Protecting Celebrities from DeepFake with Identity Consistency Transformer

Xiaoyi Dong[1],[*] Jianmin Bao[2], Dongdong Chen[3],[†] Ting Zhang[2], Weiming Zhang[1],
Nenghai Yu[1], Dong Chen[2], Fang Wen[2], Baining Guo[2]
[1]University of Science and Technology of China
[2]Microsoft Research Asia [3]Microsoft Cloud + AI
{dlight@mail., zhangwm@, ynh@}.ustc.edu.cn cddlyf@gmail.com
{jianbao, Ting.Zhang, doch, fangwen, bainguo }@microsoft.com

Abstract

In this work we propose Identity Consistency Transformer, a novel face forgery detection method that focuses on high-level semantics, specifically identity information, and detecting a suspect face by finding identity inconsistency in inner and outer face regions. The Identity Consistency Transformer incorporates a consistency loss for identity consistency determination. We show that Identity Consistency Transformer exhibits superior generalization ability not only across different datasets but also across various types of image degradation forms found in real-world applications including deepfake videos. The Identity Consistency Transformer can be easily enhanced with additional identity information when such information is available, and for this reason it is especially well-suited for detecting face forgeries involving celebrities. [1]

1. Introduction

Deepfake techniques [1–5,9,32,46,47] have been largely advanced to be able to create incredibly realistic fake images of which the face is replaced with someone else in another image. The malicious usage and spread of deepfake have raised serious societal concerns and posed an increasing threat to our trust in online media. Therefore, face forgery detection is in urgent need and has gained a considerable amount of attention recently.

Notably, the overwhelming majority among all cases of face forgeries involve politicians, celebrities and corporate leaders, as their photos and videos are easier to find on the web and hence easily manipulated to generate impressively photo-realistic deepfake. Yet previous detection algorithms make predictions about the forgery based only on the suspect images, and neglect to exploit those freely available

---

[*]Work done during an internship at Microsoft Research Asia.
[†]Dongdong Chen is the corresponding author.
[1]Code will be released at https://github.com/LightDXY/ICT_DeepFake

Figure 1. Five fake (1st row) and real (2nd row) faces and their identity consistency scores. Higher ICT score indicates the inner and outer face more likely from two people, *i.e.* DeepFake face.

data. In this paper, we argue that the images / videos available online can not only be used in generating face forgeries but also be utilized to detect them, and try to protect people whose face is accessible online and thus vulnerable for face manipulation, i.e., celebrities in a broad sense.

Recently, numerous efforts have been devoted to detecting face forgery and achieve promising detection performance. Most existing methods aim to discriminate fake images by exploiting low-level textures and searching for the underlying generation artifacts [7,10,33,35,38,42,44,45,49,52,67,68]. While deploying these techniques in real-world products, we observe two common problems: (1) deepfake detection is usually performed on suspected videos and the video frames have image degradations, such as image rescale, noise and video codec conversion; and (2) when the generated deepfake is convincingly photo-realistic, the low-level traces of forgeries become very hard to detect. These problems make the deepfake detection unstable with video input. We wish to make deepfake detection significantly more robust by making heavy use of semantically meaningful identity information.

In this paper we propose a new face forgery detection technique called Identity Consistency Transformer (ICT) based on high-level semantics. The key idea is to detect identity consistency in the suspect image, i.e., whether the inner face and the outer face belongs to the same person. This turns out to be a non-trivial task. A naïve so-
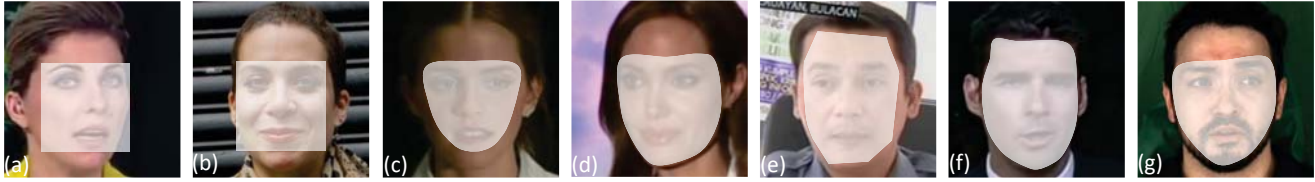
Figure 2. Forgery regions of current face forgery methods. (a) DeepFake in FF++. (b) DeepFake in Google Deepfake Detection. (c) DeepFake in CelebDF. (d) DeepFake in DeepFaceLab. (e)Face2face. (f) FSGAN (g) DF-VAE. They replace the inner face region with different shapes while keep the outer face unchanged.

lution would be to compare identity vectors extracted using the off-the-shelf face recognition model from the inner- and outer- faces, similar to the popular face verification technique. Unfortunately existing face verification methods [22, 59] tend to characterize the most discriminative region, i.e., the inner face for verification and fail to capture the identity information in the outer face. With Identity Consistency Transformer, we train a model to learn a pair of identity vectors, one for the inner face and the other for the outer face, by designing a Transformer such that the inner and the outer identities can be learned simultaneously in a seamlessly unified model. Our Identity Consistency Transformer incorporates a novel consistency loss to pull the two identities together when their labels are the same and consequently push them away when they are associated with different labels.

Our approach only assumes the identity discrepancy and thus can be trained with a large number of swapped faces obtained from swapping different identities, without any fake images generated by face manipulation methods. Empirically we show that Identity Consistency Transformer exhibits significantly superior performance in the situations where low-level texture based methods fail. Figure 1 presents five examples illustrating that state-of-the-art detection methods [7, 35, 52] fail to detect them while our method uncovers quite different identity consistency scores and thus can differentiate real and fake images.

Another advantage of Identity Consistency Transformer is that it can be easily enhanced with additional identity information when such information happens to be available, as is the case with celebrities. For celebrities, their reference images are available and with the Identity Consistency Transformer we construct an authentic reference set consisting of the extracted identity vector pairs of the celebrities and thus create a new consistency score to enhance identity consistency detection. The resulting reference-assisted ICT (ICT-Ref) achieves the state-of-the-art performance on conventional benchmark datasets, demonstrating its strong ability for additional protection for celebrities. Finally, we show in our experiments that the proposed ICT and ICT-Ref significantly improve the generalization ability in two directions: 1) across different datasets and 2) more importantly, across different image degradation forms in real-world applications including video applications.

2. Related Work

Deepfake generation. We roughly divide the face manipulation algorithms into three categories. (1) Early attempts are landmark-based approaches [11, 60] that utilize face alignment to find a source face with a similar pose to the target face and swap to the target face after adjusting the color, lighting and so on. As a consequence, these works are limited to swapping faces with similar poses. (2) To address that limitation, subsequent efforts [13, 17, 20, 26, 37, 46, 47] introduce 3D face representations, such as Face2Face approach [56], face reenactment [54] and expression manipulation [18, 40, 41]. However, these methods are unable to generate parts that do not exist in the source image, e.g., teeth, resulting in non-realistic artifacts in synthesized faces. (3) Recently GAN-based methods achieve more vivid face swapping results. Korshunava *et al*. [30] and Deep-Fakes [4] train a transformation model for each target person, while RSGAN [43], IPGAN [9] and FaceShifter [32] introduced subject-agnostic face swapping works.

Despite the high-quality of current deepfake generation methods, almost all face manipulations still pay more attention to the inner face and adopt a blending post-processing step in order to generate an uncannily realistic deepfake.

Deepfake detection. With the popularity and accessibility of deepfakes, how to detect such compelling deepfakes has become increasingly critical and attracted a huge number of detection methods [7, 8, 10, 31, 33–35, 38, 42, 44, 45, 49, 52, 53, 63, 64, 66–68] recently. Most of these methods are based on low-level textures to detect the visual artifacts, such as pixel-level artifacts [10,52], texture differences [35] and blending ghost [33]. However, such low-level textures are easily faded away when encountering various disruptive forms of image degradation, leading to the drastic drop of the performance of those low-level based methods.

In order to be more resilient, we turn to leverage more semantically meaningful traces. Along this direction, there exist some deepfake video detection methods seeking physical unnaturalness, e.g., utilizing head pose inconsistencies between the inner face and the surrounding using 3-D geometry [63] and detecting the abnormally low or high number of eye blinks [34].

In this work, we are interested in exploiting the identity information, another type of high-level semantics for more robust face forgery detection. The most related works
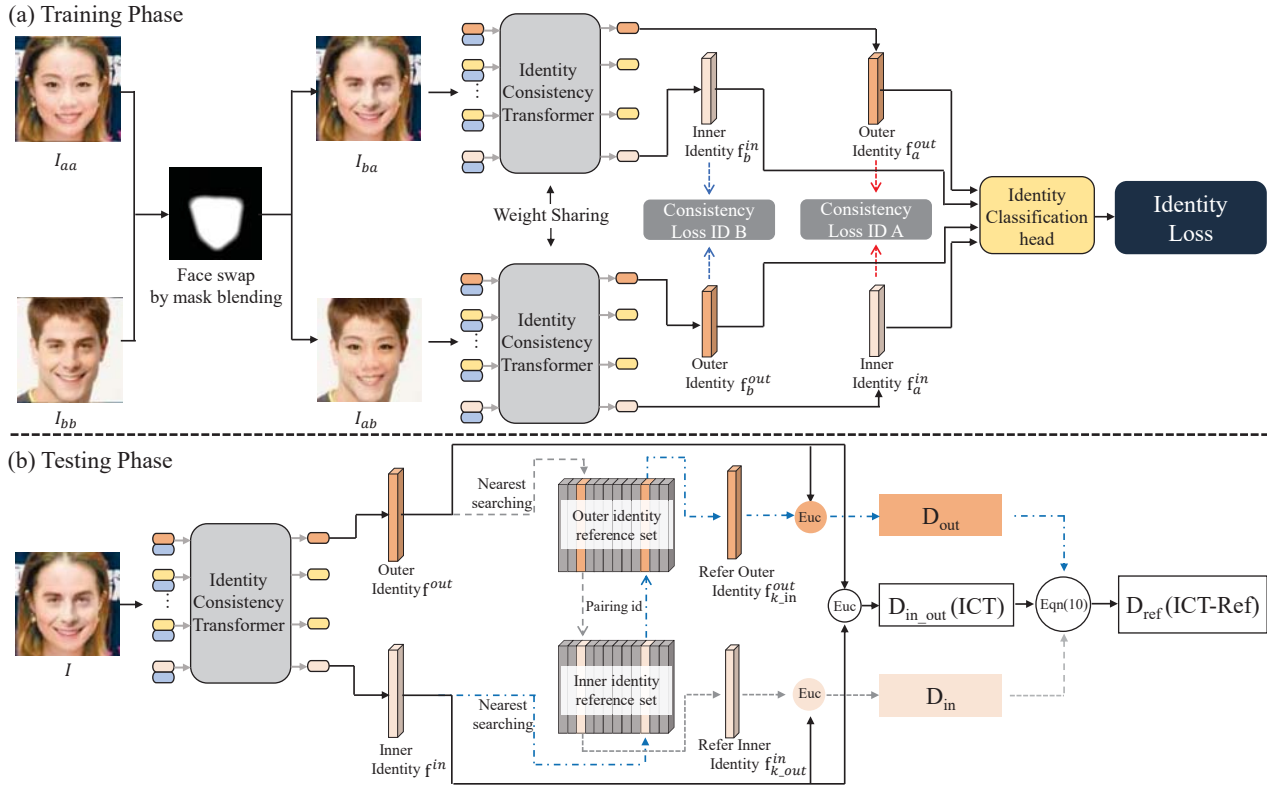
Figure 3. Illustrating the (a) training phase and (b) testing phase of our proposed Identity Consistency Transformer.

are [19] and [48]. The first work [19] proposes example-based forgery detection to detect fake videos given by a reference video, while our method is able to make predictions based only on the suspect image. Even in the reference-assisted variant, the reference is automatically retrieved, not given by a user. The second work [48] also harvests the identity consistency between the inner face and the outer face. Yet our work is different and has three advantages: 1) our model does not need any further training after obtaining the identity extraction network while their model does; 2) our model does not require any fake videos generated by face manipulation methods while their model does; 3) we introduce an extra novel and effective identity consistency loss instead of only adopting face recognition loss as in [48]. Empirically, our work performs much better results, e.g., 85.71% on Celeb-DF dataset, nearly 20% higher than the number in [48].

Transformer. Transformer [57] based entirely on a self-attention mechanism is first proposed in modeling text and has been shown enormous success in many NLP taks [12, 23, 50, 51]. Recent works [15, 16, 24, 25, 58, 61] have extended Transformer to image and also achieved excellent results in vision tasks, e.g., image classification in ViT [16, 24, 25], image inpainting in [58], image generation in iGPT [15] and video recognition [61].

We show that Transformer also works in face classification, a type of fine-grained image classification. While

it seems like a natural extension, it is actually significant. As it is generally believed that discriminating fine-grained categories depends heavily on the subtle and local differences, we demonstrate that Transformer with all global attentions can also learn semantically meaningful features for fine-grained classes. In addition to that, we specially modify Transformer by introducing a new consistency loss to make it appropriate for identity consistency detection.

3. Identity Consistency Transformer

Given an input face image I, the goal of deepfake detection is to classify the input I either as a real image or a deepfake. We start by introducing the identity extraction model and then describe how to exploit the identity consistency to differentiate the given input. An overview of our approach is illustrated in Figure 3.

3.1. Identity Extraction Model

To extract the identity information of the inner face as well as the outer face, we draw inspiration from the recent success of Transformer in application to full-sized images [15, 25], and apply a vision Transformer [25] on face classification to learn the identity information. This brings an advantage that the inner identity and the outer identity can be learned simultaneously by a uniform model instead of adopting two separate face recognition models that each is accountable for one of the two identities. In addition,
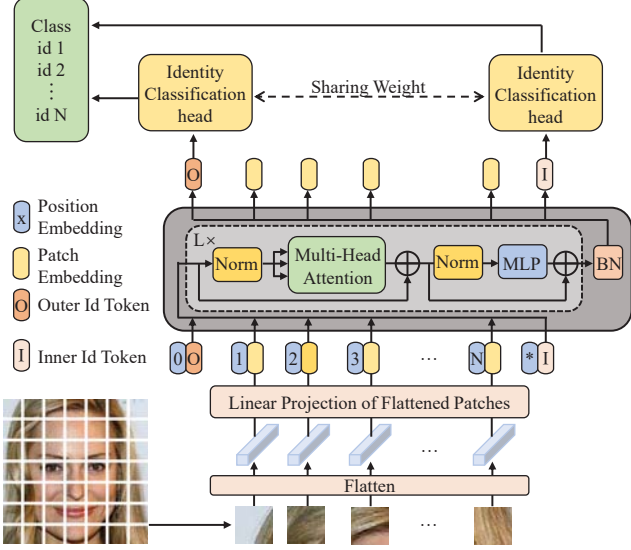
Figure 4. Architecture of the identity extraction model.

we design a new consistency loss to make the Transformer more appropriate for identity consistency detection.

Similar to the class token in [23, 25], we add two extra learnable embedding tokens, denoted as "inner token" and "outer token", to the sequence of patch embeddings. Each patch embedding is flattened from an image patch. Specifically, if H and W are the height and the width of the image, and P is the resolution of the image patch, the number of patch embeddings would be $N = HW/P^2$ and the dimension of the patch embedding is P^2C where C is the channel number of the image.

Our Transformer encoder [57] consists of multiple blocks that are stacked on top of each other. Each block includes a multi-head self-attention layer and a MLP layer with layer normalization in front. We illustrate the overview of model architecture in Figure 4. The outputs of the inner token and the outer token of the last block are regarded as the identity information, which are fed into a classification head implemented by a fully connected layer.

Training data generation. To train the identity extraction model, we first need real face images that are associated with identities. We therefore adopt the publicly available MS-Celeb-1M [27] which is a large dataset containing 10 million face images and 1 million identities. Motivated by Face X-ray [33], we create deepfakes by swapping the inner face of two real faces belonging to different identities, since our approach is based purely on the identity consistency to detect face forgeries. As a result, our network can be trained without fake images generated by any of the face manipulation methods. The generation process is similar to that in Face X-ray [33], which consists of three steps:1) find two faces that has similar poses using face landmarks; 2) generate a mask deformed from the convex hull of the face landmarks; 3) swap the inner face according to the mask after applying color correction. Except that we use local means

of the RGB channels instead of the global mean of the RGB channels in the color correction technique.

In our experiments, we only use such swapped faces during training. Formally, we denote the training dataset as $\mathcal{X} = \{I_{ij}\}_{i \neq j}$, where I_{ij} is a fake face of which the inner face come from I_i and the outer face come from I_j, accordingly y_i is the inner face label and y_j is the outer face label. **Loss functions.** We train the identity extraction model similar to general face classification. Inspired from face recognition, we use the cosine-based softmax loss proposed in ArcFace [22] rather than the vanilla cross-entropy loss,

$$\mathcal{L}_{ident} = -log \frac{e^{s \cdot cos(\theta_{i,j} + \mathbb{1}\{j=y_i\} \cdot m)}}{\sum_{j=1}^{n} e^{s \cdot cos(\theta_{i,j} + \mathbb{1}\{j=y_i\} \cdot m)}}. \quad (1)$$

Here $\theta_{i,j}$ is the angle between W_j and f_i. $W_j \in \mathbb{R}^d$ is the j-th column of weights of the last fully-connected layer in the projection head, $f_i \in \mathbb{R}^d$ is the feature of i-th sample in the input batch. $\mathbb{1}\{j = y_i\}$ is the indicator function that returns 1 when $j = y_i$ and 0 otherwise. s is a scale hyperparameter and m is an additive angular margin. By default, we set $s = 64$ and $m = 0.3$.

Note that the classification head is shared between the inner identity and the outer identity. In this way, when the inner face and the outer face belong to the same person, their corresponding identity vectors are similar because of the shared classification weights. On the other hand, the inner face and the outer face from different identities are consequently pushed away.

To further encourage this desired property, we introduce a consistency loss across the pair of the swapped faces,

$$\mathcal{L}_{consist} = \mathbb{1}\{p_{ij}^{in} = p_{ji}^{out} = y_i\} \|\mathbf{f}_{ij}^{in} - \mathbf{f}_{ji}^{out}\|_2^2 \quad (2)$$
$$+ \mathbb{1}\{p_{ij}^{out} = p_{ji}^{in} = y_j\} \|\mathbf{f}_{ij}^{out} - \mathbf{f}_{ji}^{in}\|_2^2. \quad (3)$$

Here $\mathbf{f}_{ij}^{in}$ ($\mathbf{f}_{ij}^{out}$) is the inner (outer) identity of image I_{ij} and p_{ij}^{in} (p_{ij}^{out}) is the predicted inner (outer) label of image I_{ij}. Note that the consistency loss is activated only when their corresponding predicted labels are correct so that the consistency loss is enforced on the meaningful identity vectors.

Finally, our overall loss function is,

$$\mathcal{L} = \mathcal{L}_{ident} + \eta \mathcal{L}_{consist}, \quad (4)$$

where η is the weight parameter balancing the two losses.

3.2. Identity Consistency Detection

Let T be the already learned identity extraction model. Given an image I, we denote the extracted two identity vectors are $\mathbf{f}_{in}$ and $\mathbf{f}_{out}$, i.e., $(\mathbf{f}^{in}, \mathbf{f}^{out}) = T(I)$. To discriminate the image, we directly compute the distance from inner to outer and use it as the measure,

$$D_{in\_out} = d(\mathbf{f}^{in}, \mathbf{f}^{out}). \quad (5)$$

If D_{in_out} is smaller, this suggests that the identity is more consistent and thus the suspect image is more likely to be real. One may notice that as the inner identity is extracted from the inner face while the outer identity is extracted from a different region, i.e., the outer face, it is possible that D_{in_out} may still be large even if the inner face and the outer face represent the same person. In fact, the proposed consistency loss that pulls the inner identity and the outer identity together when they come from the same person help mitigate this issue.

Reference-assisted identity consistency detection. It is worth noting that creating deepfakes usually requires real face images to take on the role of the source image or the target image. We thereupon think that why not leverage publicly available real face images in detecting deepfakes.

To achieve that, we construct a reference set that contains the identity vector pairs of all available real images. Say we have N real face images, $I_1, \cdots, I_N$, and the reference set is represented as $\mathcal{F} = \{(\mathbf{f}_n^{in}, \mathbf{f}_n^{out}) = T(I_n)\}_{n=1}^N$. We then use the inner identity of the suspect image, $\mathbf{f}^{in}$, to find its nearest neighbor in the reference set,

$$k^{in} = \operatorname{argmin}_{n \in 1, \cdots, N} d(\mathbf{f}^{in}, \mathbf{f}_n^{in}), \alpha_{in} = d(\mathbf{f}^{in}, \mathbf{f}_{k^{in}}^{in}), \quad (6)$$

where $d(\cdot, \cdot)$ is a pre-defined distance measure. This indicates that the inner face of the suspect image shares the same identity as that represented by $\mathbf{f}_{k^{in}}^{in}$ in the reference set. Then the remaining question is whether their corresponding outer faces belong to the same identity or not. Specifically, we retrieve the corresponding outer identity $\mathbf{f}_{k^{in}}^{out}$ in the reference set and compute the distance to the outer identity of the suspect image,

$$D_{out} = d(\mathbf{f}^{out}, \mathbf{f}_{k^{in}}^{out}). \quad (7)$$

The detection process of comparing the inner identity (i.e., nearest neighbor search using the outer identity) can be similarly derived, and accordingly we have

$$D_{in} = d(\mathbf{f}^{in}, \mathbf{f}_{k^{out}}^{in}), \quad (8)$$

$$k^{out} = \operatorname{argmin}_{n \in 1, \cdots, N} d(\mathbf{f}^{out}, \mathbf{f}_n^{out}), \alpha_{out} = d(\mathbf{f}^{out}, \mathbf{f}_{k^{out}}^{out}). \quad (9)$$

The final reference-assisted identity consistency detection is evaluated by the following measure that is combined from all the distances, i.e.,

$$D_{ref} = \lambda D_{in\_out} + \omega(\alpha_{in}) D_{out} + \omega(\alpha_{out}) D_{in}, \quad (10)$$

where $\omega(\cdot)$ is a weight function modulating the importance of the corresponding distance and $\lambda = 2.5 - \omega(\alpha_{out}) - \omega(\alpha_{in})$ is a balancing parameter. In our experiments, we adopt the Euclidean distance as the distance measure and $\omega(x) = \frac{1}{1+exp((x-\theta)/\tau)}$. θ is a pre-defined identity similarity threshold and τ is a scale factor. In the following experiments, we set $\theta = 0.5$ and $\tau = 0.5$ by default.

3.3. Benefits and Limitations and Impacts

First, our approach is general to different types of face manipulation methods as the detection model is not trained to any specific type of face manipulation. Second, our approach is general to different types of image degradation. This is because the identity information we used is a high-level evidence that is more resistant to various image degradation operations. Third, our method depends on the strong face recognition technique, which is a very mature technology that has been widely used in real-world applications. Therefore the extracted identity information is quite reliable and thus improves the detection performance.

On the other hand, we are aware that our framework has one main limitation: our approach is specifically targeted for detecting fake faces that exist identity inconsistency, namely face swap results. That is to say, our method may fail in detecting face reenactment results where the identity is intended to keep the same. Robust and general DeepFake detection in real world is a complex problem, it is hard to be solved with a single model, and we think our method is an important step that focusing on face swapping.

As a face-related tasks, it is inevitable to use face-related datasets. But our method only need an one-hot label for each person and no name is required. Meanwhile, our goal is protecting the identity information from malicious usage of DeepFake.

4. Experiment

Dataset. The training dataset we used is MS-Celeb-1M [27] that contains 10 million images and 1 million identities. The test datasets are conventional benchmark datasets to demonstrate the strong generalization ability of ICT. (1) FaceForensics++ [52] (FF++) is a large scale video dataset consisting of 1000 original videos and the fake videos obtained by four face manipulation methods: DeepFakes [33], Face2Face [56], FaceSwap [56] and NeuralTextures [55]. (2) DeepFake Detection [6] (DFD) including thousands of visual deep fake videos released by Google. (3) Celeb-DeepFake v1 [36] (CD1) containing 408 real videos and 795 synthesized videos with reduced visual artifacts. (4) Celeb-DeepFake v2 [36] extending CD1 which contains 590 real videos and 5639 synthesized video. (5) Deeper-Forensics (Deeper) [29] proposing a new deepfake generation method and generate 1000 fake videos based on FF++ [52]. For our reference-assisted detection, we construct a reference set by combining the real videos of all the test datasets. Specifically, we randomly sample 10 images for each identity in the dataset. In the following experiments, we use frame-level AUC(%) as the performance metric.

Implementation detail. The identity consistency Transformer is trained from scratch at a resolution of 112×112 without any data argumentation. Our Transformer consists

Method	DFD	FF++	Deeper	CD1	CD2	Avg
Multi-task [44]	65.21	72.23	65.32	72.28	61.06	65.96
MesoInc4 [7]	59.06	63.41	51.41	42.26	53.60	53.95
Capsule [45]	69.70	96.50	68.44	69.98	63.65	67.94
Xcep-c0 [52]	89.05	99.26	57.76	48.08	50.37	61.32
Xcep-c23 [52]	**95.60**	98.54	69.85	74.97	77.82	79.56
FWA [35]	80.59	74.82	45.46	72.88	64.87	67.72
DSP-FWA [35]	90.99	81.90	60.00	78.51	81.41	78.56
CNNDetect [62]	60.12	71.08	57.16	56.12	57.17	60.33
Patch-Foren [14]	49.91	73.75	55.35	59.66	57.16	55.52
FFD [21]	76.61	**92.32**	46.64	74.15	77.80	73.50
Face X-ray [33]	94.14	98.44	72.35	74.76	75.39	79.16
Two-Branch [39]*	-	-	-	-	73.41	-
PCL+I2G [65]*	-	-	-	-	81.80	-
Nirkin.*et al*. [48]*	-	99.70	-	-	66.00	-
ICT (Ours)	84.13	90.22	**93.57**	**81.43**	**85.71**	**87.01**
ICT-Ref (Ours)	93.17	98.56	99.25	96.41	94.43	96.34

Table 1. DeepFake detection AUC on unseen datasets. Here gray means close-set evaluation (training and testing on the same dataset). 'Avg' means the average open-set AUC (the close-set results and methods with only one result is not included). '*' indicates the code is not released and we report the results from the original paper. The best results among all the methods except ICT-Ref are denoted in bold.

of 12 blocks and 12 head for the multi-head self-attention. The input image is divided to 14×14 patches and we project each patch to embedding features of dimension 384. The number of training epochs is 30, and the batch size is 1024. The initial learning rate is set to 0.0005 and divided by 10 after 12, 15, 18 epochs. The loss balancing weight parameter η is set as 4 at the first epoch, and increase by 0.5 after each epoch. To further ease the training, we set the margin m as 0 at the first epoch and increase to 0.3 after 10 epochs.

4.1. Comparison with State-of-the-art Methods

We compare our approach with state-of-the-art methods from two categories: methods that are specifically designed for detecting images from certain face manipulations: Multi-task [44], MesoInc4 [7], Capsule [45], Xception-c0, c23 [52], and FWA, DSP-FWA [35], Two-Branch [39], PCL+I2G [65], Nirkin.*et al*. [48]; and methods that aim to detect general deepfakes: Face X-ray [33], FFD [21], CNNDetection [62], and Patch-Forensics [14]. Indeed, the generalization ability is the most important property of deepfake detection for real-world usage and also the biggest challenge among existing methods. We evaluate the generalization ability in two directions: across different unseen datasets and across different unseen image degradation forms. The evaluation metric is the widely used AUC (area under the Receiver Operating Characteristic curve) and we report frame-level AUC in all experiments.

Generalization ability to unseen datasets. In real-world scenarios, it is often the case that the suspect image is generated by an unseen face manipulation method, hence the

Method	Video1	Video2	Video3	Avg
Multi-task [44]	62.50	26.45	78.50	55.81
DSP-FWA [35]	23.08	33.14	51.81	36.01
Xception-c23 [52]	44.56	83.85	99.31	75.90
FFD [21]	56.04	71.16	79.07	68.75
Face X-ray [33]	66.49	94.03	87.10	82.54
ICT (Ours)	82.27	97.78	98.05	94.36
ICT-Ref (Ours)	100.0	100.0	100.0	100.0

Table 2. Deepfake detection AUC (%) on carefully crafted videos collected from Internet.

generalization ability to unseen data is of great importance for deepfake detection. In Table 1, we present the detection performance on five different datasets, where the close-set results (*i.e.*, training and testing are on the same dataset) is marked with gray. We can see that all the compared low-level based baselines perform not so good on Deeper, Celeb-DF v1 and Celeb-DF v2 datasets, which are newly released and thus exhibit less artifacts. While our approach ICT gets significant improvement and achieves high detection AUC, indicating that the identity information is a more reliable evidence than low-level textures in this case. On the other hand, our ICT-Ref variant further greatly boosts the number and achieves the state-of-the-art performance over all benchmark datasets.

Generalization ability to image degradation. Image degradation is very common in the spread of deepfake images and videos. Therefore, we further evaluate the generalization ability to different kinds of image degradation. Here we follow the image degradation strategies in [29]. Note that The 'JPEG' in [29] is indeed Pixelation, so we realize real JPEG as an extra degradation and set five severity level as jpeg quality factor 90, 70, 50, 30, 20.

The performance variation with respect to different degrees of all types of degradation is plotted in Figure 5. It can be easily seen that all the compared methods drop drastically when the degradation gets severer, while our approach still possesses a high level of detection performance. Except for the Gaussian noise where our approach also degrades when the degree gets too severe.

Real-world scenario simulation. We further showcase one typical real-world deepfake detection example by using our ICT. The real-word deepfake videos are downloaded from the YouTube channel "Ctrl Shift Face[2]", which are carefully crafted and all the faked identities are celebrities. With the results listed in Table 2, we find that the accuracy of most existing methods is relatively low. But for our ICT, we get high performance which is nearly 95% on average. The reference-assisted version even boosts the performance to 100%. The video-version results please refer to https://www.youtube.com/watch?v=zgF50dcymj8.

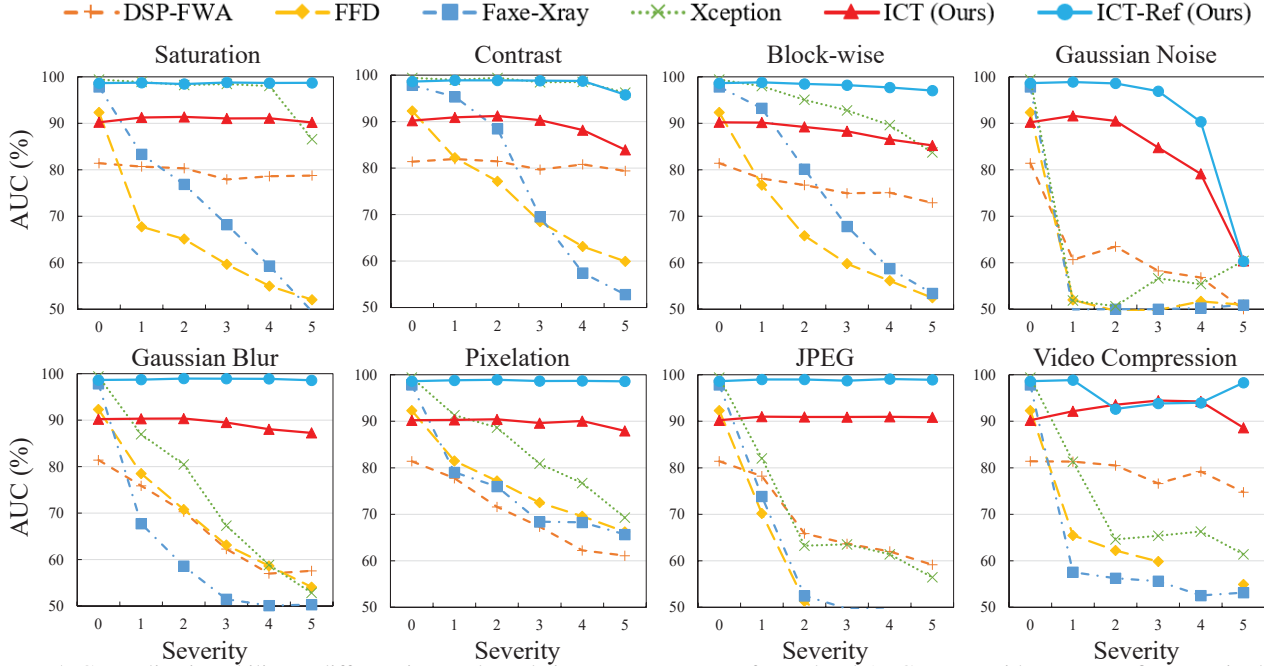[2]https : / / www . youtube . com / channel / UCKpH0CKltc73e4wh0_pgL3g

Figure 5. Generalization ability to different image degradation. Here we report frame-level AUC scores with respect to five severity level for each degradation form.

	DFD	FF++	Deeper	CD2
w/o consistency loss	60.82	62.24	52.90	51.48
w/o mask deformation	83.59	83.47	79.06	83.44
w/o color correction	80.43	90.11	92.15	82.86
ICT	84.13	90.22	93.57	85.71

Table 3. Analysis of consistency loss, mask deformation and color correction for our Identity Consistency Transformer.

4.2. Analysis of ICT

The effect of consistency loss. In our framework, we introduce a new consistency loss to further pull the inner identity and the outer identity together when their corresponding labels are the same. Here, we experiment with an alternative counterpart, i.e., ICT without consistency loss, and present the comparison in Table 3. It can be seen that the performance of our model without consistency loss greatly drops by about $23\% - 40\%$, which verifies that the proposed consistency loss plays a critical role in our identity consistency Transformer.

The effect of different consistency measure. Here we study each component in Equation 10 and experiment four kinds of different identity consistency measures: (1) the proposed ICT, i.e., D_{in_out}; (2) Only using D_{in}; (3) only using D_{out}; and (4) the proposed ICT-Ref, i.e., D_{ref}. The resulting detection AUC for the four different measures on an example dataset Celeb-DeepFake v2 [36] are 85.71%, 92.52%, 87.45% and 94.43% respectively. As expected, the reference-assisted ICT achieves the best performance by combining the other three types. It is worth noting that even only using D_{in} or D_{out} achieves better performance
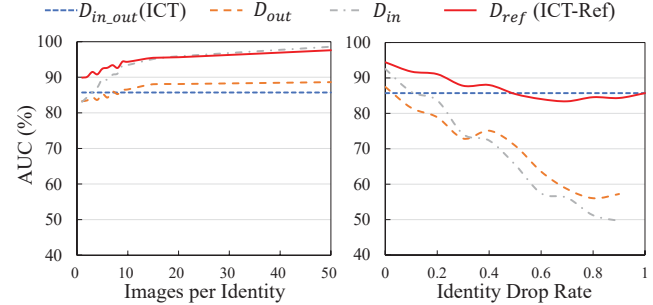

Figure 6. Illustrating the effect of the size of the reference set by considering varying the number of images for each identity (left figure) and varying the number of identities (right figure).

than ICT. This shows that we have effectively made good use of the freely available reference set and verified its crucial role in detecting deepfakes. In addition, it is surprisingly found that using D_{in} performs much better than using D_{out}, achieving about 5% gain. The reason might be that as the outer face is usually not manipulated in face swapping results, using the outer identity to retrieve the nearest neighbor (followed by comparing the inner identity) is more likely to retrieve the accurate suspect identity.

The effect of the reference set. We further analyze how the size of the reference set affects the detection performance. We consider two factors that affect the size of the reference set: (1) the number of images randomly sampled for each identity; and (2) the number of identities. The results on Celeb-DeepFake v2 [36] are presented in Figure 6. It can be seen that in general as the number of images per identity increases, the performance also improves until get saturated, and as the number of identities decreases, the performance

Model	#Params	Deeper	CD2
Res50	43.79M	Nan	Nan
Res50-split-image	43.79M	79.42	73.71
Res50-split-model	$2\times$ 43.79M	91.48	84.66
ICT	21.45M	93.57	85.71

Table 4. Comparison between the proposed Transformer-based ICT and ConvNet-based Res50.

also decline. Interestingly, we find that at some point, e.g., small number of images per identity or larger identity drop rate, the performance gets even lower than the ICT which does not rely on the reference set. The reason might be that in this case, the size of the reference set is so small that the retrieved identity is not accurate. Note that larger identity drop rate has a stronger influence on the performance, even making the ICT-Ref lower than ICT.

The effect of ViT backbone. Here we compare our ViT backbone with traditional CNN models ResNet50 [28]. We add two separate fully connected layers following the last 7×7 feature map in order to learn the inner identity and the outer identity, and finally pass the identity vector to the shared classification head (one fc similar to ours).

However, we find this CNN model (denoted as Res50) fails to converge. The reason might be that the shared backbone pulls the inner identity and the outer identity so hard that it fails to capture the difference when the two identities have different labels. While for our ICT, it has the global attention mechanism at every layer so it could split the inner and outer easily. We further study two substitute models: a) Res50-split-image, splitting images by the blending mask and letting the model learn inner (outer) identity from the cropped inner (outer) face rather than learning from the entire image; b) Res50-split-model, using two separate models to learn the inner and outer identity individually. The results are shown in Table 4. We can see that Res50-split-image still performs poorly and Res50-split-model performs slightly worse than our ICT, with the cost of $4\times$ the number of parameters and two separate models.

The saliency map of the inner identity and the outer identity. Here we show the attention map of the inner identity and the outer identity to see which part of the face contributes most to learning the inner identity and also the outer identity as well. We can see that the inner identity mainly focuses on the most discriminative inner face part, which is carefully crafted by deepfake techniques. On the contrary, the outer identity focuses on the surrounding areas, such as face contour, which is usually unchanged during the Deep-Fake generation. We further test the LFW face recognition accuracy and find the accuray of both inner and outer token is higher than 98%. With such difference between attention regions, the inner identity and the outer one are indeed semantically meaningful and thus beneficial for examining the identity inconsistency.

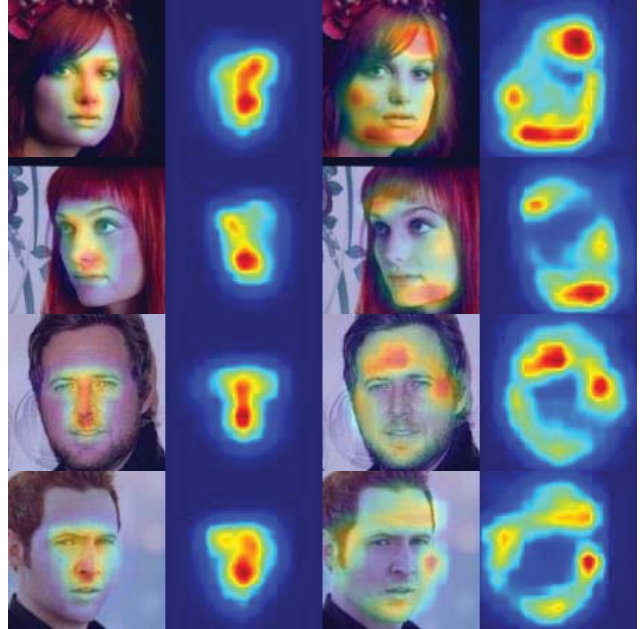**The effect of training data generation.** Our method is

Figure 7. Saliency map of the inner identity (the second column) and the outer identity (the fourth column) with different poses.

a fake-free method and only needs swapped images generated by real faces for training. Here we ablate (a) mask deformation aiming to generate different shapes of mask and (b) color correction aiming to produce a more realistic swapped face in the training data generation process. The ablation results are given in Table 3. It can be seen that the performance of our model without mask deformation drops slightly on some of the datasets while greatly on others, as is the color correction. This verifies that both techniques are important and helpful for the final model.

5. Conclusion

In this work, we propose a novel approach, identity consistency Transformer, for the detection of forged face images. We adopt Transformer to simultaneously learn the inner identity as well as the outer identity and a novel consistency loss is introduced. We show that our work based on high-level semantics is specially effective for the cases where low-level based methods fail. In addition, our approach is further enhanced by leveraging additional identity information from celebrities. Extensive experiments have been conducted to demonstrate the effectiveness of our approach. We hope that our work can encourage more works to investigate inconsistencies in high-level semantics for face forgery detection.

References

[1] https://github.com/iperov/DeepFaceLab. 1

[2] https://github.com/dfaker/df. 1

[3] https://www.malavida.com/en/soft/fakeapp/. 1

[4] https://github.com/deepfakes/faceswap. 1, 2

[5] https://github.com/shaoanlu/faceswap-GAN. 1

[6] https://ai.googleblog.com/2019/09/contributing-data-to-deepfake-detection.html. 5

[7] Darius Afchar, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. Mesonet: a compact facial video forgery detection network, 2018. 1, 2, 6

[8] Sandipan Banerjee, John S Bernhard, Walter J Scheirer, Kevin W Bowyer, and Patrick J Flynn. Srefi: Synthesis of realistic example face images. In *2017 IEEE International Joint Conference on Biometrics (IJCB)*, pages 37–45. IEEE, 2017. 2

[9] Jianmin Bao, Dong Chen, Fang Wen, Houqiang Li, and Gang Hua. Towards open-set identity preserving face synthesis, 2018. 1, 2

[10] Jawadul H Bappy, Cody Simons, Lakshmanan Nataraj, BS Manjunath, and Amit K Roy-Chowdhury. Hybrid lstm and encoder–decoder architecture for detection of image forgeries. *IEEE Transactions on Image Processing*, 28(7):3286–3300, 2019. 1, 2

[11] Dmitri Bitouk, Neeraj Kumar, Samreen Dhillon, Peter Belhumeur, and Shree K Nayar. Face swapping: automatically replacing faces in photographs. In *ACM Transactions on Graphics (TOG)*, volume 27, page 39. ACM, 2008. 2

[12] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020. 3

[13] Chen Cao, Yanlin Weng, Shun Zhou, Yiying Tong, and Kun Zhou. Facewarehouse: A 3d facial expression database for visual computing. *IEEE Transactions on Visualization and Computer Graphics*, 20(3):413–425, 2013. 2

[14] Lucy Chai, David Bau, Ser-Nam Lim, and Phillip Isola. What makes fake images detectable? understanding properties that generalize, 2020. 6

[15] Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In *International Conference on Machine Learning*, pages 1691–1703. PMLR, 2020. 3

[16] Yinpeng Chen, Xiyang Dai, Dongdong Chen, Mengchen Liu, Xiaoyi Dong, Lu Yuan, and Zicheng Liu. Mobileformer: Bridging mobilenet and transformer. *arXiv preprint arXiv:2108.05895*, 2021. 3

[17] Yi-Ting Cheng, Virginia Tzeng, Yu Liang, Chuan-Chang Wang, Bing-Yu Chen, Yung-Yu Chuang, and Ming Ouhy-oung. 3d-model-based face replacement in video. In *SIGGRAPH'09: Posters*, page 29. ACM, 2009. 2

[18] Joni Adamson Clarke. Bringing the past to lif, 1996. 2

[19] Davide Cozzolino, Andreas Rössler, Justus Thies, Matthias Nießner, and Luisa Verdoliva. Id-reveal: Identity-aware deepfake video detection. *arXiv preprint arXiv:2012.02512*, 2020. 3

[20] Kevin Dale, Kalyan Sunkavalli, Micah K Johnson, Daniel Vlasic, Wojciech Matusik, and Hanspeter Pfister. Video face replacement. In *ACM Transactions on Graphics (TOG)*, volume 30, page 130. ACM, 2011. 2

[21] Hao Dang, Feng Liu, Joel Stehouwer, Xiaoming Liu, and Anil K Jain. On the detection of digital face manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5781–5790, 2020. 6

[22] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition, 2019. 2, 4

[23] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 3, 4

[24] Xiaoyi Dong, Jianmin Bao, Dongdong Chen, Weiming Zhang, Nenghai Yu, Lu Yuan, Dong Chen, and Baining Guo. Cswin transformer: A general vision transformer backbone with cross-shaped windows. *arXiv preprint arXiv:2107.00652*, 2021. 3

[25] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3, 4

[26] Zhenglin Geng, Chen Cao, and Sergey Tulyakov. 3d guided fine-grained face manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9821–9830, 2019. 2

[27] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *European conference on computer vision*, pages 87–102. Springer, 2016. 4, 5

[28] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 8

[29] Liming Jiang, Ren Li, Wayne Wu, Chen Qian, and Chen Change Loy. Deeperforensics-1.0: A large-scale dataset for real-world face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2889–2898, 2020. 5, 6

[30] Iryna Korshunova, Wenzhe Shi, Joni Dambre, and Lucas Theis. Fast face-swap using convolutional neural networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3677–3685, 2017. 2

[31] Trung-Nghia Le, Huy H. Nguyen, Junichi Yamagishi, and Isao Echizen. Openforensics: Large-scale challenging dataset for multi-face forgery detection and segmentation in-the-wild. In *Proceedings of the IEEE/CVF International*

Conference on Computer Vision (ICCV), pages 10117–10127, October 2021. 2

[32] Lingzhi Li, Jianmin Bao, Hao Yang, Dong Chen, and Fang Wen. Faceshifter: Towards high fidelity and occlusion aware face swapping. *arXiv preprint arXiv:1912.13457*, 2019. 1, 2

[33] Lingzhi Li, Jianmin Bao, Ting Zhang, Hao Yang, Dong Chen, Fang Wen, and Baining Guo. Face x-ray for more general face forgery detection, 2020. 1, 2, 4, 5, 6

[34] Yuezun Li, Ming-Ching Chang, and Siwei Lyu. In ictu oculi: Exposing ai generated fake face videos by detecting eye blinking, 2018. 2

[35] Yuezun Li and Siwei Lyu. Exposing deepfake videos by detecting face warping artifacts, 2019. 1, 2, 6

[36] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-df: A large-scale challenging dataset for deepfake forensics, 2020. 5, 7

[37] Yuan Lin, Shengjin Wang, Qian Lin, and Feng Tang. Face swapping under large pose variations: A 3d model based approach. In *2012 IEEE International Conference on Multimedia and Expo*, pages 333–338. IEEE, 2012. 2

[38] Yaqi Liu, Qingxiao Guan, Xianfeng Zhao, and Yun Cao. Image forgery localization based on multi-scale convolutional neural networks. In *Proceedings of the 6th ACM Workshop on Information Hiding and Multimedia Security*, pages 85–90, 2018. 1, 2

[39] Iacopo Masi, Aditya Killekar, Royston Marian Mascarenhas, Shenoy Pratik Gurudatt, and Wael AbdAlmageed. Two-branch recurrent network for isolating deepfakes in videos. In *European Conference on Computer Vision*, pages 667–684. Springer, 2020. 6

[40] Iacopo Masi, Anh Tuan Tran, Tal Hassner, Jatuporn Toy Leksut, and Gerard Medioni. Do we really need to collect millions of faces for effective face recognition? In *European conference on computer vision*, pages 579–596. Springer, 2016. 2

[41] Iacopo Masi, Anh Tuan Tran, Tal Hassner, Gozde Sahin, and Gerard Medioni. Face-specific data augmentation for unconstrained face recognition. *International Journal of Computer Vision*, 127(6):642–667, 2019. 2

[42] Falko Matern, Christian Riess, and Marc Stamminger. Exploiting visual artifacts to expose deepfakes and face manipulations. In *2019 IEEE Winter Applications of Computer Vision Workshops (WACVW)*, pages 83–92. IEEE, 2019. 1, 2

[43] Ryota Natsume, Tatsuya Yatagawa, and Shigeo Morishima. Rsgan: face swapping and editing using face and hair representation in latent spaces. *arXiv preprint arXiv:1804.03447*, 2018. 2

[44] Huy H. Nguyen, Fuming Fang, Junichi Yamagishi, and Isao Echizen. Multi-task learning for detecting and segmenting manipulated facial images and videos, 2019. 1, 2, 6

[45] Huy H. Nguyen, Junichi Yamagishi, and Isao Echizen. Use of a capsule network to detect fake images and videos, 2019. 1, 2, 6

[46] Yuval Nirkin, Yosi Keller, and Tal Hassner. Fsgan: Subject agnostic face swapping and reenactment, 2019. 1, 2

[47] Yuval Nirkin, Iacopo Masi, Anh Tran Tuan, Tal Hassner, and Gerard Medioni. On face segmentation, face swapping, and

face perception. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 98–105. IEEE, 2018. 1, 2

[48] Yuval Nirkin, Lior Wolf, Yosi Keller, and Tal Hassner. Deepfake detection based on the discrepancy between the face and its context. *arXiv preprint arXiv:2008.12262*, 2020. 3, 6

[49] Yuyang Qian, Guojun Yin, Lu Sheng, Zixuan Chen, and Jing Shao. Thinking in frequency: Face forgery detection by mining frequency-aware clues, 2020. 1, 2

[50] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. 2018. 3

[51] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 3

[52] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images, 2019. 1, 2, 5, 6

[53] Branko Samarzija and Slobodan Ribaric. An approach to the de-identification of faces in different poses. In *2014 37th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, pages 1246–1251. IEEE, 2014. 2

[54] Supasorn Suwajanakorn, Steven M Seitz, and Ira Kemelmacher-Shlizerman. Synthesizing obama: learning lip sync from audio. *ACM Transactions on Graphics (ToG)*, 36(4):1–13, 2017. 2

[55] Justus Thies, Michael Zollhöfer, and Matthias Nießner. Deferred neural rendering: Image synthesis using neural textures. *ACM Transactions on Graphics (TOG)*, 38(4):1–12, 2019. 5

[56] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2387–2395, 2016. 2, 5

[57] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017. 3, 4

[58] Ziyu Wan, Jingbo Zhang, Dongdong Chen, and Jing Liao. High-fidelity pluralistic image completion with transformers. In *ICCV*, 2021. 3

[59] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition, 2018. 2

[60] Hong-Xia Wang, Chunhong Pan, Haifeng Gong, and Huai-Yu Wu. Facial image composition based on active appearance model. In *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 893–896. IEEE, 2008. 2

[61] Rui Wang, Dongdong Chen, Zuxuan Wu, Yinpeng Chen, Xiyang Dai, Mengchen Liu, Yu-Gang Jiang, Luowei Zhou, and Lu Yuan. Bevt: Bert pretraining of video transformers. *arXiv preprint arXiv:2112.01529*, 2021. 3

[62] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. Cnn-generated images are surprisingly easy to spot... for now. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 7, 2020. 6

[63] Xin Yang, Yuezun Li, and Siwei Lyu. Exposing deep fakes using inconsistent head poses, 2018. 2

[64] Hanqing Zhao, Wenbo Zhou, Dongdong Chen, Tianyi Wei, Weiming Zhang, and Nenghai Yu. Multi-attentional deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 2

[65] Tianchen Zhao, Xiang Xu, Mingze Xu, Hui Ding, Yuanjun Xiong, and Wei Xia. Learning to recognize patchwise consistency for deepfake detection. *arXiv preprint arXiv:2012.09311*, 2020. 6

[66] Yinglin Zheng, Jianmin Bao, Dong Chen, Ming Zeng, and Fang Wen. Exploring temporal coherence for more general video face forgery detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15044–15054, October 2021. 2

[67] Peng Zhou, Xintong Han, Vlad I. Morariu, and Larry S. Davis. Learning rich features for image manipulation detection, 2018. 1, 2

[68] Peng Zhou, Xintong Han, Vlad I. Morariu, and Larry S. Davis. Two-stream neural networks for tampered face detection, 2018. 1, 2