A Diffusion-Based Approach for Restoring Face-swapped Images

Yuanchen Niu, Yuanman Li*, Guijia Zhang, and Xia Li
University of Shenzhen, Shenzhen
E-mail: yuanmanli@szu.edu.cn

***Abstract*—The misuse of deepfake technology underscores the critical importance of deepfake forensics. Current research in this domain primarily employs binary classification models that are limited to detecting the authenticity of facial images, and are incapable of tracing the original face replaced in the face-swapping process. Consequently, this limitation results in the inability to provide substantial judicial evidence.To address this issue, this paper introduces a novel forensic methodology, which aims to reconstruct the original face from the forged face. Specifically, we propose a novel framework that reformulates the restoration process from the forged face to the original face as a special image editing task guided by a series of conditions. Our research demonstrates that forged face images after face-swapping still retain some features of the original face, and these features can be used to restore the original face.Furthermore, we also construct the first dataset and set up evaluation metrics for this important research problem. Extensive experiments demonstrate the qualitative and quantitative effectiveness of our proposed framework.**

I. Introduction

Recent research primarily focuses on deepfake detection technologies to defense Face-swapping Deepfake [1]–[4]. These techniques identify manipulation artifacts across spatial, frequency, and temporal domains. Deep Neural Networks (DNNs) are then used to predict binary labels (Real or Fake) to discern authenticity [5]–[7]. However, such binary classifications do not yield sufficient persuasive judicial evidence. Moreover, the original identity of the target face, altered in creating fake news, remains obscured and the truth may never be revealed. This issue is particularly prevalent in conflicts such as Russia-Ukraine and Israel-Palestine. To overcome these challenges, we introduce a novel research problem, which seeks to restore the *target face* from the *result face* in the process of face-swapping, offering a more robust forensic solution for addressing the misuse of Face-swapping Deepfake technologies. It's noteworthy that the *result face* in this context refers to the face post-forgery, while the *target face* contributes its attributes and the *source face* provides its identity. We will explore the feasibility of Face-Swapped Image Restoration and introduce a powerful framework to tackle this issue.

We posit that face-swapping deepfake is traceable for several reasons. Firstly, although deepfake algorithms are capable of creating high-quality fake faces, they inevitably left traces on the result face. We believe that target-face-relevant information can be captured from these traces and used for the restoration. Secondly, it is difficult for these algorithms to effectively decouple identity and attribute information in facial images, leading to implicit tampering traces on the synthesized face. Such traces likely retain features associated with the target face, which are beneficial for restoration. Consequently, in restoration of face-swapped images, the critical step involves independently decoupling the explicit and implicit target-relevant features within the result facial image, and then extract useful information to facilitate the restoration of the target face. Motivated by these challenges, we reformulate the restoration process of the result facial image to the target facial image as a special image-to-image editing task.

In this paper, we propose a novel conditional diffusion-based framework named DRFSI (Diffusion Restoration of Face-Swapped Image). DRFSI is able to progressively decouple traceable information related to the target face throughout the denoising process for restoration. Specifically, due to the attribute information on the result face is highly relevant with the target face and should be preserved. We use the multi-scaled attribute information of the result face, compressed by information bottleneck modules, in facilitating the restoration of the target face via the AttributeNet we designed. The AttributeNet employs control mechanisms [8] and serves as a positive guidance to assist the denoising process. These enable our model to adaptively disentangle target-face-relevant information from the result face and accomplish restoration of the target face.

Our main contributions are summarized as follows:

- We introduce a novel research problem, which aims to restore the *target face* from the *result face* in face-swapping deepfake procedure. This approach offers a new forensic method for protecting facial information.
- We propose a novel diffusion-based framework DRFSI to address the challenge of face-swapped image restoration. By designing AttributeNet, we effectively instruct our diffusion model in disentangling target-relevant information, facilitating the restoration of the concealed target face.
- Extensive qualitative and quantitative experiments demonstrate the effectiveness of our method. Moreover, our method can be generalized to unknown face-swapping deepfake algorithms effectively.

II. Related Works

Face-swapping deepfakes employ deep learning techniques to execute the face-swapping task. Current methodologies in

*Corresponding author

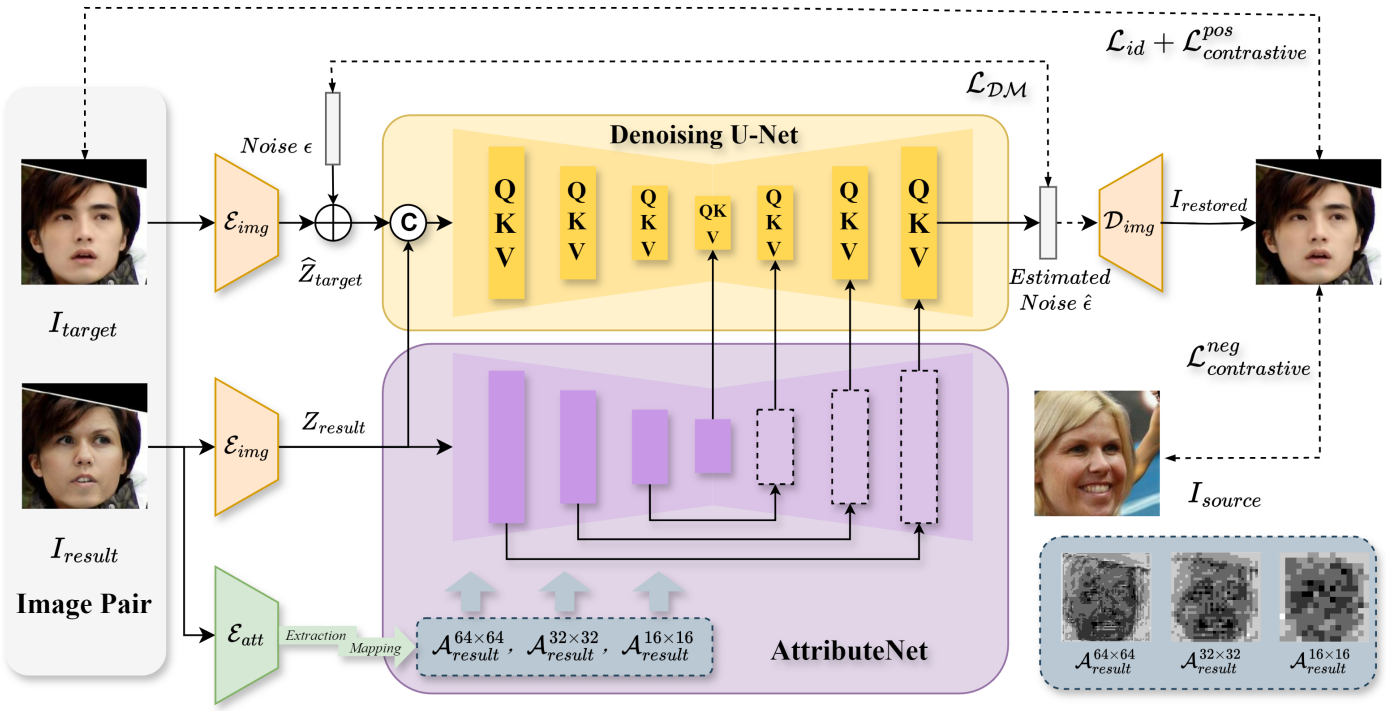

Fig. 1. **The overall pipeline of our proposed DRFSI**. For a given facial image pair comprising the target and result, we first encode target face into the latent space and generate a noisy version of latent target facial image. Subsequently, we employ a conditional diffusion model to denoise the noisy latent target facial image, using resulting facial image as conditional inputs of the AttributeNet.

face-swapping deepfakes can be categorized into two types: source identity-guided and target attribute-guided methods. Source identity-guided face swapping involves extracting identity features from the source face and integrating them into the target face. For instance, FaceShifter [9], SimSwap [10] and MobileFaceSwap [11] incorporate a source identity embedding module within the generator to preserve the source face's identity. As for target attribute-guided face swapping, it involves modifying the source face using attributes derived from the target face. This process initiates with the source face, which is then adjusted to align with the target face's attributes. For instance, FSGAN [12] applies the target face's features to direct the modifications of the source face, incorporating a hybrid network to seamlessly integrate the facial and background areas.

III. Proposed Method

A. Proposed Framework: DRFSI

As illustrated in Fig 1, DRFSI comprises two components, including Denoising Network and AttributeNet.

1) Denoising Network: DRFSI's backbone denoising network is based on the 64×64 conditional UNet model with temporal layers and inherits weights from the original SD UNet. As the input of the UNet, we follow the paradigm of diffusion-based image editing. First, we encode both the target and resulting facial images into latent space using a pre-trained autoencoder $\mathcal{E}_{img}$ to obtain Z_{result} and Z_{target}. Subsequently, a noisy version of the latent target facial image, $\widehat{Z}_{target}$, is generated as follows:

$$\begin{aligned} \widehat{Z}_{target} &= \sqrt{\overline{\alpha}_t} Z_{target} + \sqrt{1-\overline{\alpha}_t}\epsilon, \\ \overline{\alpha}_t &= \prod_{i=1}^{t}(1-\beta_i), \epsilon \sim \mathcal{N}(0, I), \end{aligned} \tag{1}$$

where β_i is derived from a predefined variance schedule β which governs the amount of noise injected at various time steps. Finally, we concatenate $\widehat{Z}_{target}$ with the latent resulting facial image Z_{result} and this concatenated feature map serves as the input for the denoising UNet.

2) Attribute Extraction: The objective of Face-swapping Deepfake is to transfer the identity of the source face onto the target face while preserving the attributes of the target face such as pose, expressions, and background. This means the result and target facial attribute information are highly relevant in Face-swapping Deepfake. Thus, we propose to utilizes the result-facial attribute information to aid in the restoration of target face.

The Information Bottleneck (IB) Principle has been effectively utilized to decouple the identity and attribute information in facial images [13]. We first average values of the same size feature maps in pre-trained IB model to obtain multi-scale attribute feature maps from result face:

$$\mathcal{A} = \left\{\mathcal{A}^{result}_{64\times64}, \mathcal{A}^{result}_{32\times32}, \mathcal{A}^{result}_{16\times16}\right\}. \tag{2}$$

Thus, when applying $\mathcal{A}$ to an encoded resulting facial image $\mathcal{Z}$, the multi-scaled facial attribute information $\mathcal{F}$ is highlighted

as follows:

$$\mathcal{F} = \mathcal{A} \odot \mathcal{Z}. \tag{3}$$

Subsequently, the highlighted attribute features $\mathcal{F}$ are integrated into the AttributeNet to facilitate the denoising restoration process via the control mechanism [8].

3) AttributeNet: As shown in Fig 1 ,the AttributeNet shares a similar structure with the encoder in the Denoising Network and serves to extract multi-scaled features from the result image. Unlike the backbone, however, AttributeNet does not employ prompt embeddings and time embeddings. Consequently, we have replaced the cross-attention layers with self-attention layers and removed the time-embedding layers. Subsequently, multi-scaled attribute feature maps $\mathcal{A}$ are extracted from the result face using a pre-trained Attribute Extractor[13]. Then, these maps are multiplied by the first feature map following each down-sampling operation within the encoder of AttributeNet. This method enables the extraction and retention of multi-scaled attribute information from the result face. Finally, the multi-scaled feature maps obtained from AttributeNet guide the denoising process in a manner analogous to that of ControlNet [8].

B. Loss Function

1) Diffusion Loss: The training objective of our diffusion-based model is similar to the conditional LDM.The Diffusion Loss can be defined as:

$$\mathcal{L}_D = E_{z_t,t,c,\epsilon}[\|\epsilon - \epsilon_\theta(z_t, t, c)\|_l^l], \tag{4}$$

where $l = 1, 2$, z_t represents the noisy version of latent target facial image, c represents the condition, ϵ represents the added noise. As the diffusion loss is actually equivalent to the reconstructive loss, we make $l = 1$ to improve visual restoration quality which is different from the common diffusion loss.

2) Identity Loss: In the restoration of face-swapped image, ensuring identity coherence between the restored and original target faces is crucial. Therefore, we use identity loss to constrain the distance in feature space between the restored target face $\hat{I}_{target}$ and the original target face I_{tar}. We compute the identity loss as:

$$\mathcal{L}_{id} = 1 - cos\left\langle \mathcal{E}_{id}(I_{target}), \mathcal{E}_{id}(\hat{I}_{target}) \right\rangle, \tag{5}$$

where $\mathcal{E}_{id}$ represents a pre-trained face recognition model, $cos\left\langle \mathcal{E}_{id}(I_{target}), \mathcal{E}_{id}(\hat{I}_{tar}) \right\rangle$ represents the cosine similarity between the original target identity and the restored target identity, $\mathcal{D}$ is the decoder of the pre-trained autoencoder.

3) Contrastive Identity Loss: We designe Contrastive Identity Loss (CIL) based on the contrastive learning principles. Here we want to ensure that the anchor sample $\mathcal{E}_{id}(\hat{I}_{target})$ is closer to the positive sample $\mathcal{E}_{id}(I_{target})$ while being distant from the negative sample $\mathcal{E}_{id}(I_{source})$. Then, we can formulate our loss function in the form of triplet loss [14] as follows:

$$\begin{aligned}\mathcal{L}_{CIL} =& 2 - \max\Big\{\big(\cos\langle \mathcal{E}_{id}(\hat{I}_{target}), \mathcal{E}_{id}(I_{target})\rangle \\ & - \cos\langle \mathcal{E}_{id}(\hat{I}_{target}), \mathcal{E}_{id}(I_{source})\rangle - \lambda\big), 0\Big\},\end{aligned} \tag{6}$$

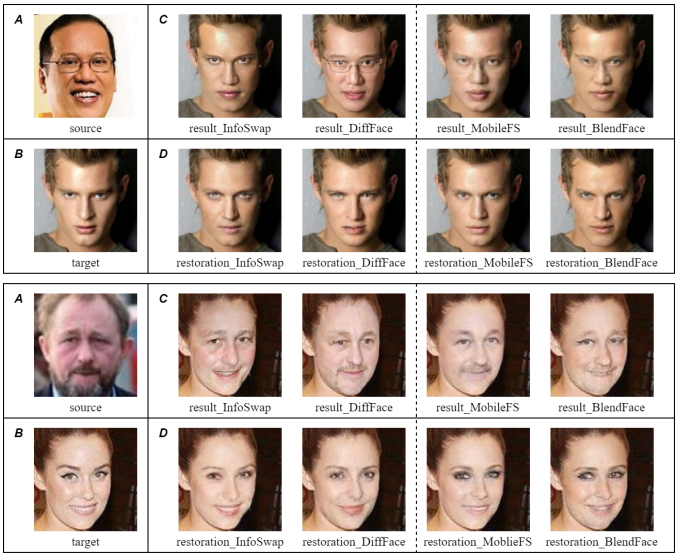

Fig. 2. **Qualitative results**. In the table above, Area ***A*** shows the source face. Area ***B*** shows the target face. Area ***C*** shows the result faces, generated by swapping the source face from ***A*** to the target face from ***B***. Correspondingly, Area ***D*** shows the target face restored by DRFSI. Specifically, the face-swapping algorithms on the left side of the dotted line were utilized for training, while those on the right side were only employed for test, indicating that our method generalizes well to unseen face-swapping deepfake algorithms.

where $\cos\langle \mathcal{E}_{id}(\hat{I}_{target}), \mathcal{E}_{id}(I_{target})\rangle$ is the cosine similarity of positive pair, and $\cos\langle \mathcal{E}_{id}(\hat{I}_{target}), \mathcal{E}_{id}(I_{source})\rangle$ is the cosine similarity of negative pair, λ is the enforced margin between positive and negative pairs.

In summary, the overall loss function for training DFDT can be written as:

$$\mathcal{L} = \mathcal{L}_D + \beta_1 \mathcal{L}_{id} + \beta_2 \mathcal{L}_{CIL}, \tag{7}$$

where β_1, β_2 are hyper parameters. We experimentally set them all to 10.

IV. Experiments

A. Experimental Settings

1) Implementation Details: In our experiments, the input images are uniformly resized to a size of 256×256 and then encoded into a $4 \times 64 \times 64$ latent space using a pre-trained autoencoder to improve the computational efficiency of the diffusion process. We train our diffusion model with batch size of 24 on 3 NVIDIA RTX4090 GPUs. We use Adam optimizer with the base learning rate of 1e-5 and the linear scaling rule.

2) Datasets: We constructed an extensive dataset based on the VGGFace2 [15], comprising 9,000 different identities, totally 100,000 source-target-result facial image pairs incorporating 4 diverse Face-swapping Deepfake algorithms (InfoSwap[13], DiffFace[16], BlendFace[17] and MobileFS[11]). In our experiment, we use two algorithms for train (InfoSwap and DiffFace), and two algorithms for generalizability test (BlendFace and MobileFS).

TABLE I
RETORATION ACCURACY (R-ACC / %) AND VERIFICATION ACCURACY (V-ACC / %) OF RESULT PAIR AND RESTORATION PAIR ON OUR TEST SET.

Methods	Result Pair		Restoration Pair	
	R-Acc	V-Acc	R-Acc	V-Acc
InfoSwap [13]	4.2	18.8	**92.8**	**84.9**
DiffFace [16]	7.2	28.8	**89.7**	**82.3**
MobileFS (test only) [11]	10.8	31.5	**88.0**	**81.4**
BlendFace (test only) [17]	7.6	35.3	**82.4**	**76.4**
Mean	7.5	28.6	**88.2**	**81.3**

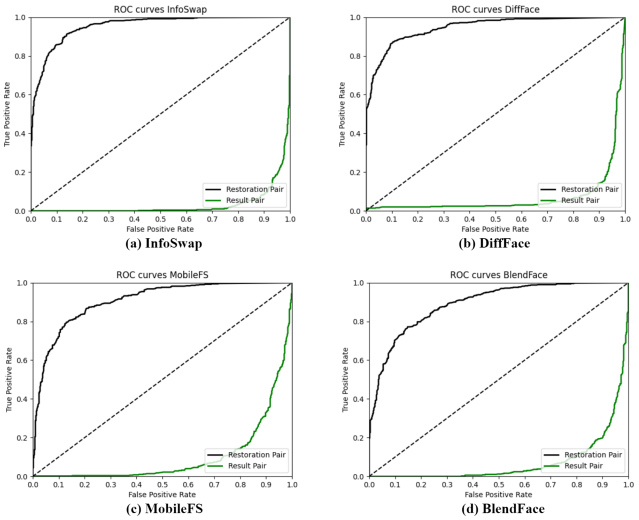

Fig. 3. The ROC curves of Result Pairs and Restoration Pairs on our test set.

B. Qualitative Results

In this section, we present the visual results of our DRFSI. As shown in Fig 2, we randomly select two cases from the test set, which two of them (left of the dotted line) used for train and another two (right of the dotted line) only used for *generalization test*. In the first case, we can find that the restored faces in area ***D*** have a certain traceability effect on the identity information of the target faces, such as eye, mouse and nose feature information, etc. At the same time, the attribute information of the target face, such as expression and demeanor, is retained. In addition, for the example where the result faces in area ***C*** have glasses (result_diffface, result_mobilefs), the target faces restored by our DTFT also correctly delete the glasses (recover_diffface, recover_mobilefs). In the second case, it's obvious that our traceability algorithm has effectively restored identity information such as the eye and mouth parts of the target face, as well as gender information.

C. Quantitative Results

1) Metrics: Given a source face, a target face and a restored face, the identity similarity between the restored face and the target face must exceed the similarity between the restored face and the source face. Then we can calculate the restoration accuracy (***R-Acc***) as follows:

$$Restoration\ Accuracy = \frac{pos}{N}, \tag{8}$$

where pos represents the number of samples that satisfy $sim\left\langle Id^{restore}, Id^{target}\right\rangle > sim\left\langle Id^{restore}, Id^{source}\right\rangle$, cosine similarity is used to evaluate the $sim\left\langle \cdot, \cdot \right\rangle$. N represents the total number of test samples.

Moreover, we evaluate our algorithm via *face verification*, which refers to assess whether two randomly selected facial images represent the same individual. For the given M sets of test facial image pairs, which include both positive and negative sample pairs, predicted labels are obtained by determining whether the identity similarity of each pair exceeds a specified threshold. Subsequently, face verification accuracy is calculated using the method for standard binary classification metrics.

2) Comparison to Result Pair: In this section, we quantitatively evaluate our method using the metrics introduced above. The comparison in Table I provides an intuitive demonstration of our DRFSI effectiveness. We define two sets of facial image pairs: the *Result Pair* consisting of source-target-result combinations, and the *Restoration Pair* consisting of source-target-restoration combinations. In Restoration Pair, the restoration and target images are considered positive sample pairs, whereas the result and source images are negative sample pairs. Conversely, in Result Pair, the result and target images are positive sample pairs, and the result and source images are negative sample pairs. Since the result and source images in the Result Pair share the same identity, similarly for the restoration and target images in the Restoration Pair, we expect much significantly lower Restoration Accuracy (R-Acc) and Verification Accuracy (V-Acc) in the Result Pair compared to the Restoration Pair. And as shown in Fig 3, the ROC curve for the Result Pair should approach the lower right corner, whereas the ROC curve for the Restoration Pair should align with the norm, approaching towards the upper left corner.

TABLE II
ABLATION EXPERIMENTS DEMONSTRATING THE EFFECT OF ATTRIBUTENET

Methods	Base		DRFSI	
	R-Acc	V-Acc	R-Acc	V-Acc
InfoSwap [13]	86.2	79.5	**92.8**	**84.9**
DiffFace [16]	86.3	78.8	**89.7**	**82.3**
MobileFS (test only) [11]	85.6	80.3	**88.0**	**81.4**
BlendFace (test only) [17]	81.2	75.0	**82.4**	**76.4**
Mean	84.8	78.4	**88.2**	**81.3**

D. Ablation Experiments

In this subsection, we conduct ablation experiments to evaluate the effectiveness of AttributeNet. We classify our framework into two configurations based on the inclusion or exclusion of the AttributeNet (AN): (1) **Base**, an unconditional diffusion model utilizing only a denoising network; (2) **DRFSI**, which combines AttributeNet with the Base model.

The experimental results presented in Table II demonstrate that the effectiveness of AN. Specifically, while an unconditional diffusion model can initially restore the target face, it falls in accurately tracing the target identity. AttributeNet leverages decoupled facial attribute maps to improve the restoration by ensuring the retention of critical target attributes through its effective spatial control. Integrating AN leads to an average improvement of 3.4% in R-Acc and 2.9% in V-Acc over Base.

V. Conclusions

In this paper, we study an important research problem, which aims to restore the original target face from a face-swapped deepfake image. Our research demonstrates that the synthetic face still retains traceable features of the original face, enabling its approximation and restoration. To address this problem, We have developed a powerful diffusion-based framework, named DRFSI. Extensive testing confirms the robust effectiveness and generalization capabilities of our framework. We hope that our efforts will inspire further research into the protection of facial privacy.

Acknowledgment

This work was supported in part by the Guangdong Basic and Applied Basic Research Foundation under Grant 2022A1515010645; in part by the Key project of Shenzhen Science and Technology Plan under Grant 20220810180617001; in part by the Open Research Project Programme of the State Key Laboratory of Internet of Things for Smart City (University of Macau) under Grant SKLIoTSC(UM)-2021-2023/ORP/GA04/2022.

References

[1] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "Faceforensics++: Learning to detect manipulated facial images," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 1–11.

[2] K. Shiohara and T. Yamasaki, "Detecting deepfakes with self-blended images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18 720–18 729.

[3] J. Li, H. Xie, J. Li, Z. Wang, and Y. Zhang, "Frequency-aware discriminative feature learning supervised by single-center loss for face forgery detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 6458–6467.

[4] Y. Zheng, J. Bao, D. Chen, M. Zeng, and F. Wen, "Exploring temporal coherence for more general video face forgery detection," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 15 044–15 054.

[5] Y. Li, L. Ye, H. Cao, W. Wang, and Z. Hua, "Cascaded adaptive graph representation learning for image copy-move forgery detection," *ACM Transactions on Multimedia Computing, Communications and Applications*,

[6] Y. Li, Y. He, C. Chen, *et al.*, "Image copy-move forgery detection via deep patchmatch and pairwise ranking learning," *IEEE Transactions on Image Processing*, pp. 1–16, 2024.

[7] Y. Tan, Y. Li, L. Zeng, J. Ye, W. Wang, and X. Li, "Multi-scale target-aware framework for constrained splicing detection and localization," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 8790–8798.

[8] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 3836–3847.

[9] L. Li, J. Bao, H. Yang, D. Chen, and F. Wen, "Advancing high fidelity identity swapping for forgery detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5074–5083.

[10] R. Chen, X. Chen, B. Ni, and Y. Ge, "Simswap: An efficient framework for high fidelity face swapping," in *Proceedings of the ACM International Conference on Multimedia*, 2020, pp. 2003–2011.

[11] Z. Xu, Z. Hong, C. Ding, *et al.*, "Mobilefaceswap: A lightweight framework for video face swapping," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, 2022, pp. 2973–2981.

[12] Y. Nirkin, Y. Keller, and T. Hassner, "Fsganv2: Improved subject agnostic face swapping and reenactment," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 1, pp. 560–575, 2022.

[13] G. Gao, H. Huang, C. Fu, Z. Li, and R. He, "Information bottleneck disentanglement for identity swapping," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 3404–3413.

[14] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 815–823.

[15] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "Vggface2: A dataset for recognising faces across pose and age," in *Proceedings of theIEEE international conference on automatic face & gesture recognition*, 2018, pp. 67–74.

[16] K. Kim, Y. Kim, S. Cho, *et al.*, "Diffface: Diffusion-based face swapping with facial guidance," *arXiv preprint arXiv:2212.13344*, 2022.

[17] K. Shiohara, X. Yang, and T. Taketomi, "Blendface: Re-designing identity encoders for face-swapping," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 7634–7644.