2024 18th International Conference on Automatic Face and Gesture Recognition (FG)

ClipSwap: Towards High Fidelity Face Swapping via Attributes and CLIP-Informed Loss

Phyo Thet Yee[1], Sudeepta Mishra[1] and Abhinav Dhall[2,1]
[1] Indian Institute of Technology Ropar, India
[2] Flinders University, Australia

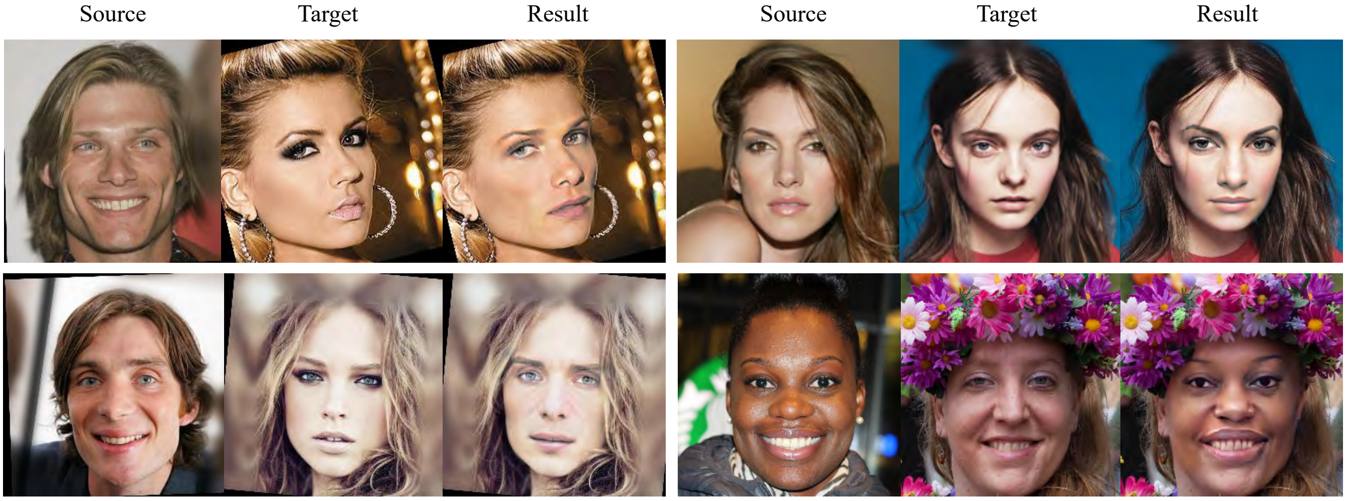

Fig. 1: Face swapped results of proposed ClipSwap method. Here the identity in the target is replaced with that of the source, while maintaining the style of the target face.

***Abstract*— This paper introduces ClipSwap, a new framework designed for high-fidelity face swapping. Earlier methods for face swapping often struggle in identity transfer due to the mismatches in attributes between the target and source images. To handle this issue, an attributes-aware face swapping approach is proposed in our work. We use a conditional Generative Adversarial Network and a CLIP-based encoder, which extracts rich semantic knowledge to achieve attributes-aware face swapping. Our framework uses CLIP embedding in the face swapping process for improving the transmission of source image's identity details to the swapped image by refining the high-level semantic attributes obtained from the source image. And source image serves as the input reference image for CLIP and ensures a more accurate and detailed identity representation in the final result. Additionally, we apply *Contrastive Loss* to guide the transformation of source facial attributes onto the swapped image from various viewpoints. We also introduce *Attributes Preservation Loss*, which penalizes the network to keep the facial attributes of the target image. Thorough quantitative and qualitative evaluations on multiple datasets illustrate the high-quality swapping results. Our proposed ClipSwap outperforms prior state-of-the-art (SOTA) methods in face swapping, particularly in terms of identity transfer and facial attribute features.**

I. INTRODUCTION

The primary goal of face swapping is to create new images which combine the unique characteristics of the source faces, including skin color and facial features, with the attributes of the target face, encompassing facial expressions, eye gaze, head poses, backgrounds, and more. This research problem has gained substantial interest in computer vision and image processing community due to numerous applications, including film industries, gaming visual, social media applications and for aiding privacy protection [19]. Recently, there have been notable advancements in face swapping techniques. However it remains challenging to accurately extract and merge identity information from source images and attributes from target images for generating a realistic high quality output [55].

There are two main types in face swapping techniques. The first one is source-oriented methods, which operate at the image level on the source face and the second one is the target-oriented methods, which function at the feature level on the target face. In source-oriented approaches [3], [4], [35], the process begins by transferring facial attributes such as pose and expression from the target to the source face, followed by blending the source face into the target face. Such methodologies are sometimes noisy, especially with extreme head poses and illumination, and they often struggle to generate the facial expression from the target. On the other hand, target-oriented methods [2], [24], [28], [20] work in the feature space and are more robust to facial attributes. These approaches rely on Generative Adversarial

979-8-3503-9494-8/24/$31.00 ©2024 IEEE

Networks (GANs). This helps in maintaining the attributes of the target image, including aspects like head pose and illumination, without the need for supplementary processing steps, such as learning perceptual and deep features during the training phase [39], [29], [8]. In this paper, we present a target-oriented approach called ClipSwap, which leverages on the semantic information through a CLIP network [40] for tackling challenges, including issues related to head pose, illumination, and semantic structure.

In this study, we use a conditional GAN in conjunction with the rich semantic knowledge embedded within the CLIP encoder. Our work provides four **main contributions**: a) we introduce a fresh perspective on the CLIP-based reference-guided face swapping. Our approach involves the transfer of source facial attributes to the target image whilst integrating semantic information that was taken from the reference image. The main objective here is to ensure the seamless preservation of the subject's identity throughout the transformation process. As far as we are aware, this is the first approach in face swapping that utilizes CLIP architecture. b) Our network uses an attribute preservation loss, which ensures the preservation of essential facial attributes during the face swapping process. This loss function guarantees that key characteristics such as facial expressions and distinctive features are maintained, thereby enhancing the overall quality of the results. c) We use a contrastive loss for optimizing predefined directions within the CLIP-space in order to control the editing process in desired directions, from various perceptual perspectives. This approach allows for more precise and controlled adjustments to the final result, ensuring high-quality outcomes. d) We use three datasets for our experiments: FaceForensic++ [44], CelebA-HQ [16] and FFHQ [12]. The outcomes indicate that our proposed ClipSwap outperforms prior SOTA methods in face swapping. Furthermore, it excels in preserving pose quality, surpassing most of the previous methods in this aspect.

II. RELATED WORKS

In recent times, notable advancements have emerged in the area of face swapping. We discuss the major relevant techniques below.

A. 3D Fitting Based Methods

Earlier face-swapping methods [49], [35] use 3D Morphable Models (3DMM) [5] and facial segmentation networks. For instance, Face2Face [49] and Nirkin et al. [35] involve the 3DMM fitting of both the source and target images, which allow the transfer of expression and pose parameters to generate swapped image. Nirkin et al. [35] gather data for training a supervised occlusion-aware face segmentation network. Methods such as the RSGAN [32], FSNet [33], and FSGAN [34] perform blending of segmented facial parts. Nevertheless, 3D-based models encounter challenges in achieving precise 3D reconstruction, often resulting in distortions and artifacts in the final swapped facial image.

B. GAN-Based Methods

Recent methods take into account end-to-end training for creating a face-swapped image based on learnt features. SimSwap [8], introduces the concept of weak feature matching, placing a stronger emphasis on preserving the source's facial expressions. Another method, MegaFS [59] is based on a pre-trained StyleGAN architecture. FaceShifter [29], on the other hand, adopts a strategy involving multi-level mixing, utilizing an encoder-decoder architecture to mitigate information loss, a challenge faced by the IP-GAN approach. In contrast, HifiFace [52] proposes a method that integrates a 3D shape model, prioritizing active shape modifications. Despite their ability to create realistic face swaps, most GAN-based methods often have trouble retaining the identity details of the source face and the structural information of the target face.

C. Contrastive Language–Image Pretraining (CLIP)

The CLIP-based embedding has become an efficient tool in different image manipulation and generation tasks such as image synthesis and content transfer. StyleMC [23] proposes an efficient and fast technique for text-driven image generation and manipulation. It combines the capabilities of CLIP and StyleGAN2, utilizing CLIP-based and identity losses to modify images based on a single text input while keeping other attributes. In contrast, [7] proposes a method for essence transfer that seamlessly transfers semantic features from a target image to a source image using StyleGAN for image generation and employs CLIP for image recognition. CF-CLIP [57] presents a method for precise text-guided image editing using CLIP. A new loss function, CLIP-based Noise Contrastive Estimation (CLIP-NCE) loss is also introduced in their work, to utilize the semantic knowledge of CLIP. HairCLIP [53] presents a new model for hair editing, enabling the manipulation of hair attributes individually or in combination, guided by text descriptions or reference images. CRFAST [27] presents a method that enables the transfer of meaningful information from a reference image to a source image. In their work, they introduce a new contrastive loss designed to thoroughly employ CLIP's rich semantic knowledge for facial features.

While CLIP-based methods have not been directly applied to face swapping in prior research, their demonstrated success in various related domains suggests untapped potential that serves as the foundation for the novel approach presented in our work. Contrary to earlier approaches, our proposed ClipSwap model integrates a GAN-based approach with the CLIP model which enables the accurate and seamless transfer of source identity information to the swapped faces while maintaining the attributes and expressions from the target faces. During training, attributes preservation loss function constraints the network to preserve target image's attributes.

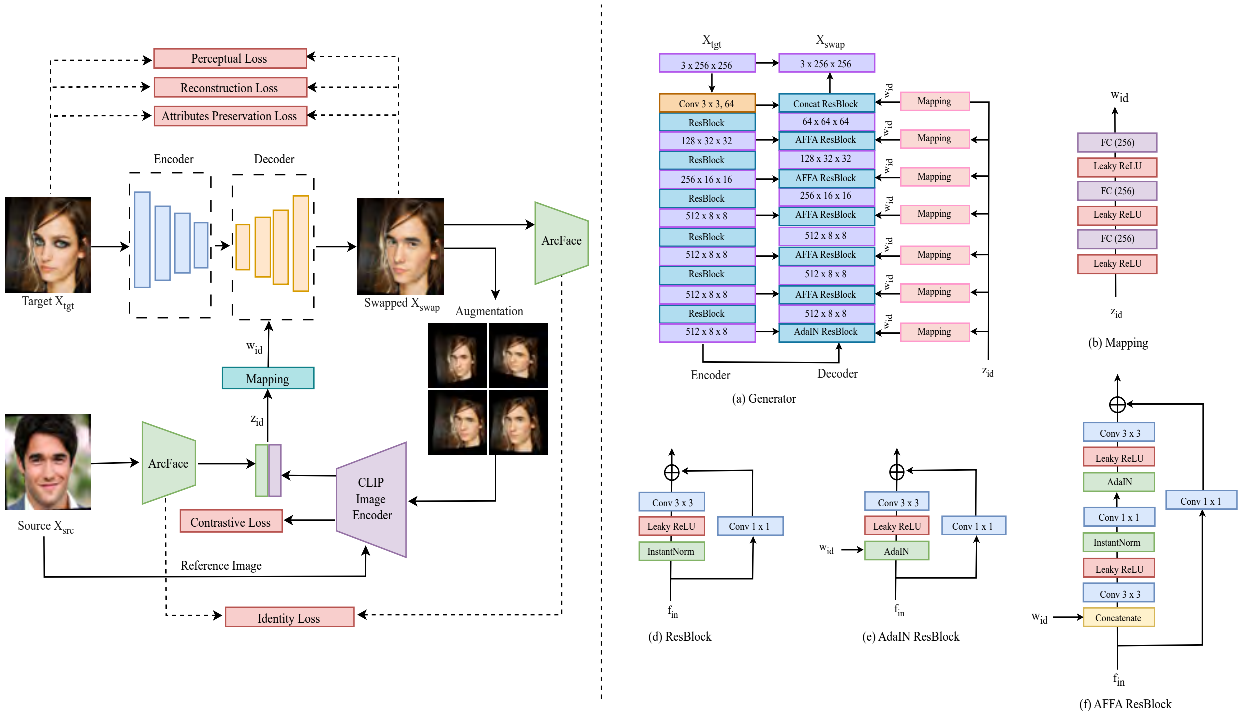

Fig. 2: Architecture of ClipSwap is built upon U-Net encoder-decoder in conjunction with CLIP image encoder and mapping network. We use pretrained ArcFace [10] identity encoder, to extract identity information. z_{id} represents the concatenated identity vectors from ArcFace [10] and CLIP image encoder [41], while w_{id} denotes the mapped identity vector. We apply perspective augmentation to the swapped image and encode it in the CLIP space to compute the CLIP-based contrastive loss. Detailed structures of the essential building blocks within our proposed ClipSwap model are shown on the right side. f_{in} is the input feature map. Each block, including ResBlock, AdaIN [14], AFFA [42], and Mapping, plays an significant role in the overall architecture.

III. METHOD

This section depicts the architecture of ClipSwap, illustrated in Fig. 2, alongside the CLIP image encoder, and the loss functions employed for model training.

A. *Network Architecture*

Our objective is to create a high fidelity swapped face X_{swap} which retains the identity of the source face, X_{src}, while also capturing the attributes of the target face, X_{tgt} (e.g, illumination, head pose, facial expression, and background components). To achieve this goal, our network is based on a conditional GAN with a generator, discriminator, and a mapping network, along with ArcFace and a CLIP image encoder, as depicted in the proposed architecture.

Identity Encoder: We use a ResNet50 backbone-equipped, pre-trained identity encoder ArcFace [10] for extracting the identity information from a source image X_{src}. An identity vector with a size of 512 is produced by ArcFace and used as an input to the model.

Attributes Encoder: Maintaining the specifics of facial attributes including head pose, facial expression, background, and illumination, demands a richer spatial representation compared to identity preservation. We input the X_{tgt} into a network similar to U-Net in order to maintain this information and define the attribute embedding through multi-level feature maps.

Generator: Inspired by the mechanism of FaceDancer [42], our generator follows an encoder-decoder architecture resembling U-Net. The encoder uses a series of blocks that gradually capture more complex details by increasing the number of filters. Additionally, the decoder uses blocks, each employing techniques such as concatenation layer, or Adaptive Instance Normalization (AdaIN) [14], [8], or an Adaptive Feature Fusion Attention (AFFA) [42] module, to utilize skip connections, as illustrated on the right side in Fig. 2. Specific details of each block are also illustrated.

Mapping Network: As demonstrated in prior works such as [18], [17], we use a mapping network, referred to as M, to augment the capabilities of G to convert the original identity distribution to a new distribution. In the mapping network, there are four fully-connected layers (FC), and all layers with the exception of the last one, employ leaky ReLU which serves as the activation function (Fig. 2).

CLIP Image Encoder: We use the CLIP model for transferring source facial attributes to the target image and integrating high-level semantic attributes from the source image. By utilizing the capabilities of this large-scale pre-trained CLIP model, we enable guided manipulation based on reference image. This methodology encourages diversity in image generation and safeguards against the emergence of unrealistic results by adjusting the direction of CLIP-space between the pair of reference image and result image.

B. *CLIP Contrastive Learning*

CF-CLIP [57] and CRFAST [27] adopt CLIP-NCE and contrastive learning for text or image-guided manipulation tasks. Inspired from their works, we also use the CLIP-based contrastive learning in our face swapping process. This aims to enhance the transfer of source's identity and other related facial attributes to the swapped image while accommodating various perspectives by optimizing the CLIP contrastive loss. Following CLIPStyler [25], swapped image is randomly augmented before computing the CLIP contrastive loss. Augmenting the swapped image before feeding it into the CLIP image encoder introduces diversity in the image. This diversity helps in capturing different variations, such as facial expressions, poses, lighting conditions, and other attributes present in the swapped image. As a result, the model learns to understand and encode different variations and features more effectively in examining the semantic information in the CLIP framework. So, this strategy greatly assists our network in moving closer to its ultimate goal. The augmentation process can be formulated as follows:

$$X_{aug} = augmentation(X_{swap}) \tag{1}$$

The contrastive learning approach we introduce seeks to improve the shared information between similar and dissimilar pairs by computing the contrastive loss, enabling a comprehensive investigation of the semantic information contained within the CLIP space. Similar to established contrastive loss approaches [38], [54], [46], it is necessary to generate query, similar, and dissimilar samples. As depicted in Fig. 3, the contrastive loss brings similar samples S^+ closer to the query Q, while moving dissimilar samples S^- more distant from it. And the query Q is defined as follows:

$$Q = CLIP(X_{aug}) - CLIP(X_{swap}), \tag{2}$$

Q denotes the "semantic direction" originating from X_{swap} to X_{aug}. Then, we establish similar samples based on the following two criteria:

$$\begin{aligned} S_1^+ &= CLIP(X_{ref}) - CLIP(X_{swap}) \\ S_2^+ &= CLIP(X_{ref}) - CLIP(X_{mean}) \end{aligned} \tag{3}$$

where S_1^+ serves to facilitate the alignment of facial identity features, directing them from the X_{ref} towards the X_{swap}. On the other hand, S_2^+ represents the direction of features extending from the X_{ref} to the X_{mean}. This mean image captures the combined, averaged semantic attributes of all faces that have been generated. S_2^+ plays a role in ensuring that the desired direction aligns with identity features of the reference image X_{ref} .

To provide comprehensive guidance and optimize CLIP space utilization, we define dissimilar samples as S^-. These samples represent directions within the CLIP embedding space that go from the swapped image to the mean image. This prevents the model from generating images with identity features differing from those observed in the reference image, such as the mean image. As a result, our design of dissimilar samples S^- ensures that the model avoids CLIP embeddings associated with face images displaying identity features distinct from those in the reference image.

$$S^- = CLIP(X_{swap}) - CLIP(X_{mean}) \tag{4}$$

The CLIP contrastive loss aims to enhance CLIP's editing capability by increasing the mutual information between selected similar pairs while decreasing it between dissimilar pairs in the CLIP's feature space. This approach provides comprehensive guidance, enabling desired adjustments from different perspectives. Specifically, the CLIP contrastive loss we present can be expressed as follows:

$$\begin{aligned} L_{contra} = &-log \frac{e^{(Q.S_1^+/\tau)}}{e^{(Q.S_1^+/\tau)} + \sum_{S^-} e^{(Q.S^-/\tau)}} \\ &-log \frac{e^{(Q.S_2^+/\tau)}}{e^{(Q.S_2^+/\tau)} + \sum_{S^-} e^{(Q.S^-/\tau)}}, \end{aligned} \tag{5}$$

where the temperature τ is set to 0.1 in this work.

C. *Loss Functions*

Alongside the CLIP contrastive loss, we incorporate other critical loss components: *identity loss*, *reconstruction loss*, *attributes preservation loss* and *perceptual loss*. To gain a thorough understanding of the influence of these loss functions on inputs and outputs, please refer to Fig. 2.

Identity Loss: The identity loss takes a pivotal role in instructing the network to retain the source face's unique identity attributes in the swapped image, ensuring a faithful transfer of identity. To ensure the preservation of the identity details of the source image during the process of swapping, we utilize the identity loss, which is calculated as follows:

$$L_{id} = 1 - \cos\left((R(X_{src}), R(X_{swap}))\right) \tag{6}$$

where R($\cdot$) represents the ArcFace [10] network and cos(.) stands for the cosine similarity.

Reconstruction Loss: The inclusion of the reconstruction loss serves as a precise objective: when the target image X_{tgt} and the source image X_{src} have the same identity, the reconstruction loss is computed to ensure that X_{swap}, the generated swapped image, is equivalent to the target image. The following is the definition of the reconstruction loss:

$$L_{rec} = \begin{cases} \|X_{tgt} - X_{swap}\| & \text{if } X_{tgt} = X_{src} \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

Perceptual Loss: To further improve both the reconstruction performance and the image's semantic understanding,

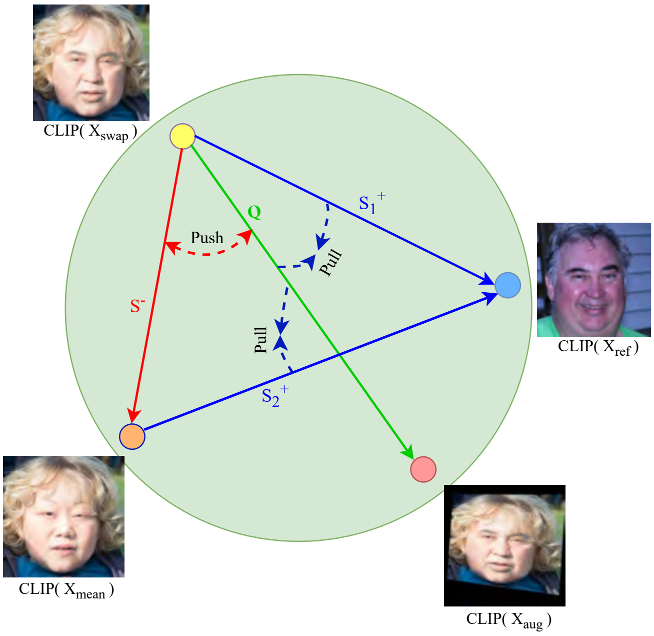

Fig. 3: Visualizing CLIP Contrastive Learning: Within the CLIP embedding space, the query's direction, leading from the swapped image X_{swap} to the augmented image X_{aug} (indicated by the green arrow), is a critical focus. Here, we make two types of similar samples, (represented by the blue arrow), S_1^+ guides the query to adjust its alignment towards the direction from X_{swap} to X_{ref}, while S_2^+ serves as a control mechanism to prevent excessive deviation. The red arrow highlights dissimilar samples originating from the swapped image to the reference image. Through the process of attracting similar pairs and pushing apart dissimilar pairs, the CLIP contrastive loss facilitates an extensive exploration of CLIP representations.

we incorporate the perceptual loss. This is motivated by the robustness of deep features in various reconstruction tasks, as evidenced by the previous work [51], [15]. The perceptual loss can be described as follows:

$$L_p = \begin{cases} \sum_{i=0}^{n} \| P^{(i)}(X_{tgt}) - P^{(i)}(X_{swap}) \| & \text{if } X_{tgt} = X_{src} \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

Attributes Preservation Loss: We introduce the attributes preservation loss, calculated as the L2 distance between the target image and the swapped image, to enhance the preservation of key attributes such as hair, hat, eyeglasses, ear, earring, etc. This approach is designed to utilize the strengths of two distinct pretrained models: a FaceNet [11] model for precise facial coordinate detection and an MXNet [31] facial attribute extraction model for comprehensive attribute analysis. By minimizing the L2 distance between the target and swapped images, we try to ensure the preservation of essential attributes. The rationale behind the attributes preservation loss function is to make the network learn how to maintain the target face's attributes in the output. This loss can be defined as

$$L_{att} = \sum_{i=1}^{N} \| A(X_{tgt}) - A(X_{swap}) \|_2^2 \tag{9}$$

Here, $A(X_{tgt})$ and $A(X_{swap})$ represent the attribute vectors extracted from the target and swapped images, respectively, for N attributes.

During the training of our ClipSwap model, a weighted sum of the losses mentioned earlier is employed, with the weights assigned as follows:

$$L = \lambda_1 L_{id} + \lambda_2 L_{rec} + \lambda_3 L_p + \lambda_4 L_{contra} + \lambda_5 L_{att}, \tag{10}$$

where λ_1, λ_2, λ_3, λ_4 and λ_5 are weights for the loss terms.

IV. EXPERIMENTS

A. *Datasets*

We train our ClipSwap network on the VGGFace2 dataset [6], which includes 3.31 million pictures of 9131 participants. Then we evaluate the network on the following three datasets:

FaceForensics++ [44] contains 1000 original conversational videos retrieved from YouTube and manipulated with five distinct methods. Following previous works [29], [52], [8], we take 10 frames from every video to produce a test dataset that contains 10K facial images in total.

CelebA-HQ [16] consists of 30K celebrity faces, each with a resolution of 1024 × 1024. We select 10K images for testing.

FFHQ [12] contains 70K face images obtained from Flickr and contains different kinds of gender, age, ethnicity, and background. We use 10K images for testing.

B. *Implementation Details*

In our experiments, we resize the images to 256 × 256, covering both the entire face and some background regions. We align all faces using five-point landmarks extracted with RetinaFace [9] to match ArcFace's input requirements. ArcFace [10] is pretrained on MS-Celeb-1M [13] using a ResNet50 backbone. The loss weights are configured as $\lambda_1 = 10$, $\lambda_2 = 5$, $\lambda_3 = 0.2$, $\lambda_4 = 0.3$ and $\lambda_5 = 10$. We train our model employing the Adam optimizer [22] with the parameters of $\beta_1 = 0$, $\beta_2 = 0.999$ and learning rate of 10^{-4}. The training process continues for 300,000 steps with a batch size of 10.

Evaluation Metrics: We employ various evaluation metrics, including identity retrieval (ID), pose error, expression error, structural similarity index method (SSIM), peak signal-to-noise ratio (PSNR), multiply accumulate operations (MACs) and number of parameters (Param.). We use the CosFace [50] encoder to evaluate the identity. In order to optimize the computing cost for some tests, we follow [59], [56] and measure the ID similarity. This similarity metric

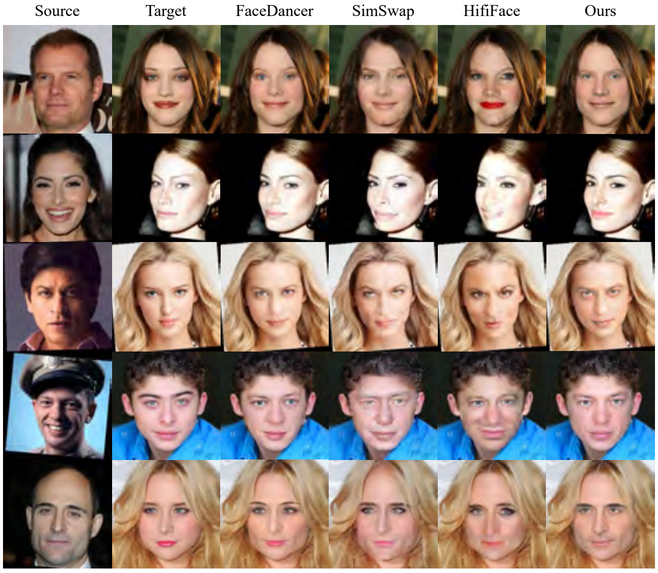

Fig. 4: Qualitative comparisons of ClipSwap (Ours) with FaceDancer [42], SimSwap [8] and HifiFace [52] on CelebA-HQ [16]. Our method achieves high-fidelity results while better preserving the source identity (e.g., eyebrow, eye, lip shape and color).

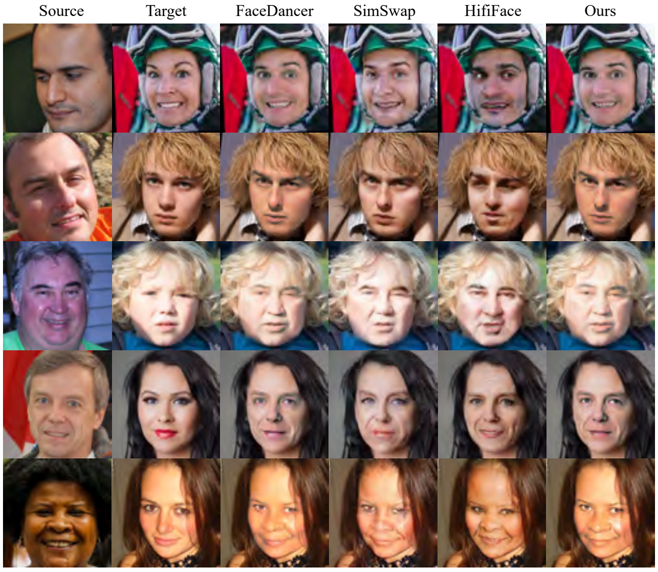

Fig. 5: Qualitative comparisons of ClipSwap (Ours) with FaceDancer [42], SimSwap [8] and HifiFace [52] on FFHQ [12]. Our method achieves high-fidelity results while better preserving the source identity.

is determined by computing the cosine similarity between swapped and their respective source faces [50]. Expression is measured by the average L_2 distance between the facial landmarks of target face and the output, as described in [19], [34]. We use dlib library [21] for the detection of facial landmarks. The pose estimator in [45] is used to compare poses, and the average L_2 error is reported.

C. Results

Qualitative Results: We conduct qualitative comparisons on both the CelebA-HQ and FFHQ datasets, wherein we evaluate our method against SOTA techniques such as FaceDancer, SimSwap, and HifiFace. We present the results in Fig. 4 and Fig. 5. It is evident that our approach yields swapped results that are of better quality. This can be attributed to the use of attribute preservation loss and use of CLIP embedding. The swapped faces in SimSwap and HifiFace exhibit artifacts and distortions (See rows 1-4 in Fig. 4 and row 1, 2 and 5 in Fig. 5). Lip color differences in HifiFace are noticeable (See the first row in Fig. 4). ClipSwap maintains pixelation artifacts, whereas SimSwap and HifiFace often lead to the creation of a smooth face or, in some cases, outright failure. While FaceDancer can generate visually aesthetic results, it frequently retains the identity details from the target, resulting in swapped outcomes that closely resemble the target, as seen in all sample images in Fig. 4 (Notice the eyebrows, eyes, lip shape and color). Additionally, the faces generated by FaceDancer often exhibit a blurred appearance (row 1, 2, 4 in Fig. 5). In contrast, due to CLIP-based reference-guided face swapping, high-level semantic facial attributes are learnt from the source image and generate in the swapped image with the help of reference image. By successfully incorporating the attributes of the target image while preserving the identities of the source image, our method outperforms in this regard. It is evident that our method retains the source image's identity features in the output.

Quantitative Results: We perform the experiment using the FaceForensics++ dataset [44], and the results are compared to other SOTA methods, such as SimSwap [8], FaceDancer [42], and HifiFace [52]. We create a test dataset of 10K frames by selecting 10 frames randomly from every video in the FaceForensic++ dataset, similar to earlier works [8], [42], [29] and [52]. The results are depicted in Table I. Our method, ClipSwap, evidently outperforms SOTA methods based on ID retrieval, pose, and SSIM metrics. Regarding expression and PSNR, we achieve comparable results to SimSwap and FaceDancer methods. Although our model uses more parameters than SimSwap and HifiFace, we perform the swapping process with lower computational costs, as measured by MACs and number of parameters.

To compare swapped results in high-resolution images, we randomly sample 10K pairs of images from both the CelebA-HQ and the FFHQ datasets to create the test datasets. We evaluate various metrics such as ID similarity, expression errors, pose errors, SSIM and PSNR. As presented in Tables II and III, our method consistently outperforms other methods across all evaluation metrics, with the exception of pose error on the CelebA-HQ test dataset, where ClipSwap achieves the second lowest pose error (2.83), following FaceDancer. This

Method	ID↑ Ret.	Pose↓	Exp↓	SSIM↑	PSNR↑	MAC↓ (G)	Param.↓ (M)
FaceDancer	98.84	2.04	7.97	0.97	**33.34**	NA	NA
SimSwap	92.83	1.94	**2.39**	0.81	24.38	55.69	**107.24**
HifiFace	98.48	2.63	NA	0.89	24.55	102.39	146.8
Ours	**98.91**	**1.73**	5.76	**0.98**	33.04	**53.56**	174.23

TABLE I: Results of a quantitative comparison on FaceForensics++ [44]. The best result is shown in bold. Upward arrow signifies that higher values correspond to improved performance, while lower values indicate the opposite.

Method	ID Sim. ↑	Pose ↓	Exp ↓	SSIM ↑	PSNR ↑
FaceDancer	0.52	**2.72**	26.07	0.92	27.89
SimSwap	0.31	3.34	30.28	0.71	23.38
HifiFace	0.29	3.69	40.63	0.85	23.65
Ours	**0.55**	2.83	**25.52**	**0.94**	**28.90**

TABLE II: Results of a quantitative comparison on CelebA-HQ [16]. The best result is shown in bold. Upward arrow signifies that higher values correspond to improved performance, while lower values indicate the opposite.

Method	ID Sim. ↑	Pose ↓	Exp ↓	SSIM ↑	PSNR ↑
FaceDancer	0.51	2.95	31.57	0.90	26.17
SimSwap	0.38	4.29	33.03	0.69	23.04
HifiFace	0.36	4.87	50.89	0.77	22.27
Ours	**0.53**	**2.84**	**30.77**	**0.91**	**26.50**

TABLE III: Results of a quantitative comparison on FFHQ [12]. The best result is shown in bold. The upward arrow signifies that higher values correspond to improved performance, while lower values indicate the opposite.

observation is supported by visual comparisons shown in Fig. 4 and Fig. 5.

D. Ablation Study

We perform both qualitative and quantitative ablation studies on CelebA-HQ dataset.

W/o CLIP Image Encoder: Our proposed model ClipSwap is compared with a baseline model that excludes the CLIP image encoder, resulting in the omission of providing the reference image to CLIP. As illustrated in Fig. 6, the images in the third column generated by this baseline model fail to maintain the source face's facial identity information. Notably, it struggles in transferring features such as eyebrows, eyes, lip shape and lip color of the source image, while retaining some identity elements of the target image. Consequently, the resulting image bears a closer resemblance to the target image. In contrast, our ClipSwap approach excels in addressing these challenges, achieving superior results through the integration of CLIP-based reference image.

W/o Attributes Preservation Loss: We also conduct a comparison between ClipSwap and a model that omits the inclusion of our proposed attributes preservation loss. Not including the attributes preservation loss leads to the exclusion of specific attribute details, which is noticeable in the images presented in the fourth column of Fig. 6. It is evident from the visual results that this model, lacking the attributes preservation loss, struggles to accurately transfer attributes such as eye gaze and eye color to the swapped image.

W/o Augmentation: We perform an experiment on the augmentation process. When we calculate the contrastive loss without using augmentation, the model lacks diverse representations of the swapped image. Consequently, it fails to generate accurate facial expression and eye gaze, as demonstrated in the fifth column of Fig. 6. This limitation arises because the model lacks sufficient positive and negative samples to understand the semantic information within the CLIP framework. Therefore, the augmentation process plays an important role in our model.

W/ ArcFace Image Encoder: To access the impact of the CLIP-based reference guided approach on our model, we conduct experiments by replacing the CLIP image encoder with ArcFace encoder. To guarantee a fair comparison, we compute the contrastive loss between the identity features of the source image, obtained by the ArcFace encoder, and both similar and dissimilar samples of the swapped image. As shown in the sixth column of Fig. 6, ArcFace provides better face shape (second row), but it fails to preserve accurate gaze direction (first and second row). Moreover, it retains some identity features of the target image, such as lip shape and color (first row). Thus, we can conclude that CLIP incorporates a broader understanding of visual content and may capture more context-aware features. Additionally, the ability to utilize the source image as a guiding reference image in CLIP provides a way to steer the generation process, ensuring that the swapped face aligns closely with the reference image.

The results shown in Table 4 confirm the aforementioned observations, validating that our proposed model consistently outperforms alternative baselines developed through various ablations. These results highlight the superior performance of our model across multiple quantitative metrics.

Method	ID Sim. ↑	Pose ↓	Exp ↓	SSIM ↑	PSNR ↑
w/o CLIP	0.52	3.13	25.67	0.93	27.78
w/o att. loss	0.54	2.99	26.31	0.94	27.90
w/o Aug.	0.47	3.99	31.41	0.90	24.03
w/ ArcFace	0.52	2.92	25.71	0.94	28.86
Ours	**0.55**	**2.83**	**25.52**	**0.94**	**28.90**

TABLE IV: Ablation Study: Quantitative comparison of the proposed ClipSwap with different configurations on the CelebA-HQ [16] dataset images.

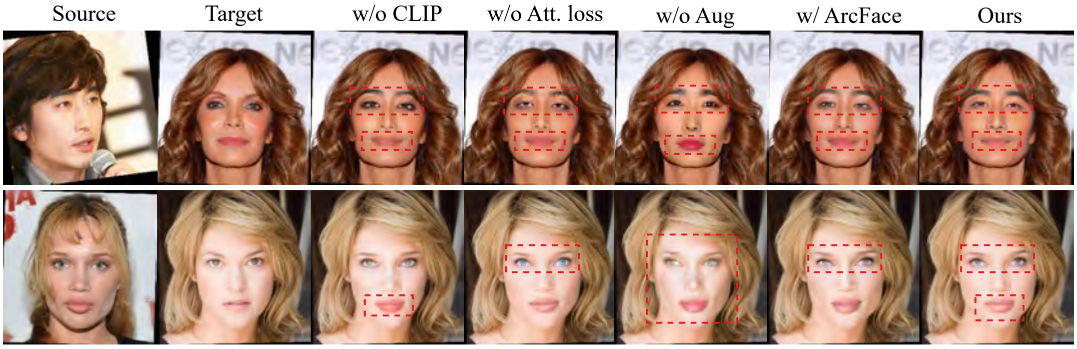

Fig. 6: Ablation Study: Qualitative comparison of ClipSwap with different configurations (such as w/o CLIP, w/o attributes preservation loss, w/o augmentaion process in computing CLIP contrastive loss and the model w/ ArcFace is for ablating CLIP image encoder) on the CelebA-HQ [16]. Please see Section D. for more details.

Failure Cases: Although ClipSwap is successful in many instances, we have observed certain failure cases where it does not perform as expected. These challenges typically occur when the face is not directly facing the camera or when the lighting conditions are extremely poor. In such cases, certain facial features become less visible or obscured, causing our model to struggle in accurately detecting and transferring the source identity onto the target face. Consequently, the resulting image tends to retain more characteristics of the target face, often introducing artifacts and failing to achieve the intended result. In Fig. 7, we illustrate some instances where our model faces these particular challenges and may not perform as well.

Fairness and Ethics: As face related machine learning applications are extremely important, they require training data that is representative of different cultures, ethnicities, and genders. Therefore, the results of face swapping technologies may not always be accurate. In terms of ethics, the use of such technologies should be regulated, and the users should be made aware of any negative use cases and their repercussions. The purpose of creating such face swapped images in this work is for several ethical applications such as the film industry [1] and computer games, primarily for generating fictional twins and enhancing gaming visuals. In addition, it can also be applied in several other domains including privacy preservation [19], [44], [43], aiding digital forensics investigation [37] and contributing to academic studies. Moreover, in areas like facial emotion recognition [26] where there is not enough data, face swapping could help as a potential method for facial data augmentation [30], [47]. It is noteworthy that improving face swapping technologies will also improve the ability to detect forged facial images [36], [48], [58]. Therefore, face swapping has received much attention in the research area of computer vision and graphics [8], [28], [52]. Additionally, we want to highlight that our research aims to contribute to ethical applications within these domains.

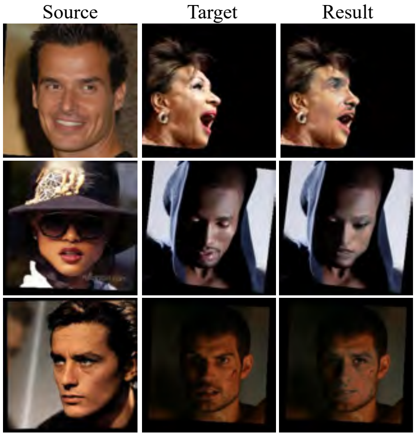

Fig. 7: Few noisy generated results from ClipSwap. Notice the head pose and poor illumination in the target images.

V. CONCLUSION

In this paper, we introduce ClipSwap, a framework for high-fidelity face swapping designed to maintain sensitive facial attributes. We present the concept of CLIP-based reference-guided face swapping, which enables the transfer of source facial attributes to the swapped image while preserving the identity of the original subject. Additionally, we use the CLIP contrastive loss to optimize CLIP-space directions, thus guiding the editing process towards desired attributes. This optimization is achieved by increasing the mutual information between the swapped and reference images. Furthermore, we ensure the faithful preservation of essential attributes between the target and swapped images through the incorporation of an attributes preservation loss. Our proposed framework consistently demonstrates superior performance in generating realistic face images, ensuring that not only facial coordinates but also critical attributes are faithfully matched. Our extensive experiments provide the validation of the effectiveness of our approach, demonstrating a clear and significant superiority over previous face swapping methods. As part of the future work, we will include hair style transfer into the network, while also enhancing its robustness to occlusion. There is also scope for video based face swapping via temporal changes.

REFERENCES

[1] O. Alexander, M. Rogers, W. Lambeth, M. Chiang, and P. Debevec. Creating a photoreal digital actor: The digital emily project. In *2009 Conference for Visual Media Production*, pages 176–187. IEEE, 2009.

[2] J. Bao, D. Chen, F. Wen, H. Li, and G. Hua. Towards open-set identity preserving face synthesis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6713–6722, 2018.

[3] D. Bitouk, N. Kumar, S. Dhillon, P. Belhumeur, and S. K. Nayar. Face swapping: automatically replacing faces in photographs. In *ACM SIGGRAPH 2008 papers*, pages 1–8. 2008.

[4] V. Blanz, K. Scherbaum, T. Vetter, and H.-P. Seidel. Exchanging faces in images. In *Computer Graphics Forum*, volume 23, pages 669–676. Wiley Online Library, 2004.

[5] V. Blanz and T. Vetter. A morphable model for the synthesis of 3d faces. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 157–164. 2023.

[6] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pages 67–74. IEEE, 2018.

[7] H. Chefer, S. Benaim, R. Paiss, and L. Wolf. Image-based clip-guided essence transfer. In *European Conference on Computer Vision*, pages 695–711. Springer, 2022.

[8] R. Chen, X. Chen, B. Ni, and Y. Ge. Simswap: An efficient framework for high fidelity face swapping. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2003–2011, 2020.

[9] J. Deng, J. Guo, E. Ververas, I. Kotsia, and S. Zafeiriou. Retinaface: Single-shot multi-level face localisation in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5203–5212, 2020.

[10] J. Deng, J. Guo, N. Xue, and S. Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019.

[11] FaceNet. https://github.com/davidsandberg/facenet.

[12] FFHQ. https://github.com/nvlabs/ffhq-dataset.

[13] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*, pages 87–102. Springer, 2016.

[14] X. Huang and S. Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*, pages 1501–1510, 2017.

[15] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 694–711. Springer, 2016.

[16] T. Karras, T. Aila, S. Laine, and J. Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

[17] T. Karras, M. Aittala, J. Hellsten, S. Laine, J. Lehtinen, and T. Aila. Training generative adversarial networks with limited data. *Advances in neural information processing systems*, 33:12104–12114, 2020.

[18] T. Karras, S. Laine, and T. Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019.

[19] I. Kemelmacher-Shlizerman. Transfiguring portraits. *ACM Transactions on Graphics (TOG)*, 35(4):1–8, 2016.

[20] J. Kim, J. Lee, and B.-T. Zhang. Smooth-swap: a simple enhancement for face-swapping with smoothness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10779–10788, 2022.

[21] D. E. King. Dlib-ml: A machine learning toolkit. *The Journal of Machine Learning Research*, 10:1755–1758, 2009.

[22] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[23] U. Kocasari, A. Dirik, M. Tiftikci, and P. Yanardag. Stylemc: multi-channel based fast text-guided image generation and manipulation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 895–904, 2022.

[24] I. Korshunova, W. Shi, J. Dambre, and L. Theis. Fast face-swap using convolutional neural networks. In *Proceedings of the IEEE international conference on computer vision*, pages 3677–3685, 2017.

[25] G. Kwon and J. C. Ye. Clipstyler: Image style transfer with a single text condition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18062–18071, 2022.

[26] G. Levi and T. Hassner. Emotion recognition in the wild via convolutional neural networks and mapped binary patterns. In *Proceedings of the 2015 ACM on international conference on multimodal interaction*, pages 503–510, 2015.

[27] A. Li, L. Zhao, Z. Zuo, Z. Wang, W. Xing, and D. Lu. Crfast: Clip-based reference-guided facial image semantic transfer. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[28] L. Li, J. Bao, H. Yang, D. Chen, and F. Wen. Faceshifter: Towards high fidelity and occlusion aware face swapping. *arXiv preprint arXiv:1912.13457*, 2019.

[29] 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*, pages 5074–5083, 2020.

[30] I. Masi, A. T. Tran, T. Hassner, J. T. Leksut, and G. Medioni. Do we really need to collect millions of faces for effective face recognition? In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*, pages 579–596. Springer, 2016.

[31] MXNet. https://github.com/tornadomeet/mxnet-face.

[32] R. Natsume, T. Yatagawa, and S. Morishima. Rsgan: face swapping and editing using face and hair representation in latent spaces. *arXiv preprint arXiv:1804.03447*, 2018.

[33] R. Natsume, T. Yatagawa, and S. Morishima. Fsnet: An identity-aware generative model for image-based face swapping. In *Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part VI 14*, pages 117–132. Springer, 2019.

[34] Y. Nirkin, Y. Keller, and T. Hassner. Fsgan: Subject agnostic face swapping and reenactment. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7184–7193, 2019.

[35] Y. Nirkin, I. Masi, A. T. Tuan, T. Hassner, and G. Medioni. On face segmentation, face swapping, and face perception. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 98–105. IEEE, 2018.

[36] Y. Nirkin, L. Wolf, Y. Keller, and T. Hassner. Deepfake detection based on discrepancies between faces and their context. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):6111–6121, 2021.

[37] M. A. Oikawa, Z. Dias, A. de Rezende Rocha, and S. Goldenstein. Manifold learning and spectral clustering for image phylogeny forests. *IEEE Transactions on Information Forensics and Security*, 11(1):5–18, 2015.

[38] A. v. d. Oord, Y. Li, and O. Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[39] T. Park, M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2337–2346, 2019.

[40] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[41] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[42] F. Rosberg, E. E. Aksoy, F. Alonso-Fernandez, and C. Englund. Facedancer: pose-and occlusion-aware high fidelity face swapping. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 3454–3463, 2023.

[43] A. Ross and A. Othman. Visual cryptography for biometric privacy. *IEEE transactions on information forensics and security*, 6(1):70–81, 2010.

[44] 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*, pages 1–11, 2019.

[45] N. Ruiz, E. Chong, and J. M. Rehg. Fine-grained head pose estimation without keypoints. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 2074–2083, 2018.

[46] G. Sharma, A. Dhall, and R. Subramanian. Mars: A multi-view contrastive approach to human activity recognition from accelerometer sensor. *IEEE Sensors Letters*, 2024.

[47] G. Sharma, C. Gupta, A. Agarwal, L. Sharma, and A. Dhall. Generat-

ing point cloud augmentations via class-conditioned diffusion model. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 480–488, 2024.

[48] K. Shiohara and T. Yamasaki. Detecting deepfakes with self-blended images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18720–18729, 2022.

[49] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, and M. Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2387–2395, 2016.

[50] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, and W. Liu. Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5265–5274, 2018.

[51] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8798–8807, 2018.

[52] Y. Wang, X. Chen, J. Zhu, W. Chu, Y. Tai, C. Wang, J. Li, Y. Wu, F. Huang, and R. Ji. Hififace: 3d shape and semantic prior guided high fidelity face swapping. *arXiv preprint arXiv:2106.09965*, 2021.

[53] T. Wei, D. Chen, W. Zhou, J. Liao, Z. Tan, L. Yuan, W. Zhang, and N. Yu. Hairclip: Design your hair by text and reference image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18072–18081, 2022.

[54] H. Wu, Y. Qu, S. Lin, J. Zhou, R. Qiao, Z. Zhang, Y. Xie, and L. Ma. Contrastive learning for compact single image dehazing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10551–10560, 2021.

[55] Y. Xu, B. Deng, J. Wang, Y. Jing, J. Pan, and S. He. High-resolution face swapping via latent semantics disentanglement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7642–7651, 2022.

[56] Y. Xu, B. Deng, J. Wang, Y. Jing, J. Pan, and S. He. High-resolution face swapping via latent semantics disentanglement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7642–7651, 2022.

[57] Y. Yu, F. Zhan, R. Wu, J. Zhang, S. Lu, M. Cui, X. Xie, X.-S. Hua, and C. Miao. Towards counterfactual image manipulation via clip. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 3637–3645, 2022.

[58] T. Zhao, X. Xu, M. Xu, H. Ding, Y. Xiong, and W. Xia. Learning self-consistency for deepfake detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 15023–15033, 2021.

[59] Y. Zhu, Q. Li, J. Wang, C.-Z. Xu, and Z. Sun. One shot face swapping on megapixels. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4834–4844, 2021.