Chapter 14
DeepFakes Detection: the DeeperForensics Dataset and Challenge

Liming Jiang, Wayne Wu, Chen Qian, and Chen Change Loy

Abstract Recent years have witnessed exciting progress in automatic face swapping and editing. Many techniques have been proposed, facilitating the rapid development of creative content creation. The emergence and easy accessibility of such techniques, however, also cause potential unprecedented ethical and moral issues. To this end, academia and industry proposed several effective forgery detection methods. Nonetheless, challenges could still exist. (1) Current face manipulation advances can produce high-fidelity fake videos, rendering forgery detection challenging. (2) The generalization capability of most existing detection models is poor, particularly in real-world scenarios where the media sources and distortions are unknown. The primary difficulty in overcoming these challenges is the lack of amenable datasets for real-world face forgery detection. Most existing datasets are either of a small number, of low quality, or overly artificial. Meanwhile, the large distribution gap between training data and actual test videos also leads to weak generalization ability. In this chapter, we present our on-going effort of constructing DeeperForensics-1.0, a large-scale forgery detection dataset, to address the challenges above. We discuss approaches to ensure the quality and diversity of the dataset. Besides, we describe the observations we obtained from organizing DeeperForensics Challenge 2020, a real-world face forgery detection competition based on DeeperForensics-1.0. Specifically, we summarize the winning solutions and provide some discussions on potential research directions.

L. Jiang · C. C. Loy (✉)
S -Lab, Nanyang Technological University, Jurong West, Singapore
e-mail: ccloy@ntu.edu.sg

L. Jiang
e-mail: liming002@ntu.edu.sg

W. Wu · C. Qian
SenseTime Research, Beijing, China
e-mail: wuwenyan@sensetime.com

C. Qian
e-mail: qianchen@sensetime.com

C. Rathgeb et al. (eds.), *Handbook of Digital Face Manipulation and Detection*,
Advances in Computer Vision and Pattern Recognition,
https://doi.org/10.1007/978-3-030-87664-7_14

14.1 Introduction

Face swapping has become an emerging topic in computer vision and graphics. Indeed, many works [1, 4, 6, 41, 53, 76] on automatic face swapping have been proposed in recent years. These efforts have circumvented the cumbersome and tedious manual face editing processes, hence expediting the advancement in face editing. At the same time, such enabling technology has also sparked legitimate concerns on its potential for being misused and abused. The popularization of "DeepFakes" on the Internet has further set off alarm bells among the general public and authorities, in view of the conceivable perilous implications. Accordingly, countermeasures to safeguard against these photorealistic fake videos become a dire need to be in place promptly, especially innovations that can effectively detect videos that have been manipulated.

Although academia and industry have contributed several effective face forgery detection methods [54, 56, 63, 64, 93, 99], some challenges could still exist. First, current face manipulation advances can produce high-fidelity fake videos, making forgery detection challenging. Besides, the generalization capability of most existing detection models is poor, particularly in *real-world* scenarios where the media sources and distortions are unknown. Meanwhile, the DeepFakes techniques will keep evolving in the future. The better face editing quality will render forgery detection more challenging, entailing the increasing importance of the model generalization.

In this chapter, we present our on-going efforts to address the challenges above. The primary difficulty in overcoming these challenges is the lack of amenable datasets. Working toward forgery detection, various groups have contributed datasets (*e.g.*, FaceForensics++ [81], Deep Fake Detection [13], and DFDC [23, 24]) comprising manipulated video footages. The availability of these datasets has undoubtedly provided essential avenues for research into forgery detection. Nonetheless, the aforementioned datasets fall short in several ways. Videos in these datasets are either of a small number, of low quality, or overly artificial. Understandably, these datasets are inadequate to train a good model for effective forgery detection in real-world scenarios. This is particularly true when current advances in human face editing are able to produce more photorealistic videos than the ones in these datasets. On another note, we observe a high similarity between training and test videos, in terms of their distribution, in certain works [57, 81]. Their actual efficacy in detecting real-world face forgery cases, which are much more variable and unpredictable, remains to be further elucidated.

We believe that forgery detection models can only be enhanced when trained with a dataset that is exhaustive enough to encompass as many potential real-world variations as possible. To this end, we propose a large-scale dataset, named DeeperForensics-1.0 [41], consisting of 60, 000 videos with a total of 17.6 million frames for real-world face forgery detection. The main steps of our dataset construction are shown in Fig. 14.1. We set forth three yardsticks when constructing this dataset: (1) *Good quality*. The dataset shall contain the videos that are more realistic

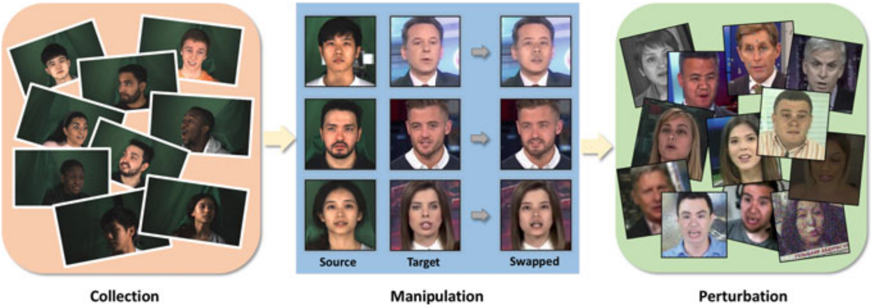

Fig. 14.1 DeeperForensics-1.0 is a new large-scale dataset for *real-world* face forgery detection

and closer to the state-of-the-art DeepFakes video distributions. (Sections 14.3.1 and 14.3.2) (2) *Large scale.* The dataset shall be made up of a large number of video sets. (Section 14.3.3) (3) *High diversity.* There shall be sufficient variations in the video footages (*e.g.*, compression, blurry, and transmission errors) to match those that may be encountered in the real world (Sect. 14.3.3).

The major challenge in the preparation of this dataset is the lack of good-quality video footages. Specifically, most publicly available videos are captured under an unconstrained environment resulting in large variations, including but not limited to suboptimal illumination, large occlusion of the target faces, and extreme head poses. Importantly, the lack of the official informed consents from the video subjects precludes the use of these videos, even for non-commercial purposes. On the other hand, while some videos of manipulated faces are deceptively real, a larger number remains easily distinguishable by human eyes. The latter is often caused by model negligence toward appearance variations or temporal differences, leading to preposterous and incongruous results.

We approach the aforementioned challenge from two perspectives. (1) Collecting fresh face data from 100 individuals with informed consents (Sect. 14.3.1). (2) Devising a novel end-to-end face swapping method, DeepFake Variational Auto-Encoder (DF-VAE), to enhance existing videos (Sect. 14.3.2). In addition, we introduce diversity into the video footages through the deliberate addition of distortions and perturbations, simulating real-world scenarios. The DeeperForensics-1.0 dataset also features a hidden test set, containing manipulated videos that achieve the high deceptive ranking in user studies. The hidden test set is richer in distribution than the publicly available training set, suggesting a better real-world forgery detection setting.

Using the introduced DeeperForensics-1.0 dataset, we organized the DeeperForensics Challenge 2020 [40] with the aim to advance the state of the art in face forgery detection. Participants in this challenge were expected to develop robust and generic methods for forgery detection in real-world scenarios. This chapter also covers details of the DeeperForensics Challenge 2020, including the platform, evaluation metric, timeline, participants, results, *etc*. The winning solutions of top-3 entries are included. We present discussions to take a closer look at the current status and possible future development of real-world face forgery detection.

14.2 Related Work

In this section, we provide an overview of the current status of relevant studies *w.r.t.*DeepFakes detection. The taxonomy of these works can be generally grouped into four paradigms, namely DeepFakes generation methods, DeepFakes detection methods, DeepFakes detection datasets, and DeepFakes detection benchmarks.

14.2.1 DeepFakes Generation Methods

The popularization of DeepFakes videos is attributed to the rapid development of generative models. Existing state-of-the-art generative models are mainly built on deep neural networks [26, 33, 48, 50, 74], showing impressive capability in capturing high-level latent representations of visual data and synthesizing new images. Two popular categories of generative models for face manipulation are auto-encoders (AE) [33, 50] and generative adversarial networks (GAN) [26].

The vanilla AE [33] reconstructs images, aiming at learning latent codes in an unsupervised manner, typically for dimensional reduction and feature learning. Auto-encoders have been widely used to generate images since the development of variational auto-encoders (VAE) [49, 50]. Extensive well-known off-the-shelf face manipulation software are based on auto-encoders, *e.g.*, DeepFakes [4] and DeepFaceLab [1, 76]. These methods tend to learn the identity information for face manipulation through the reconstruction process. However, they usually fit the specific domain and cannot scale to multiple identities. The manipulation method DF-VAE [41] for the DeeperForensics-1.0 dataset is based on variational auto-encoders. DF-VAE is an end-to-end many-to-many face swapping method, which considers style matching and temporal continuity for video manipulation.

Another category of generative models is GAN [26, 67, 79], where a generator tries to fool a discriminator by refining the synthesized images continuously until the discriminator fails to perceive them as fakes. GAN has been extensively applied in face generation [43–45], image-to-image translation [17, 38, 39, 42, 104], style transfer [36, 59], and semantic image synthesis [39, 42, 60, 75, 95]. For face manipulation, the open-source DeepFakes software, faceswap-GAN [6], is a typical GAN-based method. It exploits adversarial losses to the denoising auto-encoder and applies attention mechanisms to improve the clarity of the swapped faces. ReenactGAN [97] introduced the notion of boundary latent space for robust many-to-one face reenactment. Some recent GAN-based innovations were designed in the more challenging face manipulation context, *e.g.*, subject agnostic [72] and occlusion aware [53].

14.2.2 DeepFakes Detection Methods

The development of face forgery detection approaches is constantly evolving along with the advancement of face manipulation techniques. One of the early forgery detection methods is [103]. They proposed a two-stream network for forgery detection. The initial system was trained to detect facial expression manipulations. Later on, MesoNet was proposed in [10]. They introduced two different networks composed of few layers in order to focus on the mesoscopic properties of the images. This method was originally tested in their private database and has been proved to be an effective approach in the FaceForensics benchmark [81]. A temporal-aware framework for automatic fake video detection was discussed in [28]. They leveraged the benefits of both convolutional neural networks (CNN) and recurrent neural networks (RNN). They integrated them into a single framework and averaged the results for evaluation.

More recent forgery detection approaches mainly considered different artifacts introduced during face manipulation. Some methods were based on visual artifacts, *e.g.*, face warping artifacts [56], dissonance of saturation [65], discrepancy between the face and its context [73], region-based artifacts [87], and temporal inconsistencies [91]. Some approaches considered noises from generative models, *e.g.*, GAN fingerprints [100], convolutional traces [27], and frequency-domain clues [78]. Others exploited physiological signs as an important forgery detection basis. They utilized eye blinking [55], head poses [99], heart rate [32], and emotions [68] as important cues for effective face forgery detection.

Real-world face forgery detection, in which video sources and distortions are highly unconstrained and unpredictable, remains less explored. Some studies [16, 54, 83, 93] have started to consider the model generalization issue for forgery detection, which is crucial for real-world face forgery detection. The design of the DeeperForensics-1.0 dataset [41] and the DeeperForensics Challenge 2020 [40] aims to offer a benchmark and platform for a more systematic study about this problem.

14.2.3 DeepFakes Detection Datasets

Building a dataset for forgery detection requires a huge amount of effort on data collection and manipulation. Early forgery detection datasets comprised images captured under highly restrictive conditions, *e.g.*, MICC_F2000 [11], Wild Web dataset [101], and Realistic Tampering dataset [52].

Due to the urgent need for video-based face forgery detection, some research groups have devoted their efforts to create video forensics datasets. UADFV [99] contained 98 videos, *i.e.*, 49 real videos from YouTube and 49 fake ones generated by FakeAPP [7]. DeepFake-TIMIT [51] manually selected 16 similar looking pairs of people from VidTIMIT [82] database. For each of the 32 subjects, they generated about 10 videos using low-quality and high-quality versions of faceswap-

GAN [6], resulting in a total of 620 fake videos. Celeb-DF [57] included 408 YouTube videos, mostly of celebrities, from which 795 fake videos were synthesized. FaceForensics++ [81] is the first large-scale face forensic dataset that consisted of $4,000$ fake videos manipulated by four methods (*i.e.*, DeepFakes [4], Face2Face [86], FaceSwap [5], and NeuralTextures [85])), as well as $1,000$ real videos from YouTube. Afterward, Google joined FaceForensics++ and contributed Deep Fake Detection [13] dataset with $3,431$ real and fake videos from 28 actors. Recently, Facebook invited 66 individuals and built the DFDC preview dataset [24], which comprised $5,214$ original and tampered videos with three types of augmentations.

To build the DeeperForensics-1.0 dataset, we invite 100 actors and collect high-resolution (1920×1080) source data with these actors showing various poses and expressions under different illuminations. 3DMM blendshapes [14] are taken as a reference to supplement some exaggerated expressions. We obtain consents from all the actors for using and manipulating their faces. A newly proposed end-to-end face swapping method (*i.e.*, DF-VAE) is exploited to improve the generated video quality. Besides, seven types of perturbations at five intensity levels are applied to simulate real-world scenes better. The dataset also includes a mixture of distortions to a single video. In total, the DeeperForensics-1.0 dataset contains $60,000$ high-quality videos with 17.6 million frames.

14.2.4 DeepFakes Detection Benchmarks

The FaceForensics benchmark [81] is a popular benchmark for facial manipulation detection. The benchmark included six image-level face forgery detection baselines [10, 12, 18, 19, 25, 80]. The FaceForensics benchmark added several distortions to the videos by converting them into different compression rates. The benchmark did not include different perturbation types or a mixture of them. Celeb-DF [57] also provided a face forgery detection benchmark including seven methods [10, 18, 56, 64, 70, 99, 103] trained and tested on different datasets. In the aforementioned benchmarks, the test set usually shares a similar distribution with the training set. Such an assumption may inherently introduce biases and render the detection methods impractical for face forgery detection in real-world settings with much more diverse and unknown fake videos.

The DeeperForensics-1.0 benchmark features a challenging hidden test set with manipulated videos achieving high deceptive scores in user studies. The hidden test set is richer in distribution than the publicly available training set to better simulate the real-world distribution. The benchmark includes the entries submitted to the DeeperForensics Challenge 2020. The top-3 challenge winning solutions in this benchmark are elaborated on in Sect. 14.4.5. Temporal information—a significant cue for video forgery detection besides the single-frame quality—has been considered. In addition, readers are referred to [41] for more video-level forgery detection baselines [15, 30, 34, 89, 92] in the DeeperForensics-1.0 benchmark.

14.3 DeeperForensics-1.0 Dataset

This section introduces the DeeperForensics-1.0 dataset [41]. The dataset consists of 60, 000 videos with 17.6 million frames in total, including 50, 000 collected source videos and 10, 000 manipulated videos. Toward building a dataset that is suitable for real-world face forgery detection, DeeperForensics-1.0 is designed with the careful consideration of *quality*, *scale*, and *diversity*. In Sects. 14.3.1 and 14.3.2, we discuss the details of data collection and methodology (*i.e.*, DF-VAE) to improve the quality of data. In Sect. 14.3.3, we show our approaches to increase the scale and diversity of samples.

14.3.1 Data Collection

Source data is the first factor that highly affects *quality*. Taking results in Fig. 14.2 as an example, the source data collection increases the robustness of our face swapping method to extreme poses, since videos on the Internet usually have limited head pose variations.

We refer to the identity in the driving video as the "target" face and the identity of the face that is swapped onto the driving video as the "source" face. Different from previous works, we find that the source faces play a more critical role than the target faces in building a high-quality dataset. Specifically, the expressions, poses,

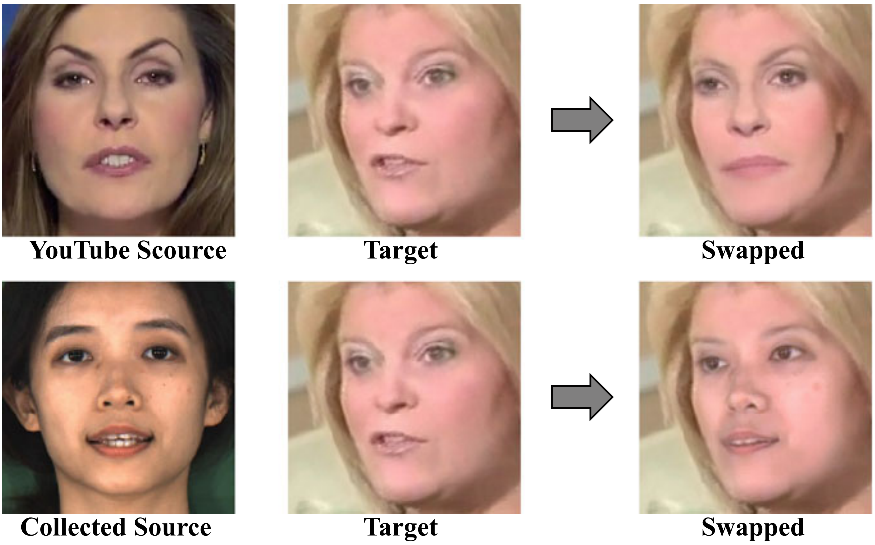

Fig. 14.2 Comparison of face swapping results using an in-the-wild YouTube video or the collected video as source data, with the same manipulation method and setting

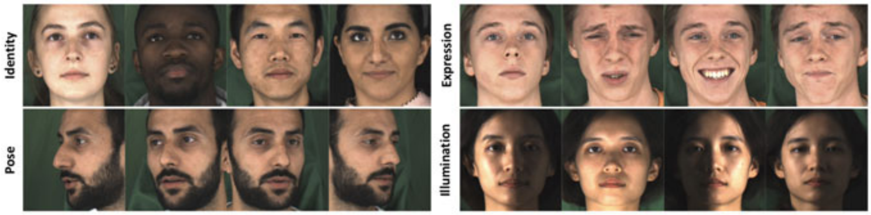

Fig. 14.3 Diversity in identities, poses, expressions, and illuminations in the collected source data of DeeperForensics-1.0

and lighting conditions of source faces should be much richer in order to perform robust face swapping. The data collection of DeeperForensics-1.0 mainly focuses on source face videos. Figure 14.3 shows the diversity in different attributes of the collected source data.

We invite 100 paid actors to record the source videos. Similar to [13, 24], we obtain consents from all the actors for using and manipulating their faces to avoid the portrait right issues. The participants are carefully selected to ensure variability in genders, ages, skin colors, and nationalities. We maintain a roughly equal proportion *w.r.t.*each of the attributes above. In particular, we invite 55 males and 45 females from 26 countries. Their ages range from 20 to 45 years old to match the most common age group appearing on real-world videos. The actors have four typical skin tones: *white*, *black*, *yellow*, and *brown*, with ratio 1:1:1:1. All faces are clean without glasses or decorations.

A professional indoor environment is built for a more controllable data collection. We only use the facial regions (detected and cropped by LAB [96]) of the source data; thus, the background is neglected. We set seven HD cameras from different angles: front, left, left-front, right, right-front, oblique-above, and oblique-below. The resolution of the recorded videos is high (1920 × 1080). The actors are trained in advance to keep the collection process smooth. We request the actors to turn their heads and speak naturally with eight expressions: neutral, angry, happy, sad, surprise, contempt, disgust, and fear. The head poses range from $-90°$ to $+90°$. Furthermore, the actors are asked to perform 53 expressions defined by 3DMM blendshapes [14] (see Fig. 14.4) to supplement some extremely exaggerated expressions. When performing 3DMM blendshapes, the actors also speak naturally to avoid excessive frames that show a closed mouth.

In addition to expressions and poses, we systematically set nine lighting conditions from various directions: uniform, left, top-left, bottom-left, right, top-right, bottom-right, top, and bottom. The actors are only asked to turn their heads under the uniform illumination, so the lighting remains unchanged on specific facial regions to avoid many duplicated data samples recorded by the cameras set at different angles. In total, the collected source data of DeeperForensics-1.0 comprise over 50, 000 videos with around 12.6 million frames.

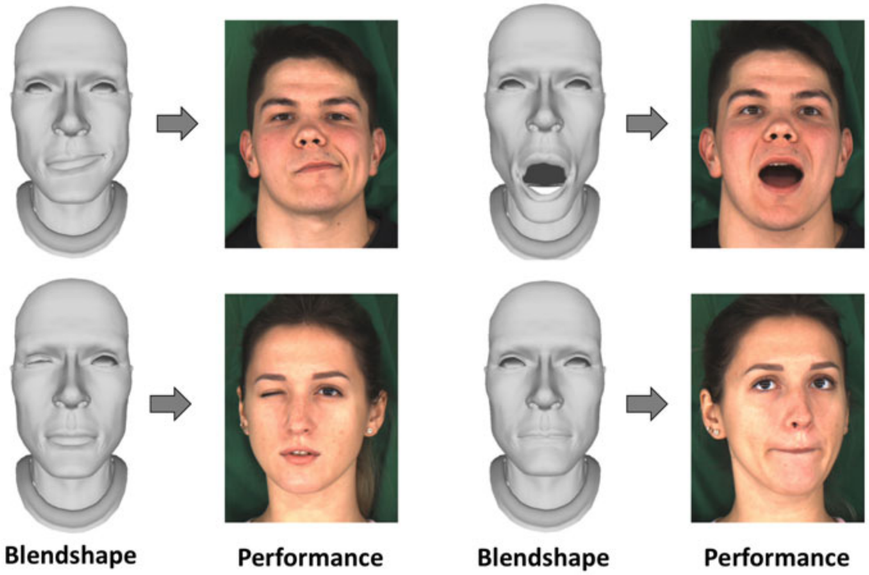

Fig. 14.4 Examples of 3DMM blendshapes and the respective collected source data in DeeperForensics-1.0

14.3.2 DeepFake Variational Auto-Encoder

To improve the *quality* of manipulated data in DeeperForensics-1.0, we consider three key requirements in formulating a high-fidelity face swapping method: (1) The method should be generic and scalable to generate a large number of videos with high quality. (2) The problem of face style mismatch caused by the appearance variations should be addressed. Some failure cases in existing datasets are shown in Fig. 14.5. (3) Temporal continuity of generated videos should be taken into consideration.

Based on the aforementioned requirements, we propose DeepFake Variational Auto-Encoder (DF-VAE), a learning-based face swapping framework. DF-VAE con-

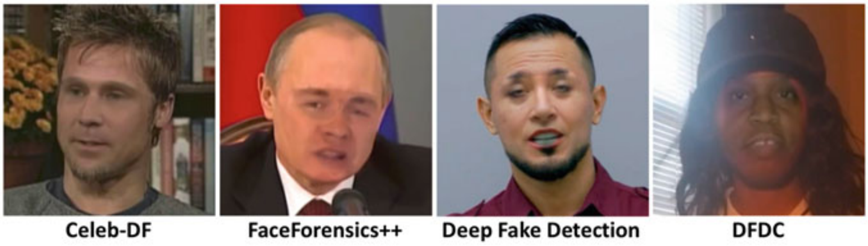

Fig. 14.5 Examples of style mismatch problems in several existing face forensics datasets

sists of three main parts, namely a structure extraction module, a disentangled module, and a fusion module. The details of DF-VAE framework are introduced in this section.

Disentanglement of structure and appearance. The first step of DF-VAE method is face reenactment—animating the source face with similar expression as the target face, without any paired data. Face swapping can be considered as a subsequent step of face reenactment that performs fusion between the reenacted face and the target background. For the robust and scalable face reenactment, we should disentangle the structure (*i.e.*, expression and pose) and appearance (*i.e.*, texture, skin color, *etc.*) representations of a face. This disentanglement is difficult since the structure and appearance representations are far from independent.

Let $\mathbf{x}_{1:T} \equiv \{x_1, x_2, ..., x_T\} \in X$ be a sequence of source face video frames, and $\mathbf{y}_{1:T} \equiv \{y_1, y_2, ..., y_T\} \in Y$ be the sequence of corresponding target face video frames. We first simplify our problem and only consider two specific snapshots at time t, x_t, and y_t. Let $\tilde{x}_t$, $\tilde{y}_t$, d_t represent the reconstructed source face, the reconstructed target face, and the reenacted face, respectively.

Consider the reconstruction procedure of the source face x_t. Let s_x denote the structure representation and a_x denote the appearance information. The face generator can be depicted as the posteriori estimate $p_\theta\left(x_t|s_x, a_x\right)$. The solution of our reconstruction goal, marginal log-likelihood $\tilde{x}_t \sim \log p_\theta\left(x_t\right)$, by a common variational auto-encoder (VAE) [50] can be written as follows:

$$\begin{aligned}\log p_\theta\left(x_t\right) = D_{KL}\left(q_\phi\left(s_x, a_x|x_t\right) \| p_\theta\left(s_x, a_x|x_t\right)\right) \\ +L\left(\theta, \phi; x_t\right),\end{aligned} \tag{14.1}$$

where q_ϕ is an approximate posterior to achieve the evidence lower bound (ELBO) in the intractable case, and the second RHS term $L\left(\theta, \phi; x_t\right)$ is the variational lower bound *w.r.t.*both the variational parameters ϕ and generative parameters θ.

In Eq. (14.1), we assume that both s_x and a_x are latent priors computed by the same posterior x_t. However, the separation of these two variables in the latent space is rather difficult without additional conditions. Therefore, DF-VAE employs a simple yet effective approach to disentangle these two variables.

The blue arrows in Fig. 14.6 demonstrate the reconstruction procedure of the source face x_t. Instead of feeding a single source face x_t, we sample another source face x' to construct unpaired data in the source domain. To make the structure representation more evident, we use the stacked hourglass networks [69] to extract landmarks of x_t in the structure extraction module and get the heatmap $\hat{x}_t$. Then we feed the heatmap $\hat{x}_t$ to the Structure Encoder E_α, and x' to the Appearance Encoder E_β. We concatenate the latent representations (small cubes in red and green) and feed it to the Decoder D_γ. Finally, we get the reconstructed face $\tilde{x}_t$, *i.e.*, marginal log-likelihood of x_t.

Therefore, the latent structure representation s_x in Eq. (14.1) becomes a more evident heatmap representation $\hat{x}_t$, which is introduced as a new condition. The unpaired

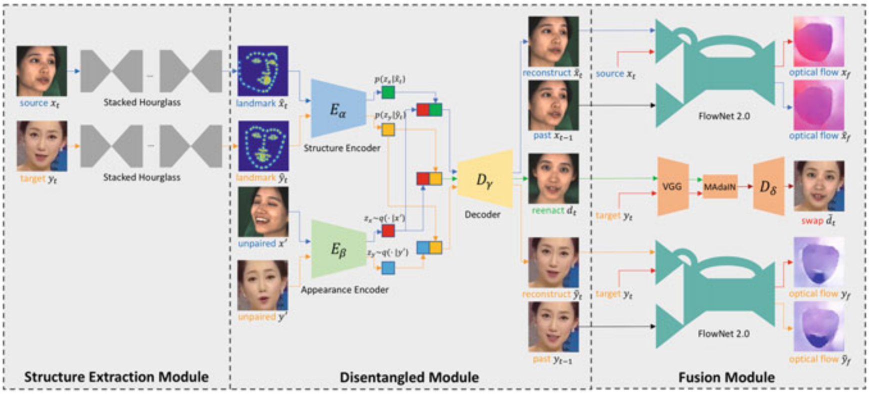

Fig. 14.6 The main framework of DeepFake Variational Auto-Encoder (DF-VAE). In training, we reconstruct the source and target faces in blue and orange arrows, respectively, by extracting landmarks and constructing an unpaired sample as the condition. Optical flow differences are minimized after reconstruction to improve temporal continuity. In inference, we swap the latent codes and get the reenacted face in green arrows. Subsequent MAdaIN module fuses the reenacted face and the original background resulting in the swapped face

sample x' with the same identity $w.r.t.x_t$ is another condition, being a substitute for a_x. Equation (14.1) can be rewritten as a conditional log-likelihood:

$$log\, p_\theta\left(x_t|\hat{x}_t, x'\right) = D_{KL}\left(q_\phi\left(z_x|x_t, \hat{x}_t, x'\right) \| p_\theta\left(z_x|x_t, \hat{x}_t, x'\right)\right) \\ +L\left(\theta, \phi; x_t, \hat{x}_t, x'\right). \tag{14.2}$$

The first RHS term KL-divergence is non-negative, we get the following:

$$\log p_\theta\left(x_t|\hat{x}_t, x'\right) \geq L(\theta, \phi; x_t, \hat{x}_t, x') \\ = \mathbb{E}_{q_\phi\left(z_x|x_t, \hat{x}_t, x'\right)}\left[-\log q_\phi\left(z_x|x_t, \hat{x}_t, x'\right) + \log p_\theta\left(x_t, z_x|\hat{x}_t, x'\right)\right], \tag{14.3}$$

and $L(\theta, \phi; x_t, \hat{x}_t, x')$ can also be written as follows:

$$L\left(\theta, \phi; x_t, \hat{x}_t, x'\right) = -D_{KL}\left(q_\phi\left(z_x|x_t, \hat{x}_t, x'\right) \| p_\theta\left(z_x|\hat{x}_t, x'\right)\right) \\ + \mathbb{E}_{q_\phi\left(z_x|x_t, \hat{x}_t, x'\right)}\left[\log p_\theta\left(x_t|z_x, \hat{x}_t, x'\right)\right]. \tag{14.4}$$

We let the variational approximate posterior be a multivariate Gaussian with a diagonal covariance structure:

$$\log q_\phi\left(z_x|x_t, \hat{x}_t, x'\right) \equiv \log \mathcal{N}\left(z_x; \mu, \sigma^2\mathbf{I}\right), \tag{14.5}$$

where **I** is an identity matrix. Exploiting the reparameterization trick [50], the non-differentiable operation of sampling can become differentiable by an auxiliary variable with independent marginal. In this case, $z_x \sim q_\phi\left(z_x|x_t, \hat{x}_t, x'\right)$ is implemented by $z_x = \mu + \sigma\epsilon$ where ϵ is an auxiliary noise variable $\epsilon \sim \mathcal{N}(0, 1)$. Finally, the approximate posterior $q_\phi(z_x|x_t, \hat{x}_t, x')$ is estimated by the separated encoders, Structure Encoder E_α and Appearance Encoder E_β, in an end-to-end training process by standard gradient descent.

We discuss the whole workflow of reconstructing the source face. In the target face domain, the reconstruction procedure is the same, as shown by orange arrows in Fig. 14.6. During training, the network learns structure and appearance information in both the source and the target domains. It is noteworthy that even if both y_t and x' belong to arbitrary identities, our effective disentangled module is capable of learning meaningful structure and appearance information of each identity. During inference, we concatenate the appearance prior of x' and the structure prior of y_t (small cubes in red and orange) in the latent space, and the reconstructed face d_t shares the same structure with y_t and keeps the appearance of x'. DF-VAE framework allows concatenations of structure and appearance latent codes extracted from arbitrary identities in inference and permits *many-to-many face reenactment*.

In summary, DF-VAE is a conditional variational auto-encoder [49] with robustness and scalability. It conditions on two posteriors in different domains. In the disentangled module, the separated design of two encoders E_α and E_β, the explicit structure heatmap, and the unpaired data construction jointly force E_α to learn structure information and E_β to learn appearance information.

Style matching and fusion. To fix the obvious style mismatch problems as shown in Fig. 14.5, we adopt a masked adaptive instance normalization (MAdaIN) module in DF-VAE. We place a typical AdaIN [35] network after the reenacted face d_t. In the face swapping scenario, we only need to adjust the style of the face area to match the original background. Therefore, we use a mask m_t to guide AdaIN [35] network to focus on style matching of the face area. To avoid boundary artifacts, we apply Gaussian Blur to m_t and get the blurred mask m_t^b.

In our face swapping context, d_t is the content input of MAdaIN, and y_t is the style input. MAdaIN adaptively computes the affine parameters from the face area of the style input:

$$\text{MAdaIN}\,(c, s) = \sigma\,(s)\left(\frac{c - \mu\,(c)}{\sigma\,(c)}\right) + \mu\,(s)\,, \tag{14.6}$$

where $c = m_t^b \cdot d_t$, $s = m_t^b \cdot y_t$. With the low-cost MAdaIN module, we reconstruct d_t again by Decoder D_δ. The blurred mask m_t^b is used again to fuse the reconstructed image with the background of y_t. At last, we get the swapped face $\overline{d}_t$.

The MAdaIN module is jointly trained with the disentangled module in an end-to-end manner. Thus, by a *single* model, DF-VAE can perform *many-to-many face swapping* with obvious reduction of style mismatch and facial boundary artifacts (see Fig. 14.7 for the face swapping between three source identities and three target identities). Even if there are multiple identities in both the source domain and the target domain, the quality of face swapping does not degrade.

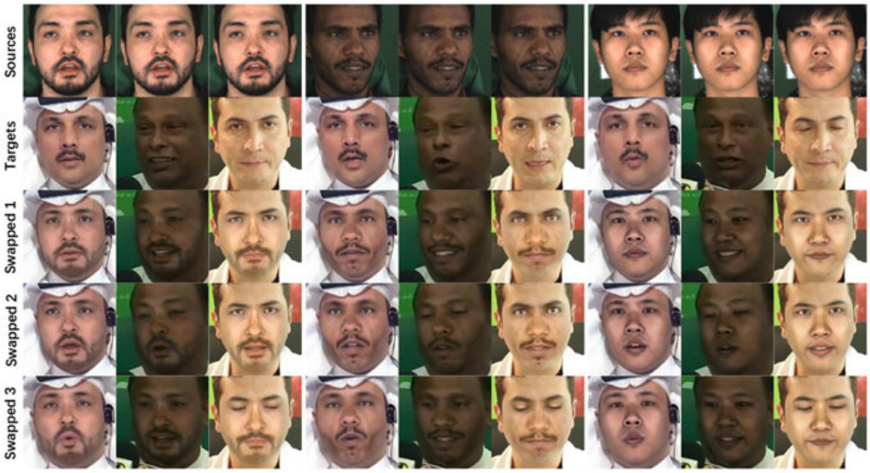

Fig. 14.7 Many-to-many (three-to-three) face swapping by a *single* model with obvious reduction of style mismatch problems. This figure shows the results between three source identities and three target identities. The whole process is end-to-end

Temporal consistency constraint. Temporal discontinuity of the fake videos generated by certain face manipulation methods leads to obvious flickering of the face area, making them easy to be spotted by forgery detection methods and human eyes. To improve temporal continuity, DF-VAE lets the disentangled module learn temporal information of both the source face and the target face.

For simplification, we make a Markov assumption that the generation of the frame at time t sequentially depends on its previous P frames $\mathbf{x}_{(t-p):(t-1)}$. We set $P = 1$ to balance quality improvement and training time.

To build the relationship between a current frame and previous ones, we further make an intuitive assumption that the optical flows should remain unchanged after reconstruction. We use FlowNet 2.0 [37] to estimate the optical flow $\tilde{x}_f$ *w.r.t.* $\tilde{x}_t$ and x_{t-1} and x_f *w.r.t.*x_t and x_{t-1}. Since face swapping is sensitive to minor facial details which can be greatly affected by flow estimation, we do not warp x_{t-1} by the estimated flow like [94]. Instead, we minimize the difference between $\tilde{x}_f$ and x_f to improve temporal continuity while keeping stable facial detail generation. To this end, we propose a new temporal consistency constraint, which can be written as follows:

$$L_{temporal} = \frac{1}{CHW} \|\tilde{x}_f - x_f\|_1, \tag{14.7}$$

where $C = 2$ for a common form of optical flow.

We only discuss the temporal continuity *w.r.t.*the source face in this section. The case of the target face is the same. If multiple identities exist in one domain, temporal information of all these identities can be learned in an end-to-end manner.

14.3.3 Scale and Diversity

The extensive data collection and the introduced DF-VAE method are designed to improve the *quality* of manipulated videos in the DeeperForensics-1.0 dataset. In this section, we mainly discuss the *scale* and *diversity* aspects.

The DeeperForensics-1.0 dataset contains 10, 000 manipulated videos with 5 million frames. We take 1, 000 refined YouTube videos collected by FaceForensics++ [81] as the target videos. Each face of our collected 100 identities is swapped onto 10 target videos; thus, 1, 000 raw manipulated videos are generated directly by DF-VAE in an end-to-end process. Thanks to the scalability and multimodality of DF-VAE, the time overhead of model training and data generation is reduced to 1/5 compared to the common DeepFakes methods, with no degradation in quality. Thus, a larger scale dataset construction is possible.

To enhance diversity, we apply various perturbations existing in real scenes. Specifically, as shown in Fig. 14.8, seven types of distortions defined in Image Quality Assessment (IQA) [58, 77] are included. Each distortion is divided into five intensity levels. We apply random-type distortions to the 1, 000 raw manipulated videos at five different intensity levels, producing a total of 5, 000 manipulated videos. Besides, an additional of 1, 000 robust manipulated videos are generated by adding random-type, random-level distortions to the 1, 000 raw manipulated videos. Moreover, in contrast to other datasets [13, 51, 57, 81, 99], each sample of another 3, 000 manipulated videos in DeeperForensics-1.0 is subjected to a mixture of more than one distortion (examples shown in Fig. 14.8). The variety of perturbations improves the *diversity* of DeeperForensics-1.0 to approximate the data distribution of real-world scenarios better.

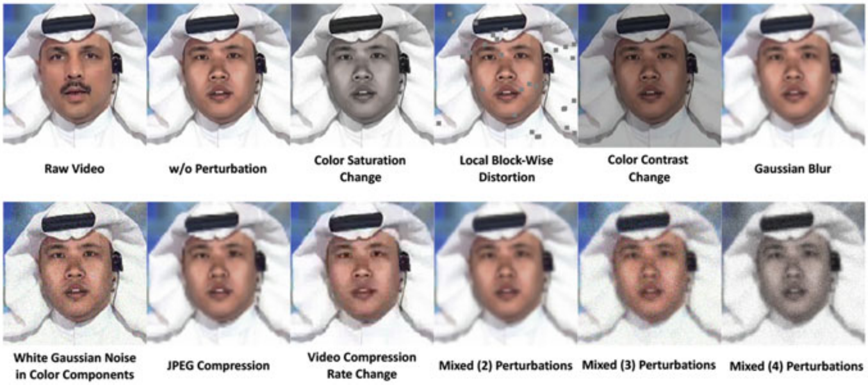

Fig. 14.8 Seven types of perturbations and the mixture of two (Gaussian blur, JPEG compression) / three (Gaussian blur, JPEG compression, white Gaussian noise in color components) / four (Gaussian blur, JPEG compression, white Gaussian noise in color components, color saturation change) perturbations in DeeperForensics-1.0.

14.3.4 Hidden Test Set

Several existing benchmarks [57, 81] have demonstrated high-accuracy face forgery detection results using their proposed datasets. However, the sources and imposed distortions of DeepFakes videos are much more variable and unpredictable in real-world scenarios. Due to the huge biases introduced by a close distribution between the training and test sets, the actual efficacy of these studies [57, 81] in detecting real-world face forgery cases remains to be further elucidated.

An indispensable component of DeeperForensics-1.0 is its introduced hidden test set, which is richer in distribution than the publicly available training set. The hidden test set suggests a better real-world face forgery detection setting: (1) Multiple sources. Fake videos in the wild should be manipulated by different unknown methods; (2) High quality. Threatening fake videos should have high quality to deceive human eyes; (3) Diverse distortions. Different perturbations should be taken into consideration. The ground truth labels are hidden and are used on the host server to evaluate the accuracy of detection models. The hidden test set will evolve by including more challenging samples along with the development of DeepFakes technology.

Overall, DeeperForensics-1.0 is a new *large-scale* dataset consisting of over $60,000$ videos with 17.6 million frames for real-world face forgery detection. *Good-quality* source videos and manipulated videos constitute two main contributions of this dataset. The *high-diversity* perturbations applying to the manipulated videos enhance the robustness of DeeperForensics-1.0 to simulate real scenes. The dataset has been released, free to all research communities, for developing face forgery detection and more general human-face-related research.[1,2]

14.4 DeeperForensics Challenge 2020

In this section, we detail the DeeperForensics Challenge 2020 on real-world face forgery detection, which aims at soliciting innovations to advance the state of the art in DeepFakes detection. The challenge uses the DeeperForensics-1.0 dataset introduced above, and the model evaluation is performed online on the current version of the hidden test set. Participants are expected to devise robust and generic methods for forgery detection in real-world scenarios. The challenge results constitute an essential part of the DeeperForensics-1.0 benchmark. We describe the detailed challenge information and summarize the winning solutions to take a closer look at the current status and possible future development of real-world face forgery detection.

[1] GitHub (dataset and code): https://github.com/EndlessSora/DeeperForensics-1.0.

[2] Project page: https://liming-jiang.com/projects/DrF1/DrF1.html.

14.4.1 Platform

The DeeperForensics Challenge 2020 is hosted on the CodaLab platform[3] in conjunction with ECCV 2020, the second Workshop on Sensing, Understanding, and Synthesizing Humans.[4] The online evaluation is conducted using Amazon Web Services (AWS).[5]

First, participants register their teams on the CodaLab challenge website. Then, they are requested to submit their models to the AWS evaluation server (with one 16 GB Tesla V100 GPU for each team) to perform the online evaluation on the hidden test set. When the evaluation is done, participants receive the encrypted prediction files through an automatic email. Finally, they submit the result file to the CodaLab challenge website.

14.4.2 Challenge Dataset

The DeeperForensics Challenge 2020 employs the DeeperForensics-1.0 dataset [41] that was proposed in CVPR 2020. The detailed information of this dataset has been provided in Sect. 14.3. The evaluation of the challenge is performed online on the current version of the hidden test set (Sect. 14.3.4).

All the participants using the DeeperForensics-1.0 dataset should agree to its Terms of Use [9]. They are recommended but not restricted to train their algorithms on DeeperForensics-1.0. The use of any external datasets should be disclosed and follow the Terms of Use.

14.4.3 Evaluation Metric

Similar to Deepfake Detection Challenge (DFDC) [2], the DeeperForensics Challenge 2020 uses the binary cross-entropy loss (BCELoss) to evaluate the performance of face forgery detection models:

$$\text{BCELoss} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \cdot \log\left(p\left(y_i\right)\right) + \left(1 - y_i\right) \cdot \log\left(1 - p\left(y_i\right)\right)\right], \tag{14.8}$$

where N is the number of videos in the hidden test set, y_i denotes the ground truth label of video i (fake: 1, real: 0), and $p\left(y_i\right)$ indicates the predicted probability that video i is fake. A smaller BCELoss score is better, which directly contributes to a

[3] Challenge website: https://competitions.codalab.org/competitions/25228.

[4] Workshop website: https://sense-human.github.io/index_2020.html.

[5] Online evaluation website: https://aws.amazon.com.

higher ranking. If the BCELoss score is the same, the one with less runtime will achieve a higher ranking. To avoid an infinite BCELoss that is both too confident and wrong, the score is bounded by a threshold value.

14.4.4 Timeline

The DeeperForensics Challenge 2020 lasted for nine weeks—eight weeks for the *development phase* and one week for the *final test phase*.

The challenge officially started at the ECCV 2020 SenseHuman Workshop on August 28, 2020, and it immediately entered the development phase. In the development phase, the evaluation is performed on the *test-dev* hidden test set, which contains 1, 000 videos representing general circumstances of the full hidden test set. The *test-dev* hidden test set is used to maintain a public leaderboard. Participants can conduct four online evaluations (each with 2.5 h of runtime limit) per week.

The final test phase started on October 24, 2020. The evaluation is conducted on the *test-final* hidden test set, containing 3, 000 videos (also including test-dev videos) with a similar distribution as test-dev, for the final competition results. A total of two online evaluations (each with 7.5 h of runtime limit) are allowed. The final test phase ended on October 31, 2020.

Finally, the challenge results were announced in December 2020. In total, 115 participants registered for the competition, and 25 teams made valid submissions.

14.4.5 Results and Solutions

Among the 25 teams who made valid submissions, many participants achieve promising results. We show the final results of the top-5 teams in Table 14.1. In the following subsections, we present the winning solutions of top-3 entries.

Table 14.1 Final results of the top-5 teams in the DeeperForensics Challenge 2020. The runtime is shown in seconds.

Ranking	TeamName	UserName	BCELoss↓	Runtime↓
1	Forensics	BokingChen	0.2674	7690
2	RealFace	Iverson	0.3699	11368
3	VISG	zz110	0.4060	11012
4	jiashangplus	jiashangplus	0.4064	16389
5	Miao	miaotao	0.4132	19823

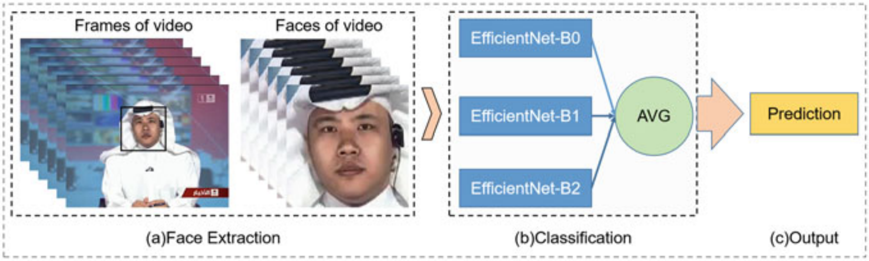

Fig. 14.9 The framework of the first-place solution in the DeeperForensics Challenge 2020

- **Solution of First Place**

As shown in Fig. 14.9, the method designed by the champion team contains three stages, namely Face Extraction, Classification, and Output.

Face Extraction. They first extract 15 frames from each video at equal intervals using VideoCapture of OpenCV. Then, they use the face detector MTCNN [102] to detect the face region of each frame and expand the region by 1.2 times to crop the face image.

Classification. They define the prediction of the probability that the face is fake as the face score. They use EfficientNet [84] as the backbone, which was proven effective in the Deepfake Detection Challenge (DFDC) [2]. The results of three models (EfficientNet-B0, EfficientNet-B1, and EfficientNet-B2) are ensembled for each face.

Output. The final output score of a video is the predicted probability that the video is fake, which is calculated by the average of face scores for the extracted frames.

Implementation Details. The team employs EfficientNet pre-trained on ImageNet as the backbone. They select EfficientNet-B0, EfficientNet-B1, and EfficientNet-B2 for the model ensemble. In addition to DeeperForensics-1.0, they use some other public datasets, *i.e.*, UADFV [99], Deep Fake Detection [13], FaceForensics++ [81], Celeb-DF [57], and DFDC Preview [24]. They balance the class samples with the down-sampling mode. The code of the champion solution has been made publicly available.[6]

– *Training*: Inspired by the DFDC winning solution, appropriate data augmentation could contribute to better results. As for the data augmentation, the champion team uses the perturbation implementation in DeeperForensics-1.0 [8] during training. They only apply the image-level distortions: color saturation change (CS), color contrast change (CC), local block-wise (BW), white Gaussian noise in color components (GNC), Gaussian blur (GB), and JPEG compression (JPEG). They randomly mix up these distortions with a probability of 0.2. Besides, they also try other data

[6] https://github.com/beibuwandeluori/DeeperForensicsChallengeSolution.

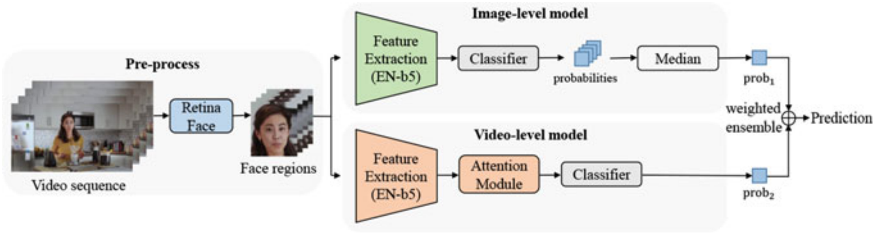

Fig. 14.10 The framework of the second-place solution in the DeeperForensics Challenge 2020

augmentation [3], but the performance improvement is slim. The images are resized to 224 × 224. The batch size is 128, and the total training epoch is 50. They use AdamW optimizer [62] with initial learning rate of 0.001. Label smoothing is applied with a smoothing factor of 0.05.

– *Testing*: The testing pipeline follows the three stages in Fig. 14.9. They clip the prediction score of each video in a range of $[0.01, 0.99]$ to reduce the large loss caused by the prediction errors. In addition to the best BCELoss score, their fastest execution speed may be attributed to the use of the faster face extractor MTCNN and the ensemble of three image-level models with fewer parameters.

- **Solution of Second Place**

Face manipulated video contains two types of forgery traces, *i.e.*, image-level artifacts and video-level artifacts. The former refers to the artifacts such as blending boundaries and abnormal textures within image, while the latter is the face jitter problem between video frames. Most previous works only focused on artifacts in a specific modality and lacked consideration of both. The team in the second place proposes to use an attention mechanism to fuse the temporal information in videos and further combine it with an image model to achieve better results.

The overall framework of their method is shown in Fig. 14.10. First, they use RetinaFace [22] with 20% margin to detect faces in video frames. Then, the face sequence is fed into an image-based model and a video-based model, where the backbones are both EfficientNet-b5 [84] with NoisyStudent [98] pre-trained weights. The image-based model predicts frame by frame and takes the median of probabilities as the prediction. The video-based model takes the entire face sequence as the input and adopts an attention module to fuse the temporal information between frames. Finally, the per-video prediction score is obtained by averaging the probabilities predicted by the above two models.

Implementation Details. The team implements the proposed method via PyTorch. All the models are trained on 8 NVIDIA Tesla V100 GPUs. In addition to the DeeperForensics-1.0 dataset, they use three external datasets, *i.e.*, FaceForensics++ [81], Celeb-DF [57], and Diverse Fake Face Dataset [21]. They used the official splits provided by the above datasets to construct the training, validation, and test sets. They balance the positive and negative samples through the down-sampling technique.

– *Training*: The second-place team uses the following data augmentations: RandAugment [20], patch Gaussian [61], Gaussian blur, image compression, random flip, random crop, and random brightness contrast. They also employ the perturbation implementation in DeeperForensics-1.0 [8]. For the image-based model, they train a classifier based on EfficientNet-b5 [84], using binary cross-entropy loss as the loss function. They adopt a two-stage training strategy for the video-based model. In stage-1, they train an image-based classifier based on EfficientNet-b5. In stage-2, they fix the model parameters trained in stage-1 to serve as face feature extractor and introduce an attention module to learn temporal information via nonlinear transformations and *softmax* operations. The input of the network is the face sequence (*i.e.*, 5 frames per video) in stage-2, and only the attention module and classification layers are trained. The binary cross-entropy loss is adopted as the loss function. The input size is scaled to 320 × 320. The Adam optimizer [47] is used with a learning rate of 0.0002, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and weight decay of 0.00001. The batch size is 32. The total number of training epochs is set to 20, and the learning rate is halved every 5 epochs.

– *Testing*: They sample 10 frames at equal intervals for each video and detect faces by RetinaFace [22] as in the training phase. Then, the face images are resized to 320 × 320. Test-time augmentation (TTA) (*e.g.*, flip) is applied to get 20 images (10 original and 10 flipped), which are fed into the network to get the prediction score. They clip the prediction score of each video to [0.01, 0.99] to avoid excessive losses on extreme error samples.

- **Solution of Third Place**

Similar to the second-place entry, the team in the third place also utilize the poor temporal consistency in existing face manipulation techniques. To this end, they propose to use a 3D convolutional neural network (3DCNN) to capture spatial-temporal features for forgery detection. The framework of their method is shown in Fig. 14.11.

Implementation Details. First, the team crops faces in the video frames using the MTCNN [102] face detector. They combine all the cropped face images into a face video clip. Each video clip is then resized to 64 × 224 × 224 or 64 × 112 × 112. Various data augmentations are applied, including Gaussian blur, white Gaussian noise in color components, random crop, random flip, *etc*. Then, they use the processed video clips as the input to train a 3D convolutional neural network (3DCNN) using the cross-entropy loss. They examine three kinds of networks, I3D [15], 3D ResNet [29], and R(2+1)D [90]. These models are pre-trained on the action recognition datasets, *e.g.*, kinetics [46]. In addition to DeeperForensics-1.0, they use three external public face manipulation datasets, *i.e.*, the DFDC dataset [23], Deep Fake Detection [13], and FaceForensics++ [81].

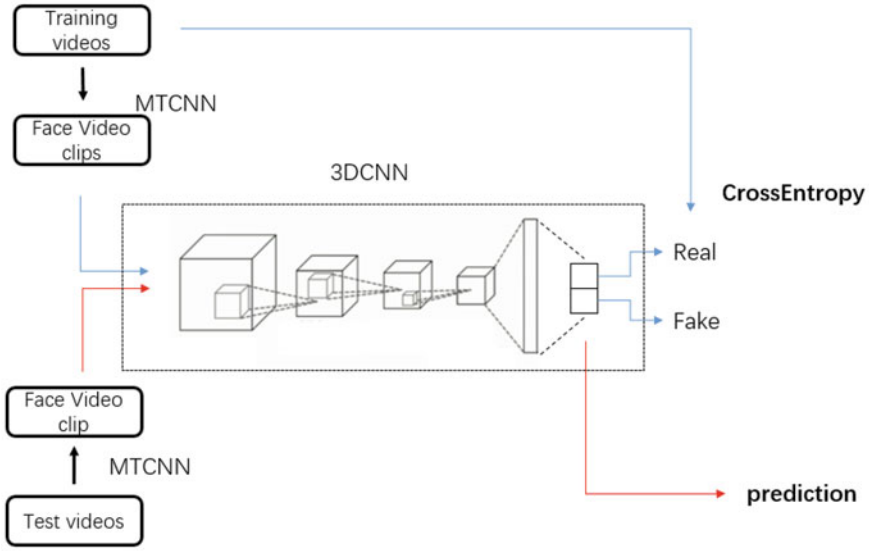

Fig. 14.11 The framework of the third-place solution in the DeeperForensics Challenge 2020

14.5 Discussion

In this chapter, we have introduced a new large-scale dataset named DeeperForensics-1.0. The dataset facilitates the research of face forgery detection in real-world scenarios. We have also presented several methods that consider different potential aspects in developing a robust face forgery detection model. Winning solutions of the DeeperForensics Challenge 2020 have achieved promising performance.

In summary, there are three key points inspired by these methods that could improve real-world face forgery detection. (1) Strong backbone. Backbone selection for a forgery detection model is important. The high-performance winning solutions are based on state-of-the-art EfficientNet. (2) Diverse augmentations. Applying appropriate data augmentations may better simulate real-world scenarios and boost the model performance. (3) Temporal information. Since the primary detection target is the fake videos, temporal information can be a critical clue to distinguish the real from the fake.

Despite the promising results, we believe that there is still much room for improvement in the real-world face forgery detection task. (1) More suitable and diverse data augmentations may contribute to a better simulation of real-world data distribution. (2) Developing a robust detection method that can cope with unseen manipulation methods and distortions is a critical problem. At this stage, we observe that the model training is data-dependent. Although data augmentations can help improve the performance to a certain extent, the generalization ability of most forgery detection models is still poor. (3) Different artifacts in the DeepFakes videos (*e.g.*, checkerboard Artifacts and fusion boundary artifacts) remain rarely explored.

14.6 Further Reading

Interested readers are referred to the following further readings:

- [41] for more detailed information about the DeeperForensics-1.0 dataset and more detection baselines in the DeeperForensics-1.0 video forgery detection benchmark.
- [40] for more detailed information about the DeeperForensics Challenge 2020.
- [23, 31, 57, 81] for other closely related DeepFakes detection datasets.
- [66, 71, 88] for surveys on DeepFakes creation and detection.

References

1. DeepFaceLab https://github.com/iperov/DeepFaceLab. Accessed 20 Aug 2019
2. Deepfake Detection Challenge. https://www.kaggle.com/c/deepfake-detection-challenge. Accessed 15 Feb 2020
3. Deepfake detection (DFDC) solution by selimsef https://github.com/selimsef/dfdc_deepfake_challenge. Accessed 30 Oct 2020
4. DeepFakes https://github.com/deepfakes/faceswap. Accessed 16 Aug 2019
5. FaceSwap https://github.com/MarekKowalski/FaceSwap. Accessed 18 Aug 2019
6. faceswap-GAN https://github.com/shaoanlu/faceswap-GAN. Accessed 16 Aug 2019
7. FakeAPP https://www.fakeapp.com. Accessed 25 July 2019
8. Perturbation implementation in DeeperForensics-1.0 https://github.com/EndlessSora/DeeperForensics-1.0/tree/master/perturbation. Accessed 30 Oct 2020
9. Terms of use: DeeperForensics-1.0 dataset https://github.com/EndlessSora/DeeperForensics-1.0/blob/master/dataset/Terms_of_Use.pdf. Accessed 21 May 2020
10. Afchar D, Nozick V, Yamagishi J, Echizen I (2018) Mesonet: a compact facial video forgery detection network. In: Proceedings of the IEEE international workshop on information forensics and security
11. Amerini I, Ballan L, Caldelli R, Del Bimbo A, Serra G (2011) A sift-based forensic method for copy-move attack detection and transformation recovery. IEEE Trans Inf Forensics Secur 6:1099–1110
12. Bayar B, Stamm MC (2016) 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
13. Blog GA Contributing data to deepfake detection research. https://ai.googleblog.com/2019/09/contributing-data-to-deepfake-detection.html. Accessed 25 Sep 2019
14. Cao C, Weng Y, Zhou S, Tong Y, Zhou K (2013) FaceWarehouse: a 3D facial expression database for visual computing. IEEE Trans Visualization Comput Gr 20:413–425
15. Carreira J, Zisserman A (2017) Quo vadis, action recognition? a new model and the kinetics dataset. In: Proceedings of the IEEE conference on computer vision and pattern recognition

16. Chai L, Bau D, Lim SN, Isola P (2020) What makes fake images detectable? understanding properties that generalize. In: Proceedings of the european conference on computer vision
17. Choi Y, Choi M, Kim M, Ha JW, Kim S, Choo J (2018) StarGAN: unified generative adversarial networks for multi-domain image-to-image translation. In: Proceedings of the IEEE conference on computer vision and pattern recognition
18. Chollet F (2017) Xception: deep learning with depthwise separable convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition
19. Cozzolino D, Poggi G, Verdoliva L (2017) Recasting residual-based local descriptors as convolutional neural networks: an application to image forgery detection. In: Proceedings of the 5th ACM workshop on information hiding and multimedia security
20. Cubuk ED, Zoph B, Shlens J, Le QV (2020) RandAugment: practical automated data augmentation with a reduced search space. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops
21. Dang H, Liu F, Stehouwer J, Liu X, Jain AK (2020) On the detection of digital face manipulation. In: Proceedings of the IEEE conference on computer vision and pattern recognition
22. Deng J, Guo J, Ververas E, Kotsia I, Zafeiriou S (2020) RetinaFace: Single-shot multi-level face localisation in the wild. In: Proceedings of the IEEE conference on computer vision and pattern recognition
23. Dolhansky B, Bitton J, Pflaum B, Lu J, Howes R, Wang M, Ferrer CC (2020) The deepfake detection challenge dataset. arXiv preprint arXiv:2006.07397
24. Dolhansky B, Howes R, Pflaum B, Baram N, Ferrer CC (2019) The deepfake detection challenge (DFDC) preview dataset. arXiv preprint arXiv:1910.08854
25. Fridrich J, Kodovsky J (2012) Rich models for steg analysis of digital images. IEEE Trans Inf Forensics Secur 7:868–882
26. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y (2014) Generative adversarial nets. In: Proceedings of the advances in neural information processing systems
27. Guarnera L, Giudice O, Battiato S (2020) Deepfake detection by analyzing convolutional traces. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops
28. Güera D, Delp EJ (2018) Deepfake video detection using recurrent neural networks. In: Proceedings of the IEEE international conference on advanced video and signal based surveillance
29. Hara K, Kataoka H, Satoh Y (2017) Learning spatio-temporal features with 3D residual networks for action recognition. In: Proceedings of the IEEE international conference on computer vision workshops
30. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition
31. He Y, Gan B, Chen S, Zhou Y, Yin G, Song L, Sheng L, Shao J, Liu Z (2021) ForgeryNet: A versatile benchmark for comprehensive forgery analysis. In: Proceedings of the IEEE conference on computer vision and pattern recognition
32. Hernandez-Ortega J, Tolosana R, Fierrez J, Morales A (2021) DeepFakesON-Phys: deepfakes detection based on heart rate estimation. In: Proceedings of the AAAI conference on artificial intelligence workshops (2021)
33. Hinton GE, Salakhutdinov RR (2006) Reducing the dimensionality of data with neural networks. Science 313:504–507
34. Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9:1735–1780
35. Huang X, Belongie S (2017) Arbitrary style transfer in real-time with adaptive instance normalization. In: Proceedings of the IEEE international conference on computer vision
36. Huang X, Liu MY, Belongie S, Kautz J (2018) Multimodal unsupervised image-to-image translation. In: Proceedings of the European conference on computer vision
37. Ilg E, Mayer N, Saikia T, Keuper M, Dosovitskiy A, Brox T (2017) Flownet 2.0: Evolution of optical flow estimation with deep networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition

38. Isola P, Zhu JY, Zhou T, Efros AA (2017) Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition
39. Jiang L, Dai B, Wu W, Loy CC (2020) Focal frequency loss for image reconstruction and synthesis. In: Proceedings of the IEEE international conference on computer vision
40. Jiang L, Guo Z, Wu W, Liu Z, Liu Z, Loy CC, Yang S, Xiong Y, Xia W, Chen B, Zhuang P, Li S, Chen S, Yao T, Ding S, Li J, Huang F, Cao L, Ji R, Lu C, Tan G (2021) DeeperForensics Challenge 2020 on real-world face forgery detection: methods and results. arXiv:2102.09471
41. Jiang L, Li R, Wu W, Qian C, Loy CC (2020) DeeperForensics-1.0: a large-scale dataset for real-world face forgery detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition
42. Jiang L, Zhang C, Huang M, Liu C, Shi J, Loy CC (2020) TSIT: a simple and versatile framework for image-to-image translation. In: Proceedings of the european conference on computer vision
43. Karras T, Aila T, Laine S, Lehtinen J (2017) Progressive growing of GANs for improved quality, stability, and variation. arXiv:1710.10196
44. Karras T, Laine S, Aila T (2019) A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition
45. Karras T, Laine S, Aittala M, Hellsten J, Lehtinen J, Aila T (2020) Analyzing and improving the image quality of StyleGAN. In: Proceedings of the IEEE conference on computer vision and pattern recognition
46. Kay W, Carreira J, Simonyan K, Zhang B, Hillier C, Vijayanarasimhan S, Viola F, Green T, Back T, Natsev P et al (2017) The kinetics human action video dataset. arXiv:1705.06950
47. Kingma DP, Ba J (2014) Adam: a method for stochastic optimization. arXiv:1412.6980
48. Kingma DP, Dhariwal P (2018) Glow: generative flow with invertible 1x1 convolutions. In: Proceedings of the advances in neural information processing systems
49. Kingma DP, Mohamed S, Rezende DJ, Welling M (2014) Semi-supervised learning with deep generative models. In: Proceedings of the advances in neural information processing systems
50. Kingma DP, Welling M (2013) Auto-encoding variational bayes. arXiv:1312.6114
51. Korshunov P, Marcel S (2018) Deepfakes: a new threat to face recognition? assessment and detection. arXiv:1812.08685
52. Korus P, Huang J (2016) Multi-scale analysis strategies in PRNU-based tampering localization. IEEE Trans Inf Forensics Secur 12:809–824
53. Li L, Bao J, Yang H, Chen D, Wen F (2020) FaceShifter: towards high fidelity and occlusion aware face swapping. In: Proceedings of the IEEE conference on computer vision and pattern recognition
54. Li L, Bao J, Zhang T, Yang H, Chen D, Wen F, Guo B (2020) Face x-ray for more general face forgery detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition
55. Li Y, Chang MC, Lyu S (2018) In ictu oculi: exposing ai created fake videos by detecting eye blinking. In: Proceedings of the IEEE international workshop on information forensics and security
56. Li Y, Lyu S (2018) Exposing deepfake videos by detecting face warping artifacts. arXiv:1811.00656
57. Li Y, Yang X, Sun P, Qi H, Lyu S (2019) Celeb-DF: a new dataset for deepfake forensics. arXiv:1909.12962
58. Lin KY, Wang G (2018) Hallucinated-iqa: No-reference image quality assessment via adversarial learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition
59. Liu MY, Breuel T, Kautz J (2017) Unsupervised image-to-image translation networks. In: Proceedings of the advances in neural information processing systems
60. Liu X, Yin G, Shao J, Wang X, Li H (2019) Learning to predict layout-to-image conditional convolutions for semantic image synthesis. In: Proceedings of the advances in neural information processing systems

61. Lopes RG, Yin D, Poole B, Gilmer J, Cubuk ED (2019) Improving robustness without sacrificing accuracy with patch gaussian augmentation. arXiv:1906.02611
62. Loshchilov I, Hutter F (2019) Decoupled weight decay regularization. In: Proceedings of the international conference on learning representations
63. Masi I, Killekar A, Mascarenhas RM, Gurudatt SP, AbdAlmageed W (2020) Two-branch recurrent network for isolating deepfakes in videos. In: Proceedings of the european conference on computer vision
64. Matern F, Riess C, Stamminger M (2019) Exploiting visual artifacts to expose deepfakes and face manipulations. In: Proceedings of the IEEE winter applications of computer vision workshops
65. McCloskey S, Albright M (2019) Detecting gan-generated imagery using saturation cues. In: Proceedings of the IEEE international conference on image processing
66. Mirsky Y, Lee W (2021) The creation and detection of deepfakes: a survey. ACM Comput Surveys 54:1–41
67. Mirza M, Osindero S (2014) Conditional generative adversarial nets. arXiv:1411.1784
68. Mittal T, Bhattacharya U, Chandra R, Bera A, Manocha D (2020) Emotions don't lie: a deepfake detection method using audio-visual affective cues. arXiv:2003.06711
69. Newell A, Yang K, Deng J (2016) Stacked hourglass networks for human pose estimation. In: Proceedings of the european conference on computer vision
70. Nguyen HH, Fang F, Yamagishi J, Echizen I (2019) Multi-task learning for detecting and segmenting manipulated facial images and videos. arXiv:1906.06876
71. Nguyen TT, Nguyen CM, Nguyen DT, Nguyen DT, Nahavandi S (2019) Deep learning for deepfakes creation and detection: a survey. arXiv:1909.11573
72. Nirkin Y, Keller Y, Hassner T (2019) FSGAN: subject agnostic face swapping and reenactment. In: Proceedings of the IEEE international conference on computer vision
73. Nirkin Y, Wolf L, Keller Y, Hassner T (2020) Deepfake detection based on the discrepancy between the face and its context. arXiv:2008.12262
74. Van den Oord A, Kalchbrenner N, Espeholt L, Vinyals O, Graves A et al (2016) Conditional image generation with PixelCNN decoders. In: Proceedings of the advances in neural information processing systems
75. Park T, Liu MY, Wang TC, Zhu JY (2019) Semantic image synthesis with spatially-adaptive normalization. In: Proceedings of the IEEE conference on computer vision and pattern recognition
76. Petrov I, Gao D, Chervoniy N, Liu K, Marangonda S, Umé C, Jiang J, RP L, Zhang S, Wu P et al (2020) DeepFaceLab: a simple, flexible and extensible face swapping framework. arXiv:2005.05535
77. Ponomarenko N, Jin L, Ieremeiev O, Lukin V, Egiazarian K, Astola J, Vozel B, Chehdi K, Carli M, Battisti F et al (2015) Image database TID2013: peculiarities, results and perspectives. Signal Process Image Commun 30:57–77
78. Qian Y, Yin G, Sheng L, Chen Z, Shao J (2020) Thinking in frequency: face forgery detection by mining frequency-aware clues. In: Proceedings of the european conference on computer vision
79. Radford A, Metz L, Chintala S (2015) Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv:1511.06434
80. Rahmouni N, Nozick V, Yamagishi J, Echizen I (2017) Distinguishing computer graphics from natural images using convolution neural networks. In: Proceedings of the IEEE workshop on information forensics and security
81. Rössler A, Cozzolino D, Verdoliva L, Riess C, Thies J, Nießner M (2019) FaceForensics++: learning to detect manipulated facial images. In: Proceedings of the IEEE international conference on computer vision (2019)
82. Sanderson C (2002) The vidtimit database. Tech. rep, IDIAP
83. Sun K, Liu H, Ye Q, Liu J, Gao Y, Shao L, Ji R (2021) Domain general face forgery detection by learning to weight. In: Proceedings of the AAAI conference on artificial intelligence

84. Tan M, Le Q (2019) EfficientNet: rethinking model scaling for convolutional neural networks. In: Proceedings of the international conference on machine learning
85. Thies J, Zollhöfer M, Nießner M (2019) Deferred neural rendering: image synthesis using neural textures. arXiv:1904.12356
86. Thies J, Zollhofer M, Stamminger M, Theobalt C, Nießner M (2016) Face2face: real-time face capture and reenactment of rgb videos. In: Proceedings of the IEEE conference on computer vision and pattern recognition
87. Tolosana R, Romero-Tapiador S, Fierrez J, Vera-Rodriguez R (2021) Deepfakes evolution: analysis of facial regions and fake detection performance. In: Proceedings of the international conference on pattern recognition workshops
88. Tolosana R, Vera-Rodriguez R, Fierrez J, Morales A, Ortega-Garcia J (2020) Deepfakes and beyond: a survey of face manipulation and fake detection. Inf Fusion 64:131–148
89. Tran D, Bourdev L, Fergus R, Torresani L, Paluri M (2015) Learning spatiotemporal features with 3D convolutional networks. In: Proceedings of the IEEE international conference on computer vision
90. Tran D, Wang H, Torresani L, Ray J, LeCun Y, Paluri M (2018) A closer look at spatiotemporal convolutions for action recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition
91. Trinh L, Tsang M, Rambhatla S, Liu Y (2021) Interpretable and trustworthy deepfake detection via dynamic prototypes. In: Proceedings of the IEEE winter conference on applications of computer vision
92. Wang L, Xiong Y, Wang Z, Qiao Y, Lin D, Tang X, Van Gool L (2016) Temporal segment networks: towards good practices for deep action recognition. In: Proceedings of the european conference on computer vision
93. Wang SY, Wang O, Zhang R, Owens A, Efros AA (2020) CNN-generated images are urprisingly easy to spot...for now. In: Proceedings of the IEEE conference on computer vision and pattern recognition
94. Wang TC, Liu MY, Zhu JY, Liu G, Tao A, Kautz J, Catanzaro B (2018) Video-to-video synthesis. arXiv:1808.06601
95. Wang, T.C., Liu, M.Y., Zhu, J.Y., Tao, A., Kautz, J., Catanzaro, B.: High-resolution image synthesis and semantic manipulation with conditional GANs. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2018)
96. Wu W, Qian C, Yang S, Wang Q, Cai Y, Zhou Q (2018) Look at boundary: a boundary-aware face alignment algorithm. In: Proceedings of the IEEE conference on computer vision and pattern recognition
97. Wu W, Zhang Y, Li C, Qian C, Loy CC(2018) ReenactGAN: learning to reenact faces via boundary transfer. In: Proceedings of the European conference on computer vision
98. Xie Q, Luong MT, Hovy E, Le QV (2020) Self-training with noisy student improves ImageNet classification. In: Proceedings of the IEEE conference on computer vision and pattern recognition
99. Yang X, Li Y, Lyu S (2019) Exposing deep fakes using inconsistent head poses. In: Proceedings of the IEEE international conference on acoustics, speech and signal processing
100. Yu N, Davis LS, Fritz M (2019) Attributing fake images to GANs: learning and analyzing GAN fingerprints. In: Proceedings of the IEEE international conference on computer vision
101. Zampoglou M, Papadopoulos S, Kompatsiaris Y (2015) Detecting image splicing in the wild (web). In: Proceedings of the IEEE international conference on multimedia & expo workshops
102. Zhang K, Zhang Z, Li Z, Qiao Y (2016) Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Process Lett 23:1499–1503
103. Zhou P, Han X, Morariu VI, Davis LS (2017) Two-stream neural networks for tampered face detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops
104. Zhu JY, Park T, Isola P, Efros AA (2017) Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision