detection Models Using Metamorphic Testing". In: *Proceedings of the 7th International Workshop on Metamorphic Testing*. MET '22. Pittsburgh, Pennsylvania: Association for Computing Machinery, 2023, 7–14. ISBN: 9781450393072. DOI: 10.1145/3524846.3527337. URL: https://doi.org/10.1145/3524846.3527337.

[114] Ying Xu, Philipp Terhörst, Kiran Raja, and Marius Pedersen. *A Comprehensive Analysis of AI Biases in DeepFake Detection With Massively Annotated Databases*. 2023. arXiv: 2208.05845.

[115] Mingxing Tan and Quoc Le. "Efficientnet: Rethinking model scaling for convolutional neural networks". In: *Proceedings of the 36th International Conference on Machine Learning*. Volume 97. PMLR. 2019, pages 6105–6114. URL: https://proceedings.mlr.press/v97/tan19a.html.

[116] Huy H Nguyen, Junichi Yamagishi, and Isao Echizen. "Use of a capsule network to detect fake images and videos". In: *arXiv preprint arXiv:1910.12467* (2019). DOI: https://doi.org/10.48550/arXiv.1910.12467.

[117] Peisong He, Haoliang Li, and Hongxia Wang. "Detection of fake images via the ensemble of deep representations from multi color spaces". In: *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE. 2019, pages 2299–2303. DOI: 10.1109/ICIP.2019.8803740.

[118] Haodong Li, Bin Li, Shunquan Tan, and Jiwu Huang. "Identification of deep network generated images using disparities in color components". In: *Signal Processing* 174 (2020), page 107616. ISSN: 0165-1684. DOI: https://doi.org/10.1016/j.sigpro.2020.107616.

[119] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. "Training Generative Adversarial Networks with Limited Data". In: *Proc. NeurIPS*. 2020. URL: https://proceedings.neurips.cc/paper/2020/file/8d30aa96e72440759f74bd2306c1fa3d-Paper.pdf.

[120] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. "Imagenet large scale visual recognition challenge". In: *International journal of computer vision* 115.3 (2015), pages 211–252. DOI: https://doi.org/10.1007/s11263-015-0816-y.

[121] Shreyank N Gowda and Chun Yuan. "ColorNet: Investigating the importance of color spaces for image classification". In: *Asian Conference on Computer Vision*. Springer. 2018, pages 581–596. DOI: https://doi.org/10.1007/978-3-030-20870-7_36.

[122] Mingxing Tan, Bo Chen, Ruoming Pang, Vijay Vasudevan, Mark Sandler, Andrew Howard, and Quoc V Le. "Mnasnet: Platform-aware neural architecture search for mobile". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pages 2820–2828. DOI: 10.1109/CVPR.2019.00293.

[123] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. "Mobilenetv2: Inverted residuals and linear bottlenecks". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pages 4510–4520. DOI: 10.1109/CVPR.2018.00474.

[124] Jie Hu, Li Shen, and Gang Sun. "Squeeze-and-excitation networks". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pages 7132–7141. DOI: 10.1109/CVPR.2018.00745.

[125] Prajit Ramachandran, Barret Zoph, and Quoc V Le. "Searching for activation functions". In: *arXiv preprint arXiv:1710.05941* (2017). URL: http://arxiv.org/abs/1710.05941.

[126] Tomáš Pevny, Patrick Bas, and Jessica Fridrich. "Steganalysis by subtractive pixel adjacency matrix". In: *IEEE Transactions on information Forensics and Security* 5.2 (2010), pages 215–224. DOI: 10.1109/TIFS.2010.2045842.

[127] Liqian Ma, Xu Jia, Qianru Sun, Bernt Schiele, Tinne Tuytelaars, and Luc Van Gool. "Pose Guided Person Image Generation". In: *Advances in Neural Information Processing Systems (NIPS)*. Volume 31. Research Collection School Of Computing and Information Systems, 2017, pages 406–416. URL: https://ink.library.smu.edu.sg/sis_research/4458.

[128] Xuan Ni, Linqiang Chen, Lifeng Yuan, Guohua Wu, and Ye Yao. "An evaluation of deep learning-based computer generated image detection approaches". In: *IEEE Access* 7 (2019), pages 130830–130840. DOI: 10.1109/ACCESS.2019.2940383.

[129] M Piaskiewicz. "Level-design reference database". In: *Accessed: Jan* 15 (2017), page 2018. URL: http://level-design.org/referencedb/.

[130] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. "Unpaired image-to-image translation using cycle-consistent adversarial networks". In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2017, pages 2242–2251. DOI: 10.1109/ICCV.2017.244.

[131] Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer. *The DeepFake Detection Challenge Dataset*. 2020. arXiv: 2006.07397 [cs.CV].

[132] CGSociety. Accessed 26 April 2022. URL: https://cgsociety.org.

[133] Laurens Van der Maaten and Geoffrey Hinton. "Visualizing data using t-SNE." In: *Journal of machine learning research* 9.11 (2008). URL: http://jmlr.org/papers/v9/vandermaaten08a.html.

[134] A. Dutta, A. Gupta, and A. Zissermann. *VGG Image Annotator (VIA)*. http://www.robots.ox.ac.uk/~vgg/software/via/. Version: 2.0.8, Accessed: 4 September 2021. 2016.

[135] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. "Grad-cam: Visual explanations from deep networks via gradient-based localization". In: *Proceedings of the IEEE international conference on computer vision*. 2017, pages 618–626. DOI: 10.1109/ICCV.2017.74.

[136] Sara Mandelli, Nicolò Bonettini, Paolo Bestagini, and Stefano Tubaro. "Training CNNs in Presence of JPEG Compression: Multimedia Forensics vs Computer Vision". In: *2020 IEEE International Workshop on Information Forensics and Security (WIFS)*. 2020, pages 1–6. DOI: 10.1109/WIFS49906.2020.9360903. URL: https://ieeexplore.ieee.org/document/9360903.

[137] Menglu Wang, Xueyang Fu, Jiawei Liu, and Zheng-Jun Zha. "JPEG Compression-Aware Image Forgery Localization". In: *Proceedings of the 30th ACM International Conference on Multimedia*. MM '22. Lisboa, Portugal: Association for Computing Machinery, 2022, 5871–5879. ISBN: 9781450392037. DOI: 10.1145/3503161.3547749. URL: https://doi.org/10.1145/3503161.3547749.

[138] Mauro Barni, Matthew C. Stamm, and Benedetta Tondi. "Adversarial Multimedia Forensics: Overview and Challenges Ahead". In: *2018 26th European Signal Processing Conference (EUSIPCO)*. IEEE. 2018, pages 962–966. DOI: 10.23919/EUSIPCO.2018.8553305. URL: https://ieeexplore.ieee.org/document/8553305.

[139] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale". In: *International Conference on Learning Representations*. 2021. URL: https://openreview.net/forum?id=YicbFdNTTy.

[140] Manjary P. Gangan, Anoop K., and Lajish V. L. "Distinguishing natural and computer generated images using Multi-Colorspace fused EfficientNet". In: *Journal of Information Security and Applications* 68 (2022), page 103261. ISSN: 2214-2126. DOI: https://doi.org/10.1016/j.jisa.2022.103261. URL: https://www.sciencedirect.com/science/article/pii/S2214212622001247.

[141] Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik-Manor. "ImageNet-21K Pretraining for the Masses". In: (2021). URL: https://openreview.net/forum?id=Zkj_VcZ6ol.

[142] *JPEG: Joint Photographic Experts Group — cs.auckland.ac.nz*. https://www.cs.auckland.ac.nz/compsci708s1c/lectures/jpeg_mpeg/jpeg.html. [Accessed 05-08-2023].

[143] *JPEG Compression Explained | Baeldung on Computer Science — baeldung.com*. https://www.baeldung.com/cs/jpeg-compression. [Accessed 05-08-2023].

[144] Manjary P Gangan, Anoop Kadan, and Lajish V L. *A Robust Approach Towards Distinguishing Natural and Computer Generated Images using Multi-Colorspace fused and Enriched Vision Transformer*. 2023. arXiv: 2308.07279.

[145] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. "End-to-end object detection with transformers". In: *European conference on computer vision*. Springer. 2020, pages 213–229. DOI: https://doi.org/10.1007/978-3-030-58452-8_13.

[146] Wenyu Lv, Shangliang Xu, Yian Zhao, Guanzhong Wang, Jinman Wei, Cheng Cui, Yuning Du, Qingqing Dang, and Yi Liu. "Detrs beat yolos on real-time object detection". In: *arXiv preprint arXiv:2304.08069* (2023). DOI: https://doi.org/10.48550/arXiv.2304.08069.

[147] Zixiao Zhang, Xiaoqiang Lu, Guojin Cao, Yuting Yang, Licheng Jiao, and Fang Liu. "ViT-YOLO: Transformer-based YOLO for object detection". In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pages 2799–2808. DOI: 10.1109/ICCVW54120.2021.00314.

[148] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, and Ping Luo. "SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers". In: *Advances in Neural Information Processing Systems*. Volume 34. Curran Associates, Inc., 2021, pages 12077–12090. URL: https://proceedings.neurips.cc/paper_files/paper/2021/file/64f1f27bf1b4ec22924fd0acb550c235-Paper.pdf.

[149] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. "Segmenter: Transformer for semantic segmentation". In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pages 7262–7272. DOI: 10.1109/ICCV48922.2021.00717.

[150] Yifan Jiang, Shiyu Chang, and Zhangyang Wang. "Transgan: Two pure transformers can make one strong gan, and that can scale up". In: *Advances in Neural Information Processing Systems* 34 (2021). URL: https://proceedings.neurips.cc/paper_files/paper/2021/file/7c220a2091c26a7f5e9f1cfb099511e3-Paper.pdf.

[151] Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. "Transformers in Vision: A Survey". In: *ACM Comput. Surv.* 54.10s (2022). ISSN: 0360-0300. DOI: 10.1145/3505244. URL: https://doi.org/10.1145/3505244.

[152] Kai Han, Yunhe Wang, Hanting Chen, Xinghao Chen, Jianyuan Guo, Zhenhua Liu, Yehui Tang, An Xiao, Chunjing Xu, Yixing Xu, et al. "A survey on vision transformer". In: *IEEE transactions on pattern analysis and machine intelligence* 45.1 (2022), pages 87–110. DOI: 10.1109/TPAMI.2022.3152247.

[153] Davide Alessandro Coccomini, Roberto Caldelli, Fabrizio Falchi, Claudio Gennaro, and Giuseppe Amato. "Cross-Forgery Analysis of Vision Transformers and CNNs for Deepfake Image Detection". In: *Proceedings of the 1st International Workshop on Multimedia AI against Disinformation*. MAD '22. Newark, NJ, USA: Association for Computing Machinery, 2022, 52–58. ISBN: 9781450392426. DOI: 10.1145/3512732.3533582.

[154] Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. "CvT: Introducing Convolutions to Vision Transformers". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2021, pages 22–31. DOI: 10.1109/ICCV48922.2021.00009.

[155] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. "Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2021, pages 9992–10002. DOI: 10.1109/ICCV48922.2021.00986.

[156] Julia Dressel and Hany Farid. "The accuracy, fairness, and limits of predicting recidivism". In: *Science Advances* 4.1 (2018), eaao5580. DOI: 10.1126/sciadv.aao5580. URL: https://www.science.org/doi/abs/10.1126/sciadv.aao5580.

[157] Kadan Anoop, P. Deepak, Bhadra Sahely, P. Gangan Manjary, and V L Lajish. *Understanding latent affective bias in large pre-trained neural language models*. 2024. DOI: https://doi.org/10.1016/j.nlp.2024.100062.

[158] Huan Tian, Tianqing Zhu, Wei Liu, and Wanlei Zhou. "Image fairness in deep learning: problems, models, and challenges". In: *Neural Computing and Applications* 34.15 (2022), pages 12875–12893. DOI: https://doi.org/10.1007/s00521-022-07136-1.

[159] Moritz Hardt, Eric Price, Eric Price, and Nati Srebro. "Equality of Opportunity in Supervised Learning". In: *Advances in Neural Information Processing Systems*. Edited by D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett. Volume 29. Curran Associates, Inc., 2016. URL: https://proceedings.neurips.cc/paper_files/paper/2016/file/9d2682367c3935defcb1f9e247a97c0d-Paper.pdf.

[160] Jonas Ricker, Simon Damm, Thorsten Holz, and Asja Fischer. "Towards the detection of diffusion model deepfakes". In: *arXiv preprint arXiv:2210.14571* (2022). DOI: https://doi.org/10.48550/arXiv.2210.14571.