A Novel Facial Manipulation Detection Method Based on Contrastive Learning

Zhiyuan Ma
College of Electrical Engineering and Control Science
Nanjing Tech University
Nanjing, China
e-mail address:
zhiyuanma741@gmail.com

Pengxiang Xu
School of Computer Science and Engineering
Nanjing University of Science and Technology
Nanjing, China
e-mail address:
xpx0517@163.com

Xue Mei*
College of Electrical Engineering and Control Science
Nanjing Tech University
Nanjing, China
*Corresponding author's e-mail address: mx@njtech.edu.cn

Jie Shen
College of Electrical Engineering and Control Science
Nanjing Tech University
Nanjing, China
e-mail address:
shenjie@njtech.edu.cn

2022 IEEE 5th International Conference on Electronics Technology (ICET) | 978-1-6654-8508-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICET55676.2022.9825156

***Abstract*—Nowadays, numerous synthesized face-swapping videos generated by face forgery algorithms have become an emerging problem, which promotes facial manipulation detection to be a significant topic. With the development of face forgery algorithms, some fake face images or videos generated by those strong forgery algorithms are very realistic, which have brought much difficulty to facial manipulation detection. In this paper, we present a novel facial manipulation detection method based on contrastive learning. We analyze the texture features of manipulated facial images and propose to compare and learn the features of the whole face and the center face in order to get more general features. We calculate the similarity and distribution distance between the whole face and the center face. The experiments implemented on FaceForensics++ dataset demonstrate that the proposed method achieves outstanding results and can learn the general features.**

Keywords—Facial Manipulation Detection, Face Forgery Detection, Contrastive Learning, Siamese Network, Deep learning

I. INTRODUCTION

Facial manipulation is a technology that uses computer graphics-based techniques or deep learning methods to process face images or videos to generate new tampered or synthetic face images or videos. In recent years, Generative Adversarial Network (GAN) has become a hot topic in deep learning, the emergence and development of GAN has greatly promoted the progress of facial manipulation technology [13]. It is difficult for human eyes to distinguish the fake face images and videos generated by GANs, even for some modest detection algorithms. Therefore, research on facial manipulation detection is extremely necessary now.

Facial manipulations can be categorized into four main different types: entire face synthesis, identity swap, attribute manipulation and expression swap [19]. Entire face synthesis generates non-existent face images, usually using GANs. Identity swap replaces the face of one person with the face of another person in order to exchange identity. Attribute manipulation consists of changing some attributes of the face such as the color and style of the hair, adding glasses, etc. Expression swap refers to changing the expression of one person according to the expression of another person without changing the identity. The most common manipulation methods are identity swap and expression swap such as Deepfakes [6], FaceSwap [10], Face2Face [15], NeuralTextures [14], etc., as shown in Fig. 1. These technologies can be used for Virtual Reality (VR), Augmented Reality (AR) and other visual tasks, they also can provide convenience and possibilities for other industries such as education, film and TV.

In recent years, with the rapid development of face forgery algorithms, it is much easier to generate hyper-realistic synthetic face images and videos than before. People could download related codes from the Internet or use mobile or computer applications to synthesis videos. However, due to the deception nature of facial manipulation technology and the use of real faces, the digital content generated by such technologies may infringe on public security and individual rights when they are used and spread illegally. People tend to be misled and confused when exposed to false information or fake news, which may cause a loss of trust to digital content on the Internet.

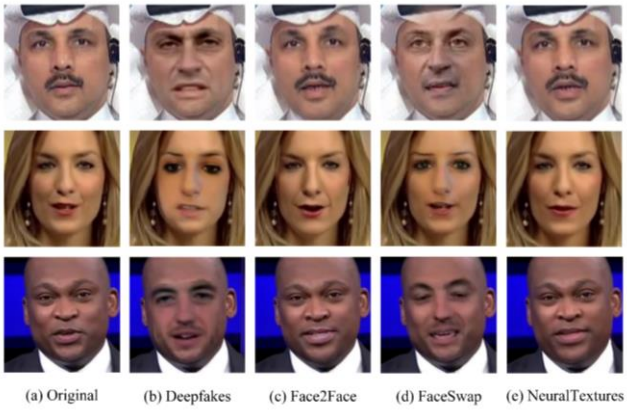

Fig. 1. The most common facial manipulation methods.

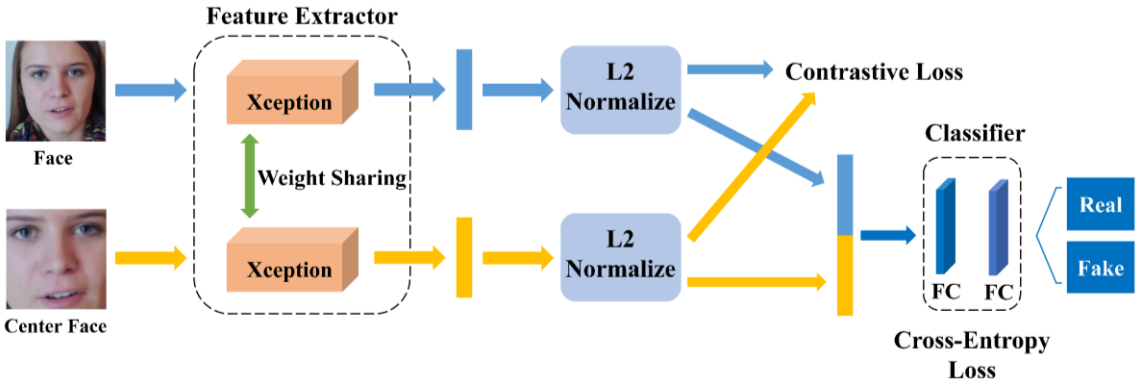

Fig. 2. The framework of the model.

In order to deal with the above problems, much effort has been made to detect manipulated facial images and videos. Many researchers have proposed various methods and achieved some good results. Some researchers tried to focus on some behaviors or areas of the face. For instance, Yuezun et al. [22] and TackHyun et al. [21] noticed that the eye blinking was discontinuous in the fake video and proposed a detection method based on this. Rössler et al. [2] trained the face images with neural networks directly for binary classification. These methods extracted different features of the face and achieved good results on specific datasets. However, most of them relied too much on datasets and could not learn more general features that can be generalized to other tasks.

In this paper, we introduce contrastive learning method into facial manipulation detection problem. Siamese network, as a classic structure of contrastive learning, is applied in this work. Contrastive loss function is additionally added into the total loss function, which can help the model calculate the similarity and distribution distance between samples. Contrastive learning method enables the model to compare different regions of the image during training, so as to get more general feature. We implement our method on FaceForensics++ dataset, the great performance shows the effectiveness of contrastive learning method.

II. Related Works

In what follows, we describe traditional methods and data driven methods to facial manipulation detection. We then elaborate on the contrastive learning on facial manipulation detection.

A. Data Driven Methods

Data driven methods usually use facial images or videos to train the neural network to obtain classification ability. Rössler et al. [2] proposed to use XceptionNet to classify the real and fake images directly and achieved good results. Xu et al. [25] regarded the multiple video frames as a set and tried to enhance the performance of the model by learning the elements that carry the same characteristics in the set. Wang et al. [3] used Siamese Network to measure the similarity between the face area and the background of the image to extract the features in the grey space. Dang et al. [11] used attention mechanisms to improve the feature maps of CNN models. Although various neural networks could help to improve the performance of the models, most of these models ignore the inherent features of fake images.

B. Traditional Methods

Before the advent of Convolutional Neural Networks (CNNs) [1], typical methods to facial manipulation detection relied on some handcrafted features to detect the unreasonable regions of the face. As mentioned above, Yuezun et al. [22] and TackHyun et al. [21] detected facial manipulation according to the inconsistency of the eye blinking. Matern et al. [9] proposed a detection method based on the discrepancy in color, reflections and other details around the eyes and mouth between real faces and fake faces. Korshunov et al. [18] proposed a detection method based on extracting the inconsistencies between lip movements and audio speech. These methods rely on hand-extracted specific features and fail to learn the more general features, which means they cannot be generalized to other datasets.

C. Contrastive Learning

Contrastive learning is a typical method of self-supervised learning and aims at learning more general features rather than pixel-level generation. Saunshi et al. [17] tried to improve the inner product of positive samples and negative samples and guarantee the generalization ability of the model. Chen et al. [20] presented a new framework named SimCLR to learn general visual representation that enables the model distinguish the similarity of objects and achieved better results on ImageNet dataset than supervised learning methods. He et al. [16] proposed Momentum Contrast (MoCo) to learn visual representation that can be well transferred to downstream tasks. MoCo method obtained good results in many visual tasks, even much better than supervised pre-trained methods. Tian et al. [24] proposed to reduce the mutual information between views in the face of contrastive learning between multiple views of the data, and keep task-relevant information intact. Contrastive learning has been proved to be effective and achieved outstanding performance on various visual tasks.

III. Prpposed Method

In this section, we introduce Siamese network, which is a classical structure in contrastive learning. Then we introduce our method in detail and analyze the application of contrastive learning in facial manipulation detection. The framework of our model is shown as Fig. 2.

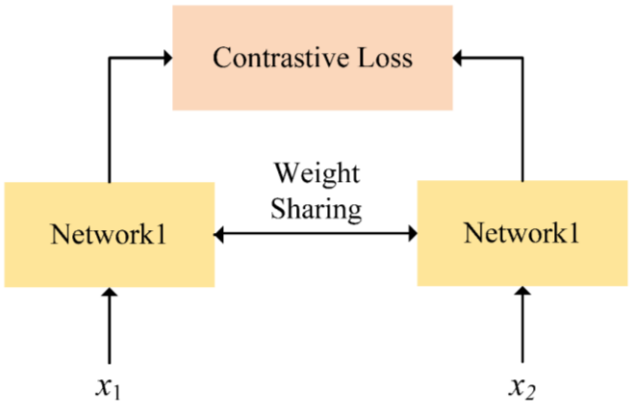

Fig. 3. The framework of Siamese Network.

A. Siamese Network

Siamese Network is often used for various contrastive learning tasks. The framework of Siamese Network is shown as Fig. 3. Siamese Network is composed of two neural networks with the same structure and parameters, which share weights. The input of the Siamese Network is two different samples, and the output is two feature vectors. Siamese Network can calculate the similarity of two samples and learn the distribution difference between different samples.

B. Contrastive Learning on Facial Manipulation Detection

1) Contrasting and learning general features

Texture feature is an important cue in image binary classification task. Neural networks can learn the texture feature of images to distinguish the authenticity of facial images. The texture distribution of real faces is consistent, while the texture distribution of fake faces is inconsistent in some regions. Most face forgery algorithms replace the expression or the whole face of one person with the expression or the whole face of another person in order to exchange identity or expression. Hence, the texture distribution of fake faces is inconsistent around the edge of face. The distribution of the part of the face outside the edge is consistent with the background of the image, while the center face within the edge is another distribution.

As mentioned in section A, Siamese Network can learn the distribution difference between different samples. Therefore, we design a model based on Siamese Network. As shown in Fig. 2, we take the whole face image and the center face as the inputs of the Siamese Network. Siamese Network can contrast the feature distribution of the whole face and the center face, and learn the distribution difference of two samples. The goal of contrastive learning is defined as follows:

$$Distance(G(x), G(x^+)) << Distance(G(x), G(x^-)) \quad (1)$$

$$Score(G(x), G(x^+)) >> Score(G(x), G(x^-)) \quad (2)$$

In this task, $G(x)$ represents a sample, that is, the real face image. $G(x^+)$ represents a sample that is similar to $G(x)$, that is, a real center face. $G(x^-)$ represents a sample that is not similar to $G(x)$, that is, a fake center face. Contrastive learning makes the distance between dissimilar samples and the score between similar samples larger, and the distance between similar samples and the score between dissimilar samples smaller. Contrastive learning compares the features of different regions of the image and learns more general features.

2) Normalization

The output of each network is a feature vector of size 2048×1. Due to the difference in backgrounds and production techniques, the norm of feature vectors output from different face images through the same network may be quite different. Therefore, we use l_2 normalization to normalize the feature vectors. The calculation formula is:

$$Y = \frac{X}{\| X \|} \quad (3)$$

where X represents the feature vector output from the model, Y represents the normalized feature vector.

3) Classifier

Two normalized feature vectors of size 2048×1 are recombined into a new vector of size 4096×1, which realizes the fusion of different features. As shown in Fig. 2, the new vector is input into the label classifier composed of two fully connected layers.

4) Loss function

Our method uses two different loss function: contrastive loss function and cross-entropy loss function. Contrastive learning can calculate the similarity of two samples. In this work, we use cosine similarity to measure the similarity of two feature vectors. Cosine similarity is defined as (6).

$$A_i = [a_i^1, a_i^2, a_i^3, ..., a_i^{2048}] \quad (4)$$

$$B_i = [b_i^1, b_i^2, b_i^3, ..., b_i^{2048}] \quad (5)$$

$$sim_i = \frac{\sum_{i=1}^{2048}(a_i \times b_i)}{\sqrt{\sum_{i=1}^{2048}(a_i^2)} \times \sqrt{\sum_{i=1}^{2048}(b_i^2)}} \quad (6)$$

A_i and B_i represents the feature vectors of the whole face and the center face of sample S_i, respectively. sim_i calculates the cosine similarity between A_i and B_i. Based on this, we design the contrastive loss function:

$$l_{con} = \frac{1}{2N}\sum_{i=1}^{N}[(1-\ell)sim_i^2 + \ell \max(margin - sim_i, 0)^2] \quad (7)$$

where ℓ represents the label of the sample, the label of real face is 0, the label of fake face is 1. *margin* represents the measurement threshold, the value of *margin* is set to 0.5 according to the experiment. When the label is 0, sim_i^2 will be minimized. When the label is 1, $\max(margin - sim_i, 0)^2$ will be minimized. Thus, the distribution of the center face and the original feature of real images can be closer, and the distribution of the center face and the original feature of fake images can be spread farther.

The dimension of the output of the label classifier is 2, which represents the predicted probability of real image and

fake image, respectively. Therefore, we use cross-entropy loss function here:

$$l_{cls} = -\sum \ell \log(p_{S_i}) \tag{8}$$

The total loss function L shown as (9), where k is an equilibrium coefficient.

$$L = l_{cls} + k \cdot l_{con} \tag{9}$$

IV. Experiments

In this section, we introduce the datasets and details of the experiments. Then we implement our method on the datasets and compare the results with the result of several typical models and the state-of-the-art results. The results achieved on the datasets show the effectiveness of our method.

A. Datasets

We implement our method on FaceForensics++ [2] to improve the effectiveness of our method. FaceForensics++ is a large-scale dataset that contains 1000 original videos and 4000 fake videos. FaceForensics++ is produced with four different facial manipulated techniques: Deepfakes [6], Face2Face [15], FaceSwap [10], and NeuralTextures [14], each of them corresponds to 1000 fake videos. The videos in FaceForensics++ are all from YouTube. FaceForensics++ provides videos with three resolutions: uncompressed raw videos (c0), high quality videos lightly compressed with 23 quantized (c23), and low quality videos heavily compressed with 40 quantized (c40).

B. Implementation Details

The dataset is divided into three parts according to training set, validation set and test set. The training set contains 720 videos, the validation set contains 140 videos, and the test set contains 140 videos. Each video is split frame by frame. We use Dlib [5] to extract the face region and scale it down to get the center face. The experiments are implemented on the images.

The experiment platform is a high-performance computer. The CPU is Intel i7 8700K, the GPU is dual NVIDIA GTX 1080Ti, and the video memory is 22G. The experiments have been implemented with Python 3.8 using PyTorch, CUDA 10.0, and cuDNN 7.6.2. We select Xception [8] as the backbone network. The networks are optimized by the Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\varepsilon = 1 \times 10^{-8}$. The learning rate is set to 0.0002, the batch size is set to 32, and the epoch is set to 15.

C. Results and Comparisons

1) Experiments on raw images

We first implement the experiments on raw images of all four subdatasets: Deepfakes (DF), Face2Face (F2F), FaceSwap (FS) and NeuralTextures (NT), and compare the accuracy (Acc) of our method with several most widely used networks: MesoNet4 [4], MesoInception4 [4], and Xception [8]. As shown in Table I, our method achieves Acc scores of 99.87%, 98.25%, 99.96%, and 97.65% on Deepfakes, Face2Face, FaceSwap, and NeuralTextures, respectively, which are all higher than other three networks.

TABLE I. Acc Score on the Raw Images

Method	DF	F2F	FS	NT
Meso4 [4]	96.37	97.95	98.17	93.30
MesoInception4 [4]	88.34	97.65	97.81	92.52
Xception [8]	98.31	97.75	98.10	96.45
Our method	**99.87**	**98.25**	**99.96**	**97.65**

TABLE II. Acc Score on the c23 Images

Method	DF	F2F	FS	NT
Meso4 [4]	89.44	94.25	95.50	78.70
MesoInception4 [4]	83.74	91.48	94.34	75.06
Xception [8]	95.15	97.07	95.96	87.99
Our method	**96.23**	**97.45**	**96.37**	**89.28**

TABLE III. Comparision with Recent Works

Method	DF	F2F	FS	NT
Wang et al. [3]	98.72	97.92	98.77	**98.18**
Sabir et al. [7]	96.9	94.35	96.3	-
Luo et al. [23]	96.1	97.59	98.22	-
OC-FakeDect-1 [12]	86.0	70.70	84.5	95.2
OC-FakeDect-2 [12]	88.5	71.2	86.3	97.9
Our method	**99.87**	**98.25**	**99.96**	97.65

2) Experiments on compressed images

Similar to the experiments on raw images, we implement our method on c23 images and compare the Acc score with several widely used Networks [4, 8]. As shown in Table II, the Acc scores of four methos on c23 images are all lower than the Acc scores on c0 images, which means the image compression has a certain influence on the detection results. As mentioned before, texture features are the main basis in this work. Image compression brings noise interference, which may lead to inconsistent and unnatural local features. Nevertheless, our method still achieves the best Acc scores on for subdatasets, which shows the effectiveness of our method, which proves our method works on both uncompressed and compressed images.

3) Comparison with recent works

In order to further demonstrate the effectiveness of our method, we compare the Acc score of our method with the recent works [3, 7, 23, 12]. As can be seen in Table III, our method achieves best Acc scores on Deepfakes, Face2Face, and FaceSwap. Wang et al. [3] achieve 98.18% Acc score on NeuralTextures, which is the best Acc score on NeuralTextures. The performance of our method exceeds most recent methods, especially on Deepfakes, Face2Face, and FaceSwap.

V. Conclusion

In this work, we analyze the common facial manipulation methods and the difference of texture features between real facial images and fake facial images. We propose that the texture features of fake facial images is inconsistent around the

face, while the texture features of real facial images in consistent throughout the image. Based on this, we design a novel facial manipulation detection method using contrastive learning. Siamese Network is applied to contrast the texture distribution between the whole face and the center face. Contrastive loss function is applied in this work in order to calculate the similarity of two samples, and make the distance of similar samples smaller and the distance of dissimilar samples larger so as to distinguish real images and fake images. We implement our method on FaceForensics++ dataset and compare the performance of our method with several widely used networks and recent works. The outstanding performance of our model improves that our method can learn the general feature. In further, we will explore the more general feature representation and apply our method to more visual tasks.

ACKNOWLEDGMENT

This work was supported by Research Center of Security Video and Image Processing Engineering Technology of Guizhou (China) under Grant SRC-Open Project ([2020]001]).

REFERENCES

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," Advances in neural information processing systems. vol. 25, 2012.

[2] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Niessner, "FaceForensics++: Learning to Detect Manipulated Facial Images," In Proceedings of the IEEE/CVF International Conference on Computer Vision. Korea, pp. 1-11, November 2019.

[3] B. Wang, Y. Li, X. Wu, Y. Ma, Z. Song, and M. Wu, "Face Forgery Detection Based on the Improved Siamese Network," Secur. Commun. Netw. vol. 2022, February 2022.

[4] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "MesoNet: a Compact Facial Video Forgery Detection Network," 2018 IEEE International Workshop on Information Forensics and Security. HongKong, pp. 1-7, December 2018.

[5] D. E. King, "Dlib-ml: A Machine Learning Toolkit," J. Mach. Learn. vol.10, pp. 1755–1758, 2009.

[6] Deepfakes. https://www.github.com/deepfakes/faceswap. Accessed: 2018-10-29.

[7] E. Sabir, J. Cheng, A. Jaiswal, W. AbdAlmageed, I. Masi, and P. Natarajan, "Recurrent Convolutional Strategies for Face Manipulation Detection in Videos," Interfaces (GUI). vol. 3, pp. 80-87, 2019.

[8] F. Chollet, "Xception: Deep Learning with Depthwise Separable Convolutions," In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, pp. 1800–1807, November 2017.

[9] F. Matern, C. Riess, and M. Stamminger, "Exploiting Visual Artifacts to Expose DeepFakes and Face Manipulations," 2019 IEEE Winter Applications of Computer Vision Workshops. Waikoloa, pp. 83-92, January 2019.

[10] FaceSwap. https://www.github.com/MarekKowalski/FaceSwap. Accessed: 2019-09-30.

[11] H. Dang, F. Liu, J. Stehouwer, X. Liu, and A. K. Jain, "On the Detection of Digital Face Manipulation," In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, pp. 5781-5790, June 2020.

[12] H. Khalid and S. S. Woo, "OC-FakeDect: Classifying Deepfakes Using One-class Variational Autoencoder," In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Seattle, pp. 656-657, June 2020.

[13] I. Goodfellow, J. P. Abadie, M. Mirza, B. Xu, D. W. Farley, S. Ozair et al., "Generative Adversarial Nets," Advances in Neural Information Processing Systems. vol. 27, 2014.

[14] J. Thies, M. Zollhöfer, and M. Nießner, "Deferred neural rendering: image synthesis using neural textures," ACM Trans. Graph. vol. 38, pp. 1-12, July 2019.

[15] J. Thies, M. Zollhöfer, M. Stamminger, C. Theobalt, and M. Nießner, "Face2Face: Real-Time Face Capture and Reenactment of RGB Videos," In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, pp. 2387–2395, June 2016.

[16] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum Contrast for Unsupervised Visual Representation Learning," In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, pp. 9729-9738, June 2020.

[17] N. Saunshi, O. Plevrakis, S. Arora, M. Khodak, and H. Khandeparkar, "A Theoretical Analysis of Contrastive Unsupervised Representation Learning," In Proceedings of the 36th International Conference on Machine Learning. vol. 97, pp. 5628-5637, June 2019.

[18] P. Korshunov and S. Marcel, "Deepfakes: a New Threat to Face Recognition? Assessment and Detection," arXiv:1812.08685. 2018.

[19] R. Tolosana, R. V. Rodriguez, J. Fierrez, A. Morales, and J. O. Garcia, "Deepfakes and beyond: A Survey of face manipulation and fake detection," Inf. Fusion. vol.64, pp. 131-148, July 2020.

[20] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A Simple Framework for Contrastive Learning of Visual Representations," In Proceedings of the 37th International Conference on Machine Learning. vol. 119, pp. 1597-1607, July 2020.

[21] T. Jung, S. Kim, and K. Kim, "DeepVision: Deepfakes Detection Using Human Eye Blinking Pattern," IEEE Access. vol. 8, pp. 83144-83154, April 2020.

[22] Y. Li, M. Chang, and S. Lyu, "In Ictu Oculi: Exposing AI Created Fake Videos by Detecting Eye Blinking," 2018 IEEE International Workshop on Information Forensics and Security. Hong Kong, pp. 1-7, December 2018.

[23] Y. Luo, F. Ye, B. Weng, S. Du, and T. Huang, "A Novel Defensive Strategy for Facial Manipulation Detection Combining Bilateral Filtering and Joint Adversarial Training," Secur. Commun. Netw. vol. 2021, August 2021.

[24] Y. Tian, C. Sun, B. Poole, D. Krishnan, C. Schmid, and P. Isola. "What Makes for Good Views for Contrastive Learning?" Advances in Neural Information Processing Systems. vol. 33, pp. 6827-6839, 2020.

[25] Z. Xu, J. Liu, W. Lu, B. Xu, X. Zhao, B. Li et al., "Detecting facial manipulated videos based on set convolutional neural networks," J. Vis. Commun. Image Represent. vol. 77, pp. 103119, May 2021.