A Comparative Review of Deep-Learning Models for Deepfakes Detection

Rémi Cogranne

A Comparative Review of Deep-Learning Models for Deepfakes Detection

Rémi Cogranne[a]

[a]Troyes University of Technology, Troyes, France

ABSTRACT

The development of generative AI has advanced significantly over the past few decades, enabling the creation of deepfake images that are increasingly difficult to distinguish from genuine photographs. The widespread availability of these models, which can be easily used, poses a substantial risk of spreading disinformation. Consequently, there is a pressing need for robust and reliable methods to identify images that have been created or altered using generative AI models. To address this need, a diverse range of methods and models have been developed. However, these approaches are often not exhaustively compared to one another, nor are they evaluated using a common reference dataset. Moreover, the majority of existing deepfake detection models rely on deep learning techniques, with numerous models available for feature extraction and detection, ranging from simple Convolutional Neural Networks (CNNs) to more advanced Vision Transformers (ViTs). To ensure the comparability and reproducibility of deepfake detection models, a standardized benchmark is urgently required to evaluate their performance across a large-scale, common dataset. Such a reference benchmark would facilitate the development of more effective and robust detection methods by providing insight into the strengths and weaknesses of existing AI-based approaches. In addition to establishing this benchmark, this paper also explores the challenges of combining different AI-based deepfake detection models and investigates various aggregation methods to further improve overall detection performance. A large-scale experiment involving almost 50 generative AI methods and over 40 deep learning-based feature extraction and detection models demonstrates the relevance of this study.

Keywords: Deepfake detection, Generative AI, Survey, Review, Benchmarking, Empirical evaluation, Artificial Inteligence, Image Forensics, Media Security

1. INTRODUCTION

Thanks to advances in artificial intelligence, AI-based image generation tools, including Deep Fakes, are now widely available to the public. These AI-generated images are so sophisticated that they can no longer be distinguished from natural photographs with the naked eye. Advanced statistical methods, particularly those based on Deep Learning, are required to detect them, but even these methods are not always accurate.

We've all seen fake images used to manipulate news on social media, and sometimes even on reputable news websites. This raises significant concerns about the credibility of digital photographs, which are often used as evidence to support factual claims. This issue is now widely recognized by policymakers, as highlighted in a report by the National Science and Technology Council.[1] As noted in a recent study,[2] the use of AI-generated images poses a particular threat in the context of foreign influence, as it enables the creation of "*seemingly-authentic and tailored messaging*."

To address these risks, researchers have developed various methods to detect images generated by AI.[3–5] However, most of these methods rely on Deep Learning, which, while effective, presents a challenge from a forensic perspective. This is because "black box methods" like Deep Learning and AI can lead to models that are difficult to explain and interpret.[6] To improve media forensic investigations, it's essential to enhance the traceability of AI decisions and develop more interpretable models."

Further author information: Send correspondence to remi.cogranne@utt.fr

This work has been funded by the French ANR PACeS project No. ANR-21-CE39-0002.

1.1 Position and Contribution of Present Paper

Recent deepfake detection methods have predominantly employed deep learning techniques, with a vast majority of approaches relying on these methods to identify and mitigate the spread of manipulated media. Even methods that focus on exploiting specific features, such as the upsampling operation of images generated by generative AI, often utilize a deep learning model as a detector. For instance, a custom CNN model was used in the method proposed by Nataraj et al.[7] to detect the specific upsampling operation, while Barni et al.[8] employed an ad-hoc CNN architecture to exploit the co-occurrence in spectral subband.

Reference and weel-know deeplearning models have been extensively used to address the problem of deepfake detection. For instance, the VGG network was used by Chang et al.[9] for detection of face images generated using GAN-based generative AI. Similarly Inception-ResNet-v2 was employed by Rajalaxmi et al.[10] for similar purposes of faces deepfake detection. Other deep learning architectures, such as ResNet50 and XceptionNet, have also been used for deepfake detection, as, for example, in the work of Coccomini et al.[11] and Wang et al.[12,13] EfficientNetB4, one of the variant of the EfficientNet architecture, has been used for fake video detection, image detection generated by GANs, and detection of images generated by diffusion models, as demonstrated in the work of Bonettini et al.,[14] Mandelli et al.,[15] and Mandelli et al.,[16] respectively.
The recent introduction of Vision Transformers has revolutionized the field of deep learning-based detection and feature extraction, and has been quickly adopted for deepfake detection, as, for instance, in the work of Wodajo et al.[17] which proposed a custom model. Additionally, an approach combining EfficientNet and transformers with an attention-based mechanism was leveraged by Coccomini et al.[18] for the detection of video deepfakes. Other transformers-based deeplearning architectures, such as MobileNetv2 block (MNV2)[19] for detecton of deepfake images generated by diffusion-based models. Similarly, RegNet was used in[20] for Deepfake detection with application on faces manipulation and maxVIT was used in conjunction with ResNet in[21] for exposing deepfakes in general. More recently, a Challenge on DeepFake Analysis and Detection (DFAD 2023) has been organized during the International Conference on Computer Vision (ICCV 2023) in Paris and the winner of this challenge used the well-celebrated swin transformer as originally proposed in.[22,23]

From the previous quick review of the state-of-the-art, one can conclude that that deeplearning-based method have been extensively used for all possible applications of deepfake detection. This trends raises fundamental questions for instance about the generalizability and robustness of these approaches and highlights the fundamental need for a more comprehensive comparison of these models in the context of deepfake detection Indeed, a thorough analysis of the state-of-the-art and detailed comparison of these models, including their strengths, weaknesses, and limitations. Moreover, the lack of a general comparison of deep learning for deepfake detection in general is striking. Most of the works we cited above are based on their own dataset on a very limited number of generative AI models. A more comprehensive evaluation of deep learning models against different deepfake generetion models would enable researchers to identify the most effective models and would possibly provide valuable insights into the most effective approaches for deepfake detection. Ultimately, the lack of a reference benchmark and comparison framework prevents the development of more effective and robust deepfake detection systems, and highlights the need for a more systematic and comprehensive evaluation of deep learning models in this context.

2. PRACTICAL METHDOLOGY AND EXPERIMENTAL FRAMEWORK

In our experimentation we generated 12.000 images from a wide range of generative AI models. For a numerical comparison of the performance of deep-learning based method for deepfake detection our goal was to create a dataset of deepfake images as large as possible. In total we ended up with a total of 48 classes of image based on 47 generative AI models plus one real photograph. For the real photograph we used a mix of ALASKA image dataset[24,25] and images collected from FlickR photo sharing platform and from pexels free stocks images website. The total number of real photograph was $200,000$ in the training set and $20,000$ in the testing set. The list of text-to-image generative models we used in detailed in the Table 1. We have tried to be as exhaustive as possible incling the first GAN-based image generator such as Glide or Lafite and the latest diffusion-tranformer based image generator such as stable diffusion 3.5 and FLUX.1. We used the generative AI-models as proposed in the Huggingface website (see Table 1 for the precise reference for all generators). The image were generated using a random choice from the One Million Random Midjourney Prompts which were collected from the Discord

channel interface for Midjourney. The guidance scales and number of diffusions steps were also randomized and all the other settings of the generator were left to defaulf. For each text-to-image generator we used 10.000 image for training and 2,000 images for testing.

Regarding the classification, all the models we used were obtained from the library timm[26] from which we loaded model pretrained over imagenet dataset because this generally greately improve training convergence. All the models were trained using a decreasing learning rate following stochastic gradient descent with warm restarts (SGDR).[27–29] Given the size of training dataset with 670.000 images, we used "only" 35 epochs. The initial value for the learning rate was obtained using from the method initially proposed in.[30] In brief, it essentially consists in a 1-cycle training of the deep-learning method over so-called mini-batches : the learning rate is gradually increased, at each mini-batch, from a very low initial value to a final high value. Throughout this process, the loss is measured at each iteration in order to find the largest value, with a margin value before the loss begins to diverge.
The learning rate scheduler is thus based on this initial guess and then slowing decreased over one cycle. The initial cycle length is set to 5, it is doubled for every cycle and the initial learning rate is divided by two after each cycle. We used three cycles for a total of 35 epochs, which is far enough for convergence, as shown in the next Section 3 and especially in the Figure 1.

Another important factor that we noticed in the importance of data augmention to prevent overfitting of the model and ensure a better generalization even though testing and training sets were generated in the very same manner ; there are in fact a random split from the same dataset. This fact has also been repported in.[18,31] In our case we carried out data augmentation by adding the following operation : mirroring, flipping along x and y axis, gaussien i.i.d. noise addition (with standard deviation between 0.05 and 0.15), multiplication noise addition (with factor between 0.975 and 1.025) resizing (with rescaling factor between 0.95 and 1.05) and rotation (with angle between -5 and 5 degrees). Each operation was applied independently, using the Albumentations library, and with a probability $p = 0.25$ for each. While each operation within the overall data augmentation process does not modify the image significanlty, and, from a quick search we did not found interest in applying more important operations, we have found that these processings were enough to greatly improve the testing accuracy.

2.1 List of Deep-Learning Models for Deepfakes Detection

The list of all 41 models we used for deepfake detection is given in the table 3 along with the year it was released, the link to the arxiv publication and the number of trainable parameter which gives a rough idea of the model complexity.
Without describing in detail the architecture of all these models we would like to briefly justify the choice behind this selection. The familly of Efficient Net models was included because it is widely recognized as a reference architecture and because it has been often used in the field of deepfake detection.[14–16] The idea behind efficient net is to find a tradeoff between increase in the deepth of the network, its resolution and the width of the architecture and to jointly optimize those parameters.
We also included NFNet and Mix-Net because these two are somewhat derivations from the EfficientNet approach and because, given the popularity of these models, it is unsurprising that they have been used for deepfake detection.[32]
Similarly, we used also several reference and well-known CNN models such as VGG, densenet, ResNet and Inception because both of their popularity and because of their used to address the problem of deepfake detection, see for instance[9,10,33,34] ; the interested reader is also referred to the survey papers.[35,36]
With a different goal in mind, we included the mobileNet family as it constitutes the reference lightweight and efficient Convolutional Neural Network (CNN) models. It has been been design for the objective deployment of deeplearning object detection and recognition, under challenging limitation of resource-constrained environment such as micro-controller with limited memory, energy, and power.

We also wanted to included the latest deeplearning features extraction and classification models based on vision transformer. These methods indeed have substantially improved the performance of deep learning models in general. Note that these methods have already been widely adopted for the problem of deepfake detection, see for instance the survey papers.[37,38] We included the instance of the well-known RegNet familly because of its popularity which makes it an unsurprisingly choice for deepfake detection.[20] Similarly we included the Twins

Type	Name	URL / Source
Transformers	FLUX.1-dev	HuggingFace
	FLUX.1-schnell	HuggingFace
	Lumina-T2I	HuggingFace
	Unidiffuser-v1	HuggingFace
	Stable Diffusion 3	HuggingFace
	Stable Diffusion 3.5	HuggingFace
Diffusion Models	Kandinsky v2.1	HuggingFace
	Kandinsky v2.2	HuggingFace
	Kandinsky v3	HuggingFace
	Pixart-α	HuggingFace
	Pixart-Σ	HuggingFace
	DreamLike 2.0	HuggingFace
	Playground 2.0	HuggingFace
	Playground 2.5	HuggingFace
	Stable Diffusion 1.5	HuggingFace
	Stable Diffusion 2.1	HuggingFace
	Stable Diffusion XL	HuggingFace
	Stable Diffusion XL-Turbo	HuggingFace
	Stable Diffusion Refiner	HuggingFace
	Stable Cascade	HuggingFace
	CogView 3	HuggingFace
	ConrolNetXS	HuggingFace
	Kolors	HuggingFace
	Wuerstchen v2	HuggingFace
	DeepFloyd-IF	HuggingFace
	Latent-Diff	GitHub
	ShifftedDiff	GitHub
GANs	StyleGAN3	GitHub
	StyleGAN2	GitHub
	GigaGAN	GitHub
	Glide	GitHub
	Lafite	GitHub
	Dall-E Mini	HuggingFace
	Dall-E Mega	HuggingFace
Finetuning / Variations	Stable-Diff Turbo (from SD 2.1)	HuggingFace
	animagine XL 3.0 (from SDXL)	HuggingFace
	lcm-LoRa SDXL Turbo	HuggingFace
	Fire Generation (from SDXL-Turbo)	HuggingFace
	DreamBooth (from SDXL-Turbo)	HuggingFace
	SDXL Turbo DPO LoRA (from SDXL-Turbo)	HuggingFace
	Pixart-α-LCM	HuggingFace
	Pixart-α-ControlNet	HuggingFace
	Pixart-α-DMD	HuggingFace
	Kolors-PAG	HuggingFace
	Jovie-Midjourney (from FLUX)	HuggingFace
	flux-schnell-realism (from FLUX)	HuggingFace
	Jovie-Retro (from FLUX)	HuggingFace

Table 1: List of generators and their accessibility.

Name	year	URL (arXiv)	# Parameters
EfficientNet2 m	2021	arXiv Link	∼ 20.2 M
EfficientNet2 s	2021	arXiv Link	∼ 52.9 M
MixNet	2019	arXiv Link	∼ 5.8 M
NFNet	2021	arXiv Link	∼ 32.8 M
TinyNet	2020	arXiv Link	∼ 4.9 M
ConvNeXt-V2	2023	arXiv Link	∼ 87.7 M
ResNeXt	2016	arXiv Link	∼ 192 M
ResNetClip 101	2021	arXiv Link	∼ 42.6 M
Res2NeXt 50	2019	arXiv Link	∼ 22.7 M
ResNet 101	2015	arXiv Link	∼ 42.6 M
SEResNext	2021	arXiv Link	∼ 91.6 M
VGG19	2014	arXiv Link	∼ 139 M
DenseNet	2016	arXiv Link	∼ 18.1 M
Xception	2016	arXiv Link	∼ 40.4 M
Inception-V3	2015	arXiv Link	∼ 21.9 M
Inception-ResNet-V2	2016	arXiv Link	∼ 54.4 M
Lambda ResNet50	2021	arXiv Link	∼ 19.6 M
MobileNet-V2	2018	arXiv Link	∼ 4.4 M
MobileNet-V3	2019	arXiv Link	∼ 4.2 M
MobileNet-V4	2024	arXiv Link	∼ 3.6 M
MobileViT-V2	2022	arXiv Link	∼ 4.4 M
EfficientViT b0	2023	arXiv Link	∼ 2.2 M
EfficientViT b3	2023	arXiv Link	∼ 46.2 M
MaxViT	2022	arXiv Link	∼ 119 M
CAFormer	2022	arXiv Link	∼ 95.8 M
ConvFormer base	2022	arXiv Link	∼ 96.9 M
ConvFormer medium	2022	arXiv Link	∼ 54.8 M
DaViT	2022	arXiv Link	∼ 87.0 M
mViT-v2	2021	arXiv Link	∼ 50.7 M
Next-ViT	2022	arXiv Link	∼ 43.8 M
RepViT	2023	arXiv Link	∼ 7.8 M
RepVGG	2021	arXiv Link	∼ 86.6 M
XCiT small	2021	arXiv Link	∼ 25.8 M
XCiT medium	2021	arXiv Link	∼ 83.8 M
RegNetY 160	2020	arXiv Link	∼ 80.7 M
RegNetY 320	2020	arXiv Link	∼ 141 M
MambaOut	2024	arXiv Link	∼ 81.9 M
BEiTv2	2022	arXiv Link	∼ 85.3 M
Twins pcpvt	2021	arXiv Link	∼ 43.3 M

Table 2: List of DL-based detectors.

Name	Test Accuracy	Epoch duration
EfficientNet b0	64.38	230
EfficientNet b3	69.47	780
EfficientNet2 m	85.30	1021
EfficientNet2 s	85.21	584
MixNet	83.89	566
NFNet	84.31	576
ECA-NFNet	85.18	1393
TinyNet	83.87	346
ConvNeXt-V2	85.34	2681
ResNeXt	56.35	4143
Res2NeXt 50	84.17	711
ResNet 101	84.39	854
SEResNext	85.08	2625
VGG19	75.13	806
DenseNet	84.36	834
Xception	84.58	1604
Inception-V3	81.66	360
Inception-ResNet-V2	83.11	725
Lambda ResNet50	84.65	664
MobileNet-V2	83.57	323
MobileNet-V3	82.36	436
MobileNet-V4	79.53	555
MobileViT-V2	84.65	463
EfficientViT b0	76.58	256
EfficientViT b3	83.25	1100
MaxViT	84.13	3350
CAFormer	85.96	3681
ConvFormer base	85.44	3216
ConvFormer medium	85.53	6102
DaViT	85.23	2497
mViT-v2	85.09	1893
Next-ViT	84.78	1076
RepViT	84.31	345
RepVGG	83.77	1120
XCiT small	85.12	3131
XCiT medium	85.59	11429
RegNetY 160	84.63	1690
RegNetY 320	85.59	2730
MambaOut	86.05	2916
BEiTv2	76.35	1587
Twins	83.87	964

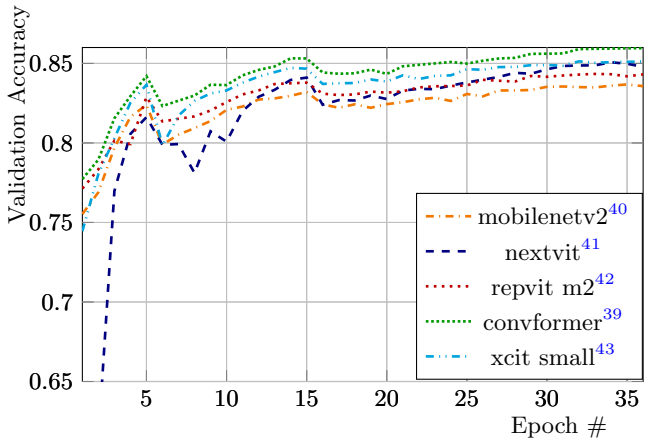
Table 3: List of DL-based detectors.

models as it is based on swin model which is well-known and has already been used for deepfake detection.[22,23] While we included lightweights models such as mobileNet, we included some representative of the MetaFormer models,[39] namely CAFormer and ConvFormer, as they are popular models as well as examples of reather heavy-weights vision transformers models for classification.

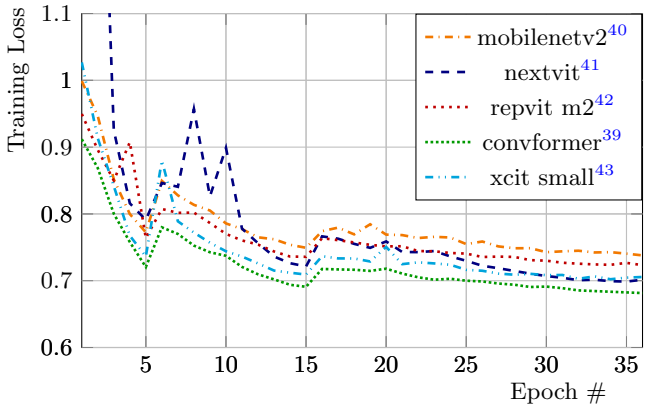
With this wide selection we also wanted to show as much diversity as possible, for example by including the very latest deplearning models as well as older ones, some very light and some very heavy.

3. NUMERICAL RESULTS AND ANALYSIS

The main results of the present study is present in Table 3. It presents the detection accuracy of all models on the problem of deepfake models identification. Note for speeding up the training process and in under to get



(a) Validation loss (Binary Cross-Entropy) over iterative training.



(b) Accuracy (average of true-positive and true-negative rates) over iterative training process.

Figure 1: Comparison of the accuracy of the proposed model with several of its possible competitots. The test detection accuracy (left) and the validation loss functions (right) are repported during the training process, more precisely plotted as a function of the epoch number.

results which are easier to read, the training and the testing was carried out of images of size 128×128 and converted in grayscale. The accuracy may seem rather small if one consider that papers on deepfake detection ofent report detection accuracy higher than 95%. However in our case we trained the model to perform a multi-class classification taks with almost 50 classes corresponding each to an generative AI model, see Table 1. Keeping in mind the small image size we used, the detection results are actually very impressive.

While the number of 36 epochs may seem rather small we would like to convince the reader that this is enough is our case by presenting in the figure 1 the evolution of the training model. First of all, the figure on the left-hand side show the evoluation of the testing accuracy as a function of the training epoch. On this figure, one can note that very negligible amount of improvement after epoch 30 even for the heaviest model such as ConvFormer and Xcit which are supposedly the longest and the hardest to train.
The figure on the right-hand side the training loss as a function of the epoch index. Note the very small amount of the during the last cycle, after epoch number 15 and especially during its second hand. The small number of epoch needed to train the models can be explained by large size of the training dataset which is made of $670,000$ number of images.

Overall, it is striking to note in Table 3 that Vision-Transformers generally achieve the best overall results. In other words the five deeplearning models that achieves the best performance are all amongst the recent vision-transformer-based models: MambaOut, CAFormer, ConvFormer, RegNet and XCiT.
However, one can also note that, unsurprisingly, that the deeplearning models with the highest accuracy are generally amongst the most ressource-consumming to train. Alternatively, some models seems to be very good choice in terms of tradeoff between detection accuracy and training ressources especially RepViT, EfficientNet v2 and MobileNet V2 whose training time can be 20 times faster than ConvFormer for only 1% of loss in terms of detection accuracy.

3.1 Ensemble Classifier methods using the a variability of Deep Leaning Models

Last bu not least we have tried five different methods for aggregating all the results form all the classifier in order to further improve the detection accuracy with an ensemble classifier. First we tried as a reference comparison to average the "soft output" all the classifiers. This methods is certainly not the most effective as it gives the same important to week classifier and to the best ones. nevertheless, we included it in the results as a benchmark to show the improvement in terms of accuracy. Insprired by the work we also tried including the average score of

average / sum	top-10	Linear model	non-linear model	Fully connected
83.93	85.51	87.27	88.69	88.72

Table 4: List of DL-based detectors.

the classifier that gives the largest "soft output". The idea behind this methodology is to aggregate the score of the k classifiers who make the most certain decision in favour of a deepfake image. On the opposite, a classifier whose score is so-so shall not be considered.
Then we used three classical methods, first we trained a linear classifier: with the "soft outputs" from all the models it computes a weightes sum from all classifiers. The results we present were obtained with a linear SVM which we implemented using the scikit-learn python package, but the others linear classifiers we used gave similar results.
We implemented a non-linear SVM classifier. Even though the number of deep learning models is relatively small, the relatively high number of images, together with the cross-validation for the hyperparameters of the kernel width and regularization parameter made the training very long hence our choice to use a Nyström approximation[44] ; empirically we determine that 200 components was a good tradeoff.
Last but not least we implemented a 2 layers fully connected classifier with pytorch and we used the well-known Xavier, or Glorot, initialization.[45]

As expected, the results presented in the Table 4 show that the average of all deep learning models is a poor choice as it performs worse than many deep learning models alone. Again, this results is not very surprising as such an aggregation function gives the same importance to all classifiers, the worst as well as the best. The Top-k classifier outputs with $k = 10$ gave slightly better results. However, we can note that, here it is not satisfying at all as the accuracy of the aggregation of all deep learning models gives is worse than the best models alone. Interestingly the linear classifier, though very simple, seems a much better choice as it improves the accuracy of almost 1.3% which is not a negligible improvement comparing the accuracy of deplearning models. The non-linear SVM classifier, or more precisely its Nyström approximation, as well as the fully connected classifier allow improving the accuracy futher by approximatively 1.4%.
Unsurprisingly the non-linear methods provides the largest improvement in terms of classification accuracy. However, the improvement shall be balanced by the relatively large number of deep learning models we used. Indeed, while a 2.7% improvement of the balanced accuracy is a rather very good results, it is striking that we needed almost 40 deep learning models to reach that results which is a rather high number of classifier with regards to the improvement. We can conclude that the classifier are, most of the time, very much correlated and likely to provide decision the similar decisions.

4. CONCLUSION AND FUTURE WORKS

The present paper proposes a large scale benchmark for the problem of deepfake detection. Indeed this large scale is based on a wide range of almost 50 different generative AI models used in the training and testing set. Additionnaly, the paper proposes a numerical comparison of over 40 models for deepfake detection ranging from the simplest and oldest one, such as densenet, resnet and vgg because they are popular and often used in deepfake detection, up to the latest vision-transformers based deep learining classification models.
We show that, generally speaking, vision-transformers performs extremely well for this task of weak signal detection. Additionnaly, we also demonstrate empirically that larger model provide the best overall performance. Nevertheless some models are good tradeoff in terms of deepfake detection accuracy and training complexity such as EfficientNet v2 and MobiletViT.
We are strongly believed that the study presented in the present paper will serves for the design of future deepfake detection methods and will serves as a reference benchmark for the evaluation of future deeplearning methods.

Upon acceptance of the present paper we will released the dataset of deepfake image along with the weights of the trained models to further serve the future researches in the field of deepfake detection.
We also believe that our large dataset of deepfake images can be used to study the problem of generalization, when testing over out-of-distribution image.

REFERENCES

[1] Science, N. and Council, T., "Roadmap for researchers on priorities related to information integrity research and development," (2022).

[2] Haines, A., "An update on foreign threats to the 2024 elections, senate select committee on intelligence," (May 2024).

[3] Nguyen, T. T., Nguyen, Q. V. H., Nguyen, D. T., Nguyen, D. T., Huynh-The, T., Nahavandi, S., Nguyen, T. T., Pham, Q.-V., and Nguyen, C. M., "Deep learning for deepfakes creation and detection: A survey," *Computer Vision and Image Understanding* **223**, 103525 (2022).

[4] Masood, M., Nawaz, M., Malik, K. M., Javed, A., Irtaza, A., and Malik, H., "Deepfakes generation and detection: state-of-the-art, open challenges, countermeasures, and way forward," *Applied Intelligence* **53**, 3974–4026 (jun 2022).

[5] Pan, D., Sun, L., Wang, R., Zhang, X., and Sinnott, R. O., "Deepfake detection through deep learning," in [*2020 IEEE/ACM International Conference on Big Data Computing, Applications and Technologies (BDCAT)*], 134–143 (2020).

[6] Siegel, D., Krätzer, C., Seidlitz, S., and Dittmann, J., "Forensic data model for artificial intelligence based media forensics-illustrated on the example of deepfake detection," *Electronic Imaging* **34**, 1–6 (2022).

[7] Nataraj, L., Mohammed, T. M., Chandrasekaran, S., Flenner, A., Bappy, J. H., Roy-Chowdhury, A. K., and Manjunath, B., "Detecting gan generated fake images using co-occurrence matrices," *arXiv preprint arXiv:1903.06836* (2019).

[8] Barni, M., Kallas, K., Nowroozi, E., and Tondi, B., "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)*], 1–6, IEEE (2020).

[9] Chang, X., Wu, J., Yang, T., and Feng, G., "Deepfake face image detection based on improved vgg convolutional neural network," in [*2020 39th chinese control conference (CCC)*], 7252–7256, IEEE (2020).

[10] Rajalaxmi, R., Sudharsana, P., Rithani, A., Preethika, S., Dhivakar, P., and Gothai, E., "Deepfake detection using inception-resnet-v2 network," in [*2023 7th International Conference on Computing Methodologies and Communication (ICCMC)*], 580–586, IEEE (2023).

[11] Coccomini, D. A., Esuli, A., Falchi, F., Gennaro, C., and Amato, G., "Detecting images generated by diffusers," *PeerJ Computer Science* **10**, e2127 (2024).

[12] Wang, S.-Y., Wang, O., Zhang, R., Owens, A., and Efros, A. A., "Cnn-generated images are surprisingly easy to spot... for now," (2020).

[13] Gragnaniello, D., Cozzolino, D., Marra, F., Poggi, G., and Verdoliva, L., "Are gan generated images easy to detect? a critical analysis of the state-of-the-art," in [*2021 IEEE international conference on multimedia and expo (ICME)*], 1–6, IEEE (2021).

[14] Bonettini, N., Cannas, E. D., Mandelli, S., Bondi, L., Bestagini, P., and Tubaro, S., "Video face manipulation detection through ensemble of cnns," in [*2020 25th International Conference on Pattern Recognition (ICPR)*], 5012–5019 (2021).

[15] Mandelli, S., Bonettini, N., Bestagini, P., and Tubaro, S., "Detecting gan-generated images by orthogonal training of multiple cnns," in [*2022 IEEE International Conference on Image Processing (ICIP)*], 3091–3095, IEEE (2022).

[16] Mandelli, S., Bestagini, P., and Tubaro, S., "When synthetic traces hide real content: Analysis of stable diffusion image laundering," in [*IEEE International Workshop on Information Forensics and Security (WIFS)*], (2024).

[17] Wodajo, D. and Atnafu, S., "Deepfake video detection using convolutional vision transformer," *arXiv preprint arXiv:2102.11126* (2021).

[18] Coccomini, D. A., Messina, N., Gennaro, C., and Falchi, F., "Combining efficientnet and vision transformers for video deepfake detection," in [*International conference on image analysis and processing*], 219–229, Springer (2022).

[19] Xu, Q., Wang, H., Meng, L., Mi, Z., Yuan, J., and Yan, H., "Exposing fake images generated by text-to-image diffusion models," *Pattern Recognition Letters* **176**, 76–82 (2023).

[20] Dang, M., "Efficient vision-based face image manipulation identification framework based on deep learning," *Electronics* **11**(22), 3773 (2022).

[21] Zhou, Y., Fan, B., K. Atrey, P., and Ding, F., "Exposing deepfakes using dual-channel network with multi-axis attention and frequency analysis," in [*Proceedings of the 2023 ACM Workshop on Information Hiding and Multimedia Security*], 169–174 (2023).

[22] Coccomini, D. A., Caldelli, R., Falchi, F., Gennaro, C., and Amato, G., "Cross-forgery analysis of vision transformers and cnns for deepfake image detection," in [*Proceedings of the 1st International Workshop on Multimedia AI against Disinformation*], 52–58 (2022).

[23] Coccomini, D. A., Caldelli, R., Falchi, F., and Gennaro, C., "On the generalization of deep learning models in video deepfake detection," *Journal of Imaging* **9**(5), 89 (2023).

[24] Cogranne, R., Giboulot, Q., and Bas, P., "The alaska steganalysis challenge: A first step towards steganalysis," in [*Proceedings of the ACM Workshop on Information Hiding and Multimedia Security*], 125–137 (2019).

[25] Cogranne, R., Giboulot, Q., and Bas, P., "Alaska# 2: Challenging academic research on steganalysis with realistic images," in [*2020 IEEE International Workshop on Information Forensics and Security (WIFS)*], 1–5, IEEE (2020).

[26] Wightman, R., "Pytorch image models." https://github.com/huggingface/pytorch-image-models (2019).

[27] Smith, L. N., "No more pesky learning rate guessing games," *CoRR, abs/1506.01186* **5**, 575 (2015).

[28] Loshchilov, I. and Hutter, F., "Sgdr: Stochastic gradient descent with warm restarts," *arXiv preprint arXiv:1608.03983* (2016).

[29] Smith, L. N., "Cyclical learning rates for training neural networks," in [*2017 IEEE winter conference on applications of computer vision (WACV)*], 464–472, IEEE (2017).

[30] Smith, L. N. and Topin, N., "Super-convergence: Very fast training of neural networks using large learning rates," in [*Artificial intelligence and machine learning for multi-domain operations applications*], **11006**, 369–386, SPIE (2019).

[31] Guarnera, L., Giudice, O., Guarnera, F., Ortis, A., Puglisi, G., Paratore, A., Bui, L. M., Fontani, M., Coccomini, D. A., Caldelli, R., et al., "The face deepfake detection challenge," *Journal of Imaging* **8**(10), 263 (2022).

[32] Deb, D., Liu, X., and Jain, A. K., "Unified detection of digital and physical face attacks," in [*2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG)*], 1–8 (2023).

[33] Seo, R., Kuribayashi, M., Ura, A., Mallet, A., Cogranne, R., Mazurczyk, W., and Megías, D., "Toward universal detector for synthesized images by estimating generative ai models," in [*Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*], (2024).

[34] Malik, A., Kuribayashi, M., Abdullahi, S. M., and Khan, A. N., "Deepfake detection for human face images and videos: A survey," *IEEE Access* **10**, 18757–18775 (2022).

[35] Pan, D., Sun, L., Wang, R., Zhang, X., and Sinnott, R. O., "Deepfake detection through deep learning," in [*2020 IEEE/ACM International Conference on Big Data Computing, Applications and Technologies (BDCAT)*], 134–143, IEEE (2020).

[36] Rana, M. S., Nobi, M. N., Murali, B., and Sung, A. H., "Deepfake detection: A systematic literature review," *IEEE access* **10**, 25494–25513 (2022).

[37] Wang, Z., Cheng, Z., Xiong, J., Xu, X., Li, T., Veeravalli, B., and Yang, X., "A timely survey on vision transformer for deepfake detection," *arXiv preprint arXiv:2405.08463* (2024).

[38] Ghita, B., Kuzminykh, I., Usama, A., Bakhshi, T., and Marchang, J., "Deepfake image detection using vision transformer models," in [*2024 IEEE International Black Sea Conference on Communications and Networking (BlackSeaCom)*], 332–335, IEEE (2024).

[39] Yu, W., Si, C., Zhou, P., Luo, M., Zhou, Y., Feng, J., Yan, S., and Wang, X., "Metaformer baselines for vision," *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2023).

[40] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., and Chen, L.-C., "Mobilenetv2: Inverted residuals and linear bottlenecks," in [*Proceedings of the IEEE conference on computer vision and pattern recognition*], 4510–4520 (2018).

[41] Li, J., Xia, X., Li, W., Li, H., Wang, X., Xiao, X., Wang, R., Zheng, M., and Pan, X., "Next-vit: Next generation vision transformer for efficient deployment in realistic industrial scenarios," *arXiv preprint arXiv:2207.05501* (2022).

[42] Wang, A., Chen, H., Lin, Z., Han, J., and Ding, G., "Repvit: Revisiting mobile cnn from vit perspective," in [*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*], 15909–15920 (2024).

[43] Ali, A., Touvron, H., Caron, M., Bojanowski, P., Douze, M., Joulin, A., Laptev, I., Neverova, N., Synnaeve, G., Verbeek, J., et al., "Xcit: Cross-covariance image transformers," *Advances in neural information processing systems* **34**, 20014–20027 (2021).

[44] Rasmussen, C. E., "Gaussian processes in machine learning," in [*Summer school on machine learning*], 63–71, Springer (2003).

[45] Glorot, X. and Bengio, Y., "Understanding the difficulty of training deep feedforward neural networks," in [*Proceedings of the thirteenth international conference on artificial intelligence and statistics*], 249–256, JMLR Workshop and Conference Proceedings (2010).