MMD based Discriminative Learning for Face Forgery Detection

Jian Han and Theo Gevers

University of Amsterdam, Amsterdam, the Netherlands
{j.han, th.gevers}@uva.nl

Abstract. Face forensic detection is to distinguish manipulated from pristine face images. The main drawback of existing face forensics detection methods is their limited generalization ability due to differences in domains. Furthermore, artifacts such as imaging variations or face attributes do not persistently exist among all generated results for a single generation method. Therefore, in this paper, we propose a novel framework to address the domain gap induced by multiple deep fake datasets. To this end, the maximum mean discrepancy (MMD) loss is incorporated to align the different feature distributions. The center and triplet losses are added to enhance generalization. This addition ensures that the learned features are shared by multiple domains and provides better generalization abilities to unseen deep fake samples. Evaluations on various deep fake benchmarks (DF-TIMIT, UADFV, Celeb-DF and FaceForensics++) show that the proposed method achieves the best overall performance. An ablation study is performed to investigate the effect of the different components and style transfer losses.

1 Introduction

With the rapid development of face manipulation and generation, more and more photo-realistic applications have emerged. These modified images or videos are commonly known as deep fakes [1]. Even human experts find it difficult to make a distinction between pristine and manipulated facial images. Different generative methods exist nowadays to produce manipulated images and videos. In fact, it's easy to generate new types of synthetic face data by simply changing the architectural design or hyper parameters. Attackers don't need to have profound knowledge about the generation process of deep fake (face) attacks [2]. Therefore, it is of crucial importance to develop robust and accurate methods to detect manipulated face images.

Face forensic detection is to distinguish between manipulated and pristine face images. Using the same pair of subjects, different manipulation methods may generate significantly different outcomes (see Fig. 1). If the same modification method is applied on different pairs of data, the results can have quite diverse artifacts due to the variations in pose, lighting, or ethnicity. Because these artifacts do not exist in all samples, simple artifacts-based detection systems are not sufficiently robust to unseen artifacts in the test set. Other methods

choose to exploit cues which are specific for the generative network at hand [3, 4]. When the dataset is a combination of multiple domains of deep fake data like FaceForensics++, each category of manipulated face images can be a different domain compared to the rest of the data. Performance may be negatively affected by this cross-domain mismatch. In summary, the major challenges of face forensics detection are: 1) The difference among positive and negative samples is much smaller than the difference among positive examples. 2) The artifacts including imaging variations and face attributes do not persist across all generated results for a single generation method.

Our paper focuses on detecting manipulated face images which are produced by generative methods based on neural networks. To distinguish real and fake face images is equivalent to performance evaluations of different generative methods. To this end, the maximum mean discrepancy (MMD) is used to measure different properties and to analyze the performance of different generative adversarial networks [5]. The face manipulation process requires a pair of faces from source and target subjects. This process resembles the neural style transfer operation between content and reference images [6]. The final results are contingent on the source and target images. Inspired by [7], we use a MMD loss to align the extracted features from different distributions. A triplet loss is added to maximize the distance between real and fake samples and to minimize the discrepancies among positive samples. Center loss is further integrated to enhance the generalization ability. In order to fully investigate the performance of the proposed method, we evaluate our method on several deep fake benchmarks: DF-TIMIT [8], UADFV [9], Celeb-DF [10], and FaceForensics++ [1].

Our main contributions are:

- We propose a deep network based on a joint supervision framework to detect manipulated face images.
- We systematically examine the effect of style transfer loss on the performance of face forgery detection.
- The proposed method achieves the overall best performance on different deep fake benchmarks.

2 Related Work

2.1 Face Manipulation Methods

In general, face manipulation methods can be classified into two categories: facial reenactment and identity modification [1]. Deep fake has become the name for all face modification methods. However, it is originally a specific approach based on an auto-encoder architecture. Face swap represents methods that use information of 3D face models to assist the reconstructing process. Face2Face [11] is a facial reenactment framework that transfers the expressions of a source video to a target video while maintaining the identity of the target person. NeuralTextures [12] is a GAN-based rendering approach which is applied to the transformation of facial expressions between faces.

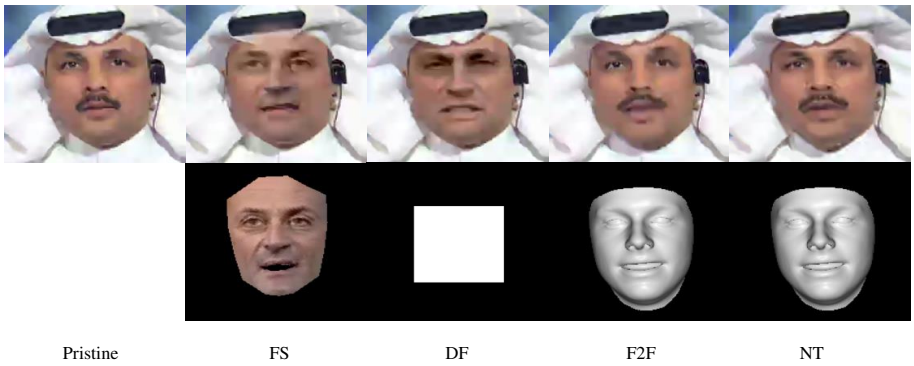

Fig. 1. Visualization of a number of samples from FaceForensics++. The first row shows pristine and generated face images and the second row contains face masks used to add the modifications. "DF": "DeepFakes"; "NT": "Neural Textures"; "FS": "Face Swap"; "F2F": "Face2Face". Although NT and F2F share the same face mask, NT only modifies the region around the mouth.

2.2 Face Forgery Detection

A survey of face forensics detection can be found in [13]. Several methods are proposed to detect manipulated faces [14–17]. While previous literature often relies on hand-crafted features, more and more ConvNet-based methods are proposed. There are two main directions to detect face forgery.

The most straightforward approach is data-driven. Forensic transfer [18] uses the features learned from face forensics to adapt to new domains. [19, 20] combine forgery detection and location simultaneously. [21] uses a modified semantic segmentation architecture to detect manipulated regions. Peng et al. [22] provide a two stream network to detect tampered faces. Shruti et al. [2] concentrate on detecting fake videos of celebrities. Ghazal et al. [23] proposes a deep network based image forgery detection framework using full-resolution information. Ekraam et al. [24] propose a recurrent model to include temporal information. Irene et al. [25] propose an optical flow based CNN for deep fake video detection.

Another category of deep network-based methods is to capture features from the generation process. The features include artifacts or cues introduced by the network [26]. The Face Warping Artifacts (FWA) exploit post processing artifacts in generated videos [9]. Falko et al. [27] use visual artifacts around the face region to detect manipulated videos. [28] proposes a capsule network based method to detect a wide range of forged images and videos. Xin et al. [29] propose a fake video detection method based on inconsistencies between head poses. [30, 31] exploit the effect of illumination. [3] monitors neuron behavior to detect synthetic face images. [4] uses fingerprints from generative adversarial networks to achieve face forensic detection. And [32] proposes a framework based on detecting noise from blending methods.

2.3 Domain Adaptation

Domain adaptation has been widely used in face-related applications. It aims to transfer features from a source to a target domain. The problem is how to measure and minimize the difference between source and target distributions. Several deep domain generalization methods are proposed [33, 34] to improve the generalization ability. Rui et al. [35] propose a multi-adversarial based deep domain generalization with a triplet constraint and depth estimation to handle face anti-spoofing. Maximum mean discrepancy (MMD) [36] is a discrepancy metric to measure the difference in a Reproducing Kernel Hilbert Space. [37] uses MMD-based adversarial learning to align multiple source domains with a prior distribution. [7] considers neural style transfer as a domain adaptation task and theoretically analyzes the effect of the MMD loss.

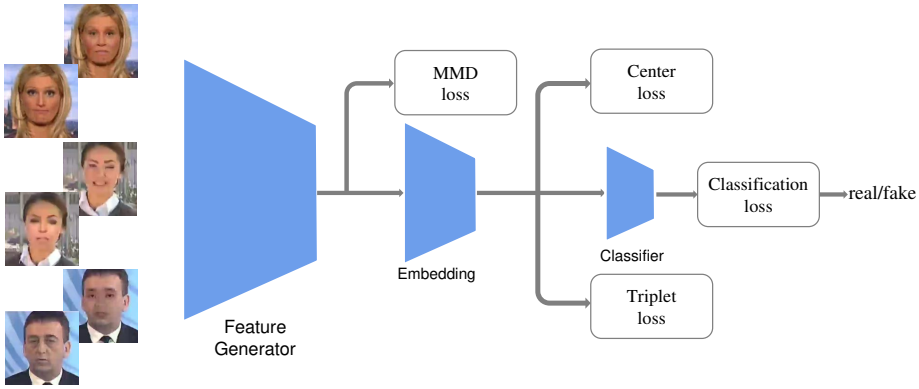

Fig. 2. Overview of the proposed method. Inputs of the network are frames of manipulated face videos. Deep network is used to extract features. Here we use the cross-entropy loss for binary classification. A MMD loss is added to learn a generalized feature space for different domains. Moreover, the triplet and center losses are integrated to provide a discriminative embedding.

3 Method

3.1 Overview

Most current forensic detection approaches fail to address unique issues in face manipulation results. The learned features may not generalize well to unseen deep fake samples. Some approaches choose to extract features from the modification process (e.g., detecting artifacts in manipulated results). Nevertheless, artifacts are dependent on the discrepancy between source and target face images. The discrepancy may originate from differences in head pose, occlusion,

illumination, or ethnicity. Therefore, artifacts may differ depending on the discrepancies between source and target face images. Other methods choose to exploit characteristic cues induced by different generative models. However, any minor changes in the architecture or hyper-parameter setting may negatively influence the forgery detection performance. In contrast, our aim is a generic approach to forgery detection.

To this end, we propose a ConvNet-based discriminative learning model to detect forgery faces. A maximum mean discrepancy (MMD) loss is used to penalize the difference between pristine and fake samples. As a result, extracted features are not biased to the characteristics of a single manipulation method or subject. A center loss is introduced to guide the network to focus on more influential regions of manipulated faces. Furthermore, a triplet loss is incorporated to minimize the intra-distances. We consider the task as a binary classification problem for each frame from real or manipulated videos. In Fig. 2, we provide an overview of the proposed framework. Input images are pristine and fake face samples from a deep fake dataset.

3.2 MMD based Domain Generalization

Maximum mean discrepancy (MMD) measures the difference between two distributions. MMD provides many desirable properties to analyze the performance of GAN [5]. In this paper, we use MMD to measure the performance of forgery detection as follows. Suppose that there are two sets of sample distributions P_s and P_t for a single face manipulation method. The MMD between the two distributions is measured with a finite sample approximation of the expectation. It represents the difference between distribution P_s and P_t based on the fixed kernel function k. A lower MMD means that P_s is closer to P_t. MMD is expressed by

$$MMD^2(P_s, P_t) = \mathbb{E}_{x_s, x_s^{'} \sim P_s, x_t, x_t^{'} \sim P_t} \left[k(x_s, x_s^{'}) - 2k(x_s, x_t) + k(x_t, x_t^{'}) \right] \quad (1)$$

where $k\left(a, b\right) = \left\langle \phi\left(a\right), \phi\left(b\right) \right\rangle$ denotes the kernel function defining a mapping. ϕ is an explicit function. s, t denote the source and target domains respectively. x_s and x_t are data samples from the source and target distributions.

MMD is used to measure the discrepancies among feature embeddings. The MMD loss has been used in neural style transfer tasks [7]. Different kernel functions (Gaussian, linear, or polynomial) can be used for MMD. The MMD loss is defined by:

$$L_{mmd} = \frac{1}{W_k^l} \sum_{i=1}^{M} \sum_{j=1}^{N} \left(k(f_i^l, f_i^l) + k(r_j^l, r_j^l) - 2k(f_i^l, r_j^l) \right) \quad (2)$$

where W_k^l denotes the normalization term based on the kernel function and the feature map l. M and N are numbers of fake and real examples in one batch, respectively. f and r are the features for fake and real examples. The MMD loss

can supervise the network to extract more generalized features to distinguish real from fake samples. It can be considered as an alignment mechanism for different distributions.

3.3 Triplet Constraint

For several manipulated videos which are generated from the same video, the background of most samples is the same. Face reenactment methods can manage to keep the identity of the original faces constant. Meanwhile, modifications from generative methods become more and more subtle. This makes the negative examples look more similar than the positives ones for the same subject. As shown in Fig. 1, images which are generated from F2F and NT are nearly the same as the original image. Therefore, intra-distances between positive samples are larger than their inter-distances.

To learn a more generalized feature embedding, a triplet loss is added to architecture. Illustration of the triplet process can be found in Fig. 3. It is introduced in [38, 39] and used in various face related applications. We aim to improve the generalization ability by penalizing the triplet relationships among batches. The triplet loss is defined by

$$L_{triplet} = \|g(x_i^a) - g(x_i^p)\|_2^2 - \|g(x_i^a) - g(x_i^n)\|_2^2 + \alpha \tag{3}$$

where α denotes the margin and i represents the batch index. x_i^a, x_i^p and x_i^n are anchor, positive samples, and negative samples respectively. They are selected online in each batch. g is the embedding learned from the network. The triplet loss can force the network to minimize the distance between an anchor and a positive sample and maximize the distance between the anchor and a negative sample. It can also contribute to higher robustness when the input is an unseen facial attribute or identity.

3.4 Center Loss

Different modification methods may select different face regions for manipulation (see Fig. 1). When this region is very small compared to the entire image, the majority of the features may exclude information about the manipulation. Our aim is that the network focuses on influential regions around faces instead of the background. To extract more discriminative embeddings, the center loss is used. It has been applied to face recognition [40], and proven effective in measuring intra-class variations. The center loss is defined by:

$$L_{center} = \sum_{k=1}^{M} \|\theta_k - c_k\|^2 , \tag{4}$$

where θ is extracted from feature maps by global average pooling. c denotes the center of feature. Theoretically, the feature center needs to be calculated

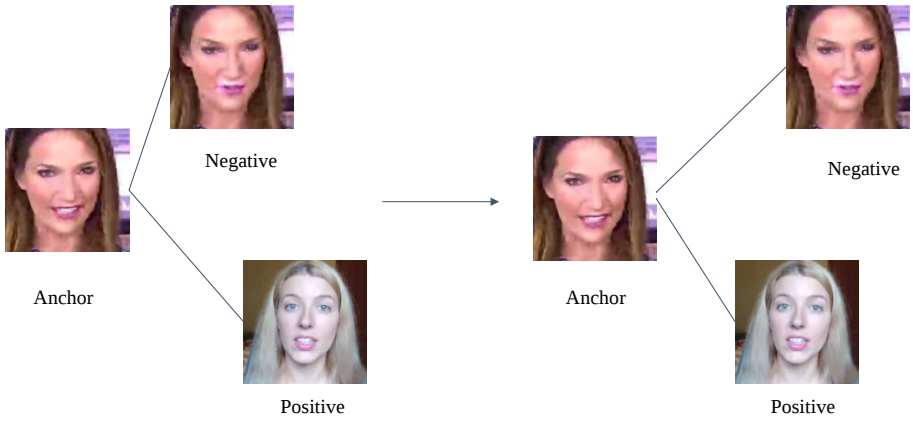

Fig. 3. Visualization of the triplet loss. Images are from FaceForensics++. Normally, the pristine and manipulated images from the same subject look similar. Through the triplet loss, we attempt to minimize the distance between positive examples while maximizing the distance between positive and negative examples.

based on the entire dataset. From [41], a more practical way is used to iteratively update the feature center:

$$c_{k+1} = c_k + \gamma(\theta_k - c_k) \tag{5}$$

where γ defines the learning rate of the feature center $c_k \in R^{N \times S}$. k denotes the iteration index, N is the batch size, and S is the dimension of the embedding. This iterative learning process provides a more smooth prediction for the feature center.

Our final loss function is given by:

$$L = L_{cls} + \lambda_1 L_{mmd} + \lambda_2 L_{triplet} + \lambda_3 L_{center}, \tag{6}$$

where λ_1, λ_2 and λ_3 are balancing factors. L_{cls} is a cross-entropy loss for binary classification.

4 Experiments

4.1 Implementation Details

Our implementation is based on TensorFlow. Adam is used for optimization. We use dlib to detect the face bounding boxes for each frame of the videos. The cropped face image is 300 x 300. All images contain a single face. For all the experiments, we follow subject-exclusive protocol meaning that each subject exists in one split of the dataset. Inception v4 [42] is used as the backbone architecture. The pre-trained model on Imagenet is used. The batch size is 16. Learning rate is 10^{-5}. The training set has a balanced distribution of real and

fake data. λ_1 , λ_2, and λ_3 are set to 0.1, 0.05, 10. For the MMD loss, we use a Gaussian kernel function. The kernel bandwidths σ are 1 and 10. Feature maps $Mixed\,3a$, $Mixed\,4a$, $Mixed\,5a$ from the Inception net are used to calculate the MMD loss. For the triplet loss, we use the implementation of [38]. The margin is 2. Triplets are generated online. For every batch, we select hard positive/negative examples. For center loss, γ is 0.095. The dimension for embedding is 1024. The center is randomly initialized. For the experiments on FaceForensics++, our settings are aligned with [1]. As for experiments on Celeb-DF, our settings follow [10].

4.2 Evaluation Results

We evaluate our approach on several deep fake datasets. A summary of the datasets is shown in Table 1. Visualization of the data samples for each dataset are given in Fig. 1 and 4.

Table 1. Main contrasts of several deep fake datasets. "DF": "DeepFakes"; "NT": "Neural Textures"; "FS": "Face Swap"; "F2F": "Face2Face". The "deep fakes" is an overarching name representing a collection of these methods. For each dataset, the manipulation algorithm and process can be different.

Dataset	UADFV [9]	DF-TIMIT [8]	FaceForensics++	Celeb-DF [10]
Number of videos	98	300	5000	6000
Number of frames	34k	68k	2500k	2341k
Method	DF	FS	FS, DF, F2F, NT	DF

Fig. 4. Visualization of data samples from DF-TIMIT (high quality), UADFV and Celeb-DF. For each pair of images, the left one is the real image and the right one is the modified image.

Results on UADFV and DF-TIMIT. Both UADFV [9] and DF-TIMIT [8] are generated by identity swap methods. UADFV has 49 manipulated videos

of celebrities obtained from the Internet. DF-TIMIT is created based on VidTIMIT [43] under constrained settings. DF-TIMIT has two different settings: low- and high-quality. We choose to evaluate our method on the high-quality subset because it is more challenging. Some pristine videos of the same subjects are selected from VidTIMIT to compose a balanced training dataset. We compare our method with Mesonet [44], XceptionNet [45], Capsule [28], and DSP-FWA [10]. We provide a brief introduction of each method below. Mesonet is based on Inception Net, which is also used in our architecture. XceptionNet is a deep network with separable convolutions and skip connections. Capsule uses a VGG network with capsule structures. DSP-FWA combines a spatial pyramid pooling with FWA [9]. In Table 2, we report all the performances following the same setting as in [10]. Two datasets are relatively small and not very challenging for forensic detection.

Table 2. Evaluation on UADFV [9], DF-TIMIT [8], FF-DF, Celeb-DF [10]. Each datset is evaluated separately. The metric is the Area Under Curve (AUC) score. "FF-DF" is the deep fake subest from FaceForensics++. We follow the same setting of [10].

AUC	UADFV	DF-TIMIT	FF-DF	Celeb-DF
MesoNet [44]	82.1	62.7	83.1	54.8
Xception [45]	83.6	70.5	93.7	65.5
Capsule [28]	61.3	74.4	96.6	57.5
DSP-FWA[10]	97.7	99.7	93.0	64.6
Ours	**98.1**	**99.8**	**97.2**	**88.3**

Results on Celeb-DF. Celeb-DF [10] is one of the largest deep fake video datasets. It is composed of more than 5,000 manipulated videos taken from celebrities. Data is collected from publicly available YouTube videos. The videos include a large range of variations such as face sizes, head poses, backgrounds and illuminants. In addition, subjects show large variations in gender, age, and ethnicity. The generation process of fake faces focuses on reducing the visual artifacts and providing a high-quality synthetic dataset. Table 2 shows that the area under curve (AUC) score of the proposed method on Celeb-DF outperforms all other approaches. The experimental settings remain the same as [10]. Compared to other deep fake datasets, Celeb-DF has fewer artifacts and better quality. The majority of failure cases are false positives. Typical false positives are shown in Fig. 5. Most cases have relatively large poses.

Results on FaceForensics++. FaceForensics++ [46, 1] is one of the largest face forgery dataset. It includes pristine and synthetic videos manipulated by Face2Face [11], NeuralTextures [12], Deepfakes and Faceswap. The modified

Fig. 5. Visualization of false negative predictions of Celeb-DF from our method. These cropped images are from frames of deep fake videos.

videos are generated from a pair of pristine videos. In Fig. 1, we plot all pristine and manipulated examples from one subject within the same frame. Even though the pair of source and target are the same, different methods lead to different results. The performance on raw and high-quality images from FaceForensics++ are already good (accuracy exceeding 95%); we therefore focus on the performance on low quality images. For all experiments, we follow the same protocol as in [1] to split the dataset into a fixed training, validation, and test set, consisting of 720, 140, and 140 videos respectively. All the evaluations are based on the test set.

We compare our method with Mesonet [44], XceptionNet [45], and other methods [47, 16]. In Table 3, we report the performance while training all the categories together with our pipeline. The total f1 score is 0.89. In Table 4, we show the performance while training each category separately. In general, a more balanced prediction is obtained among the pristine and generated examples. The overall performance is better than the other methods. As expected, training FaceForensics++ separately (Table 4) results in a better performance than combined training (Table 3). This is because each generation method is seen as a different domain to the rest. The modified face images contain different types of artifacts and features. When training entirely, manipulated faces from facial reenactment method is extremely similar to real faces. The forgery detector tends to confuse real faces with deep fake data. Our method successfully improves the performance on pristine face without impairing the performance on each deep fake category. When training each category separately, the main challenge becomes the image variations like blur. A number of false negative predictions are shown in Fig. 6. In general, the performance degrades significantly when the face is blurry or the modification region is relatively tiny.

4.3 Analysis

Performance on a single type of deep fake dataset is better than on a dataset containing multiple domains. This is because the extracted features for different manipulated results are diverse. In general, face reenactment may have fewer artifacts than identity modification methods because the transfer of the expressions may require less facial alternations. It results in better performance on detecting identity modification results. We further calculate the prediction accuracy based on each video in the test set of FaceForensics++. On average, the prediction

Fig. 6. Visualization of false negative predictions of FaceForensics++ for the proposed method. These cropped images are frames taken from the deep fake videos.

Table 3. Evaluation on the test set of FaceForensics++. The training and test set includes all the categories of manipulated dataset. "DF": "DeepFakes", "NT": "Neural Textures", "FS": "Face Swap", "F2F": "Face2Face".

Accuracy	DF	F2F	FS	NT	Real	Total
Rahmouni et al [16]	80.4	62.0	60.0	60.0	56.8	61.2
Bayar and Stamm [47]	86.9	83.7	74.3	74.4	53.9	66.8
MesoNet [44]	80.4	69.1	59.2	44.8	77.6	70.5
XceptionNet [45]	93.4	**88.1**	**87.4**	78.1	75.3	81.0
Ours	**98.8**	78.6	80.8	**97.4**	**89.5**	**89.7**

Table 4. Evaluation on each category of the FaceForensics++ test set. Each category has a balanced distribution between pristine data and fake data.

Accuracy	DF	F2F	FS	NT
Bayar and Stamm [47]	81.0	77.3	76.8	72.4
Rahmouni et al [16]	73.3	62.3	67.1	62.6
MesoNet [44]	89.5	84.4	83.6	75.8
XceptionNet [45]	94.3	**91.6**	93.7	82.1
Ours	**99.2**	89.8	**94.5**	**97.3**

for pristine and fake videos is higher than 80%. Although most of the datasets have many frames, the number of videos is relatively small. In most videos, faces have a limited range of variations like pose, illumination, or occlusion. This can also cause the network to predict negatives when pristine face images are relatively blurry or partially occluded. Also, the number of different subjects for the deep fake dataset is relatively small compared to other face-related datasets. This leads to biased results when testing an unseen identity with unique facial attributes.

In our framework, we combine several losses to jointly supervise the learning process of the network. MMD loss can be considered as aligning the distributions of different domains. The style of each image can be expressed by feature distributions in different layers of deep network. Network is constrained to learn a more discriminative feature embedding through different domains. Center loss forces network to concentrate on more influential features rather than background noise. Triplet loss can be considered as an additional constraint to reduce the intra-distance effectively among positive examples.

Table 5. Performance comparison with different components of our method. We evaluate our method on test set of FaceForensics++. Metric is f1 score.

Method	Data augmentation	MMD	Center	Triplet	F1
Basic					0.826
Ours	✓				0.846
Ours	✓	✓			0.881
Ours	✓	✓	✓		0.889
Ours	✓	✓		✓	0.887
Ours	✓	✓	✓	✓	0.897

4.4 Ablation Study

To investigate the effect of each component of our method, we evaluate the performance of the proposed method with different components, see Table 5. We start with the baseline architecture and add different components separately.

Table 6. Performance comparison with different style transfer losses. We evaluate our method on test set of FaceForensics++.

Style Loss	GRAM	BN	MMD
F1	0.862	0.887	0.897

Comparison with Other Style Transfer Losses. First, we test our method with other neural style transfer losses for distribution alignment. Here, we choose the GRAM matrix-based style loss and batch normalization (BN) statistics matching [48]. The GRAM-based loss is defined by

$$L_{GRAM} = \frac{1}{W_l} \sum \left(G_R^l - G_F^l \right)^2, \tag{7}$$

where the Gram matrix G^l is the inner product between the vectorized feature maps in layer l. G_R and G_F are GRAM matrix for real and fake samples respectively. W_l is the normalization term.

The BN style loss is described by

$$L_{BN} = \frac{1}{W_l} \sum \left[\left(\mu_{F_l} - \mu_{R_l} \right)^2 + \left(\sigma_{F_l} - \sigma_{R_l} \right)^2 \right], \tag{8}$$

where μ and σ is the mean and standard deviation of the vectorized feature maps. μ_{R_l} and σ_{R_l} are corresponding to real face samples. From Table 6, performance of the network with the MMD loss outperforms other types of losses.

Comparison with Different Kernel Functions. A different kernel function k can provide different mapping spaces for the MMD loss. In Table 7, we investigate the effect of different kernel functions. Linear and polynomial kernel functions are defined as $k(a, b) = a^T b + c$, $k(a, b) = (a^T b + c)^d$, respectively. We choose $d = 2$ for polynomial kernel function. The Gaussian kernel outperforms other kernels for the MMD loss.

Table 7. Performance comparison with different kernel functions in the MMD loss. We evaluate our method on the test set of FaceForensics++.

Kernel Function	Polynomial	Linear	Gaussian
F1	0.841	0.876	0.897

Comparison with Different Feature Maps. Different levels of feature maps capture different type of style information. We further examine how different combinations of feature maps influence the face forensic detection performance. In Table 8, we illustrate the performances of using multiple sets of feature maps. Feature maps $Mixed\,3a$, $Mixed\,4a$, $Mixed\,5a$ are slightly better than other options.

5 Conclusions

This work focused on face forgery detection and proposed a deep network based architecture. Maximum mean discrepancy (MMD) loss has been used to learn

Table 8. Performance comparison with different combinations of feature maps from our method. We evaluate our method on the test set of FaceForensics++.

Feature map	F1
$Mixed\,3a$	0.884
$Mixed\,4a$	0.881
$Mixed\,5a$	0.883
$Mixed\,3a, Mixed\,4a$	0.890
$Mixed\,4a, Mixed\,5a$	0.891
$Mixed\,3a, Mixed\,4a, Mixed\,5a$	0.897

a more generalized feature space for multiple domains of manipulation results. Furthermore, triplet constraint and center loss have been integrated to reduce the intra-distance and to provide a discriminative embedding for forensics detection.

Our proposed method achieved the best overall performance on UADFV, DF-TIMIT, Celeb-DF and FaceForensics++. Moreover, we provided a detailed analysis of each component in our framework and exploited other distribution alignment methods. Extensive experiments showed that our algorithm has high capacity and accuracy in detecting face forensics.

References

1. Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., Nießner, M.: FaceForensics++: Learning to detect manipulated facial images. In: International Conference on Computer Vision (ICCV). (2019)
2. Agarwal, S., Farid, H., Gu, Y., He, M., Nagano, K., Li, H.: Protecting world leaders against deep fakes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. (2019) 38–45
3. Wang, R., Ma, L., Juefei-Xu, F., Xie, X., Wang, J., Liu, Y.: Fakespotter: A simple baseline for spotting ai-synthesized fake faces. arXiv preprint arXiv:1909.06122 (2019)
4. Yu, N., Davis, L.S., Fritz, M.: Attributing fake images to gans: Learning and analyzing gan fingerprints. In: Proceedings of the IEEE International Conference on Computer Vision. (2019) 7556–7566
5. Xu, Q., Huang, G., Yuan, Y., Guo, C., Sun, Y., Wu, F., Weinberger, K.: An empirical study on evaluation metrics of generative adversarial networks. arXiv preprint arXiv:1806.07755 (2018)
6. Gatys, L.A., Ecker, A.S., Bethge, M.: Image style transfer using convolutional neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2016) 2414–2423
7. Li, Y., Wang, N., Liu, J., Hou, X.: Demystifying neural style transfer. arXiv preprint arXiv:1701.01036 (2017)
8. Korshunov, P., Marcel, S.: Deepfakes: a new threat to face recognition? assessment and detection. arXiv preprint arXiv:1812.08685 (2018)
9. Li, Y., Lyu, S.: Exposing deepfake videos by detecting face warping artifacts. arXiv preprint arXiv:1811.00656 **2** (2018)

10. Li, Y., Yang, X., Sun, P., Qi, H., Lyu, S.: Celeb-df: A new dataset for deepfake forensics. ArXiv (2019)
11. Thies, J., Zollhofer, M., Stamminger, M., Theobalt, C., Nießner, M.: Face2face: Real-time face capture and reenactment of rgb videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2016) 2387–2395
12. Thies, J., Zollhöfer, M., Nießner, M.: Deferred neural rendering: Image synthesis using neural textures. arXiv preprint arXiv:1904.12356 (2019)
13. Tolosana, R., Vera-Rodriguez, R., Fierrez, J., Morales, A., Ortega-Garcia, J.: Deepfakes and beyond: A survey of face manipulation and fake detection. arXiv preprint arXiv:2001.00179 (2020)
14. Dang-Nguyen, D.T., Boato, G., De Natale, F.G.: Identify computer generated characters by analysing facial expressions variation. In: 2012 IEEE International Workshop on Information Forensics and Security (WIFS), IEEE (2012) 252–257
15. Conotter, V., Bodnari, E., Boato, G., Farid, H.: Physiologically-based detection of computer generated faces in video. In: 2014 IEEE International Conference on Image Processing (ICIP), IEEE (2014) 248–252
16. Rahmouni, N., Nozick, V., Yamagishi, J., Echizen, I.: Distinguishing computer graphics from natural images using convolution neural networks. In: 2017 IEEE Workshop on Information Forensics and Security (WIFS), IEEE (2017) 1–6
17. Wang, S.Y., Wang, O., Owens, A., Zhang, R., Efros, A.A.: Detecting photoshopped faces by scripting photoshop. In: Proceedings of the IEEE International Conference on Computer Vision. (2019) 10072–10081
18. Cozzolino, D., Thies, J., Rössler, A., Riess, C., Nießner, M., Verdoliva, L.: Forensictransfer: Weakly-supervised domain adaptation for forgery detection. arXiv preprint arXiv:1812.02510 (2018)
19. Nguyen, H.H., Fang, F., Yamagishi, J., Echizen, I.: Multi-task learning for detecting and segmenting manipulated facial images and videos. arXiv preprint arXiv:1906.06876 (2019)
20. Songsri-in, K., Zafeiriou, S.: Complement face forensic detection and localization with faciallandmarks. arXiv preprint arXiv:1910.05455 (2019)
21. Huang, Y., Juefei-Xu, F., Wang, R., Xie, X., Ma, L., Li, J., Miao, W., Liu, Y., Pu, G.: Fakelocator: Robust localization of gan-based face manipulations via semantic segmentation networks with bells and whistles. arXiv preprint arXiv:2001.09598 (2020)
22. Zhou, P., Han, X., Morariu, V.I., Davis, L.S.: Two-stream neural networks for tampered face detection. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), IEEE (2017) 1831–1839
23. Mazaheri, G., Mithun, N.C., Bappy, J.H., Roy-Chowdhury, A.K.: A skip connection architecture for localization of image manipulations. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. (2019) 119–129
24. Sabir, E., Cheng, J., Jaiswal, A., AbdAlmageed, W., Masi, I., Natarajan, P.: Recurrent convolutional strategies for face manipulation detection in videos. Interfaces (GUI) **3** (2019) 1
25. Amerini, I., Galteri, L., Caldelli, R., Del Bimbo, A.: Deepfake video detection through optical flow based cnn. In: Proceedings of the IEEE International Conference on Computer Vision Workshops. (2019) 0–0
26. Kumar, P., Vatsa, M., Singh, R.: Detecting face2face facial reenactment in videos. In: The IEEE Winter Conference on Applications of Computer Vision. (2020) 2589–2597

27. Matern, F., Riess, C., Stamminger, M.: Exploiting visual artifacts to expose deepfakes and face manipulations. In: 2019 IEEE Winter Applications of Computer Vision Workshops (WACVW), IEEE (2019) 83–92
28. Nguyen, H.H., Yamagishi, J., Echizen, I.: Capsule-forensics: Using capsule networks to detect forged images and videos. In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE (2019) 2307–2311
29. Yang, X., Li, Y., Lyu, S.: Exposing deep fakes using inconsistent head poses. In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE (2019) 8261–8265
30. De Carvalho, T.J., Riess, C., Angelopoulou, E., Pedrini, H., de Rezende Rocha, A.: Exposing digital image forgeries by illumination color classification. IEEE Transactions on Information Forensics and Security **8** (2013) 1182–1194
31. Carvalho, T., Faria, F.A., Pedrini, H., Torres, R.d.S., Rocha, A.: Illuminant-based transformed spaces for image forensics. IEEE transactions on information forensics and security **11** (2015) 720–733
32. Li, L., Bao, J., Zhang, T., Yang, H., Chen, D., Wen, F., Guo, B.: Face x-ray for more general face forgery detection. arXiv preprint arXiv:1912.13458 (2019)
33. Tzeng, E., Hoffman, J., Saenko, K., Darrell, T.: Adversarial discriminative domain adaptation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2017) 7167–7176
34. Li, D., Yang, Y., Song, Y.Z., Hospedales, T.M.: Deeper, broader and artier domain generalization. In: Proceedings of the IEEE International Conference on Computer Vision. (2017) 5542–5550
35. Shao, R., Lan, X., Li, J., Yuen, P.C.: Multi-adversarial discriminative deep domain generalization for face presentation attack detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2019) 10023–10031
36. Gretton, A., Borgwardt, K.M., Rasch, M.J., Schölkopf, B., Smola, A.: A kernel two-sample test. Journal of Machine Learning Research **13** (2012) 723–773
37. Li, H., Jialin Pan, S., Wang, S., Kot, A.C.: Domain generalization with adversarial feature learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2018) 5400–5409
38. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2015) 815–823
39. Parkhi, O.M., Vedaldi, A., Zisserman, A.: Deep face recognition. (2015)
40. Wen, Y., Zhang, K., Li, Z., Qiao, Y.: A discriminative feature learning approach for deep face recognition. In: European conference on computer vision, Springer (2016) 499–515
41. Hu, T., Xu, J., Huang, C., Qi, H., Huang, Q., Lu, Y.: Weakly supervised bilinear attention network for fine-grained visual classification. arXiv preprint arXiv:1808.02152 (2018)
42. Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.A.: Inception-v4, inception-resnet and the impact of residual connections on learning. In: Thirty-first AAAI conference on artificial intelligence. (2017)
43. Sanderson, C., Lovell, B.C.: Multi-region probabilistic histograms for robust and scalable identity inference. In: International conference on biometrics, Springer (2009) 199–208
44. Afchar, D., Nozick, V., Yamagishi, J., Echizen, I.: Mesonet: a compact facial video forgery detection network. In: 2018 IEEE International Workshop on Information Forensics and Security (WIFS), IEEE (2018) 1–7

45. Chollet, F.: Xception: Deep learning with depthwise separable convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2017) 1251–1258
46. Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., Nießner, M.: Faceforensics: A large-scale video dataset for forgery detection in human faces. arXiv (2018)
47. Bayar, B., Stamm, M.C.: A deep learning approach to universal image manipulation detection using a new convolutional layer. In: Proceedings of the 4th ACM Workshop on Information Hiding and Multimedia Security. (2016) 5–10
48. Li, Y., Wang, N., Shi, J., Liu, J., Hou, X.: Revisiting batch normalization for practical domain adaptation. arXiv preprint arXiv:1603.04779 (2016)