HIGH SCHOOL EDITION
Journal of Student Research

Volume 12 Issue 1 (2023)

Disentanglement of Latent Factors of Real and Fake Appearance for Deepfake Face Manipulation Detection

Suh-Yoon Hong[1], Dayul Park[1] and GeunJung Yi[#]

[1]Cheongshim International Academy, Republic of Korea
[#]Advisor

ABSTRACT

A deepfake video is a video in which generative models are used to alter the facial features to make the subject appear to be a different person. There are various ways to utilize such content, including those that are positive such as entertainment. However, it is also very easy to exploit deepfake videos for harmful use, including for spreading fake news or creating unwanted content. Thus there have been numerous attempts to detect whether a video has been manipulated using deepfake technology so as to prevent further harm. Previous approaches for achieving this purpose have attempted to detect discrepancies in the video frames through the use of techniques such as exploiting the temporal consistency between each frame with convolutional neural networks. Though this has produced adequate results, its accuracy is insufficient for real-world use. In this paper, we propose a novel method of using a convolutional neural network based autoencoder to detect whether a video is pristine or deepfake. Our method successfully disentangles latent factors of real and fake appearance to increase the classification accuracy while maintaining a relatively low time complexity, enhancing real-world applicability. Results from extensive experimentation show significant improvement from state-of-the-art-methods by upwards of 18.51%.

Introduction

The word “Deepfake” is a compound word of “deep learning” and “fake”. A deepfake video is a video of a person whose face has been digitally altered using generative models so that they appear to be someone else [1]. Having been initially developed on the basis of GAN (Generative Adversarial Network) technology, it has now advanced to a point in which the deepfake video created is almost indistinguishable from a genuine video. In addition, with the distribution of open source video editing software, its use has been rapidly expanding over a variety of fields including special effects for video production and analysis of medical diagnostic imaging.

Although deepfake has potential for positive use, it also has a high risk of being exploited. It is often used to create adult content as well as fake news. For example, many celebrities are suffering and experiencing damage to their reputation due to the unapproved use of their faces or bodies in such videos. Not only is deepfake technology being used against celebrities, but it is also being exploited for domestic abuse and revenge pornography against ordinary people. Multiple cases have also been reported of fake news generated using deepfake technology causing confusion and misunderstandings among the public, such as a deepfake video to deceive investors [2] as well as a deepfake video of the president of Ukraine falsely asking the Ukranians to lay down their arms [3]. As such, preventing the exploitation of deepfake technology through the identification of deepfake videos has become imperative. There have been numerous deepfake detection studies proposed.

The early stages of deepfake detection methods showed the feasibility of exploiting convolutional neural networks for deepfake classifiers. Güera et al. proposed a LSTM (Long Short-Term Memory) based deepfake video classifier to analyze the sequential video frames [4]. Lima et al. showed utilizing discrete fourier transformed signals can help with increasing the accuracy of trained models [5]. Zhao et al. proposed a multi-attention based deepfake classifier [6]. This method aggregates the low–level textural feature and the high-level semantic features guided by the multi-attention map for fine-grained classification. The accuracy of these previous methods are heavily based on the training dataset samples as their models handle hand-crafted features which are not robust against recently developed high quality deepfake techniques.

To solve this problem, we propose a novel latent factor disentanglement method for deepfake detection systems. The proposed system is composed of an autoencoder and a classifier. The autoencoder disentangles the latent factors of real and fake appearance from the input face image. The classifier takes these disentangled latent factors as input to classify whether the input image is real or not. The proposed system achieves an accuracy of 98.81% and 82.18 on Celeb-DF [7] (Li, et al. 2020) and AI-Hub deepfake dataset [8] which are publicly available.

Related Works

Convolutional Neural Networks

CNN (Convolutional Neural Network) is similar to a traditional neural network in that it consists of neurons that optimize themselves through learning. The main difference between the layers of convolutional and neural networks is whether to keep the input dimension shape or not. In particular, the dimension of the CNN input image maintains spatial structure using convolution operations and it dramatically reduces the number of operations compared to the fully connected neural network. These CNN layers enable the trained network to extract more rich features compared to the layers of a traditional neural network since they preserve the spatial structure of the input image. Modern convolutional neural networks, such as Alexnet [9] (Krizhevsky, et al. 2017), VGG [10] (Simonyan, et al. 2014) and Resnet [11] (He, kaiming, et al. 2016), show remarkable performance and achieve state-of-the-art accuracy in many image classification problems. In this paper, we exploit Resnet to develop the proposed baseline model.

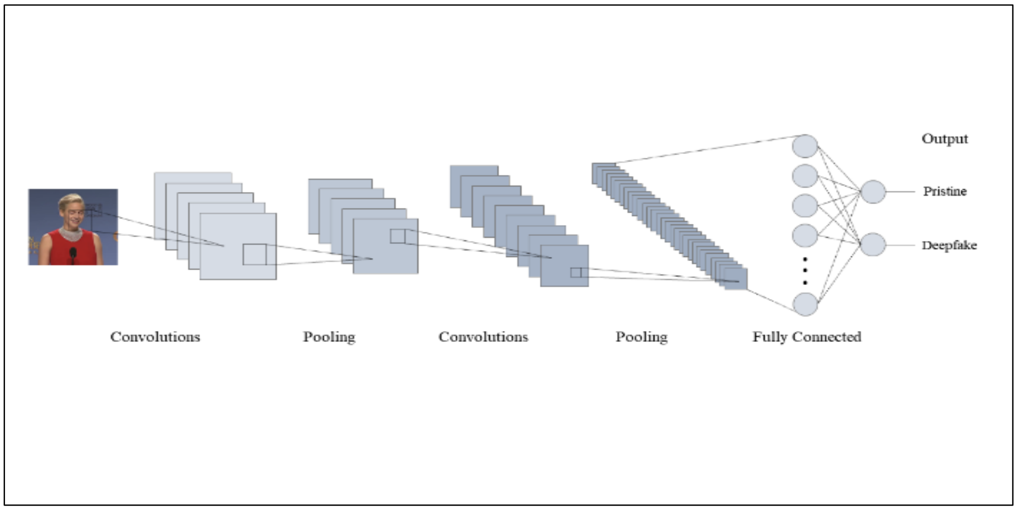

Figure 1. Construction of the Convolutional Neural Network

Image Classification

Image classification models assign an element of a predefined set of labels for a given image. This problem is one of the core areas of computer vision, and despite the simplicity of the model, it has various applications. For example, facial mask detection problems, which determine whether the individual wears a mask or not, can be easily solved by applying classification models [12] (Chavda, et al. 2021).

Another popular example of image classification is face verification. Face verification aims to verify whether the input face image is a match or not. Face verification systems have to confirm that the physical face of the user matches the one in the preinputted identification face image. Numerous face verification researches are developed heavily based on image classification models [13] (Sun, et al. 2013). In this paper, we consider deep fake detection as a classification problem. The system can assign the inputted image to the categories of modulated or real.

Generative Model

Generative models aim to learn the patterns or distributions in input image samples and synthesize new examples that plausibly could have been drawn from the dataset samples. There are three main approaches to generate images: VAE (Variational AutoEncoder) [14] (Kingma 2013), diffusion-based models [15] (Ho, et al. 2020), and GAN (Generative Adversarial Networks) [16] (Goodfellow, et al. 2020). VAE are widely being used in many images generative fields as their models are easy to train and can represent the useful latent features. However, their methods tend to produce blurry results which are easily distinguishable from real images. Diffusion models are currently actively being studied in image generation fields, but they still have technical issues to solve in order to produce high quality images. For this reason, most deepfake techniques mainly rely on the GANs which produce the cleanest and clearest image quality. In Particular, recent GANs can produce seamless synthesized face images that are indistinguishable from real face images. Therefore, this paper focuses on detecting deep fake images generated via GANs.

Method

Previous methods have shown that it is feasible to distinguish whether an image has been modified using the CNNs (Convolutional Neural Network) [17] (Badale, et al. 2018). However, these methods tend to yield poor results as their networks input entangled features into the classifier to produce the result. In general, face images contain various factors such as gender, age, ethnicity, illumination conditions, and background. In addition to the aforementioned factors, manipulated images have deepfake-related factors which can be extracted from manipulated regions of the image that are not seamless or natural. Therefore, in order to increase the accuracy, it is important to disentangle those factors from common factors that are irrelevant. The proposed system disentangles the deepfake-related factors by jointly training the autoencoder and the deepfake classifier.

Approach

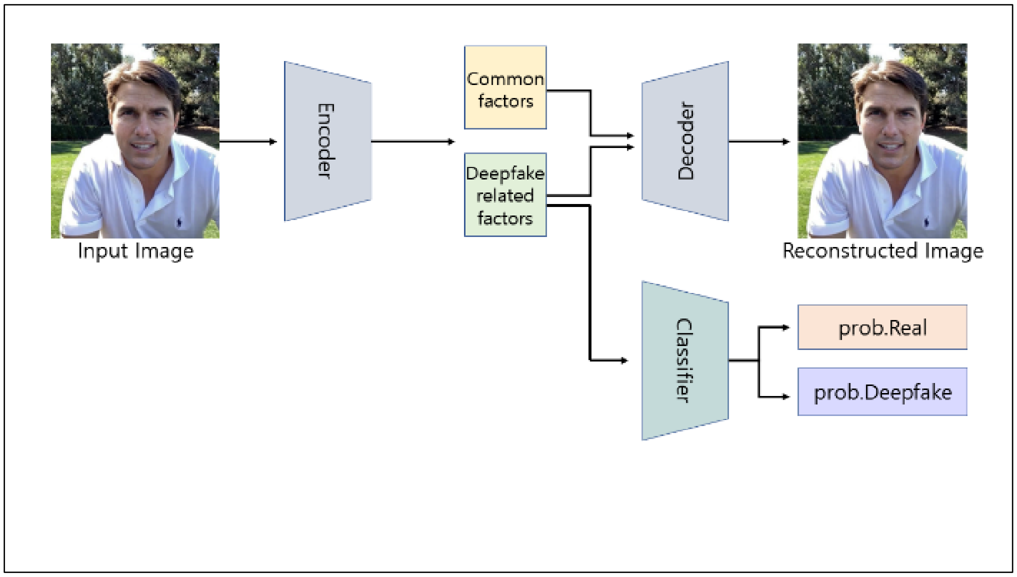

Figure 2. Overall architecture of the proposed deepfake detection system

Figure 2 represents the overall architecture of the proposed deepfake detection system. The proposed system is composed of an encoder, a decoder, and a deepfake classifier. The original input image, whether it is real or fake, is first inputted into the encoder. The encoder then executes two functions: compressing and disentangling the image features. The encoder generates compressed, disentangled image features consisting of data with common and deepfake related factors as the output. The compressed features are subsequently inputted through the decoder and classifier simultaneously. In the case of the decoder, latent factors with both common and deepfake related factors are inputted, and a reconstructed image is produced as a result. On the other hand, only latent factors with deepfake related features are inputted to the deepfake classifier.

As the decoder tries to reconstruct the original inputted image with the extracted features from the encoder, the encoder aims to extract the essential factors contained in the input image such as gender, age, ethnicity, illumination conditions, and background. Among the extracted features, the proposed deep fake classifier takes the deepfake-related features and produces the probability of the image being deepfake, and finally a result of either pristine or deepfake. This training strategy allows the trained model to successfully disentangle the deepfake-related features (features that signify whether an image is a deepfake) and make the model robust against real world samples. The effectiveness of this approach is studied further in chapter 4.3 in detail.

Network Architecture

For developing the encoder of the proposed method, we exploited Resnet18 [11] among various state-of-the-art networks including AlexNet [9] (Krizhevsky, et al. 2017), VGG [10] (Simonyan, et al. 2014), and RexNet [18] (Han, et al. 2021). Heuristically, we found Resnet18 to be most ideal as it has opposite depth which is deep enough to yield comparable accuracy while maintaining manageable time complexity, striking a balance between the trade-off of accuracy and time complexity. To implement the proposed decoder, we choose DenseNet [19] (Huang, et al. 2017) with the modification of replacing the downsampling layer with the upsampling layer in order to reconstruct the original input image. For the deep fake classifier, we use two linear layers. From substantial experimental results, it was proven that two linear layers architecture was most suitable as there was no significant improvement in accuracy even when there was an additional increase in depth.

Loss Function

The loss function is an indicator that represents the difference between the predictions and its corresponding ground truth. The loss function is an indicator of how poor the current model is processing data. In general, loss function must be defined for deep learning network training. The difference between the prediction and the ground truth is called loss, and learning proceeds in a way that reduces this loss.

In order to train the proposed method, we used two different loss functions: the cross-entropy loss function and the L1 loss function. The cross-entropy loss function computes the logarithmic loss through measuring the difference between the discovered probability distribution with the predicted distribution. The loss is a value between 0 and infinity, with an ideal model having a loss value of 0. (As such, the general goal is to get the loss of the model to be as close to 0 as possible). The cross-entropy loss function is often used for classification problems to measure the performance of the model [9-11]. We use the cross-entropy loss function to train the proposed deepfake classifier. Equation 1. is used to calculate the cross-entropy loss.

Equation 1:

$$L_{CE} = -\log \hat{y}$$

Where, ŷ denotes the softmax probability of the prediction scores.

The L1 loss function quantifies how similar two images are by computing the sum of the absolute difference between the actual value and the prediction. In the proposed method, the L1 loss function is used to train the autoencoder by comparing the pixel values of the input image and the reconstructed image. The L1 loss is calculated using Equation 2.

Equation 2:

$$L_1 = \frac{1}{W \times H} \sum_{x=1}^{W} \sum_{y=1}^{H} \left| I(x,y) - \hat{I}(x,y) \right|$$

Where, W and H represent the width and height of the image, respectively. I(x,y) and î(x,y) denote the pixel intensity of x and y coordinates for the ground truth and the reconstructed image. Finally, the overall loss function is defined as Equation 3.

Equation 3:

$$L = L_{CE} + L_1$$

Implementation Details

In this chapter, we explain the detailed information on how the proposed system is trained. We set the batch size as 32, and train the system for 200 epochs. The initial learning rate is set to 0.0001 and decreased by one tenth at 120 and 160 epoch. For example, when the epoch is from 0 to 119, the value of learning rate is 0.0001, and from 120 to 160 epoch it is set to 0.00001 and 0.000001 till the train finishes. For the optimizer, we use Adam [20] (Kingma, et al. 2014) with default parameter settings.

Experimental Results

In this chapter, we conduct experiments to prove the performance of the proposed method by comparing the results of the previous methods and that of proposed methods. For the quantitative evaluation metric, we use accuracy that is frequently used in other classification methods [21] (Hossin, et al. 2015). The metric has been widely utilized by state-of-the-art research as it is effective in numerically comparing the various classification methods. Accuracy is a metric value between 0 and 1, calculated by dividing the number of correct predictions by the total number of predictions made. The value 0 signifies all of the predictions being incorrect while 1 signifies all of the predictions being correct.

Datasets

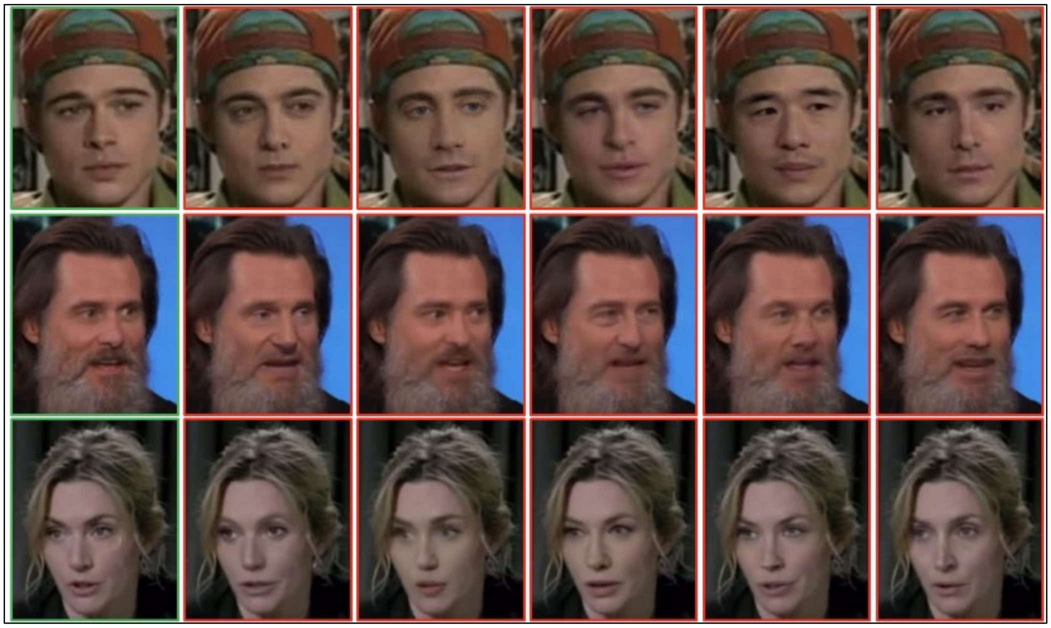

Figure 3. Snippet of Celeb-DF [7] (Li, et al. 2020) dataset

To train the proposed system, we use two types of datasets which are Celeb-DF [7] (Li, et al. 2020) and AI-Hub deepfake modulation dataset [8]. Both dataset are publicly available online. Figure 3. is a representative snapshot of the samples of Celeb-DF dataset.

First column in the figure shows pristine samples while other columns represent the corresponding deepfake modulated samples. The dataset consists of a total of 6,229 videos, 590 of which are original videos from YouTube, while 5,229 of the samples are the corresponding deep fake videos. The subjects of these videos are of a variety of ages, ethnicities, and genders, indicating that the dataset is varied.

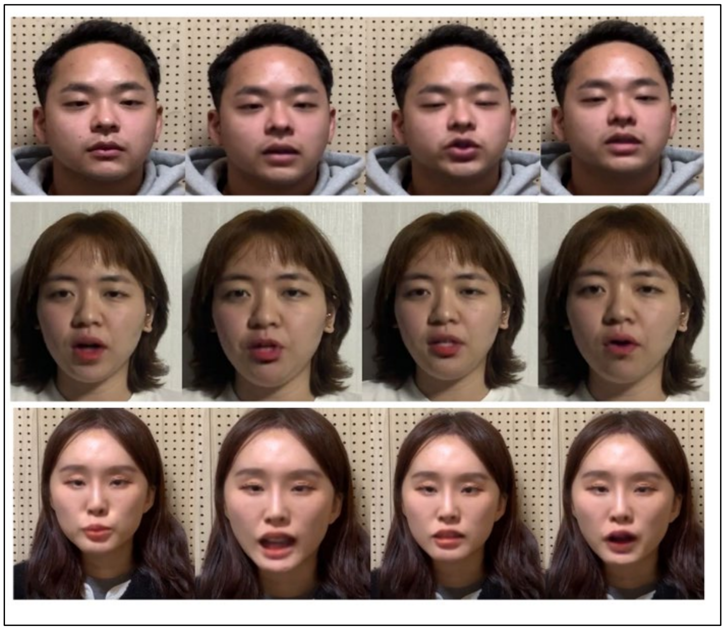

Figure 4. Snippet of AIHub deepfake modulation [8] dataset

Figure 4 represents the snapshot of the AI-Hub deepfake modulation dataset. The dataset is collected from 10 individuals with various types of deepfake modulation techniques. There are a total 150 videos containing 60K image frames captured from various lighting conditions. The individuals were asked to express different emotions as shown in figure 4. The first column in the figure shows the original images while the other columns represent modulated images. As deep fake technology has developed, modulated images have become more seamless and natural which makes the detection problem more challenging.

Qualitative Evaluation

In this paper, we evaluate the proposed method on both dataset and compare the accuracy to the existing state-of-the-art methods. For the comparison methods, we chose three different state-of-the-art deepfake detection methods [4, 22-23] that have comparable performance and were published relatively recently. For fair comparison, we trained the comparison with the aforementioned dataset with the same training protocol.

Table 1. Comparison results on Celeb-DF

Method	Accuracy (%)
Güera et al [4]	76.25
Carreira et al [22]	92.28
Tran et al [23]	97.49
Ours	98.81

Table 2. Comparison results on AI-Hub deepfake modulation dataset

Method	Accuracy (%)
Güera et al [4]	63.67

Carreira et al [22]	77.20
Tran et al [23]	80.08
Ours	82.18

Table 1 shows the comparison between the accuracy of state of the art methods and the proposed method for the Celeb-DF dataset. While Güera et al. [4] (Güera et al. 2018), Carreira et al. [22] and Tran et al. [23] (Tran et al 2018) present an accuracy of 76.25, 92.28, and 97.49% respectively, the proposed algorithm achieved an accuracy of 98.81%. Compared to Güera et al, it was shown to have 22.56% higher accuracy. Likewise, it showed higher performance compared to the methods of Carreira et al. and Tran et al., with the proposed method outperforming them by 6.53% and 1.32% respectively.

Table 2 represents the comparison between the state-of the art methods and that of the proposed method for the AI-Hub dataset. The methods presented by Güera et. al, Carreira et al., and Tran et al. each show an accuracy of 63.67, 77.20, and 80.08%. On the other hand, the proposed system shows an accuracy rate of 82.18%, which is 15.51, 4.98, 2.10 % higher compared to the performances exhibited by the systems of the comparison methods.

Overall, the proposed method outperforms other state-of-the-art methods consistently in both datasets.

In addition to evaluation on deepfake dataset manufactured specifically for experimental purposes, we also conduct additional experiments using real world deepfake samples. We collected four deepfake samples which were collected from various internet platforms such as YouTube and news articles. The first sample is a deepfake video of Barack Obama, the former president of the United States [24]. In this video, Obama is seen to make absurd statements. The second sample is a deepfake video of Morgan Freeman, in which a realistic Morgan Freeman questions whether he is in fact Morgan Freeman, casting doubt into our ability to perceive reality. In the third sample, a deepfake version of Korea's Former Minister of Small and Medium-sized Enterprises and Startups is shown to be introducing herself. The fourth sample is a deepfake of the Brazilian entrepreneur and television host Silivo Santos is delivering news about a new law making changes to traffic codes. For the experiments, we extract a total of 100 frames with 0.5 second interval from the video samples.

Table 3. Comparison results on the real world samples

	Sample1 [24]	**Sample2 [25]**	**Sample3 [26]**	**Sample4 [27]**
(Carreira et al. 2017)	53	64	65	71
(Tran et al. 2018)	71	75	84	79
Ours	76	82	92	84

As shown in Table 3, the proposed method shows a higher accuracy compared to the comparison methods. Overall, the proposed method outperforms all state-of-the-art methods in detecting deepfakes for real world samples. We attribute this superiority to the proposed autoencoder-based representation learning. By splitting the extracted features into deepfake-related and common factors the trained encoder successfully learns to extract the disentangled factors which directly affect the final accuracy. The effectiveness of the proposed idea is explained in the next chapter.

Ablation Study

In this chapter, we examine how each proposed idea contributes to the final accuracy of the proposed model by conducting ablation study. The causal relationship of the proposed idea can be effectively understood through the ablation study, which creates reliable knowledge. Ablation study is to compare the model that does not include the model when it wants to confirm how the proposed element affects the model. In machine learning, ablation study can be defined as a scientific experiment to obtain insights on the effect on overall performance by removing building blocks of the machine learning system.

Table 4. Ablation study results

Model	Accuracy (%)
w/o autoencoder	79.07
w/o latent factor split	80.64
Full model	82.18

For the first ablation model, we omit the decoder in the proposed system. This model is considered a typical deterministic convolutional neural network. To train the second ablation model, we feed the entire latent factor to both decoder and classifier rather than splitting the code into deepfake-related and common factors. Table 4 summarizes two ablation study results. The full model achieves an accuracy of 82.18 model while each first and second ablation model results 79.07 and 80.64, respectively. This results clearly shows each proposed idea affects the accuracy of the trained model.

Conclusion

In this paper, we proposed a novel representation learning for latent factor disentanglement for a deepfake detection. The proposed system consists of the encoder, decoder, and deepfake classifier, for which we employed ResNet-18, the modified DenseNet, and two linear layers respectively. The proposed method differentiates itself from state-of-the-art methods in that it disentangles deepfake-related latent factors from latent factors that are extraneous such as skin color, age, or background. Collectively, the experimental results exhibited higher performance of the proposed method compared to previous existing models for both the Celeb-DF and AI-Hub deepfake modulation datasets by 22.56% and 18.51% respectively. The results also demonstrated the effectiveness of the proposed method in detecting deepfakes in real-world scenarios. As such, we highlighted the broad applicability of the proposed algorithm and its potential for detecting deepfake content on the internet. In the future, we plan to create an application for determining whether a video has been modified using deepfake technology, and hope that it will be able to contribute to reducing the adverse effects of deepfake technology.

Acknowledgments

I would like to thank my advisor for the valuable insight provided to me on this topic.

References

Blitz, M. J. (2018). Lies, line drawing, and deep fake news. Okla. L. Rev., 71, 59.

Deepfakes: What are they, and why are they dangerous?[Website]. (2022, Otc 3).
https://wyche.com/insights/blog/posts/deepfakes-what-are-they-and-why-are-they-dangerous

Deepfake video of Volodymyr Zelensky surrendering surfaces on social media[Website]. (2022 Oct 3). https://www.youtube.com/watch?v=X17yrEV5sl4

Güera, D., & Delp, E. J. (2018, November). Deepfake video detection using recurrent neural networks. In 2018 15th IEEE international conference on advanced video and signal based surveillance (AVSS) (pp. 1-6). IEEE. https://doi.org/10.1109/AVSS.2018.8639163

de Lima, O., Franklin, S., Basu, S., Karwoski, B., & George, A. (2020). Deepfake detection using spatiotemporal convolutional networks. arXiv preprint arXiv:2006.14749. https://doi.org/10.48550/arXiv.2006.14749

Zhao, H., Zhou, W., Chen, D., Wei, T., Zhang, W., & Yu, N. (2021). Multi-attentional deepfake detection. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 2185-2194). https://doi.org/10.1109/cvpr46437.2021.00222

Li, Y., Yang, X., Sun, P., Qi, H., & Lyu, S. (2020). Celeb-df: A large-scale challenging dataset for deepfake forensics. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 3207-3216). https://doi.org/10.1109/cvpr42600.2020.00327

Deepfake modulated video[Website]. (2022 Oct 3). https://www.aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=realm&dataSetSn=55

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). Imagenet classification with deep convolutional neural networks. Communications of the ACM, 60(6), 84-90. https://doi.org/10.1145/3065386

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556. https://doi.org/10.1109/cvpr.2016.90

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778). https://doi.org/10.1109/cvpr.2016.90

Chavda, A., Dsouza, J., Badgujar, S., & Damani, A. (2021, April). Multi-stage CNN architecture for face mask detection. In 2021 6th International Conference for Convergence in Technology (i2ct) (pp. 1-8). IEEE. https://doi.org/10.1109/i2ct51068.2021.9418207

Sun, Y., Wang, X., & Tang, X. (2013). Hybrid deep learning for face verification. In Proceedings of the IEEE international conference on computer vision (pp. 1489-1496). https://doi.org/10.1109/iccv.2013.188

Kingma, D. P., & Welling, M. (2013). Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114. https://doi.org/10.48550/arXiv.1312.6114

Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems, 33, 6840-6851.
https://doi.org/10.48550/arXiv.2006.11239

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2020). Generative adversarial networks. Communications of the ACM, 63(11), 139-144.
https://doi.org/10.48550/arXiv.1406.2661

Badale, A., Castelino, L., Darekar, C., & Gomes, J. (2018). Deepfake detection using neural networks. In 15th IEEE international conference on advanced video and signal based surveillance (AVSS).
https://doi.org/10.7717/peerj-cs.881

Han, D., Yun, S., Heo, B., & Yoo, Y. (2021). Rethinking channel dimensions for efficient model design. In Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition (pp. 732-741).
https://doi.org/10.48550/arXiv.2007.00992

Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4700-4708).
https://doi.org/10.48550/arXiv.1608.06993

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.
https://doi.org/10.48550/arXiv.1412.6980

Hossin, M., & Sulaiman, M. N. (2015). A review on evaluation metrics for data classification evaluations. International journal of data mining & knowledge management process, 5(2), 1.
https://doi.org/10.5121/ijdkp.2015.5201

Carreira, J., & Zisserman, A. (2017). Quo vadis, action recognition? a new model and the kinetics dataset. In proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 6299-6308).
https://doi.org/10.1109/CVPR.2017.502.

Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., & Paluri, M. (2018). A closer look at spatiotemporal convolutions for action recognition. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (pp. 6450-6459). https://doi.org/10.1109/CVPR.2018.00675

You Won’t Believe What Obama Says In This Video![Website]. (2022 Oct 3).
https://www.youtube.com/watch?v=cQ54GDm1eL0

This is not Morgan Freeman - A Deepfake Singularity[Website]. (2022 Oct 3).
https://www.youtube.com/watch?v=oxXpB9pSETo

박영선인데 박영선 아니다…"영상을 믿지 마세요" (2019.11.13/뉴스데스크/MBC)[Website]. (2022 Oct 3).
https://www.youtube.com/watch?v=hqZhH9Qr4B0

Silvio Santos apresentando o Jornal Nacional[Website]. (2022 Oct 3). https://www.youtube.com/watch?v=VDqTIThdj1s