Diffusion-Inspired Dynamic Models for Enhanced Fake Face Detection

Mamadou Dian Bah[a], Rajjeshwar Ganguly[b] and Mohamed Dahmane[c]
Computer Research Institute of Montreal (CRIM), Canada

Keywords: DeepFake Detection, Dynamic Model, Media Forensics, U-Net.

Abstract: The conventional convolutional U-Net model was originally designed to segment images. Enhanced versions of the architecture with timestep embedding and self-attention inspired by transformers were proposed in the literature for image classification tasks. In this paper, we investigate a U-Net–encoder architecture for deepfake detection that involves two key features, the use of self-attention blocks to capture both local and global content representation, and the integration of timestep embedding to capture the dynamic perturbation of the input data. The model is trained and evaluated on FF++ dataset, comprising of real and deepfake synthesized videos. Notably, compared to traditional models pretrained on ImageNet, our model demonstrates superior performance. The experimental results highlight the effectiveness of our approach in achieving improved classification results for the challenging task of distinguishing real and deepfake images. The achieved performances suggest that the model aims to leverage both spatial information and dynamic perturbation for improved detection performance.

1 INTRODUCTION

The remarkable progress in artificial neural network (ANN)-based technologies and particularly in generative AI has played a crucial role in manipulating multimedia content. These advancements have made it increasingly feasible to generate highly realistic synthetic images, surpassing previous capabilities. In recent years, generative models have witnessed remarkable advancements in their ability to generate human-like natural language (Brown et al., 2020), high-quality synthetic images (Karras et al., 2020), and diverse human speech and music. These models find utility in various domains, such as image generation from text prompts and feature representation learning.

As a result, numerous captivating applications have emerged in the realms of entertainment and education. Through all this capability, the term deepfake has emerged and refers to multimedia content generated or altered by artificial intelligence models (Gomes et al., 2020; Lattas et al., 2020). Noteworthy examples include FaceApp (FaceApp, 2023), a popular application that leverages an autoencoder-decoder architecture to seamlessly swap faces between two images. DeepFaceLab (DeepfakeVFX.com, 2023), an enhanced framework specifically designed for face-swapping, pushes the boundaries of deepfake technology. Another notable application is Face Swapping GAN (FSGAN) (Nirkin et al., 2019), an improved version of deepfake that employs Generative Adversarial Networks (GANs) to refine segmentation masks, resulting in remarkably higher quality output videos. Additionally, several other tools are employed for deepfake content generation. DiscoFaceGAN (Deng et al., 2020), based on StyleGAN structure (Karras et al., 2019), generates synthetic face images of virtual individuals with distinct characteristics, while FaceShifter enables high-fidelity face swapping. The recent emergence of a new generation of models, the so–called Denoising Diffusion Models (DDMs), has raised great concern for the spread of fake data, as they proved capable of generating even more realistic and convincing fakes than their predecessors, Generative Adversarial Networks (GANs). Models like Stable Diffusion (Rombach et al., 2022) and DALL-E 2 (Ramesh et al., 2022) are some of the best image generators available and are renowned state-of-the-art Diffusion Models (DMs) that excel in text-to-image translation.

Given the significant threats posed by deepfakes, such as the spread of misinformation, damage to reputation, and invasion of privacy, it is crucial to de-

[a] https://orcid.org/0000-0002-0733-8587
[b] https://orcid.org/0009-0002-8893-7756
[c] https://orcid.org/0000-0002-2670-1433

Bah, M., Ganguly, R. and Dahmane, M.
Diffusion-Inspired Dynamic Models for Enhanced Fake Face Detection.
DOI: 10.5220/0012391300003660

In *Proceedings of the 19th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2024) - Volume 4: VISAPP*, pages 532-538
ISBN: 978-989-758-679-8; ISSN: 2184-4321

velop advanced technologies for detecting deepfake content. Humans often struggle to identify manipulated videos, particularly in terms of spatial aspects, and lack the ability to pinpoint the specific locations and techniques used for manipulation. This highlights the necessity for high-quality algorithms capable of detecting manipulated videos on a large scale.

In this paper, we tackle the problem of detecting facial manipulation. In particular, we focus on all the manipulation techniques reported in (Rossler et al., 2019) (i.e., deepfakes, Face2Face, FaceSwap and NeuralTextures). We consider exploring a U-Net–encoder architecture for deepfake detection that involves two key features, the use of self-attention blocks to capture both local and global spatial information, and the integration of a predicted timestep embedding to capture the dynamic perturbation of the input data.

The rest of the paper is organized as follows. Section 2 provides a literature review of commonly used algorithms for deepfake detection. Section 3 presents details on the proposed approach. Section 4 provides the results and discussions. Finally, Section 5 concludes with implications, limitations, and suggestions for future research.

2 RELATED WORKS

Deep learning techniques have emerged as the dominant approach for deepfake detection, as evidenced by a comprehensive analysis of 122 studies conducted by Rana et al. (Rana et al., 2022). Approximately 77% of these studies employed deep learning models, specifically Convolutional Neural Networks (CNNs) (Tariq et al., 2018) and Recurrent Neural Networks (RNNs) mostly used for fake videos detection. These models have shown great promise in effectively detecting deepfake content.

In (Afchar et al., 2018) MeSoNet is a CNN architecture which is used to detect Face2Face and deepfakes manipulations. XceptionNet which uses depthwise separable convolutional layers with residual connections (Chollet, 2017) has given the best result in (Rossler et al., 2019) in Faceforensics++ (FF++) dataset and is the most used baseline.

Although the use of deeplearning different approach is used based on the knowledge of human face. Authors in (Haliassos et al., 2021) proposes a method of detecting high-level semantic anomalies in mouth motion, leveraging the hypothesis that most video generators display a degree of high-level semantic irregularities near the mouth. In (Zhao et al., 2021a; Zhao et al., 2021b), fine-grained classification is applied to distinguish subtle differences in visual appearance and patterns. The authors propose FakeBuster in (Hubens et al., 2021) to address the issue of detecting face modification in video sequences using recent facial manipulation techniques. In (Ismail et al., 2021), the YOLO face detector is used to extract the face area from video frames, while the InceptionResNetV2 CNN is utilized to extract features from these faces.

However detecting deepfakes in videos solely based on counterfeit images can be difficult due to the temporal features of videos and variation in frame resolution. In (Ranjan et al., 2020), the CNN-LSTM combo is used to identify and classify the videos as fake or real.

FSSPOTTER (Chen et al., 2020), for instance, uses spatial and temporal clues to detect swapped faces in videos. These features are fed into the XGBoost, which works as a recognizer on the top level of the CNN network. Physiological signals are also used for deepfake detection. DFT-MF (Elhassan et al., 2022) is a deep-fake detection model that uses deep learning approaches to detect deepfake videos by isolating, analyzing, and verifying lip/mouth movement. Eye blinking based signal detection is also used to determine if a video is real or a deepfake generated. Deep Vision detects deepfake videos by focusing on eye blink patterns. FakeCatcher (Ciftci et al., 2020) is a method that addresses the challenge of detecting deepfakes by exploiting the fact that biological signals obtained from facial regions are not positionally and temporally well-preserved in synthetic content of portrait videos.

In a recent study, U-Net has garnered significant attention due to its computational and efficiency advantages in segmentation and feature extraction (Ronneberger et al., 2015). Eff-YNet, introduced in (Tjon et al., 2021), is a noteworthy example of the synergy between an EfficientNet encoder and a U-Net structure. This fusion enables the model to effectively perform both classification and segmentation tasks on deepfake videos. Similarly, the work presented in (Bhilare et al., 2022) aligns with this theme, where authors introduce U-YNet. This model integrates segmentation and classification capabilities by utilizing a U-Net Encoder and Decoder to generate segmentation maps, with a classification branch seamlessly integrated at the end of the U-Net Encoder. These innovations highlight the versatility and computational efficiency of U-Net in addressing a wide array of computer vision challenges. Moreover, in the recent approach for deepfake generation U-Net is one of the most important components of diffusion model (Dhariwal and Nichol, 2021) because it facilitates the

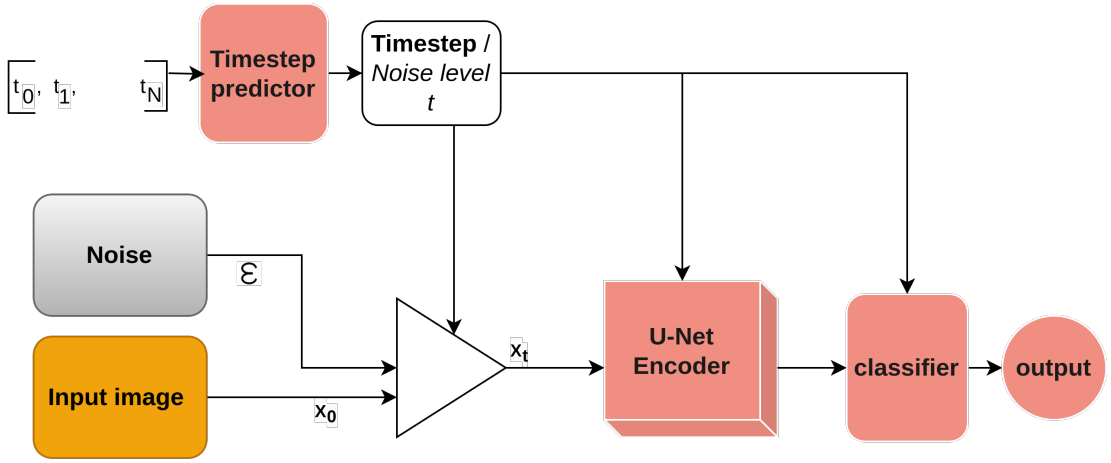

Figure 1: Architecture of proposed method. This model is refered as ClOneSelect.

actual diffusion process and improves sample quality for denoising score matching (Jolicoeur-Martineau et al., 2020). In (Dhariwal and Nichol, 2021), the authors found that a pre-trained diffusion model can be conditioned using the gradients of a classifier, further underscoring the significance of U-Net as classifier in guiding and improving the diffusion process for image generation tasks. In our research, we investigate the use of a U-Net encoder in conjunction with techniques derived from the diffusion approach to detect deepfake generated faces. Results is compared to the traditional method such as Xception (Rossler et al., 2019) and EfficienceNet (Bonettini et al., 2021) used in other works.

3 METHOD

The model consists of 3 blocks: one block containing a layer that predicts the timestep t of the input image, a block that adds noise ε to the image x_0 based on the timestep, an encoder that takes an image as input and provides vector of 1000 values, and finally, the classifier layer which maps 1000 to 1. In Figure 1, we present the architecture of our method.

3.1 Timestep Predictor

The forward process in a diffusion model works by gradually adding noise to an initial clean or observed image to generate a sequence of increasingly noisier images. This process simulates the generation of samples in a probabilistic model where the final sample represents the output of interest.

Diffusion models generate samples by reversing a gradual process of introducing noise. Essentially, the sampling process begins with noisy x_T and progressively creates less noisy samples x_{T-1}, x_{T-2}, ... until reaching a final sample x_0. Each timestep, denoted as t, corresponds to a specific noise level, and x_t can be viewed as a combination of a signal x_0 with some noise ε, where the signal-to-noise ratio is determined by the timestep t. In Figure 2, a sample of images is shown at different timestep using a linear scheduler for adding noise.

The authors of guided diffusion (Radford et al., 2021) demonstrated the effectiveness of incorporating a projection of the timestep embedding into each residual block of the U-Net encoder. This addition serves to guide the diffusion model towards the desired class, effectively transforming the architecture as if we have multiple models in one. This adaptation occurs in accordance with the timestep, allowing for dynamic adjustments in classification.

In our context, to take advantage of all those models, we have incorporated a Linear layer that takes the input of 1000 possible timesteps and selects a single value t, representing the most suitable timestep for improved classification. This value t represents the impact of the noise ε to be applied to the image x.

3.2 U-Net Encoder

The UNet model uses a stack of residual layers and downsampling convolutions. In addition, they use a global attention layer at different 32×32, 16×16 and 8x8 resolutions with a single head, and add a projection of the timestep embedding t into each residual block. Timestep t embedding capture the dynamic perturbations present in the input data. By learning embeddings specific to each timestep, the model gains the ability to make accurate predictions regarding whether an image belongs to the ”fake” or

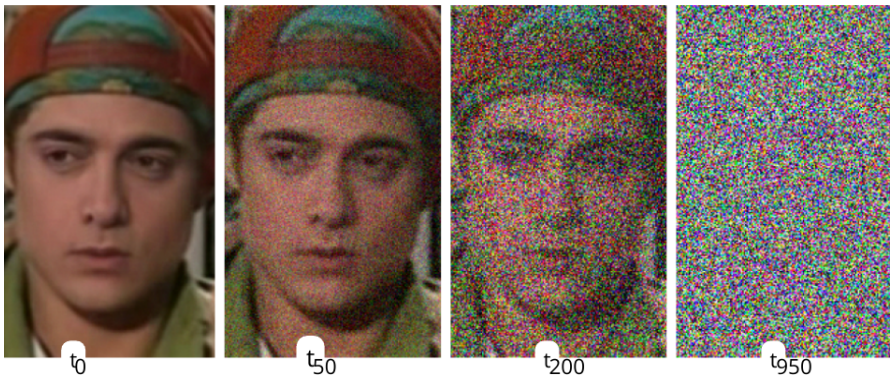

Figure 2: Example of images with corresponding timestep using a linear scheduler.

"real" class, taking into account the level of the introduced noise. This integration empowers the model to dynamically adapt its processing strategy, effectively leveraging the unique characteristics and dynamic variations observed at different time points within the input data. The model, denoted as C, using the noisy images x_t for $t \in \{0, ..., T\}$. Each timestep t corresponds to a certain noise level where T refers to the highest noise level. This encoder is trained to predict the class label of the input image x according to the level t of the added noise. Where x_0 is the image at timestep $t = 0$.

The output of the U-Net encoder architecture is the downsampling trunk of the model with an attention pool at the 8x8 layer to produce the final output of 1000 which correspond to the thousand classes of ImageNet. In our case, we added a classifier layer to map the 1000 to 1. The encoder can be represented as in Equation 1.

$$\hat{y} = C(x_t; \theta_C) \tag{1}$$

where $\hat{y}$ represents the predicted class label, x_t is the input image at timestep t, and θ_C denotes the optimizable parameters of the classifier network.

4 EXPERIMENT AND RESULTS

To assess the effectiveness of our proposed approach, we conducted extensive experiments on the FF++ dataset (Rossler et al., 2019). The FF++ dataset stands out as a large-scale facial manipulation dataset created through state-of-the-art video editing techniques. In particular, it combines classical computer graphics methods, such as Face2Face (Thies et al., 2016) and FaceSwap, with learning-based strategies, including deepfakes and NeuralTextures (Thies et al., 2019). Each of these methods was applied to 1000 high-quality pristine videos downloaded from YouTube,

Table 1: Performance Evaluation on FF++ Dataset. Methods labeled with "ClOne" are proposed by us.

Method	AUC
EfficienceNet-B4 (Bonettini et al., 2021)	96
Xception (Rossler et al., 2019)	94.98
ClOne	95.26
ClOneMixXcept	95.72
ClOneMixB4	96.68
ClOneSelect	96.21

carefully selected to ensure that the subjects faces were nearly frontal and free from occlusions. These video sequences consist of at least 280 frames each. Ultimately, this dataset comprises over 1.8 million images obtained from 4000 manipulated videos.

To ensure a fair comparison, in our experiment we adopted the evaluation protocol defined in (Rossler et al., 2019).

We used similar splits, selecting 720 videos for training, 140 for validation and 140 for test from the pool of original sequences taken from YouTube. The corresponding fake videos are assigned to the same split. We primarily focus on the subject face region for analysis. We use the BlazeFace extractor for preprocessing, extracting the best-confidence face from each frame. Our network input image shape is 256 × 256. During training and validation, we enhance the model robustness with data augmentation, including downscaling, flipping, brightness, contrast, hue, saturation adjustments, noise addition, and JPEG compression.

A total of 230302 frames were extracted from the dataset for training, 26879 frames were set aside for validation and 26879 for testing purposes.

During experiments, we utilized the Adam optimizer with an initial learning rate of 0.0001 and cross-entropy loss. Batch size of 32 is used and the model is validated every 500 iteration on 6000 sample randomly selected in validation set. Our model was exposed to diverse and challenging conditions during

training, allowing it to learn and adapt to various types and levels of noise commonly encountered in real-world scenarios. To optimize the training process, we performed fine-tuning on the U-Net encoder, which was initially pretrained on ImageNet with the specific purpose of guiding the conditional image synthesis in diffusion models (Radford et al., 2021).

To evaluate the performance of our proposed method, we conducted various tests including comparisons with various models and cross-data validation. Our method was compared with Xception (Rossler et al., 2019) and EfficienceNet-B4 (Bonettini et al., 2021), both employed in deepfake detection.

Furthermore, we conducted an ablation study of the model. Initially, we removed the timestep predictor and trained the model with only a timestep of 0. In other words, the model was directly trained with the original image as input, without adding noise during the training process. We refer to this model as ClOne throughout the paper. Additionally, we developed hybrid versions by combining Xception and EfficienceNet-B4 with the U-Net encoder model, referred to as ClOneMixXcept and ClOneMixB4, respectively. Hybrid models performed feature extraction from the initial U-Net encoder blocks and used them as input for Xception. Similar to ClOne, the image was not corrupted by noise.

The results presented in Table 1 demonstrate that ClOneMixB4 and ClOneSelect models present the highest AUC values respectively 96.68 and 96.21. ClOne also performs well, achieving an AUC of 95.26. EfficienceNet-B4 gives the third best result an AUC of 96 and ClOneMixXcept achieves an AUC of 95.72. These results underscore the performance of the proposed models.

We performed cross-dataset evaluation on Celeb-DF (V2) (Yuezun Li and Lyu, 2020), an extensive dataset designed to mimic the visual quality of online videos. Unlike its predecessor, Celeb-DF (V1), which contained only 795 deepfake videos, this updated version includes 590 original videos from YouTube, spanning various ethnicities. Additionally, Celeb-DF (V2) encompasses 5639 corresponding deepfake videos, making it a valuable resource for evaluation and analysis. For testing purposes, 16,565 frames were selected from a subset of 518 designated as test videos.

In Table 2, the results indicate that ClOne and ClOneMixXcept, show better performance compared to Xception. This suggests that the proposed approach is robust to detect manipulated content in the Celeb-DF (V2) dataset. Although ClOneSelect performs slightly below EfficienceNet-B4 in terms of AUC, it still demonstrates better performance than Xception and ClOneMixB4 in this dataset.

Table 2: Cross-dataset evaluation on Celeb-DF (V2) Dataset. Methods with "ClOne" are proposed by us.

Method	AUC
EfficienceNet-B4 (Bonettini et al., 2021)	77.66
Xception (Rossler et al., 2019)	73.06
ClOne	75.69
ClOneMixXcept	75.69
ClOneMixB4	74.06
ClOneSelect	74.85

5 CONCLUSION

In this paper we showcased the remarkable performance of the U-Net model with attention and timestep embedding in distinguishing between real and deepfake images. By capturing both local and global information and considering the dynamic perturbation of the input data, our model outperformed the well-established deepfake detection methods Xception and EfficienceNet-B4 on the FF++ dataset in terms of AUC.

As a future perspective, we intend to add a branch of noise prediction from the input image in a multitask learning setting. Additionally, we aim to explore the utilization of other datasets to enhance the model's generalization capabilities.

ACKNOWLEDGEMENTS

The work was partially supported by a NSERC (Natural Sciences and Engineering Research Council of Canada) Discovery Grant under grant agreement No RGPIN-2020-05171.
We gratefully acknowledge the support of the Computer Research Institute of Montreal (CRIM), the Ministère de l'Économie et de l'Innovation (MEI) of Quebec, and The Natural Sciences and Engineering Research Council of Canada (NSERC).

REFERENCES

Afchar, D., Nozick, V., Yamagishi, J., and Echizen, I. (2018). Mesonet: a compact facial video forgery detection network. In *2018 IEEE international workshop on information forensics and security (WIFS)*, pages 1–7. IEEE.

Bhilare, O., Singh, R., Paranjape, V., Chittupalli, S., Suratkar, S., and Kazi, F. (2022). Deepfake cli: Accelerated deepfake detection using fpgas. In *Interna-*

tional Conference on Parallel and Distributed Computing: Applications and Technologies, pages 45–56. Springer.

Bonettini, N., Cannas, E. D., Mandelli, S., Bondi, L., Bestagini, P., and Tubaro, S. (2021). Video face manipulation detection through ensemble of cnns. In *2020 25th international conference on pattern recognition (ICPR)*, pages 5012–5019. IEEE.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language models are few-shot learners.

Chen, P., Liu, J., Liang, T., Zhou, G., Gao, H., Dai, J., and Han, J. (2020). Fsspotter: Spotting face-swapped video by spatial and temporal clues. In *2020 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6.

Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258.

Ciftci, U. A., Demir, I., and Yin, L. (2020). Fakecatcher: Detection of synthetic portrait videos using biological signals. *IEEE transactions on pattern analysis and machine intelligence*.

DeepfakeVFX.com (2023). DeepFaceLab - DeepfakeVFX.com — deepfakevfx.com. https://www.deepfakevfx.com/downloads/deepfacelab/. [Accessed 21-Jun-2023].

Deng, Y., Yang, J., Chen, D., Wen, F., and Tong, X. (2020). Disentangled and controllable face image generation via 3d imitative-contrastive learning.

Dhariwal, P. and Nichol, A. (2021). Diffusion models beat gans on image synthesis.

Elhassan, A., Al-Fawa'reh, M., Jafar, M. T., Ababneh, M., and Jafar, S. T. (2022). Dft-mf: Enhanced deepfake detection using mouth movement and transfer learning. *SoftwareX*, 19:101115.

FaceApp (2023). FaceApp: Face Editor — faceapp.com. https://www.faceapp.com/. [Accessed 21-Jun-2023].

Gomes, T. L., Martins, R., Ferreira, J., and Nascimento, E. R. (2020). Do as i do: Transferring human motion and appearance between monocular videos with spatial and temporal constraints.

Haliassos, A., Vougioukas, K., Petridis, S., and Pantic, M. (2021). Lips don't lie: A generalisable and robust approach to face forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5039–5049.

Hubens, N., Mancas, M., Gosselin, B., Preda, M., and Zaharia, T. (2021). Fake-buster: A lightweight solution for deepfake detection. In *Applications of Digital Image Processing XLIV*, volume 11842, pages 146–154. SPIE.

Ismail, A., Elpeltagy, M., S. Zaki, M., and Eldahshan, K. (2021). A new deep learning-based methodology for video deepfake detection using xgboost. *Sensors*, 21(16):5413.

Jolicoeur-Martineau, A., Piché-Taillefer, R., des Combes, R. T., and Mitliagkas, I. (2020). Adversarial score matching and improved sampling for image generation. *CoRR*, abs/2009.05475.

Karras, T., Laine, S., and Aila, T. (2019). A style-based generator architecture for generative adversarial networks.

Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., and Aila, T. (2020). Analyzing and improving the image quality of stylegan.

Lattas, A., Moschoglou, S., Gecer, B., Ploumpis, S., Triantafyllou, V., Ghosh, A., and Zafeiriou, S. (2020). Avatarme: Realistically renderable 3d facial reconstruction "in-the-wild".

Nirkin, Y., Keller, Y., and Hassner, T. (2019). Fsgan: Subject agnostic face swapping and reenactment.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. (2021). Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. (2022). Hierarchical text-conditional image generation with clip latents.

Rana, M. S., Nobi, M. N., Murali, B., and Sung, A. H. (2022). Deepfake detection: A systematic literature review. *IEEE Access*.

Ranjan, P., Patil, S., and Kazi, F. (2020). Improved generalizability of deep-fakes detection using transfer learning based cnn framework. In *2020 3rd international conference on information and computer technologies (ICICT)*, pages 86–90. IEEE.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. (2022). High-resolution image synthesis with latent diffusion models.

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer.

Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., and Nießner, M. (2019). Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1–11.

Tariq, S., Lee, S., Kim, H., Shin, Y., and Woo, S. S. (2018). Detecting both machine and human created fake face images in the wild. In *Proceedings of the 2nd international workshop on multimedia privacy and security*, pages 81–87.

Thies, J., Zollhöfer, M., and Nießner, M. (2019). Deferred neural rendering: Image synthesis using neural tex-

tures. *Acm Transactions on Graphics (TOG)*, 38(4):1–12.

Thies, J., Zollhofer, M., Stamminger, M., Theobalt, C., and Nießner, M. (2016). Face2face: Real-time face capture and reenactment of rgb videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2387–2395.

Tjon, E., Moh, M., and Moh, T.-S. (2021). Eff-ynet: A dual task network for deepfake detection and segmentation. In *2021 15th International Conference on Ubiquitous Information Management and Communication (IMCOM)*, pages 1–8.

Yuezun Li, Xin Yang, P. S. H. Q. and Lyu, S. (2020). Celeb-df: A large-scale challenging dataset for deepfake forensics. In *IEEE Conference on Computer Vision and Patten Recognition (CVPR)*.

Zhao, H., Zhou, W., Chen, D., Wei, T., Zhang, W., and Yu, N. (2021a). Multi-attentional deepfake detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2185–2194.

Zhao, L., Zhang, M., Ding, H., and Cui, X. (2021b). Mff-net: deepfake detection network based on multi-feature fusion. *Entropy*, 23(12):1692.