INSTITUTE FOR HOMELAND SECURITY

Sam Houston State University

DETECTING DEEPFAKES UNDER ANTI-FORENSICS ATTACKS

Institute for Homeland Security

Sam Houston State University

Qingzhong Liu
Naciye Celebi
Bing Zhou

Zhongxue Chen

Detecting Deepfakes under Anti-forensics Attacks

Qingzhong Liu[1], Naciye Celebi[1], Bing Zhou[1] and Zhongxue Chen[2]

[1] Department of Computer Science, Sam Houston State University, Huntsville, TX 77341, USA
[2] College of Health Solutions. Arizona State University, Phoenix, AZ 85004, USA
[1]{liu, nxc038, zhou}@shsu.edu; [2]zhongxue.chen@asu.edu

Abstract. While AI is vastly evolving, wherein deepfake techniques may be used to generate more realistic faces, voices, and videos, many deepfake-based fraudulent cases are increasingly occurring. To combat deepfake-based forgery, several methods have been proposed wherein the most astonishing methods are based on convolution neural network (CNN). However, most intelligent detection systems are underrepresenting in exposing the deepfake images under anti-forensics attacks, e.g., rescaling the image, inserting noises, and compressing the image again. To our knowledge, it still falls short of an intelligent detection system being able to detect deepfake and other advanced image forgery together. Additionally, it falls short of a comprehensive comparison study on the latest deep learning models for the deepfake detection.

In this study, we apply the latest deep learning models for deepfake detection under pos anti-forensics processing mixed with seam-carving and copy-move forgery images in JPEG. Our study shows that different deep learning models have different distinction capability. Experimental results show that some latest deep learning models are effective in detecting deepfake images under post anti-forensics processing in JPEG images, they are also performing well in detecting seam-carving and copy-move forgery. Our study also shows that it is relatively easy to detect deepfake compared to the detection of seam carving forgery detection under anti-forensics processing in JPEG images.

Keywords: Deepfake, adversarial, anti-forensics, seam-carving, copy-move, image forgery, deep learning, JPEG

1 Introduction

The rapid advancements in artificial intelligence and deep learning technologies have given rise to sophisticated image manipulation. As these techniques become more pervasive, AI-based cybercrime and fraudulent cases are frequently occurring, posing substantial ethical and security challenges that need to be addressed urgently.

Deepfake technologies have demonstrated the ability to generate hyper-realistic synthetic images and videos that are nearly indistinguishable from the real ones (Grother et al. 2019). These technologies exploit deep learning frameworks such as generative adversarial networks (GANs) (Goodfellow et al. 2014) to create or modify content in a manner that seems incredibly real (Lim and Suh 2020). While initially developed for benign purposes such as entertainment and art, these technologies are increasingly used maliciously for spreading misinformation, identity theft, and breaching privacy (Bossier et al. 2019). Many deepfake tools can manipulate multimedia content effortlessly, paving the way for the dissemination of disinformation across digital platforms. For example, Midjourney, an AI image generator that creates realistic deepfakes, has been scrutinized recently for having a policy showing deference to China's government (foxnews 2023). Amidst the high-profile diplomatic talks, a picture on social media, showing Russia's president bowing down to China' leader and kissing his hand, has sparked various reactions and debates online, with many questions the authenticity of the photograph (timesnownews 2023).

Multimedia content has become a standard form of evidence across various legal sectors, emphasizing the need for the content to be authentic and verifiable. The introduction of user-friendly manipulation tools such as Zao (Zao App 2019), REFACE (Reface App 2019), FaceApp (FaceApp 2017), Audacity (audacityteam 2021), and Soundforge (Migax 2021), has amplified the perceived realism of fabricated data, thereby complicating the verification process of content's authenticity and integrity.

Similarly, seam-carving and copy-move manipulation techniques have shown their potential to tamper with digital media. Seam carving is a technique used for content-aware image resizing (Avidan and Shamir 2007). It preserves the most important visual features while altering the image's size or aspect ratio, which could be used for concealing essential elements in an image. Copy-move forgeries, on the other hand, involve duplicating or moving a part of an image within the same image, making it a potent tool for deceitful purposes (Huang and Ciou 2019).

Deepfakes, along with other forms of manipulation, e.g., seam-carving and copy-move, are projected to become commonplace tools of disinformation, potentially undermining trust in state institutions, electronic media, and more due to the public's difficulty discerning between genuine and manipulated videos. It seems that we are now immersed in a "post-truth" era where malevolent actors utilize misinformation to shape public perception, resulting in threats as severe as election tampering, public defamation, and even the incitement of conflict. The manipulated media could be leveraged to spread disinformation globally, posing a serious threat to the integrity of news and information if not properly addressed.

2 Relevant Study

Deepfake is generated and evolved with deep learning techniques. To detect deepfakes, one of the prominent research directions relies on deep learning itself. Various studies have proposed novel approaches using convolutional neural networks (CNNs), recurrent neural networks (RNNs), and/or other deep learning architectures to distinguish between real and manipulated media. For instance, Zhang and Liu (2020) exposed a popular open-source video forgery library called “DeepFaceLab” and FaceForensics dataset by making use of deep learning. Li et al. (2020) introduced a deep learning framework based on two-stream networks to effectively detect deepfake videos by capturing both spatial and temporal information. Another study by Zhou et al. (2020) employed a combination of CNN and RNN models to detect deepfake images by learning discriminative features from visual content. Lyu (2020) discussed a few of the deepfake challenges as well as the research opportunities.

To enhance the performance of deepfake detection algorithms, researchers have explored the utilization of additional modalities, such as audio and metadata. Liu et al. (2020) proposed a multimodal fusion framework that combines visual features from deepfake images with audio features extracted from corresponding audio recordings. By incorporating audio information, their method achieved improved accuracy in detecting deepfake media. Similarly, Nguyen et al. (2021) proposed a method that utilizes metadata information, such as device-specific traces and compression artifacts, to identify deepfake videos. Unfortunately, these methods become void while untouched media contents are processed by regular multimedia processing, and processed media could be falsely detected as deepfake.

While generative adversarial networks (GANs) are used for creating deepfakes, GAN-based artifacts are also remained with manipulated multimedia data. Several studies have focused on exploiting artifacts and inconsistencies introduced by GAN-based generation processes to detect deepfakes. For example, Sabir et al. (2021) proposed a deepfake detection method that leverages the concept of residual noise patterns in GAN-generated images. Their approach effectively identified manipulated images by analyzing residual artifacts left by the GAN generator. Additionally, Zhang et al. (2021) proposed a deep learning-based approach that analyzes the distribution of face landmarks to distinguish between real and deepfake videos.

By providing interpretability to the detection models, it becomes easier to identify the specific visual cues or features that contribute to the decision-making process. For instance, Jiang et al. (2022) introduced an explainable deepfake detection framework that combines the interpretability of saliency maps with the discriminative power of CNNs. Their method not only achieved high detection accuracy but also provided insights into the regions of interest in deepfake images.

Recently, Masood et al. (2023) made a comprehensive survey on deepfake, primarily focusing on the detection of deepfake images and videos. It provides a comprehensive review and detailed analysis of existing tools and machine learning-based approaches for deepfake generation, and the methodologies used to detect such manipulations in both audio and video. For each category of deepfake, it provides information related to manipulation approaches, current public datasets, and key standards for the evaluation of the performance of deepfake detection

techniques, along with their results. Additionally, it also discusses open challenges and enumerates future directions to guide researchers.

Unfortunately, although many publications are available in tackling with deepfake detection, it still falls short of the investigation to expose the deepfake forgery followed by anti-forensics processing, e.g., image crop, rescale, noise addition, and JPEG recompression, which aims to remove or cover the deepfake-based artifact traces, in such conditions, many deepfake detection approaches become powerless.

To our knowledge, there is no investigation aiming to detect deepfake forgery and other types of image forgery, for example, seam-carving, and copy-move manipulation together under anti-forensics adversarial processing.

On the other side, it has seen significant progress in the development of deep learning models such as BEIT (Bao et al 2021), ConvNext (Liu et al. 2022), FlexViT (Beyer et al. 2023). These models could provide profound insights into image manipulations and might have significantly enhanced our ability to detect multiple types of image forgery, mitigating their impact on security, privacy, and misinformation spread. However, most of these models have not been fully investigated in image forgery detection.

In response to these challenges, it has shown the potential of leveraging transfer learning and adversarial training to improve the performance of detection models. In this study, we apply the SOTA deep learning models to expose deepfake under anti-forensics adversarial manipulation mixed with seam-carving and copy-move forgery in JPEG images. The remainder of the paper is organized in this way: next section is a brief introduction on the recent advance in deep learning CNN models, followed by our proposed study, experimental results, and conclusions.

3 Latest Deep Learning Models

With the advance in deep learning, Bao et al. (2021) furthered this advancement with their BEIT model, employing the Bidirectional Encoder Representation from Transformers (BERT) for self-supervised vision representation. The BEIT model introduced a novel masked image modeling task for pretraining vision Transformers, showing impressive results on image classification and semantic segmentation tasks. BEITv2,

a model proposed by Peng et al. (2022), enhanced Masked Image Modeling (MIM) from pixel-level to semantic-level, playing a significant role in the field.

Fang et al. (2023) introduced the EVA model, a Vanilla ViT pre-training technique, which has set new benchmarks across various vision downstream tasks.

Dosovitskiy et al.'s Vision Transformer (ViT) (2020) applies transformer architecture directly to sequences of image patches, leading to substantial results when pre-trained on large amounts of data. This architecture provides a significant shift in image recognition benchmarks, which have been previously dominated by convolutional networks.

Liu et al. (2022) reexamined the design spaces of Convolutional Neural Networks (ConvNets) and tested their potential. Their work, leading to the ConvNeXt model family, has demonstrated that these purely ConvNet models can compete favorably with Transformers in terms of accuracy and scalability while maintaining the simplicity and efficiency of standard ConvNets.

Beyer et al. (2023) proposed FlexiViT, a novel approach in the Vision Transformers landscape, which allows the model to perform efficiently across different computing budgets at deployment time. By simply randomizing the patch size at training time, FlexiViT achieves competitive performance across a wide range of tasks, making it a substantial contribution in the Vision Transformers domain.

Tu et al. (2022) proposed Multi-Axis Vision Transformer, MaxViT, which introduces an efficient and scalable attention model, which includes blocked local and dilated global attention. This design enables global-local spatial interactions on arbitrary input resolutions with only linear complexity. MaxViT performs exceptionally on various vision tasks, demonstrating the potential of their proposed model as a universal vision module.

The EfficientNetV2 model, developed by Tan and Le (2021), enhances both the training speed and the parameter efficiency of convolutional networks. This performance optimization was achieved using a blend of training-aware neural architecture search and scaling, together with a search space enriched with advanced operations such as Fused-MBConv. The EfficientNetV2 models outperform previous

architectures, offering faster training times and up to 6.8 times smaller model sizes. A unique feature of EfficientNetV2 is its use of progressive learning, where the image size is gradually increased during training. This approach, while generally enhancing training speed, often results in a slight decrease in accuracy. To offset this, Tan and Le designed an improved method of progressive learning that adaptively adjusts regularization (e.g., data augmentation) in conjunction with image size.

The concept of parameter transfer between different architectures is explored by Czyzewski (2021). In this work, a simple algorithm is proposed to leverage a computationally inexpensive injection technique, which does not require data, for parameter transfer. The primary aim of this algorithm is to expedite the training process of neural networks from scratch. The findings indicate that transferring knowledge from any architecture is superior to traditional initialization methods such as Kaiming and Xavier initialization. This method not only converges faster but also provides a simple drop-in replacement for classical initialization techniques. The procedure involves two main steps: matching, where the layers of the pre-trained model are matched with the target model, and injection, where the tensor is transformed into the desired shape. Additionally, this work introduces the Transfer Learning by Injection (TLI) score as a measure to compare the similarity between current state-of-the-art architectures on ImageNet.

Dai et al. (2021) proposed the Neural Architecture-Recipe Search (NARS). Unlike traditional Neural Architecture Search (NAS) methods which typically search for architectures under a single set of training hyperparameters (also known as a training recipe), NARS aims to simultaneously find both the architectures and their corresponding training recipes. The approach makes use of an accuracy predictor that can evaluate both architecture and training recipes jointly. This dual-capability predictor is crucial in guiding sample selection and ranking. With the goal of compensating for the increased search space, it also utilizes the "free" architecture statistics such as FLOP count to pretrain the predictor, thereby enhancing its sample efficiency and prediction reliability. The trained predictor is then utilized in fast evolutionary searches to produce architecture-recipe pairs for various resource constraints, resulting in the creation of FBNetV3, within the state-of-the-art compact neural

networks that exceed the performance of both automatically and manually designed competitors. In terms of practical examples, FBNetV3 achieves parity with both EfficientNet and ResNeSt on ImageNet accuracy while using up to 2.0x and 7.1x fewer FLOPs, respectively. Furthermore, FBNetV3 also exhibits notable performance gains in downstream object detection tasks, improving mean average precision (mAP) despite having 18% fewer FLOPs and 34% fewer parameters than equivalent EfficientNet-based models.

Tan and Le (2019) designed MixConv, a novel approach for mixed depth-wise convolutional kernels. While depthwise convolution has gained popularity in efficient ConvNets, the importance of kernel size is often overlooked. The authors systematically investigate the impact of different kernel sizes and discover that combining multiple kernel sizes can yield improved accuracy and efficiency. Based on this observation, they proposed MixConv, which naturally integrates multiple kernel sizes within a single convolutional operation. MixConv serves as a straightforward replacement for vanilla depthwise convolution, enhancing the accuracy and efficiency of existing MobileNet models for ImageNet classification and COCO object detection tasks. To showcase the effectiveness of MixConv, the authors integrated it into the AutoML search space and introduced a new family of models called MixNets. These models outperform previous mobile models, including MobileNetV2, ShuffleNetV2, MnasNet, ProxylessNAS, and FBNet, achieving superior results. Notably, MixNet-L achieves a new state-of-the-art 78.9% top-1 accuracy on ImageNet under typical mobile settings (<600M FLOPS).

Howard et al. (2019) introduced MobileNetV3 as the next generation of MobileNets. They employed a combination of complementary search techniques and innovative architectural designs to optimize MobileNetV3 specifically for mobile phone CPUs. The development of MobileNetV3 involves hardware-aware network architecture search (NAS) and the utilization of the NetAdapt algorithm. These approaches are followed by architectural advancements, resulting in improved performance and state-of-the-art outcomes. The authors introduce two variants of MobileNetV3: MobileNetV3-Large and MobileNetV3-Small, designed to address high and low resource usage scenarios, respectively. The MobileNetV3 models are further adapted for object detection and

semantic segmentation tasks. For semantic segmentation, the authors propose a novel efficient segmentation decoder known as Lite Reduced Atrous Spatial Pyramid Pooling (LR-ASPP). Experimental results demonstrate the superior performance of MobileNetV3 across various mobile classification, detection, and segmentation tasks. MobileNetV3-Large surpasses MobileNetV2 with a 3.2% accuracy improvement on ImageNet classification, while reducing latency by 20%. MobileNetV3-Small achieves a 6.6% accuracy gain compared to a comparable MobileNetV2 model at similar latency. In terms of detection, MobileNetV3-Large exhibits over 25% faster speed than MobileNetV2 on COCO detection with similar accuracy. Additionally, MobileNetV3-Large LR-ASPP achieves a 34% speed enhancement over MobileNetV2 R-ASPP for Cityscapes segmentation tasks.

By incorporating these models into comprehensive detection frameworks, it is possible to enhance the capabilities of deepfake image detection, seam-carving detection, and copy-move detection under anti-forensics adversarial manipulation, ultimately contributing to the development of robust forensic tools for combating visual manipulation.

4 Detection Approaches

4.1 Deep Learning Models

The ImageNet Large Scale Visual Recognition Challenge (Russakovsky et al. 2014) evaluates algorithms for object detection and image classification at large scale. It contains hundreds and thousands of images, which have been instrumental in advancing computer vision and deep learning research. PyTorch Image Models (timm) is a library for state-of-the-art image classification, containing a collection of image models, optimizers, schedulers, augmentations, and training/validating scripts with ability to reproduce ImageNet training result (Wightman 2019). It contains validation and benchmark results for the models in this collection. We selected the following top-ranking models in the classification of ImageNet dataset, a total of 19 models, listed in Table 1, for our study in detecting deepfake, seam-carving, and copy-move forgery in JPEG images together.

Table 1. PyTorch Image Models (timm) models (Wightman 2019) in our study.

MODEL-NO	MODEL NAME
1	eva_large_patch14_196.in22k_ft_in22k_in1k
2	beitv2_large_patch16_224.in1k_ft_in22k_in1k
3	vit_large_patch14_clip_224.openai_ft_in12k_in1k
4	beit_large_patch16_224.in22k_ft_in22k_in1k
5	convnext_large.fb_in22k_ft_in1k
6	vit_base_patch8_224.augreg2_in21k_ft_in1k
7	flexivit_large.1200ep_in1k
8	beit_base_patch16_224.in22k_ft_in22k_in1k
9	vit_relpos_base_patch16_clsgap_224.sw_in1k
10	gc_efficientnetv2_rw_t.agc_in1k
11	tf_efficientnet_b2.ns_jft_in1k
12	efficientnetv2_rw_t.ra2_in1k
13	vit_srelpos_medium_patch16_224.sw_in1k
14	fbnetv3_g.ra2_in1k
15	vit_small_r26_s32_224.augreg_in21k_ft_in1k
16	mixnet_xl.ra_in1k
17	tf_mixnet_l.in1k
18	mobilenetv3_large_100.miil_in21k_ft_in1k
19	tinynet_a.in1k

4.2 Optimizer and Loss Function

The pre-trained pytorch image models (Wightman 2019) were trained with stochastic gradient descent (SGD) optimizer. Since we apply fine-tune method to the pre-trained models and retrain the models by using our image forgery data. Additionally, it shows that SGD optimizer performs well and stable compared other optimizers (Liu and Chen et al. 2022), therefore, in this study, we adopt SGD optimizer, and describe below.

Consider the object function,

$$Q(w) = \frac{1}{n}\sum_{i=1}^{n} Q_i(w) \tag{1}$$

Where the parameter w that minimizes Q(w) is to be estimated. Each summand function Q_i is typically associated with the i-th observation in the dataset for training.

To minimize the object function, a standard gradient descent method is performed in the following iterations:

$$w := w - \eta \nabla Q(w) = w - \frac{\eta}{n} \sum_{i=1}^{n} \nabla Q_i(w) \tag{2}$$

Where η is a step size or called the learning rate.

Stochastic gradient descent with momentum remembers the update Δw at each iteration, and determines the next update as a linear combination of the gradient and the pervious update:

$$\begin{aligned} \Delta w &:= \alpha \Delta w - \eta \nabla Q_i(w) \\ \mathrm{w} &:= \mathrm{w} + \Delta \mathrm{w} \end{aligned} \tag{3}$$

That leads to:

$$w := w - \eta \nabla Q_i(w) + \alpha \Delta \mathrm{w} \tag{4}$$

Where the parameter w which minimize Q(w) is to be estimated, η is a step size or called the learning rate, and αis an exponential decay factor between 0 and 1 that determines the relative contribution of the current gradient and earlier gradients to the weight change.

Cross entropy is frequently used for loss function. For a multiple-class classification, the loss function is given by,

$$loss(x, y) = -\sum x \log(y) \tag{5}$$

Where y is the predicted probability and x is probability of true label.

4.3 Image Forgery Datasets

4.3.1 Seam-carving image dataset and processing

We adopt the seam-carving image dataset that was used in the previous study (Liu and Chen 2015, Liu 2017, Liu 2019, and Celebi et al. 2022), wherein the 7837 untouched JPEG images in quality Q75, and 7837 seam-carving manipulated JPEG images in quality Q75 are utilized. Each JPEG image is 512×512. We process these JPEG images in the following ways.

1). Crop each 512×512 JPEG image, the center 256×256 is saved in JPEG at the quality Q75. We obtain the images in class label 0 and class label 1 in Table 2.

2) Rescale each 512×512 JPEG image to 256×256 and resave the processed image data in JPEG at the quality Q75. We obtain the images in class label 2 and class label 3 in Table 2.

3) Randomly crop each 512×512 JPEG image from the four corners, rescale each cropped image to 256×256, and resave the processed image

data in JPEG at the quality Q75. We obtain the images in class label 4 and class label 5 in Table 2.

4.3.2 Copy-move image dataset and processing

The dataset (Mahfoudi et al. 2019) contains about 19000 copy-move forgeries, which are available under the copymove_img directory. Each copy-move is accompanied by two binary masks. One under the probe_mask subdirectory indicates the location of the forgery and one under the donor_mask indicates the location of the source within the image. Based on the probe-mask data, we extract the original object image, and pasted object image, each image is rescaled to the size 256×256, and saved in JPEG at the quality Q75, assigned the class labels 6 and 7 in Table 2.

4.3.3 Deepfake image dataset and processing

A Kaggle deepfake image dataset contains manipulated images and real images (Le et al. 2021). The manipulated images are the faces which are created by various means. The source for this dataset is https://zenodo.org/record/5528418#.YpdlS2hBzDd. Each image is a 256×256 image of a human face either real or fake. All these images are saved in JPEG at quality Q75, assigned the class label 8 (real) and class label 9 (fake) in Table 2.

Additionally, we randomly crop each JPEG image from the four corners, rescale the image to 128×128, then rescale to 256×256, and finally saved in JPEG at quality Q75, assigned to the class label 10 (real) and class label 11(fake) in Table 2.

4.3.4 Noise Addition

In addition to the above processing, we also randomly add gaussian, speckle, and passion noise to the class labels 4, 5, 10, and 11 in Table 2, these images are saved as the new class labels 4, 5, 10, and 11 in Table 3.

Table 2. Image classes under post anti-forensics in JPEG (type 1).

CLASS	QUANTITY	IMAGE PROCESSING
0	7837	Untouched, center crop 256×256, recompression

1	7837	Seam-carving, center crop 256×256, recompression
2	7837	Untouched, rescale to 256×256, recompression
3	7837	Seam-carving, rescale to 256×256, recompression
4	7837	Untouched, random crop, rescale to 128×128, and rescale to 256×256, then recompression
5	7837	Seam-carving, random crop, rescale to 128×128, and rescale to 256×256, then recompression
6	19086	Untouched, rescale to 256×256, then recompression
7	18923	Copy-moved, rescale to 256×256, then recompression
8	70001	Real, rescale to 256×256, recompression
9	70001	fake, rescale to 256×256, recompression
10	31120	Real, random crop, rescale to 128×128, and rescale to 256×256, then recompression
11	21978	fake, random crop, rescale to 128×128, and rescale to 256×256, then recompression

Table 3. Image classes under post anti-forensics with noise addition in classes 4, 5, 10, and 11 in JPEG (type 2).

CLASS	QUANTITY	IMAGE PROCESSING
0	7837	Untouched, center crop 256×256, recompression
1	7837	Seam-carving, center crop 256×256, recompression
2	7837	Untouched, rescale to 256×256, recompression
3	7837	Seam-carving, rescale to 256×256, recompression

4	7837	Untouched, random crop, rescale to 128×128, and rescale to 256×256, noise addition, then recompression
5	7837	Seam-carving, random crop, rescale to 128×128, and rescale to 256×256, noise addition, then recompression
6	19086	Untouched, rescale to 256×256, then recompression
7	18923	Copy-moved, rescale to 256×256, then recompression
8	70001	Real, rescale to 256×256, recompression
9	70001	fake, rescale to 256×256, recompression
10	31226	Real, random crop, rescale to 128×128, and rescale to 256×256, noise addition, then recompression
11	32169	fake, random crop, rescale to 128×128, and rescale to 256×256, noise addition, then recompression

Some examples from each class in Table 2 are given in Figure 1, and the noisy examples from Table 3 are given in Figure 2.

Figure 1. Image examples from Table 2. They are sequentially extracted from class 0 to class 11, in the order from the left to right, top to the bottom.

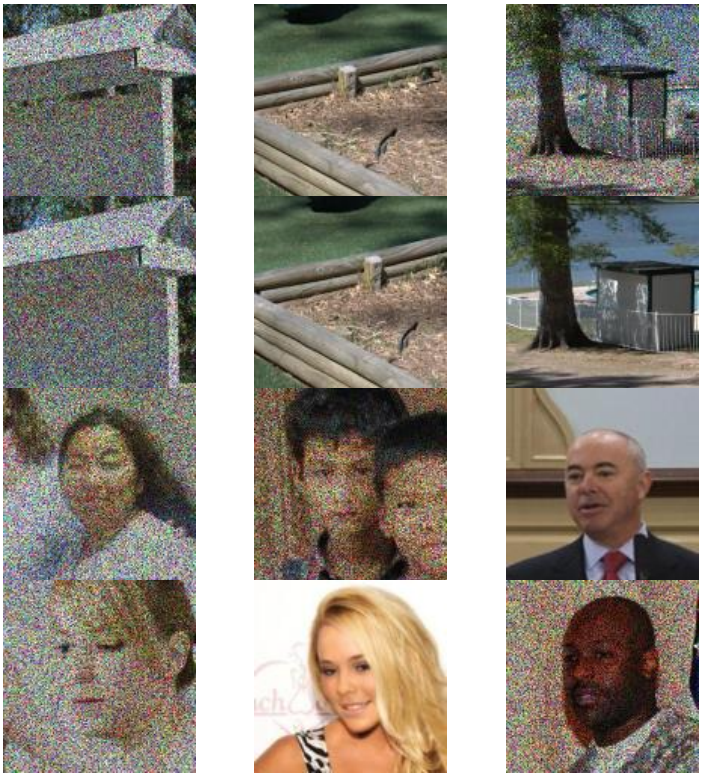

Figure 2. Image examples from Table 3. The first-row images are extracted from class 4 (untouched), the second row from class 5 (seam-carving), the third from class 10 (real), and the fourth from class 11 (fake).

5 Experimental Results

In each experiment, we randomly selected 64% from each class for training and 16% for validation, and the remaining 20% for testing. The same training, validation, and testing datasets are applied to each fine-tuning deep learning model. We selected the initial learning rate 0.001 and the cross-entropy loss. The dataset contains imbalanced classes, while we randomly selected samples for training, different weights are assigned to the twelve classes, computing by the total number of all class images over the number of each class. We ran 10 experiments for each model. Table 4 shows the mean testing accuracy in correctly detecting the 12 class images, which are described in Table 2. In Table 4, the top 3 testing accuracy values are highlighted in bold.

Table 4. Mean detection accuracy on the 12 class data sets (type 1).

MODEL NO	MEAN TESTING ACCURACY (%)
1	**91.63**
2	90.96
3	86.41
4	88.74
5	**93.91**
6	90.82
7	**91.41**
8	88.43
9	90.98
10	90.50
11	90.65
12	90.48
13	90.50
14	**91.25**
15	85.17
16	89.29
17	89.01
18	87.88
19	87.94

Table 5 shows the mean testing accuracy in correctly detecting the 12 class images, described in Table 3.

Table 5. Mean detection accuracy on the 12 class data sets (type 2).

MODEL-NO	MEAN TESTING ACCURACY (%)
1	88.80
2	88.78
3	86.13
4	88.94
5	**93.00**
6	89.50
7	**91.07**
8	88.31
9	89.98
10	**90.28**
11	89.50
12	89.89
13	89.84
14	**90.47**
15	82.96
16	87.64
17	87.35
18	86.29
19	86.71

Comparing the results, among the 19 models, it is clear that the model convnext_large.fb_in22k_ft_in1k outperforms others, followed by flexivit_large.1200ep_in1k and fbnetv3_g.ra2_in1k model. Even under rescaling and noise addition, most models are effective in distinguishing deepfake images from the real images. However, it is more challenging in distinguishing seam-carving images from untouched under anti-forensics manipulations, especially in classifying class 5 and class 4.

6 Conclusions

In this study, we apply nineteen deep learning models within the state-of-the-art for deepfake detection under post anti-forensics processing mixed with seam-carving and copy-move forgery images in JPEG. Our study shows that different deep learning models have different capabilities. Experimental results show that some latest deep learning models are

effective in detecting deepfake images under post anti-forensics processing they are also performing well in detecting seam-carving and copy-move forgery, however, it is more challenging to detect seam carving forgery detection under anti-forensics processing in JPEG images. Among the nineteen deep learning models, in general, a convnext large model outperforms others, which is more effective in distinguishing the three different types of forgery under anti-forensics manipulations in JPEG images.

Acknowledgements

The support for this study from the SHSU Institute of Homeland Security is highly appreciated.

References

audacityteam (2021). Audacity Software. Retrieved from https://www.audacityteam.org/

Bao, H., Dong, L., & Wei, F. (2021). BEiT: BERT Pre-Training of Image Transformers. ArXiv, abs/2106.08254.

Beyer, L., Izmailov, P., Kolesnikov, A., Caron, M., Kornblith, S., Zhai, X., Minderer, M., Tschannen, M., Alabdulmohsin, I., Pavetic, F. (2023). FlexiViT: One Model for All Patch Sizes. In Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR), 2023.

Celebi NH, Hsu TL, Liu Q (2022). A comparison study to detect seam carving forgery in JPEG images with deep learning models. Journal of Surveillance, Security and Safety. 2022; 3(3):88-100. http://dx.doi.org/10.20517/jsss.2022.02

Czyzewski, M.A. (2021). Transfer Learning Between Different Architectures Via Weights Injection. ArXiv, abs/2101.02757.

Dai, X., Wan, A., Zhang, P., Wu, B., He, Z., Wei, Z., Chen, K., Tian, Y., Yu, M., Vajda, P., Gonzalez, J. E. (2021). FBNetV3: Joint Architecture-Recipe Search using Predictor Pretraining, IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Pages: 16271-16280, DOI Bookmark: 10.1109/CVPR46437.2021.01601

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2020). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. ArXiv, abs/2010.11929.

FaceApp. (2017). FaceApp. Retrieved from https://www.faceapp.com/

Fang, Y., Wang, W., Xie, B., Sun, Q., Wu, L., Wang, X., Huang, T., Wang, X., & Cao, Y. (2023). EVA: Exploring the Limits of Masked Visual Representation Learning at Scale. In Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR), 2023, https://openaccess.thecvf.com/content/CVPR2023/papers/Fang_EVA_Exploring_the_Limits_of_Masked_Visual_Representation_Learning_at_CVPR_2023_paper.pdf

Foxnews (2023) https://www.foxnews.com/tech/ai-image-generator-midjourney-bans-deepfakes-china-xi-jinping-minimize-drama, accessed on May 26, 2023.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. Advances in Neural Information Processing Systems, 27.

Grother, P., Ngan, M., & Hanaoka, K. (2019). Face Recognition Vendor Test (FRVT) Part 3: Demographic Effects. NIST Interagency Report, 8280. https://doi.org/10.6028/NIST.IR.8280

Howard, A., Sandler, M., Chu, G., Chen, L.-C., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V., Le, Q. V., Adam, H. (2019). Searching for MobileNetV3. 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 1314-1324.

Huang, HY., Ciou, AJ (2019). Copy-move forgery detection for image forensics using the superpixel segmentation and the Helmert transformation. J Image Video Proc. 2019, 68 (2019). https://doi.org/10.1186/s13640-019-0469-9

Independent (2023). https://www.independent.co.uk/news/ap-deepfake-porn-new-york-meta-b2320775.html, accessed on May 26, 2023.

Jiang, Z., Bai, X., & Creusot, C. (2022). Explainable Deepfake Detection via Saliency-Guided CNNs. IEEE Transactions on Information Forensics and Security, 17, 598-612.

Le, T., Nguyen, H.H., Yamagishi, J., & Echizen, I. (2021). OpenForensics: Large-Scale Challenging Dataset For Multi-Face Forgery Detection And Segmentation In-The-Wild. 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 10097-10107.

Li Y, Chang MC, Lyu S and Bao C (2020). Two-Stream Temporal Convolutional Networks for Deepfake Detection. arXiv preprint arXiv:2011.10278.

Lim, K. and Suh, J. (2020). Detection of Deepfake Images Based on Pairwise Learning. arXiv preprint arXiv:2011.04098.

Liu Q (2017). An approach to detecting JPEG down-recompression and seam carving forgery under recompression anti-forensics. Pattern Recognition, vol. 65, pp. 35-46, ISSN 0031-3203, https://doi.org/10.1016/j.patcog.2016.12.010.

Liu Q (2019). An Improved Approach to Exposing JPEG Seam Carving Under Recompression. IEEE Transactions on Circuits and Systems for Video Technology, vol. 29, no. 7, pp. 1907-1918, July 2019, doi: 10.1109/TCSVT.2018.2859633.

Liu, C., Cao, X., Luo, J., & Liu, Q. (2020). Multimodal Deepfake Detection Using Audio-Visual Features. arXiv preprint arXiv:2008.10625.

Liu Q and Chen Z (2015). Improved Approaches with Calibrated Neighboring Joint Density to Steganalysis and Seam-Carved Forgery Detection in JPEG Images. ACM Trans. Intell. Syst. Technol. 5, 4, Article 63 (January 2015), 30 pages. https://doi.org/10.1145/2560365

Liu Q, Chen Z and Liu HC (2022). A Comparison Study to Detect COVID-19 Chest X-Ray Images with SOTA Deep Learning Models. Proceedings of the 1st Workshop on Healthcare AI and COVID-19, ICML 2022, in Proceedings of Machine Learning Research 184:146-153.

Liu, Z., Mao, H., Wu, C., Feichtenhofer, C., Darrell, T., & Xie, S. (2022). A ConvNet for the 2020s. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 11966-11976.

Lyu S (2020). Deepfake Detection: Current Challenges and Next Steps. 2020 IEEE International Conference on Multimedia & Expo Workshops (ICMEW), London, UK, 2020, pp. 1-6, doi: 10.1109/ICMEW46912.2020.9105991.

MAHFOUDI G, TAJINI B, RETRAINT F, MORAIN-NICOLIER F, DUGELAY JL and PIC M (2019). DEFACTO: Image and Face Manipulation Dataset. 2019 27th European Signal Processing Conference (EUSIPCO), A Coruna, Spain, 2019, pp. 1-5, doi: 10.23919/EUSIPCO.2019.8903181.

Magix. (2021). Sound Forge Audio Studio. Retrieved from https://www.magix.com/us/music/sound-forge/sound-forge-audio-studio/

Masood, M., Nawaz, M., Malik, K.M. et al. (2023). Deepfakes generation and detection: state-of-the-art, open challenges, countermeasures, and way forward. Appl Intell 53, 3974–4026. https://doi.org/10.1007/s10489-022-03766-z

Nguyen, H. H., Tran, L. T., & Nguyen, T. Q. (2021). Deepfake Video Detection using Device-Specific Traces and Compression Artifacts. IEEE Transactions on Information Forensics and Security, 16, 6629-6644.

Peng, Z., Dong, L., Bao, H., Ye, Q., & Wei, F. (2022). BEiT v2: Masked Image Modeling with Vector-Quantized Visual Tokenizers. ArXiv, abs/2208.06366.

Reface App. (2019). Reface App. Retrieved from https://www.reface.app/

Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M. (2019). FaceForensics++: Learning to Detect Manipulated Facial Images. International Conference on Computer Vision (ICCV).

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M.S., Berg, A.C., & Fei-Fei, L. (2014). ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision, 115, 211-252.

Sabir, E., Cholakkal, H., & Marcel, S. (2021). Detection of Deepfake Images Using Inconsistencies Arising from Residual Noise. IEEE Transactions on Information Forensics and Security, 16, 2627-2641.

Shai Avidan and Ariel Shamir. 2007. Seam carving for content-aware image resizing. In ACM SIGGRAPH 2007 papers (SIGGRAPH '07). Association for Computing Machinery, New York, NY, USA, 10–es. https://doi.org/10.1145/1275808.1276390

Tan, M., & Le, Q.V. (2019). MixConv: Mixed Depthwise Convolutional Kernels. ArXiv, abs/1907.09595.

Tan, M., & Le, Q. V. (2021). EfficientNetV2: Smaller Models and Faster Training. ICML 2021: 10096-10106

Timesnownews (2023). https://www.timesnownews.com/technology-science/exposed-ai-generated-viral-photo-of-putin-bowing-to-xi-jinping-fact-or-fiction-article-98912114, accessed on May 26, 2023.

Tu, Z., Talebi, H., Zhang, H., Yang, F., Milanfar, P., Bovik, A.C., & Li, Y. (2022). MaxViT: Multi-Axis Vision Transformer. European Conference on Computer Vision 2022. https://doi.org/10.48550/arXiv.2204.01697

Wightman R (2019). PyTorch Image Models, GitHub repository, doi :10.5281/zenodo.4414861.

Zao App. (2019). Zao App. Retrieved from https://www.zaoapp.net/

Zhang, Z., Liu, Q. (2020). Detect Video Forgery by Performing Transfer Learning on Deep Neural Network. In: Liu, Y., Wang, L., Zhao, L., Yu, Z. (eds) Advances in Natural Computation, Fuzzy Systems and Knowledge Discovery. ICNC-FSKD 2019. Advances in Intelligent Systems and Computing, vol 1075. Springer, Cham. https://doi.org/10.1007/978-3-030-32591-6_44

Zhang, C., Yao, Z., Zhang, Z., Xu, C., & Fu, Y. (2021). DeepFake Video Detection Based on Human Facial Landmark Distribution and Dual-Stream Spatio-Temporal Networks. IEEE Transactions on Information Forensics and Security, 16, 1615-1629.

Zhou, Y., Ye, Y., & Qiu, G. (2020). Deepfake Image Detection via Recurrent Neural Networks. IEEE Transactions on Information Forensics and Security, 15, 2447-2462.

INSTITUTE FOR HOMELAND SECURITY

Sam Houston State University

The Institute for Homeland Security at Sam Houston State University is focused on building strategic partnerships between public and private organizations through education and applied research ventures in the critical infrastructure sectors of Transportation, Energy, Chemical, Healthcare, and Public Health.

The Institute is a center for strategic thought with the goal of contributing to the security, resilience, and business continuity of these sectors from a Texas Homeland Security perspective. This is accomplished by facilitating collaboration activities, offering education programs, and conducting research to enhance the skills of practitioners specific to natural and human caused Homeland Security events.

Institute for Homeland Security
Sam Houston State University

Liu, Q., Celebi, N., Zhou, B., Chen, Z. (2023) Detecting Deepfakes under Anti-forensics Processing Mixed with Seam Carving and Copy-move Forgery in JPEG Images by Using SOTA Deep Learning Models. (Report No. IHS/CR-2023-1003). The Sam Houston State University Institute for Homeland Security.
https://doi.org/10.17605/OSF.IO/82UTW