OC-SAN: Unsupervised Deepfake Detection for Specific Individual Protection Based on Deep One-Class Classification

Yun Cao[1,2], Yanfei Tong[1,2], Han Bao[3(✉)], Xin Liao[4], and Meineng Zhu[5]

[1] Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100085, China
{caoyun,tongyanfei}@iie.ac.cn
[2] School of Cyber Security, University of Chinese Academy of Sciences, Beijing 100049, China
[3] Zhejiang University, Hangzhou 310058, China
baohan21@zju.edu.cn
[4] Hunan University, Changsha 410082, China
xinliao@hnu.edu.cn
[5] Beijing Institute of Electronics Technology and Application, Beijing 100091, China
zmneng@163.com

Abstract. In recent years, the misuse of Deepfake technology has become a critical security concern, especially for high-profile individuals such as politicians and celebrities. The key figure targeted forgeries are usually been carefully crafted with a sophisticated process, leaving behind no manipulation-specific artifacts. The inadequacy of current supervised binary classification methods for addressing these specific individual protection (SIP) challenges is evident for two primary reasons. Firstly, sufficient fake samples cannot be ensured to train generic and robust classifiers. Secondly, the efficacy of detection varies significantly across different identities. To address these problems, we formulate forgery detection as a one-class anomaly detection problem and propose a detection network, OC-SAN (One-Class Style Auto-encoder Network), to offer tailored protection for each individual. Borrowing from style transfer literature, we conceptualize facial features as a fusion of coarse-grained (identity) and fine-grained (personalized/style) elements. The fundamental concept driving OC-SAN is that a facial image can be well restored when its two features belong to the same identity. As a one-class method, OC-SAN can be trained only with authentic samples, indicating good generalization performance in real-world scenarios without relying on prior knowledge of forgery methods. Extensive experiments have been conducted to demonstrate the superiority of OC-SAN in SIP tasks, when compared to other state-of-the-art forgery detection methods.

Keywords: Deepfake detection · forgery detection · video forensics

Z. Lin et al. (Eds.): PRCV 2024, LNCS 15040, pp. 327–341, 2025.
https://doi.org/10.1007/978-981-97-8792-0_23

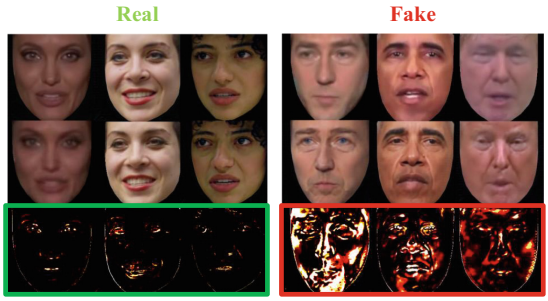


Fig. 1. Reconstruction effects of real and fake facial images. The top row shows images to test, the middle row shows the OC-SAN reconstructed images, and the bottom row visualizes the reconstruction errors to indicate forgery.

1 Introduction

Recent advances in deep learning have made it possible to create "Deepfake" videos that can swap faces in a way that is difficult to detect. These methods generally fall into two categories: facial expression manipulation and identity manipulation based on face-swapping techniques [1]. The forged videos often feature high-profile individuals such as politicians and celebrities, raising significant security concerns as they can be exploited for malicious purposes in public information campaigns. Recognizing the paramount importance of safeguarding individuals of high prominence, this paper directs its focus towards the specific individual protection (SIP) task. The objective is to enhance the safeguarding of key figures by offering tailored detection models.

Generally speaking, prevailing Deepfake detection methods formulate the detection task as a binary classification problem that uses a supervised model. However, these methods prove less adept when applied to the SIP tasks. Crafting convincing forgeries targeting key figures usually involves a blend of methods that rarely exhibit recognizable patterns. In contrast, the detection effects of these supervised learning methods heavily rely on the amount and richness of the training data representing both classes [1]. Consequently, these methods tend to suffer from over-fitting and their effectiveness is limited to the manipulation methods they are specifically trained for [2]. What's more, it has been reported that, the supervised detection models are sensitive to the identity information of the images [3]. So that it is difficult for these models to maintain effectiveness across diverse individuals.

To overcome forgery data scarcity limitations and enhance the detection accuracies, we formulate the SIP task as a one-class anomaly detection problem and propose a detection network called OC-SAN (One-Class Style Auto-encoder Network) to provide individualized protection. The primary challenge in achieving one-class detection for key figures lies in enabling the network to autonomously discern the distinctions introduced by Deepfake artifacts. Referring to the literature on style transfer, we decompose the facial features into coarse-grained

(identity) and fine-grained (personalized/style) elements and assume that certain manipulations may introduce inconsistencies between these features. In light of this conceptualization, OC-SAN is designed to facilitate unsupervised feature extraction at different levels and reconstruct the test sample to determine its authenticity based on the reconstruction error. The main contributions are summarized as follows:

- We emphasize that to perform SIP tasks, a good forgery detector should possess two key attributes: (1) generalizable to unseen forgery patterns with little or no forgery samples and (2) highly effective to a specific identity while disregarding other identities. In line with these requirements, we propose a novel unsupervised detection network named OC-SAN which can be trained only with real samples and offer personalized detection models for specific individuals.
- We conceptually categorize facial features into coarse-grained and fine-grained components. Utilizing an enhanced DSVDD objective, OC-SAN is trained to extract the two features from an input facial image and subsequently reconstruct it. As shown in Fig. 1, if the two features are incompatible with each other, the reconstruction error would expose the fact of forgery.
- Comprehensive comparative experiments have been conducted to assess OC-SAN's effectiveness. Experimental results demonstrate that our method outperforms the current supervised methods and unsupervised approaches and achieves state-of-the-art detection performance in the SIP tasks.

The rest of the paper is structured as follows: In Sect. 2, we discuss related work. In Sect. 3, we present our proposed OC-SAN in detail. In Sect. 4, comparative experiments are conducted to demonstrate the effectiveness of our approach. Finally, the conclusion is drawn in Sect. 5.

2 Related Work

Conventional binary-classification-based detection methods generally require a large amount of both real and fake face images for training, and easy to encounter over-fitting to manipulation-specific artifacts. Efforts have been made to address this challenge from diverse perspectives.

2.1 Unsupervised One-Class Detection Methods

In addition to binary classification modeling, an alternative approach is to formulate forgery detection as a one-class anomaly detection problem which uses an unsupervised model trained only on real samples to detect fake ones by treating them as anomalies [4]. Typical methods include Support Vector Machine (SVM)-based methods [5,6] and Deep Support Vector Data Description (DSVDD)-based methods [1,4,7]. Inspired by SVM, SVDD was proposed to solve the data domain description problem [8]. Since deep-learning uses neural networks and some nonlinear activation functions to fit a mathematically complex function [9], it has

been used to facilitate the mapping function in SVDD models, hence comes DSVDD. Broadly, there are two categories within DSVDD methods: one-class DSVDD [4] and reconstruction DSVDD [1,10,11]. The former employs the latent feature layer to construct a hypersphere, imposing mathematical constraints on the feature extraction network [4]. In contrast, the latter employs an Autoencoder (AE) structure to compute the reconstruction loss, emphasizing dissimilarities between characteristics of different samples during evaluation.

Nevertheless, it has turned out that existing one-class DSVDD methods fall short in addressing SIP tasks. SVM-based methods often suffer heavily from the problem of overfitting, and cannot deal with unseen patterns. DSVDD-based methods are usually good at extracting common facial features but personalized/style features which are much more subtle.

2.2 Supervised Detection Without Fake Images

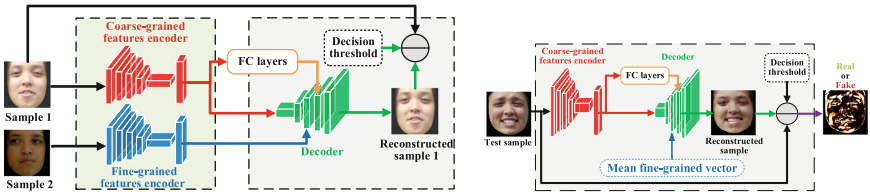
To overcome the limitations regarding fake data scarcity and dependency, one trend of the supervised methods is to manipulate real data to simulate the forgery effects. Notably, Face X-ray uses different real images to create blended facial images for training and is capable of showing the blending boundary for a forged image [2]. Inspired by this approach, ICT (Identity Consistency Transformer) creates forgeries by swapping the inner face of two real faces belonging to different identities, and detects a suspect face by finding identity inconsistency in inner and outer face regions [12]. The self-supervised Learning method [13] and SBI (Self-Blended Images) method [14] try to enrich the "diversity" of forgeries, encouraging the classifiers to learn generic and robust representations without over-fitting to manipulation specific artifacts. In this direction, the ID-unaware Deepfake Detection Model exploited the implicit identity leakage phenomenon, yielding a state-of-the-art detection effect [3].

Nevertheless, it has been highlighted that supervised detection methods, including SBI, may exhibit sensitivity to the identity information of the images [3], thereby impeding their overall generalization capabilities. Our experiments corroborate this observation, revealing that both SBI and the ID-unaware model exhibit variable detection efficacy across different individuals, which indicates such methods are not fully capable of the SIP tasks.

3 Method

In this section, we begin with our design motivation and an overview of the proposed method. Drawing inspiration from the literature on image style transfer, we conceptualize facial features as a fusion of coarse-grained (identity) and fine-grained (personalized/style) elements. The former can be construed as the facial "content", e.g., pose and face shape, while the latter can be construed as the facial "style", e.g., noise pattern, texture, and wrinkle. Importantly, for a specific individual, its coarse-grained features may exhibit variability across

different images, whereas its fine-grained feature is anticipated to remain consistent. The underlying assumption is that these two features remain compatible unless the image has undergone forgery.



(a) Training Phase (b) Testing Phase

Fig. 2. The architecture diagram of the proposed OC-SAN.

As illustrated in Fig. 2, OC-SAN acts fundamentally as an image restoration network. During the training phase (Fig. 2 (a)), two encoders receive distinct facial images as inputs, and their outputs are linked to different layers of the decoder. Drawing inspiration from the StyleGAN architecture [15], the coarse-grained features serve as the foundational input, forming the basis of a facial image, while the fine-grained features progressively contribute visual details layer by layer. Incorporating a one-class loss, this new architecture aims to autonomously learn and extract coarse-grained and fine-grained features in an unsupervised manner. During the testing phase (Fig. 2 (b)), OC-SAN extracts the coarse-grained feature from the input image and combines it with the pre-acquired fine-grained feature to reconstruct the facial image. The fundamental principle guiding OC-SAN lies in the observation that any incompatibility between these two features becomes apparent in the form of reconstruction errors. These errors are then utilized to determine the authenticity of the test sample.

For a specific individual A, OC-SAN uses only real samples to train the corresponding one-for-one protection model M_A which is characterized by a four-tuple $\{f_A^F, \Phi_A^C, \Phi_A^D, T_A\}$. Here f_A^F represents A's stable fine-grained features, Φ_A^C and Φ_A^D denote the coarse-grained features encoder and the decoder respectively, and T_A denote the classification threshold.

3.1 Training with Real Samples

To train M_A, a certain quantity of A's real facial images $X = [x_0, x_1, \ldots, x_{N-1}]$ should be collected in advance.[1] As depicted in Fig. 2 (a), in each iteration, two distinct images, x_i and x_j $(i \neq j)$, are randomly chosen and input into Φ_A^C and Φ_A^F for feature extraction.

In this paper, both encoders are implemented using the MnasNet. The coarse-grained feature $f_{x_i}^C$ is fed to Φ_A^D as the initial input while The fine-grained feature

[1] In order to reduce over-fitting for network, N should be at least 5,000.

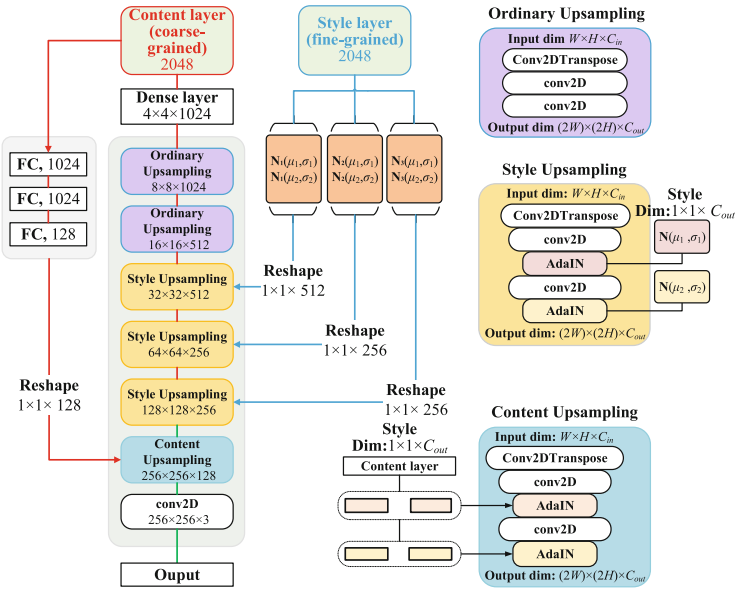


Fig. 3. The network structure of the decoder.

$f_{x_j}^F$ is used to control the image restoration process at "style" layers by adding details step by step. To be more specific, as depicted in Fig. 3, the decoder is a cascaded network of several up-sampling layers including two ordinary up-sampling blocks, three style up-sampling blocks, and one content up-sampling block. Similar to the structure of style transfer network, $f_{x_i}^C$ is used as the initial input for it represents the facial "content". Meanwhile, $f_{x_j}^F$ is parsed to three style up-sampling blocks since the network will focus on generating fine-grained features if the feature map sizes are larger than 16×16 [15]. To reduce the number of parameters, these three blocks incorporate two AdaIN modules each to receive the image "style" after up-sampling. Since the target style of AdaIN is determined by its variance and mean, we adopt an enhanced re-parameterization block based on VAE [16] to unwarp $f_{x_j}^F$ and fit the size of feature maps. So totally six re-parameterization blocks are used to map $f_{x_j}^F$ in six kinds of Gaussian distributions. Next, to speed up training convergence, $f_{x_i}^C$ is parsed to the content up-sampling block after passing a series of fully connected layers. This block is used to reduce the random influences, e.g., illumination and color conditions.

After training, the parameters of Φ_A^C and Φ_A^D can be set as fixed values, and the center value of $f_{x_j}^F$ is computed and assigned to f_A^F:

$$f_A^F = \frac{1}{n} \sum_{j=1}^{n} f_{x_j}^F, \tag{1}$$

where n is the iteration number. Next, the determination of T_A will be discussed.

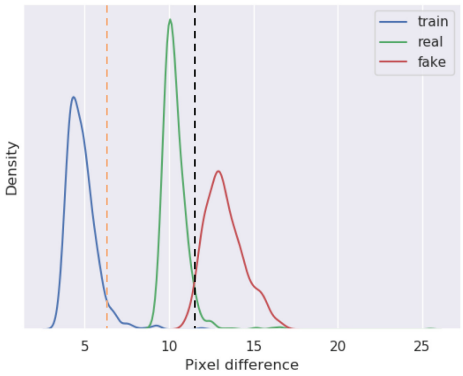


Fig. 4. Probability density diagrams of APD for the three scenarios.

3.2 Decision Threshold

In this paper, the reconstruction error is used as the anomaly score to distinguish real and fake images. Throughout the training phase, we compute the Average Pixel Differences (APDs) between the input and reconstructed facial images and depict the corresponding probability densities in Fig. 4. Through empirical analysis, we identify the 95% confidence interval (the orange dotted line) and set its **2-fold** value (the black dotted line) as the decision threshold T_A.

To authenticate an image y attributed to individual A, as illustrated in Fig. 2 (b), Φ_A^C is first used to extract f_y^C, then Φ_A^D reconstructs $\hat{y}$ with f_y^C and f_A^F. Finally "abnormal" data can be detected if corresponding APD surpasses T_A.

3.3 Loss Function

To ensure the effects of feature extraction and image restoration, four types of losses are considered in OC-SAN, i.e., the reconstruction loss $\mathcal{L}_r$, the perceptual loss $\mathcal{L}_p$, the one-class loss $\mathcal{L}_o$ and the Kullback-Leibler divergence loss $\mathcal{L}_{kl}$, and the loss function is formulated as:

$$\mathcal{L} = \alpha\mathcal{L}_r + \beta\mathcal{L}_p + \gamma\mathcal{L}_o + \rho\mathcal{L}_{kl}, \tag{2}$$

where α, β, γ and ρ are weights of losses set as hyper-parameters.

The Reconstruction Loss. In general image generation tasks, MSE is widely used as the criterion to control the visual quality. OC-SAN also adopts a per-pixel loss between the input and the restored images as the reconstruction loss:

$$\mathcal{L}_r = \frac{1}{C \times H \times W} \|\hat{x} - x\|_2^2, \tag{3}$$

where $C \times H \times W$ is the image size.

The Perceptual Loss. For the face generation tasks, only using MSE would cause image blurring. It is mainly because such per-pixel loss functions depend only on low-level pixel information [17]. So other than $\mathcal{L}_r$, OC-SAN also incorporates the perceptual loss $\mathcal{L}_p$ which measures image similarities more robustly than per-pixel losses. The perceptual loss uses a trained image classification model, e.g., VGG [18], to obtain the intermediate expression features of the generated image through the model, and constrains the high-frequency information to achieve style transformation. For more details, please refer to [17].

The One-Class Loss. OC-SAN adopts DSVDD to train the encoder Φ_A^F to converge the extracted fine-grained features. We have noticed that the original DSVDD objective function uses a fixed constant as the mean of the network representations [4]. Therefore, OC-SAN cannot adopt that objective function since our convergence objective, i.e., the fine-grained feature of the specific individual is unknown at the beginning.

Our solution is a modified DSVDD objective defined as:

$$\min_{\Phi^F} \frac{1}{n} \sum_{j=1,k=1,j\neq k}^{n} \left\| f_{x_j}^F - f_{x_k}^F \right\|^2 + \frac{\lambda}{2} \sum_{\ell=1}^{L} \left\| W^\ell \right\|^2, \tag{4}$$

where Φ^F is the feature extraction network with L layers. With this clustering-based training strategy, the extracted fine-grained features of training samples are expected to be attributed to the same category. It has the same effect as mapping sample features into one hypersphere as normal DSVDD. In this sense, the one-class loss is defined as:

$$\mathcal{L}_o = \left\| f_{x_j}^F - f_{x_k}^F \right\|^2 (j \neq k). \tag{5}$$

The Kullback-Leibler Divergence Loss. In a typical VAE, Kullback-Leibler divergence (KLD) is often used as the objective function to fit the approximated posterior of latent space:

$$\begin{aligned} L(\Phi^F, \Phi^D, x) =& D_{KL}\left(q_{\Phi^F}(z \mid x) \| p_{\Phi^D}(z)\right) \\ & - E_{q_{\Phi^F}(z|x)}\left(p_{\Phi^D}(x \mid z)\right), \end{aligned} \tag{6}$$

where x is the input distribution of f_A^F and z is the latent distribution. In our network, we need to get the posterior of f_A^F to fit each AdaIN module. Hence KLD is used to force f_A^F to approach a normal Gaussian distribution $N(0, I)$ in the latent space. Hence the KLD loss $\mathcal{L}_{kl}$ is defined as:

$$\mathcal{L}_{kl} = \sum_{i=1}^{n} D_{KL}[N(\mu_i(f_A^F), \sigma_i(f_A^F)), N(0, I)]. \tag{7}$$

where μ and σ are the mean and variance for each distribution determined by f_A^F. Similar to VAE, this loss makes the fine-grained features more robust when reconstructing images with different facial expression and accelerate convergence at the same time.

4 Experiments

4.1 Experimental Settings

Comparison Methods. In our experiments, three relevant unsupervised OC-based forgery detection methods, i.e., one-class DSVDD [4], soft-boundary DSVDD [4], OC-FakeDect-2 [1] and three state-of-the-art supervised methods with good generalization ability, i.e., ICT-Ref [12], SBI [14] and ID-unaware [3] are leveraged for performance comparison. Note that, all the chosen comparison methods use only real images to train their detection models, which are expected to be applicable to the SIP tasks. Particularly, it is noticed that the two DSVDD methods are based on cifar-10 [19] and MNIST [20] which cannot deal with facial images with the size as large as 256×256. In order to maintain their classification effects, MnasNet is used for feature extraction instead.

Dataset. We evaluate our proposed OC-SAN's intra-testing performance on the Faceforensic++ dataset [21], and conduct cross-testing by training on our collected internet videos and testing on Celeb-DF [22]. In practice, all the used video clips have to be pre-processed to get the masked facial images for training and testing. We first decode video clips into frames and extract facial images from each frame. Then the mask function in OpenFace[2] [23] is leveraged to remove the background around the face. Finally, all the masked facial images are resized to 256×256 to limit the transformation variables. What's more, to improve the robustness of the trained model, data augmentation methods [24] are used to enrich training samples, including horizontal flip, adding Gaussian noise, using Gaussian blur, adjusting image brightness, hue and contrast, etc. All the original samples and the enhanced samples are used for training.

Implementation Details. In our experiment, the loss function hyper-parameters in (2) are configured as 1:0.001:1:0.05. For $\mathcal{L}_p$, β^1 and β^2 are set to 1:0.15. The proposed network and training/testing scripts are implemented based on Keras-TensorFlow. The network is initialized by he-normal [25] for each layer. We adopt the Adam optimizer for all the loss which is widely used in deep learning tasks [26]. The learning rates of the reconstruction and DSVDD training are all set to 1e-5. Regarding the comparison methods, the number of training epochs is set to 40, which is twice as much as OC-SAN's. To assess the detection performance, precision, recall, F1, and accuracy scores serve as the evaluation indices. Both the reference implementation of the proposed OC-SAN and the used dataset are available on GitHub.

[2] An open-source toolkit possesses functions including facial landmark detection, head pose estimation, facial action unit recognition, etc.

4.2 Experimental Results

Intra-Testing. To verify the basic classification effectiveness, we compare OC-SAN with the other three OC-based methods and ICT-Ref. Both training and testing samples come from the Faceforensic++ dataset. We select 5 actors as the individuals to be protected and use their videos of quality c23 for training and testing. For real samples, 80% are used for training and the rest 20% for testing. Meanwhile, all fake ones are used as abnormal samples for testing.

The obtained precision, recall, and F1 scores are detailed in Table 1. The results highlight that the two DSVDD methods, i.e., one-class DSVDD and soft-boundary DSVDD, are almost incapable of verifying the authenticity while the others have certain detection abilities. This suggests that such DSVDD methods cannot be directly applied for the SIP tasks, since these methods are incapable of finding the convergence center as the fine-grained feature of a specific individual. Conversely, with a modified clustering-based training strategy, OC-SAN can map the extracted fine-grained features into one hypersphere and obtain the center vector with semantic information.

Table 1. Intra-testing in terms of precision (Pre.), recall (Rec.), and F1 scores.

Actor	O-C DSVDD			S-B DSVDD			OC-FakeDect-2			ICT-Ref			OC-SAN		
	Pre.	Rec.	F1	Pre.	Rec.	F1	Pre.	Rec.	F1	Pre.	Rec.	F1	Pre.	Rec.	F1
01	0.463	0.161	0.235	0.476	0.345	0.167	0.979	0.923	0.944	0.862	0.847	0.854	0.980	0.979	**0.980**
02	0.493	0.426	0.202	0.505	0.577	0.429	0.899	0.904	0.901	0.845	0.867	0.857	0.931	0.930	**0.930**
03	0.498	0.499	0.138	0.492	0.505	0.026	0.809	0.771	0.787	0.840	0.910	**0.889**	0.873	0.868	0.869
04	0.491	0.441	0.109	0.506	0.505	0.024	0.868	0.838	0.844	0.829	0.826	0.828	0.869	0.865	**0.863**
05	0.504	0.501	0.020	0.618	0.508	0.232	0.972	0.972	**0.970**	0.881	0.929	0.904	0.965	0.954	0.957

For a better illustration, Fig. 5 shows the reconstructed images compared with source images related to "Actor 01". As evident in Fig. 5 (a) and (b), the reconstructed images for training samples and testing real samples look-alike with small reconstruction errors. Conversely, for fake videos, the reconstructed images (Fig. 5 (c)) exhibit noticeable flaws, accompanied by more pronounced APD hot-maps. In OC-SAN, successful reconstruction mainly relies on the consistency between the extracted coarse-grained features and the pre-trained fine-grained features of the specific individual. When miss-match, collapse happens.

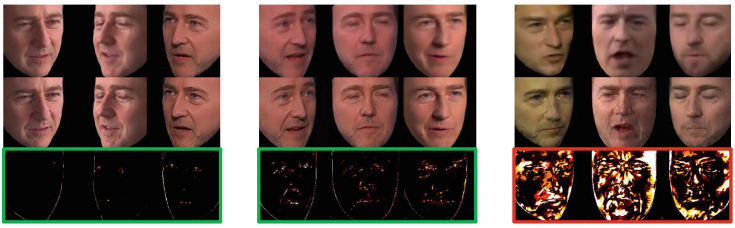
(a) Real (training) (b) Real (testing) (c) Fake (testing)

Fig. 5. Reconstruction effects in intra-testing (Actor 01). The first row shows the source images, and the second row shows the reconstructed images. The last row shows reconstruction errors.

Cross-Testing. To assess the generalization ability of our method on unseen forgery patterns, we conducted cross-dataset experiments, comparing OC-SAN with OC-FakeDect-2, ICT-Ref, SBI, and ID-unaware. Initially, two celebrities, E. Norton and A. Jolie, are chosen as the specific individuals to be protected. For OC-SAN, we collected relevant video clips through the internet and trained two one-for-one protection models M_{Norton} and M_{Jolie}. Their associated real samples and fake samples in Celeb-DF are then used as the normal and the abnormal samples for testing. Referring to Table 2, OC-FakeDect-2 exhibits significantly inferior detection performance compared to the other four methods. In the case of ICT-Ref, despite displaying comparable detection ability, the achieved accuracies rely on the selection of classification threshold which may vary with testing samples. Regarding SBI and ID-unaware, their detection abilities appear to vary considerably among different individuals. In comparison, OC-SAN demonstrates better stability, leveraging customized protection models capable of reproducing unique features during reconstruction. For instance, as illustrated in Fig. 6, E. Norton has specific styles of beard and eyebrow which are different from fake samples, and such detailed differences can be visualized as evidence of forgery.

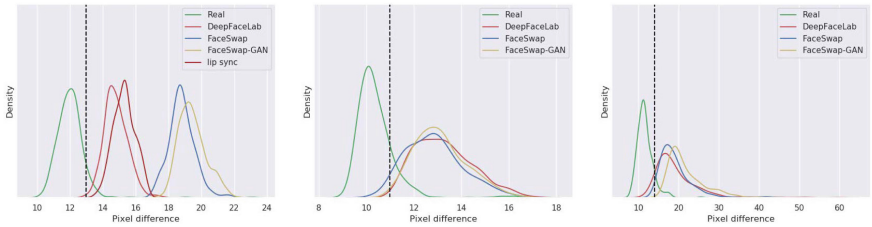
Table 2. Cross-testing in terms of precision (Pre.), recall (Rec.), and F1 scores by testing on Celeb-DF.

Celebrity	OC-FakeDect-2			ICT-Ref			SBI			ID-unaware			OC-SAN		
	Pre.	Rec.	F1	Pre.	Rec.	F1	Pre.	Rec.	F1	Pre.	Rec.	F1	Pre.	Rec.	F1
E.Norton	0.535	0.658	0.584	0.703	0.867	0.776	**0.990**	**0.887**	**0.935**	0.987	0.741	0.846	0.969	0.812	0.882
A.Jolie	0.522	0.539	0.524	0.772	**0.962**	**0.857**	0.719	0.866	0.786	**0.892**	0.641	0.746	0.808	0.818	0.810



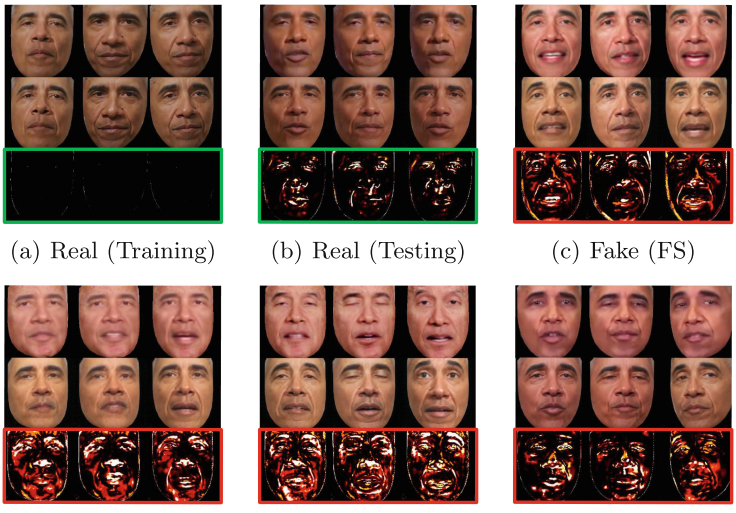
(a) Real (training) (b) Real (testing) (c) Fake (testing)

Fig. 6. Reconstruction effects in cross-testing by testing on Celeb-DF (E. Norton).



(a) B. Obama (b) D. Trump (c) V. Putin

Fig. 7. Probability density diagrams of APD for the three politicians.



(a) Real (Training) (b) Real (Testing) (c) Fake (FS)

(d) Fake (FS-G) (e) Fake (DFL) (f) Fake (LS)

Fig. 8. Reconstruction effects in cross-testing by testing on different forgery patterns (B. Obama).

We further conduct a more challenging cross-testing by comparing OC-SAN with SBI and ID-unaware. It's noteworthy that all these methods are trained using only real samples. In this test, three politicians, B. Obama, D. Trump, and V. Putin, are designated as the three specific individuals to be protected. All real samples, both for training and testing, are sourced from the Internet. To generate fake samples for testing, four commonly used generation methods, FaceSwap (FS), FaceSwap-GAN (FS-G), DeepFaceLab (DFL), and Lip-Sync (LS) [27] are utilized. For OC-SAN, we trained three one-for-one protection models M_{Obama}, M_{Trump}, and M_{Putin}, and the obtained decision thresholds are 13, 11, and 14 respectively. During the testing phase, the probability density distributions of APD between original and reconstructed samples are shown in Fig. 7, with the decision thresholds marked as black dotted lines. The detection accuracies are recorded in Table 3. Generally, OC-SAN shows a better performance for high-quality forgeries (FS-G and LS). Besides, OC-SAN exhibits a more consistent detection ability cross different individuals, even when confronted with unseen forgery patterns. What's more, Fig. 8 shows the APD hot-maps of fake samples are notably more conspicuous than those of real samples, regardless of the generation method employed.

Table 3. Cross-testing in terms of ACC. by testing on unseen patterns.

Politician	SBI					ID-unaware					OC-SAN				
	Real	FS	FS-G	DFL	LS	Real	FS	FS-G	DFL	LS	Real	FS	FS-G	DFL	LS
B. Obama	0.998	**1.0**	0.591	**1.0**	0.752	**0.999**	**1.0**	0.711	0.972	0.260	0.912	0.998	**0.999**	0.967	**0.991**
D. Trump	**0.907**	**1.0**	0.985	0.990	–	0.735	**1.0**	**0.997**	**0.998**	–	0.844	0.996	0.972	0.981	–
V. Putin	0.616	**0.999**	0.624	**0.999**	–	0.775	0.972	0.488	0.997	–	**0.918**	0.958	**0.998**	0.926	–

5 Conclusion

This paper introduces an innovative unsupervised detection method, OC-SAN, specifically designed for the SIP tasks. Following training, OC-SAN acquires the capability to extract coarse-grained facial features and successfully captures the stable fine-grained feature representing the distinctive "style" of the safeguarded individual. Given a test sample, OC-SAN extracts the coarse-grained feature, uses it with the pre-acquired fine-grained feature to reconstruct the input sample, and makes a judgment according to the degree of reconstruction error. As a one-class method, OC-SAN's detection performance does not rely on any prior knowledge of forgery methods, because it is trained only with authentic samples. Comparative experimental results have demonstrated that OC-SAN has a good generalization performance in dealing with unseen forgery patterns and is capable of offering one-for-one protections.

Acknowledgments. This work was supported by National Natural Science Foundation of China (Grant Nos. U22A2030, 61972142), National Key R&D Program of China (Grant No. 2022YFB3103500), Hunan Provincial Funds for Distinguished Young Scholars (Grant No. 2024JJ2025).

References

1. Khalid, H., Woo, S.S.: Oc-fakedect: classifying deepfakes using one-class variational autoencoder. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp. 656–657 (2020)
2. Li, L., Bao, J., Zhang, T., Yang, H., Chen, D., Wen, F., Guo, B.: Face x-ray for more general face forgery detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5001–5010 (2020)
3. Dong, S., Wang, J., Ji, R., Liang, J., Fan, H., Ge, Z.: Implicit identity leakage: the stumbling block to improving deepfake detection generalization. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3994–4004 (2023)
4. Ruff, L., Vandermeulen, R., Goernitz, N., Deecke, L., Siddiqui, S.A., Binder, A., Müller, E., Kloft, M.: Deep one-class classification. In: International Conference on Machine Learning, pp. 4393–4402 (2018)
5. Agarwal, S., Farid, H., Gu, Y., He, M., Nagano, K., Li, H.: Protecting world leaders against deep fakes. In: CVPR Workshops, pp. 38–45 (2019)
6. Vert, R., Vert, J.-P.: Consistency and convergence rates of one-class svms and related algorithms. J. Mach. Learn. Res. **7**(May), 817–854 (2006)
7. Oza, P., Patel, V.M.: One-class convolutional neural network. IEEE Signal Proc. Lett. **26**(2), 277–281 (2018)
8. Tax, D.M.J., Duin, R.P.W.: Support vector domain description. Pattern Recogn. Lett. **20**(11–13), 1191–1199 (1999)
9. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature **521**(7553), 436–444 (2015)
10. Akcay, S., Atapour-Abarghouei, A., Breckon, T.P.: Ganomaly: semi-supervised anomaly detection via adversarial training. In: Asian Conference on Computer Vision, pp. 622–637. Springer (2018)
11. Schlegl, T., Seeböck, P., Waldstein, S.M., Langs, G., Schmidt-Erfurth, U.: f-anogan: fast unsupervised anomaly detection with generative adversarial networks. Med. Image Anal. **54**, 30–44 (2019)
12. Dong, X., Bao, J., Chen, D., Zhang, T., Zhang, W., Yu, N., Chen, D., Wen, F., Guo, B.: Protecting celebrities from deepfake with identity consistency transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9468–9478 (2022)
13. Chen, L., Zhang, Y., Song, Y., Liu, L., Wang, J.: Towards good generalizations for deepfake detection, self-supervised learning of adversarial example (2022)
14. Shiohara, K., Yamasaki, T.: Detecting deepfakes with self-blended images. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 18699–18708 (2022)
15. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. IEEE Trans. Pattern Anal. Mach. Intell. **43**(12), 4217–4228 (2021)
16. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv:1312.6114 (2013)

17. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: European Conference on Computer Vision, pp. 694–711. Springer (2016)
18. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556 (2014)
19. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images (2009)
20. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proc. IEEE **86**(11), 2278–2324 (1998)
21. Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., Nießner, M.: Faceforensics++: learning to detect manipulated facial images. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1–11 (2019)
22. Li, Y., Sun, P., Qi, H., Lyu, S.: Celeb-DF: a large-scale challenging dataset for DeepFake forensics. In: IEEE Conference on Computer Vision and Patten Recognition (CVPR), Seattle, WA, United States (2020)
23. Amos, B., Ludwiczuk, B., Satyanarayanan, M., et al.: Openface: a general-purpose face recognition library with mobile applications. CMU School Comput. Sci. **6**(2) (2016)
24. Zhao, H., Cui, H., Zhou, W.: Ws-dan. https://github.com/cuihaoleo/kaggle-dfdc. Accessed 10 Sep 2020
25. He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: surpassing human-level performance on imagenet classification. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1026–1034 (2015)
26. Miyato, T., Kataoka, T., Koyama, M., Yoshida, Y.: Spectral normalization for generative adversarial networks. arXiv:1802.05957 (2018)
27. Suwajanakorn, S., Seitz, S.M., Kemelmacher-Shlizerman, I.: Synthesizing obama: learning lip sync from audio. ACM Trans. Graph. (TOG) **36**(4), 1–13 (2017)