Computer Vision and Image Understanding
Authorship Confirmation

Please save a copy of this file, complete and upload as the "Confirmation of Authorship" file.

As corresponding author I, Yu Ding, hereby confirm on behalf of all authors that:

1. This manuscript, or a large part of it, has not been published, was not, and is not being submitted to any other journal.
2. If presented at or submitted to or published at a conference(s), the conference(s) is (are) identified and substantial justification for re-publication is presented below. A copy of conference paper(s) is(are) uploaded with the manuscript.
3. If the manuscript appears as a preprint anywhere on the web, e.g. arXiv, etc., it is identified below. The preprint should include a statement that the paper is under consideration at Computer Vision and Image Understanding.
4. All text and graphics, except for those marked with sources, are original works of the authors, and all necessary permissions for publication were secured prior to submission of the manuscript.
5. All authors each made a significant contribution to the research reported and have read and approved the submitted manuscript.

Signature Yu Ding Date August 18, 2022

List any pre-prints:

Relevant Conference publication(s) (submitted, accepted, or published): Accepted by IEEE Conference on Games 2022.

Justification for re-publication: This paper is an extension of our conference (IEEE Conference on Games, title: "Paste You Into Games: Identity and Expression Consistency in Swapping Character Face") version. We have conducted more quantitative and qualitative experiments and provided a more detailed analysis of our method. The main contribution is to propose a new solution for the game character face swapping task. We collect a game character face dataset and employ a fine-tuning strategy to solve the cross-domain face swapping problem. Moreover, we propose a novel expression embedding loss to enforce the expression consistency and adopt compound identity embeddings, which can ease the feature bias in each single face recognition model to obtain a better identity consistency. The visualized results and the qualitative and quantitative comparisons reveal the significance and effectiveness of our proposed method. The content of our paper is very suitable for the journal.

Graphical Abstract (Optional)

To create your abstract, please type over the instructions in the template box below. Fonts or abstract dimensions should not be changed or altered.

Type the title of your article here
Author's names here

ELSEVIER

Customizing the appearance of game characters according to individual preferences is an important application in the gaming industry. The traditional solutions such as manual editing within the game engine require much time and some professional experience. To resolve, this paper proposes a novel face swapping framework to swap real faces onto game characters while maintaining the art style of the game. This approach helps players speed up the character customization process, as they can input any face and quickly see the results of swapping it onto the character. Specifically, our framework extracts a more robust identity embedding from compound face identity models. Then the identity embedding is sent into AdaIN for feature fusion and changes the identity attribute for the output of the decoder. To maintain the expression consistency between the swapped and target faces, we utilize a novel expression embedding loss which effectively constrains the fine-grain expression similarity. To reduce the cross-domain gap between human and game faces, we also construct a game face dataset and propose to use fine-tuning to improve the image quality of cross-domain face swapping. Extensive qualitative and quantitative experiments indicate that our method achieves leading swapping results in both pure natural human faces and game faces datasets.

Research Highlights (Required)

To create your highlights, please type the highlights against each `\item` command.

It should be short collection of bullet points that convey the core findings of the article. It should include 3 to 5 bullet points (maximum 85 characters, including spaces, per bullet point.)

- We propose the first work for game character face swapping task. By collecting a game character face dataset and applying a fine-tuning strategy, we manage to translate the face swapping model from the normal human face to the game face domain.
- We propose an identity compound strategy to improve the identity consistency between the source and the swapped face images while preserving the subject's attributes to be harmonic with the game character.
- We apply an expression consistency metric for the face swapping task. By our designed expression embedding loss, we can enforce the generated faces to keep expression similarities with the target faces as close as possible.
- We conduct full experiments on both human and game character faces. The quantitative and qualitative experiments demonstrate that our results outperform the previous face swapping methods in terms of cross-domain translation capability, expression similarities, and identity consistency.

Computer Vision and Image Understanding
journal homepage: www.elsevier.com

Face Identity and Expression Consistency For Game Character Face Swapping

Hao Zeng[a], Wei Zhang[a], Keyu Chen[a], Zhimeng Zhang[a], Lincheng Li[a], Yu Ding[a,**]

[a]*NetEase Fuxi AI Lab, Hangzhou, China*

ABSTRACT

Customizing the appearance of game characters according to individual preferences is an important application in the gaming industry. The traditional solutions such as manual editing within the game engine require much time and some professional experience. To address this problem, this paper proposes a novel face-swapping framework to swap real faces onto game characters while maintaining the art style of the game. This approach helps players speed up the character customization process, as they can input any face and quickly see the results of swapping it onto the character. Specifically, our framework extracts a more robust identity embedding from compound face identity models. Then the identity embedding is sent into AdaIN for feature fusion and changes the identity attribute for the output of the decoder. To maintain the expression consistency between the swapped and target faces, we utilize a novel expression embedding loss which effectively constrains the fine-grain expression similarity. To reduce the cross-domain gap between human and game faces, we also construct a game face dataset and propose to use fine-tuning to improve the image quality of cross-domain face swapping. Extensive qualitative and quantitative experiments indicate that our method achieves leading swapping results in both pure natural human faces and game faces datasets.

1. Introduction

Game CG videos are animations related to scene characters or plots in the game produced with the support of computer graphics (CG) technology. CG videos are essential for game promotion. To achieve personalized promotion, we propose to use face swapping to generate customized identity-specific CG videos. Given a template video (Target), we can replace the game charactor face in the video with a human face (Source) to obtain a personalized video (Result), and the result face is similar in appearance (identity) to the source face but maintains the attributes of the target faces. However, there are still several challenges when directly applying the existing face swapping methods to game characters. Firstly, the game character face swapping needs to be identity-agnostic, while many works (DeepFakes, 2019; MarekKowalski, 2021) are identity-specific. Secondly, the game character face swapping is a cross-domain problem, but some identity-agnostic methods (Li et al., 2019; Chen et al., 2020; Xu et al., 2021) only focus on human faces. Third, the game character face and the human face are from two different domains, so the identity embedding extracted by a single human-based face recognition model cannot provide sufficient identity constraints. In addition, the existing methods without fine-grained expression constraints will cause inconsistent expressions in generated videos.

Considering the above problems, we propose a new face swapping method and make efforts in three aspects to better generalize the existing methods to the game character faces: solving the cross-domain problem, preserving the identity consistency (with the source face), and the expression consistency (with the target face).

To resolve the cross-domain problem, we first train our face swapping model on the human data and then fine-tune the model on the game face dataset. Since there exists few game face datasets as large as real human face datasets, training directly on the game dataset will greatly reduce the robustness and generalization of the model, and fine-tuning can make full use of the knowledge learned by the model on the human dataset and does not require a large amount of game data.

In the aspect of identity, prior works (Li et al., 2019; Chen et al., 2020) utilize the face recognition models to ensure identity consistency. However, the single extracted identity embed-

**Corresponding author: Tel.: +0-000-000-0000; fax: +0-000-000-0000;
e-mail: dingyu01@corp.netease.com (Yu Ding)

ding lacks robustness. As we observed in Figure 9, the face identity embedding can be easily affected by facial attributes (e.g., expressions). Therefore, we propose the compound multiple identity embeddings, aiming to promote the stability of the identity embedding and achieve better identity consistency.

As for the expression, previous methods (Xu et al., 2021; Li et al., 2019) either use the landmarks or the implicit attribute constraints to ensure the consistency of expressions. However, these expression representations cannot capture subtle facial movements and complicated expressions. This work proposes an effective operation to introduce an existing effective expression embedding from DLN(Zhang et al., 2021b) which represents the identity-invariant and fine-grained human expressions. We use it as an expression loss to constrain the expression consistency. Although this embedding is designed for other recognition or detection tasks, our work is the first to employ it in the face swapping task and achieve effective results. This paper is an extension of our previous work(Zeng et al., 2022).

In summary, our main contributions are as follows:

- We propose the first work for game character face swapping task. By collecting a game character face dataset and applying a fine-tuning strategy, we manage to translate the face swapping model from the normal human face to the game face domain.
- We propose an identity compound strategy to improve the identity consistency between the source and the swapped face images while preserving the subject's attributes to be harmonic with the game character.
- We apply an effective expression embedding as our expression loss to keep expression similarities with the target faces as close as possible. Compared to some other expression constraints, it can effectively perceive fine-grained expression changes.
- We conduct full experiments on both human and game character faces. The quantitative and qualitative experiments demonstrate that our results outperform the previous face swapping methods in terms of cross-domain translation capability, expression similarities, and identity consistency.

2. Related Work

2.1. Face Swapping

Face swapping aims to replace a target image's facial identity with another one. Research works on this problem can be divided into pixel-based methods, 3DMM-based methods, and GAN-based methods. The most straightforward solution for face swapping is to replace the inner face part in pixel space (Bitouk et al., 2008; Lin et al., 2012; Chen et al., 2019a). However, the manipulated image patches usually suffer from attribute mismatch. 3DMM-based methods (Blanz et al., 2004; Thies et al., 2016; Nirkin et al., 2018) generate the face region by 3D fitting and then the source faces and target backgrounds can be blended via inverse rendering. More recently, there occurs many GAN-based methods (DeepFakes, 2019; MarekKowalski, 2021; Petrov et al., 2020; Natsume et al., 2018; Natsume, Ryota and Yatagawa, Tatsuya and Morishima, Shigeo, 2018; Nirkin et al., 2019; Li et al., 2019; Chen et al., 2020; Zhu et al., 2020; Xu et al., 2021; Wang et al., 2021; Zhu et al., 2021; Gao et al., 2021). Specifically, the most popular methods like Deepfakes (DeepFakes, 2019) and its variants(Petrov et al., 2020; MarekKowalski, 2021) need to be trained pairwise. FSGAN(Nirkin et al., 2019) first animate the source face by reenactment and then blend it into the background with an in-painting and blending network. FaceShifter (Li et al., 2019) generates a swapped face with high-fidelity and can handle the occlusions with a second-stage refinement network. SimSwap (Chen et al., 2020) proposes a weak feature matching loss to improve the facial attributes consistency. FaceController (Xu et al., 2021) proposes a unified framework for identity swapping and attribute editing and is the first work that use 3D parameters and identity embedding to represent facial identity. Later, HifiFace (Wang et al., 2021) solved the shape inconsistency problem in face swapping by 3D shape-aware identity to control the face shape with geometric supervision. InfoSwap (Gao et al., 2021) disentangle identity and identity-irrelevant information by optimizing an information bottleneck tradeoff and achieving better identity-discriminative face swapping results. MegaFS (Zhu et al., 2021) proposes the first-megapixel level method and can achieve 1024×1024 face swapping. FlowFace (Zeng et al., 2023) and FlowFace++(Zhang et al., 2023) propose to perform the swapping of face outline shape. The work (Liu et al., 2023) carries out facial semantic region-based editing for face swapping. The work (Jiang et al., 2023) makes efforts to identity-preserving by constructing a series of identity-preserving bases in respect of pose, expression, and illumination.

The methods mentioned above are based on human data, so they cannot be applied directly to the game data. Most methods (Li et al., 2019; Chen et al., 2020; Xu et al., 2021; Wang et al., 2021; Zhu et al., 2021) transfer identity with a single identity embedding which may be affected by other facial attributes such as facial expressions. In addition, implicit attribute constraints(Li et al., 2019; Chen et al., 2020) or facial landmark loss(Xu et al., 2021; Wang et al., 2021; Zhu et al., 2021) are not able to capture subtle expressions, causing the problem of expression consistency.

2.2. Expression Representation

Facial expression plays a vital role in human social communication (Li and Deng, 2020). However, due to its complicated natural and subtle movement, it is non-trivial to represent the accurate expressions of human faces and thus prevent the downstream tasks such as face editing and manipulation. Ekman et al. first propose to use seven basic categories to represent the human facial emotion (Ekman and Friesen, 1971). These categories are the most commonly-used emotion labels in many facial expression recognition methods (Li et al., 2018; Chen et al., 2019b; Wang et al., 2020; Mollahosseini et al., 2017; Kervadec et al., 2018; Happy and Routray, 2014; Kim et al., 2017; Khan et al., 2017; Nguyen et al., 2019). Nevertheless, these

categorical expressions are still a block to many fine-grained expression-related applications, being inadequate to characterize all the facial expressions and distinguish those facial expressions labeled in the same category. Later it is improved by learning a low-dimension nonlinear manifold embedded in a face image space (Vemulapalli and Agarwala, 2019). Compared with discrete expression categories, compact representations can describe more fine-grained expressions. In fact, the facial expression compact representation (Zhang et al., 2021b) has also been validated for the recognition of discrete facial expressions and action units (An et al., 2022; Zhang et al., 2021a, 2022).

In this work, we turn to address the expression consistency manner into face swapping framework and adopt a novel expression representation (Zhang et al., 2021b) which extracts a continuous space based on expression similarities.

3. Method

3.1. Framework Overview

The framework is shown in Figure 1. Given one source face image I_s and one target face image I_t, it performs face swapping and generates I_o reflecting the attribute information (expression, skin color etc.) of I_t but the identity information of I_s. The framework is built upon a generative adversarial network (GAN) (Goodfellow et al., 2014) with end-to-end training. Specially, our framework contains five components: facial image encoder(E_f), facial image decoder(D_f), identity embedding module(E_{id}), expression embedding module(E_{exp}) and a multi-scale discriminator (Isola et al., 2017). The face image encoder E_f is designed to extract multi-scale features of facial attributes (reflecting expression, pose, lighting, etc.) f^i_{attr} from the target image I_t. The compound identity embeddings $\{f^1_{id}, f^2_{id}\}$ are extracted with the identity embedding module E_{id}. The face image decoder D_f is fed with f^i_{attr} and $\{f^1_{id}, f^2_{id}\}$. The discriminator of GAN is a multi-scale discriminator (Isola et al., 2017) to make the swapped face image realistic. In face swapping, E_f extracts attribute-related $f_{attr} = \{f^1_{attr}, f^2_{attr}, f^3_{attr}, f^4_{attr}\}$ from a target image I_t. Next, E_{id} extracts compound identity embeddings $\{f^1_{id}, f^2_{id}\}$ from a source image I_s. Then, D_f is fed with f_{attr} and $\{f^1_{id}, f^2_{id}\}$, and render facial semantics into the swapped face I_o.

3.2. Game Faces Collection

The game faces in our dataset are collected from two sources. On the one hand, we render face images using a computer graphic engine. We collect 30 3D models of game characters. Each model is driven to perform 1000 different facial expressions by traversing its expression blendshape and pose parameters. The generated expressions are then rendered into the corresponding images separately by the Unity game engine (Unity, 2021). This way, we get about 30,000 rendered game face images with desired expression diversity. On the other hand, since the rendered images lack identity diversity, we crawl more game images from the Internet. To reduce labor costs, we design an automated filtering method based on face detection, and only the images with detected faces are retained. To improve the detection accuracy, we use two different face detection models (Zhang et al., 2016, 2017) for cross-validation. Finally, we obtain about 80,000 game face images with better identity diversity.

Finally, we collect a game face dataset containing about 110,000 images. These images are used to finetune and evaluate our model for cross-domain face swapping.

3.3. Identity Consistency

We propose to use compound identity embeddings instead of single identity embedding to provide the identity information of I_s. The compound identity embeddings are offered by multiple pre-trained face recognition models. They provide raw and sufficient identity information from different identity recognition models for the decoder D_f. This allows for avoiding the bias from a specific identity recognition model. Therefore, the compound strategy allows for refining identity information through D_f. Section 4.4 provides more analysis and discussion about the bias of an identity embedding.

To inject the identity information into the decoder D_f, we replace the normalization layer in the original residual block (He et al., 2016) with the adaptive instance normalization (AdaIN) (Liu et al., 2017b) as other methods(Li et al., 2019; Chen et al., 2020) done, and then the identity information (embedding) is mapped to two modulation vectors (γ_{id} and β_{id}) of the normalization layer in AdaIN with two fully connected layers. The difference is that the identity information in our method is from two different models, so we first compound the two identity embeddings with a multi-layer perceptron. The formulation is written as:

$$c^i_{id} = P^i(\left[f^1_{id}, f^2_{id}\right]), \tag{1}$$

where P^i represents the perceptron in the i_{th} block of the decoder. Then the compound identity c^i_{id} is then injected into the intermediate feature map with AdaIN:

$$Z^i_{id} = F^i_\gamma(c^i_{id})\left(\frac{Z^i - \mu(Z^i)}{\sigma(Z^i)}\right) + F^i_\beta(c^i_{id}), \tag{2}$$

where F^i_* are two fully connected layers in the i_{th} block of the decoder and Z^i is the input feature map of the i_{th} block.

To recover the details lost due to downsampling, we design a *Detail Recover Block* (DRB) following (Li et al., 2019). In the i_{th} Detail Recover Block, we first obtain the identity-injected feature Z^i_{id} through the injection method in the AdaIN residual block, and then the corresponding attribute feature f^i_{attr} is used to merge with Z^i_{id} adaptively with attention. The difference with (Li et al., 2019) is that we not only use spatial attention but also channel attention. Specifically, we first inject the f^i_{attr} into Z^i with spatially-adaptive normalization (SPADE) (Park et al., 2019), we slightly modify the SPADE by replacing the batch normalization (Ioffe and Szegedy, 2015) with instance normalization (Ulyanov et al., 2016) for comparable performance but fewer parameters.

$$Z^i_{attr} = T^i_\gamma(f^i_{attr})\left(\frac{Z^i - \mu(Z^i)}{\sigma(Z^i)}\right) + T^i_\beta(f^i_{attr}), \tag{3}$$

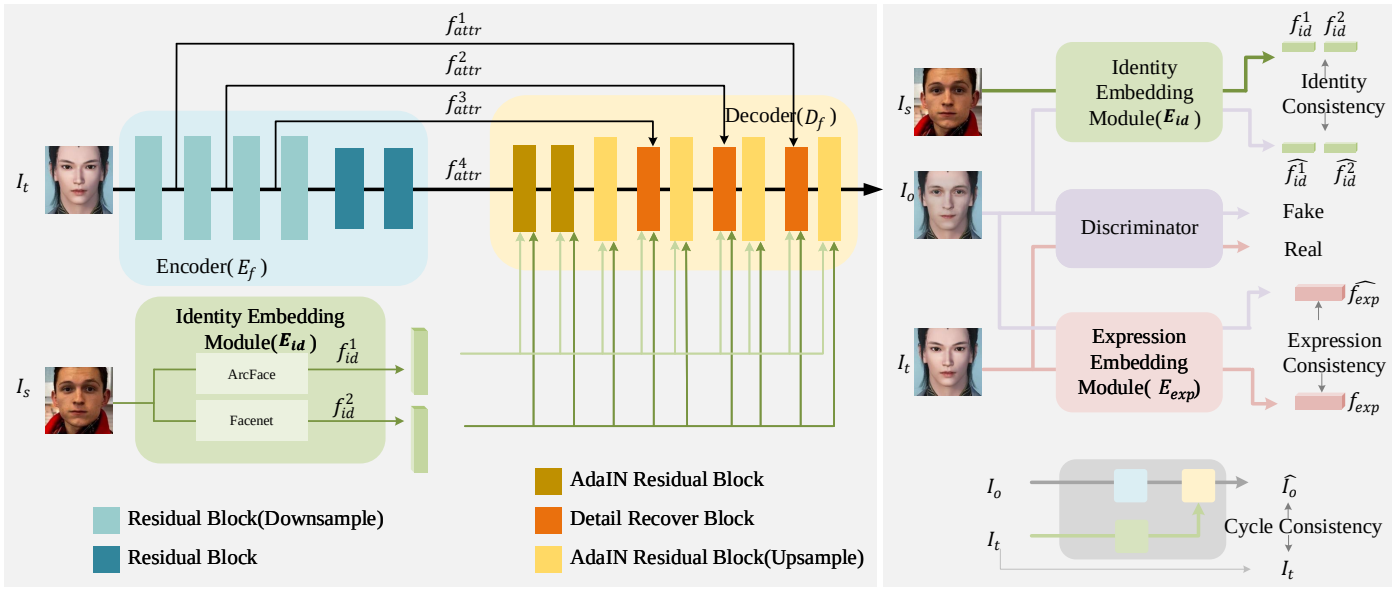


Fig. 1. Architecture of our proposed framework. The framework mainly consists of five components: facial image encoder(E_f), facial image decoder(D_f), identity embedding module(E_{id}), expression embedding module(E_{exp}) and a multi-scale discriminator (Isola et al., 2017) and is trained end-to-end.

where T^i_* are two convolutional layers used to compute modulation parameters γ_{attr} and β_{attr} of the normalization layer in SPADE. However, unlike the vector form parameters in AdaIN, γ_{attr} and β_{attr} are tensors with the same spatial dimension as Z^i.

Then, we generate the spatial attention mask M^i_s and the channel attention maskM^i_c from the original input Z^i with a convolutional block attention module (CBAM) (Woo et al., 2018).

$$(M^i_s, M^i_c) = CBAM^i(Z^i). \tag{4}$$

In each convolutional block attention module, the spatial attention is first computed as following:

$$M^i_s = \sigma\left(CONV([AP(\mathbf{Z^i}); MP(\mathbf{Z^i})])\right), \tag{5}$$

then the channel-wise attention is produced in a similar way:

$$M^i_c = \sigma(MLP(AP(\mathbf{Z^i})) + MLP(MP(\mathbf{Z^i}))), \tag{6}$$

where σ is the sigmoid function, AP and MP represent the average-pooling layer and the max-pooling layer, respectively. The difference is that the pooling operation in eq:ms is along the spatial dimension, while the pooling operation in eq:mc is along the channel dimension. The larger value in M^i_s indicates that the corresponding area is more related to the identity, and the larger value in M^i_c indicates that the corresponding feature (channel) is more related to the identity.

Finally, the attribute-injected feature Z^i_{attr} and the identity-injected feature Z^i_{id} are fused using the two attention masks:

$$\hat{Z}_i = M^i_s \times M^i_c \times Z^i_{id} + (1 - M^i_s \times M^i_c) \times Z^i_{attr}, \tag{7}$$

where $\hat{Z}_i$ is the output of the i_{th} block. With the help of the two masks, features not related to the identity are recovered by attribute features.

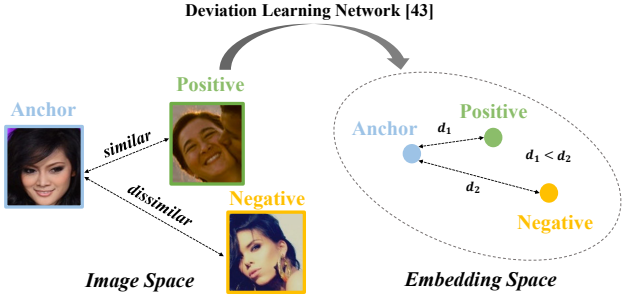


Fig. 2. Triplet loss leads to a continuous embedding space for expression representation.

3.4. Expression Consistency

Some previous methods (Li et al., 2019; Chen et al., 2020) treat the expression as the same as other attributes and achieve consistency of expressions through an implicit constraint on attribute features. Some other methods use facial landmarks (Nirkin et al., 2019; Xu et al., 2021; Zhu et al., 2021) to characterize and constrain expressions. We argue that attribute features cannot obtain some subtle expressions since they contain many other attributes like pose, skin color, etc. The facial landmark is related to the identity, which may harm the identity consistency.

To avoid problems in previous methods and achieve better expression consistency, our E_{exp} leverages an existing expression embedding technique (Zhang et al., 2021b). As shown in Figure 2, the expression embedding model is trained on FEC dataset (Vemulapalli and Agarwala, 2019) that contains a large number of expression triplets from multiple identities. Therefore, the expression embedding can represent fine-grained and identity-invariant facial expressions.

In training, E_{exp} first extracts expression embeddings of target and swapped images. Then an expression loss is defined

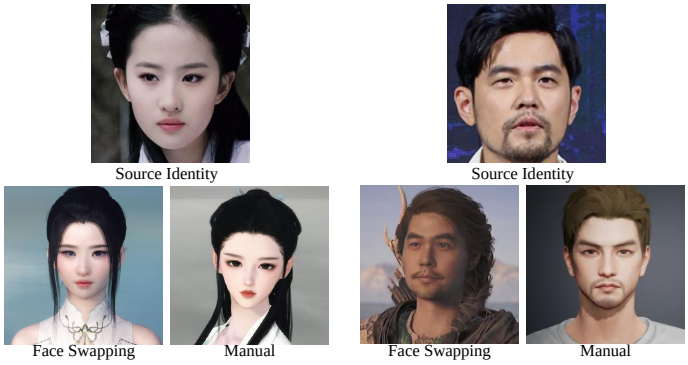


Fig. 3. Comparison of game character results generated by our face swapping method and manual method.

by calculating the euclidean distance between two expression embeddings. In this way, we can achieve better expression consistency. Although the used expression embedding is designed to represent the expressions in the previous works on recognition or detection, no other work in the face-swapping field has realized the effectiveness of this feature. Incorporating this expression constraint into our work is one of the key factors that contribute to our good performances.

3.5. Loss Function

This section details the supervision in the training, including reconstruction loss, identity loss, expression loss, and cycle consistency loss.

Reconstruction Loss: During training, we make I_s, and I_t the same with a certain probability and expect the generated image I_o to be as same as the input. So we introduce a pixel-wise reconstruction loss following (Li et al., 2019). The reconstruction loss is written as

$$\mathcal{L}_{rec} = \|I_o, I_t\|_2 , \tag{8}$$

where $\|*\|_2$ denotes the euclidean distance. In our experiment, we set the probability to 0.25.

Identity Loss: An identity loss is usually used in face swapping tasks. The loss enforces D_f to acquire identity information from injected compound identity embeddings. Due to compound identity embeddings for injection, the identity loss is also based on two face recognition models (ArcFace and FaceNet):

$$\mathcal{L}_{id} = \sum_{k=1}^{K} \lambda_k (1 - cos(E_{id}(I_o), E_{id}(I_s)), \tag{9}$$

where λ_k represents the relative weight of each face recognition model and $cos(*, *)$ denotes the cosine similarity of two identity embeddings. In our experiments, we set K=2 and $\lambda_1 = 10$, λ_2=5 for ArcFace and FaceNet respectively.

Expression Loss: To make the expression of the swapped face I_o more consistent with the target face, we adopt an expression loss that penalizes the $\mathcal{L}_2$ distance of two expression embeddings.

$$\mathcal{L}_{exp} = \left\|E_{exp}(I_o), E_{exp}(I_t)\right\|_2 . \tag{10}$$

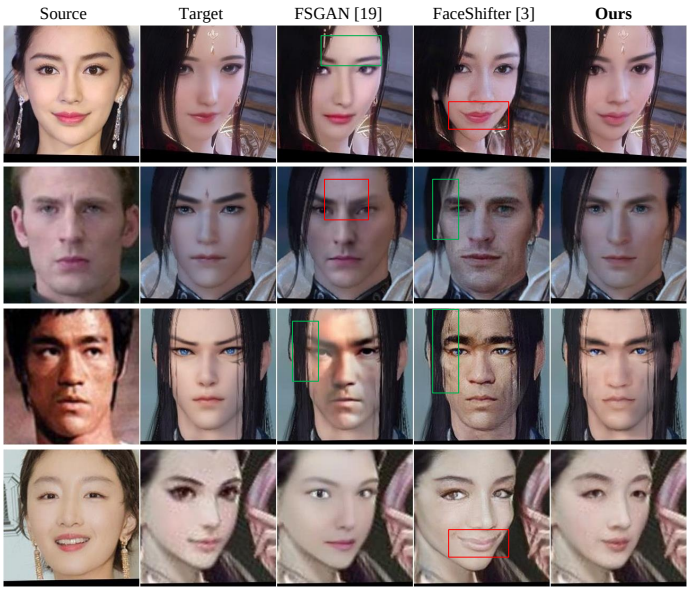


Fig. 4. Game character face swapping comparison with FSGAN (Nirkin et al., 2019) and FaceShifter (Li et al., 2019). Some expression errors and occlusion errors are marked with red and green boxes, respectively.

The expression loss encourages the generator to learn to acquire expression-related information from target faces other than some unrelated disturbance like identity.

Cycle Consistency Loss: In addition to expression and identity, it is also important to guarantee that the swapped face properly preserves the attributes of the target face. To do this, we introduce a cycle consistency loss (Choi et al., 2018):

$$\mathcal{L}_{cycle} = \left\|\hat{I}_o, I_t\right\|_1 , \tag{11}$$

where $\hat{I}_o = D_f(E_f(I_o), E_{id}(I_t))$ and $\|*\|_1$ denotes the $\mathcal{L}_1$ distance. This objective encourages the generator to learn to preserve the original attribute of I_t while only changing its identity.

GAN Loss. To make the synthesized facial images more realistic, adversarial training is employed. Specifically, we adopt Hinge loss (Lim and Ye, 2017) as the adversarial loss, denote as L_{adv}.

Full objective: Our full objective can be summarised as:

$$\mathcal{L} = \mathcal{L}_{adv} + \mathcal{L}_{id} + \lambda_{exp}\mathcal{L}_{exp} + \lambda_{rec}\mathcal{L}_{rec} + \lambda_{cycle}\mathcal{L}_{cycle}, \tag{12}$$

where λ_{exp}, λ_{rec}, λ_{cycle} are hyperparameters for each term.

4. Experiments

4.1. Datasets and Settings

We construct two datasets for our task, one is called HFDataset for the human face swapping, and the other is called GFDataset for game face swapping. HFDataset is a combination of three public datasets including CelebA-HQ (Karras et al., 2017), FFHQ (Karras et al., 2019), and VGGFace2 (Cao et al., 2018). As for GFDataset, images are collected by the method described in Section 3.2. For each image in the two datasets, we aligned and cropped the face to 256 × 256 with a face detector (Zhang et al., 2017). HFDataset is only used for

Table 1. Quantitative Comparison on Game Character Faces.

Method	ID Retrieval Accuracy (%) ↑			FID ↓	Expression Error ↓		
	CosFace	ArcFace2	SphereFace		DFER	FECNet	DLN
FSGAN	49.21	52.37	55.61	71.32	4.13	0.28	0.45
FaceShifter	95.73	98.54	97.72	66.49	4.09	0.35	0.41
Ours	**98.76**	**99.68**	**99.15**	**28.95**	**3.69**	**0.21**	0.25

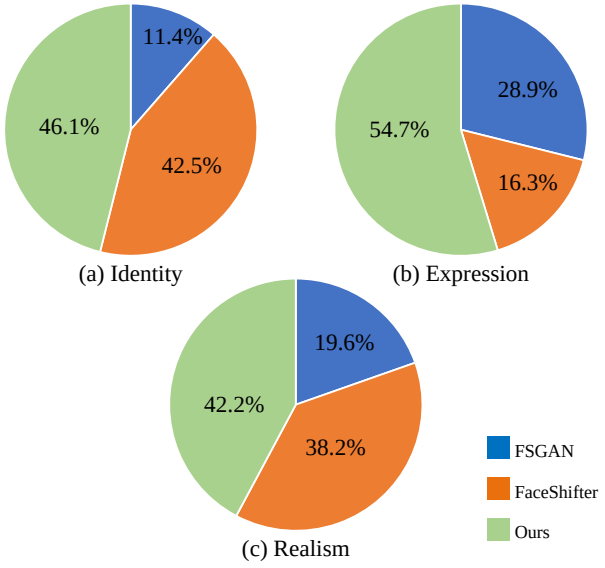


Fig. 5. Subjective comparison with FSGAN (Nirkin et al., 2019) and FaceShifter (Li et al., 2019) on the game character test set.

Table 2. FLOPs and the number of parameters comparisons of our method and some other available methods. Some of the compared methods only have the unofficial code.

Method	FLOPs (G)	Parameters (M)
SimSwap (Chen et al., 2020)	70.59	107.24
HifiFace (Wang et al., 2021)	71.60	146.78
FSGAN (Nirkin et al., 2019)	305.48	**75.96**
FaceShifter (Li et al., 2019)	**41.63**	228.23
Ours	59.30	102.63

training, and GFDataset is split into finetuning set and evaluation set.

Our model is trained on HFDataset from scratch and finetuned on GFDataset. The framework is implemented with PyTorch (Paszke et al., 2019). We adopt Adam (Kingma and Ba, 2014) optimizer with β_1=0 and β_2=0.999 and the learning rate is set to 0.0001. We set λ_{exp}=5, λ_{rec}=10, λ_{cycle}=10 for our full pipeline and our model is trained first about 550K steps and then finetuned about 200K steps with a batch size of 4.

As shown in Table 2, our framework has a computing complexity of 59.30G floating point operations (FLOPs) and 102.63M parameters. Compared to other available methods, our method has comparable requirements in terms of parameters and computational time, relatively lower FLOPs and a relatively smaller number of parameters.

4.2. Comparison on Game Character Faces

We conduct qualitative and quantitative comparisons with the existing methods to validate our method on game character faces.

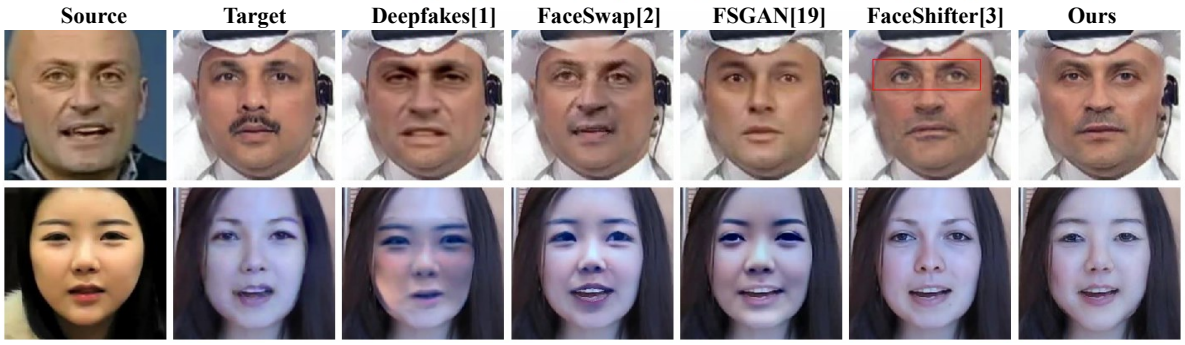
*4.2.1. **Qualitative Comparison***

We first compare our method with the Manual method. The Manual method usually tasks a skilled game player with several hours to edit the hundreds of face parameters to create a character that looks like the source face. As shown in Figure 3, our face swapping method can produce comparable results to the Manual method in less than one second. As for face swapping methods, we compare our method with FSGAN (Nirkin et al., 2019) and FaceShifter (Li et al., 2019). We first obtain the official pre-trained model of FSGAN and then reproduce the first stage of FaceShifter. As shown in Figure 4, FSGAN suffers from unpleasant illumination and face color since FSGAN adopts a blending model to fuse the swapped face with the background. When the source and target face have considerable differences in texture, lighting, or skin color (just like the difference between a game character face and a human face), such a fusion method will cause this attribute mismatch. Moreover, the swapped faces of FSGAN look less similar to the source face than our method. FaceShifter also has problems in cross-domain face swapping, and the expression is affected by the source face. It can also be observed that FaceShifter, without its refinement network can not handle the occlusions well, but our method can, even if we are a one-stage method.

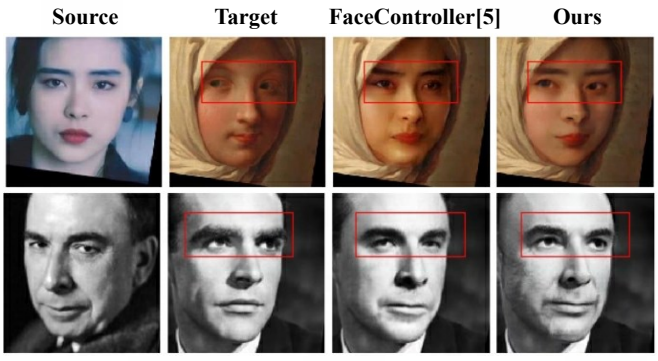
*4.2.2. **Quantitative Comparison***

We further perform the quantitative comparison with FSGAN and FaceShifter on the game character faces. We construct a test set that contains 10K human-game face pairs for human-to-game face swapping. Three types of evaluation metrics are taken into account, including identity retrieval accuracy, expression error, and Frechét inception distance (Heusel et al., 2017).

ID retrieval accuracy is used to estimate the identity consistency. Specifically, a face recognition model extracts identity embedding from swapped images. Then, identity retrieving is performed in the corresponding test set with the nearest cosine distance of identity embedding. In our work, the estimation relies on three face recognition models including CosFace (Wang et al., 2018; Wang, 2018), ArcFace2 (Deng et al., 2019; TreB1eN, 2018) and SphereFace (Liu et al., 2017a; Liu, 2018). As shown in Table 1, our method obtains the highest accuracy, which proves the use of compound identity embeddings guarantees robustness in identity transferring and contributes to identity consistency.



(a) Comparison with Deepfakes, FaceSwap, FSGAN and FaceShifter on FaceForensics++.



(b) Comparison with FaceController.

(c) Comparison with SimSwap.

Fig. 6. Comparison with existing works on natural faces. (a) Comparison with Deepfakes (DeepFakes, 2019), FaceSwap (MarekKowalski, 2021), FSGAN (Nirkin et al., 2019) and FaceShifter (Li et al., 2019) on FaceForensics++ (Rössler et al., 2019). (b) Comparison with FaceController (Xu et al., 2021). (c) Comparison with SimSwap (Chen et al., 2020). (b) and (c) show images that are cropped from their published paper. As observed, our method preserves the identity of the source image and the subtle expressions of the target image better than the other methods. Some expression errors are marked with red boxes.

Table 3. Quantitative Comparison on FF++.

Method	ID Retrieval Accuracy(%)↑			Pose	Expression Error↓		
	CosFace	ArcFace2	SphereFace	Error↓	DFER	FECNet	FECNet
Deepfakes	83.70	81.79	87.18	4.30	5.02	0.56	0.73
FaceSwap	71.45	64.04	77.01	2.65	4.35	0.42	0.58
FSGAN	48.90	49.37	53.85	2.72	4.02	0.29	0.42
FaceShifter	86.83	90.77	81.37	2.71	4.03	0.36	0.49
Ours	**97.66**	**98.84**	**98.31**	**2.56**	**3.61**	**0.21**	**0.28**

Expression error is used to evaluate the expression distance between the swapped and the target faces. We metric this error by computing the euclidean distance between the swapped face expression embedding and the target face expression embedding. The expression embedding can be obtained from a discrete facial expression recognition model (DFER) (WuJie, 2020) trained on (Goodfellow et al., 2013; Lucey et al., 2010) or two continuous facial expression encoder models (DLN (Zhang et al., 2021b) and FECNet (Vemulapalli and Agarwala, 2019)). As shown in Table 1, our method obtains the lowest expression errors in three expression metrics, illustrating our superiority in expression consistency.

Frechét inception distance is used to measure the discrepancy between two sets of images. We use the final average pooling features of an Inception-V3 (Szegedy et al., 2016) pre-trained on ImageNet (Krizhevsky et al., 2012) to compute FID. We can observe from Table 1 that we obtain a lower FID than FSGAN and FaceShifter. This proves that our method can better preserve the game domain feature.

4.2.3. Subjective Comparison

To further illustrate the effectiveness of our method, we conduct a user study on our game test set (10K human-game face pairs) with FSGAN and FaceShifter. Thirty participants are asked to complete the questionnaire in terms of identity consistency, expression consistency, or image realism. Each metric contains 30 cases, and each participant needs to choose the best result under each metric.

Figure 5 demonstrates the results of the subjective comparison in the user study. Our method outperforms the baselines in

terms of identity consistency, expression consistency, and image realism. These results further validate the performance of our method.

4.3. Comparison on Human Faces

To further validate our contributions to identity consistency and expression consistency, we conduct comparison experiments with more face swapping methods on human faces and report the comparative results below, including qualitative and quantitative comparisons. Particularly, for a fair comparison, models used in this section are only trained on human data without fine-tuning on game character images.

4.3.1. ***Qualitative Comparison***

We first compare with Deepfakes (DeepFakes, 2019), FaceSwap (MarekKowalski, 2021), FSGAN (Nirkin et al., 2019) and FaceShifter (Li et al., 2019) on the FaceForensics++ (FF++) (Rössler et al., 2019) dataset. FF++ contains the face swapping results of Deepfakes, FaceSwap, and FaceShifter. We obtain the results of FSGAN by applying the official pre-trained model to FF++. As can be observed in Figure 6, the comparisons illustrate that our work has obvious benefits in terms of expression preservation and identity consistency. Specifically, without any constraint on identity or attributes (expression, etc.), the results of Deepfakes and FaceSwap cannot preserve identity well and suffer a severe mismatch in attributes (expressions, etc.). Results generated by FSGAN lose similarity with the source face and also suffer from inconsistent lighting and skin color. FaceShifter performs very well in terms of image quality and attributes consistency but cannot preserve the target expressions, such as gaze direction. Differently, our method handles all of the above problems well.

Since we cannot acquire codes or released results of FaceController (Xu et al., 2021) and SimSwap (Chen et al., 2020), comparisons with them are conducted on the images cropped from their papers. Figure 6(b), (c) illustrate the qualitative results, and the comparisons show that besides the comparable image quality, our method preserves the identity of the source image and the subtle expressions of the target image better. As seen from the red box in Figure 6(b), (c), the results of the two compared methods contain some unwanted subtle expressions such as wrong gaze direction and disappearing frown.

A common problem can also be observed from the above comparisons: the swapped faces of the six baselines are affected by the expression of the source face to some extent which is sufficient proof of the point we mentioned that the face identity embedding could be easily affected by facial attribute information.

4.3.2. ***Quantitative Comparison***

The quantitative comparisons only involve the four results-available or codes-available methods Deepfakes, FaceSwap, FSGAN, and FaceShifter. Following (Li et al., 2019), we construct the test set of 10000 images by sampling ten frames from each of 1000 resulting videos of FF++ for the above four methods and ours, respectively. The quantitative comparisons rely on these five test sets. To measure our proposed method's effectiveness in identity consistency and expression consistency, we adopt the identity retrieval accuracy and expression error as in Section 4.2.2.

The quantitative results are shown in Table 3. Similar to the game character face swapping experiment results, we also get the highest identity retrieval accuracy, lowest pose error, and lowest expression error for human face swapping. This means that our method with compound identity is more robust in identity transferring than single-identity-based methods (FaceShifter) and much better than those methods (Deepfakes, FaceSwap) without any identity constraint. Moreover, a fine-grained expression constraint contributes more to expression preservation than implicit constraint methods.

4.4. Ablation Study

We conduct several ablation settings on the game dataset to demonstrate the effectiveness of our framework.

Cross-domain Finetune: To verify the effectiveness of the finetune strategy, we train a model without finetuning (w/o FT). As shown in Figure 8 and Table 5, it can be observed that the finetune strategy significantly improves the quality of the generated image in both visual quality and FID. This proves that the finetune strategy contributes to cross-domain face swapping.

Expression Embedding Loss: To demonstrate the effectiveness of our expression embedding loss, we conduct an experiment setting without the expression loss (w/o Exp). Quantitative results in Table 4 show that the expression error rises a lot when using no expression loss. Observing Figure 7, the swapped faces without the expression loss tend to be influenced by the expression of the source face (marked in red boxes). Both quantitative and qualitative results validate the effectiveness of our fine-grained expression constraints in training. It also reflects that the identity embeddings still contain expression information.

Compound Identity: To evaluate the effectiveness of the compound strategy proposed by us. We trained another two models called *Single ArcFace* and *Single FaceNet*. The comparison results are shown in Table 4, comparing their ID Retrieval accuracy and Expression Error, the compound identity embeddings outperform the single identity embedding. This validates that compound identity embeddings can alleviate the effect of expression leaked in identity embedding and provide more robust identity information.

To further verify this conclusion, we conduct a visualization experiment by observing the distribution of 8 basic expressions from one identity, as shown in Figure 9. It can be observed that eight points with one color referring to one identity are closed to each other but have distances from each other. This explains that the single identity embedding contains other facial attributes, and thus the swapped expression may be affected by the source face.

It can also be found that the expression variances from different face recognition models are inconsistent. This is supported by the observation in Figure 9 (b) and Figure 9 (c), where the same color clusterings (marked in red and black circles) with FaceNet and ArcFace are shaped differently. So the compound identity embeddings can alleviate attributes' effect due to the inconsistency of expression variance and reduce the expression error.

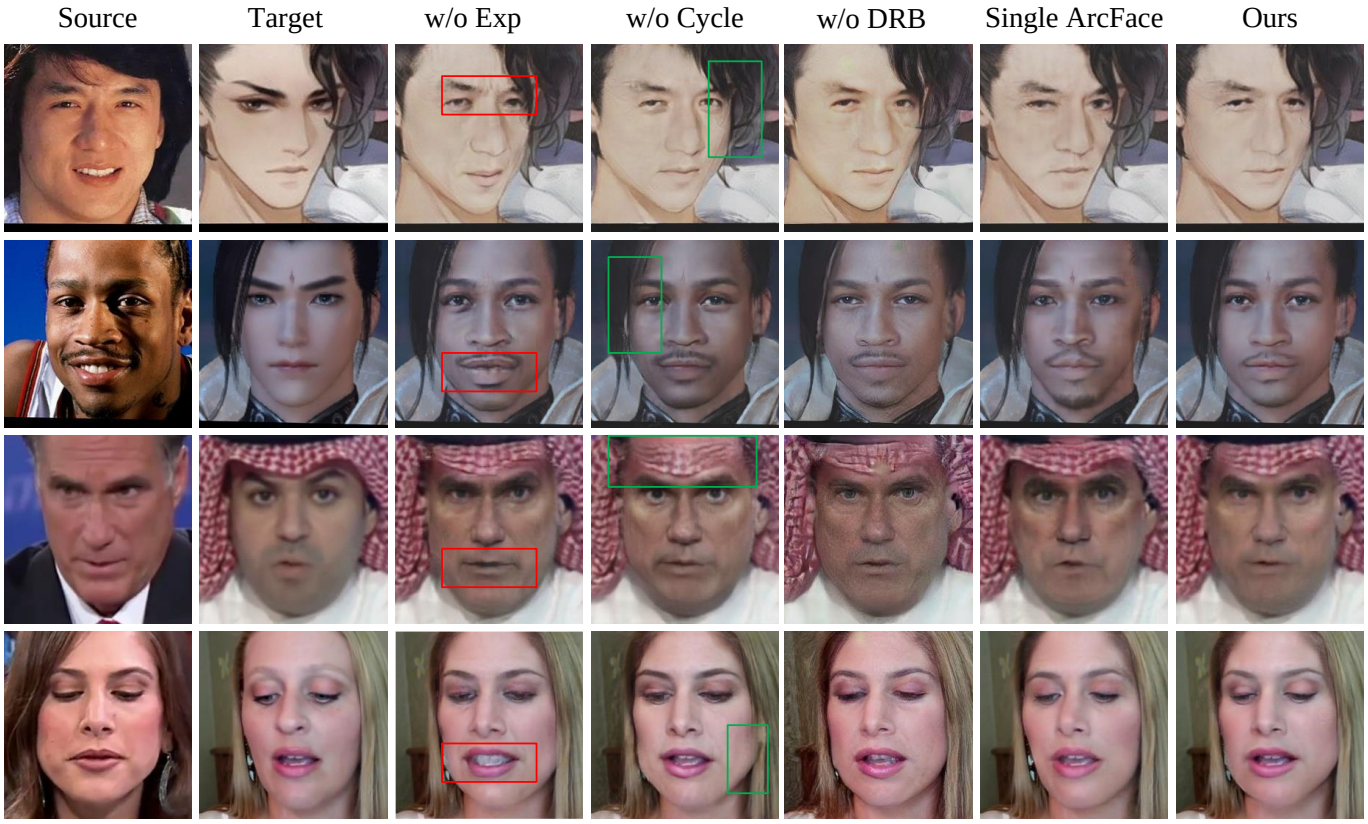


Fig. 7. Ablation study for each component in our framework. Some expression errors and occlusion errors are marked with red and green boxes, respectively. Please zoom in for more details.

Fig. 8. Ablation results for the finetune strategy.

Moreover, some expressions even change the identity of one person to another, as shown in Figure 9 (b). Therefore, a single identity embedding is not robust enough to achieve proper identity transfer, as shown in Figure 9 (a). However, our compound identity embeddings can avoid this problem.

Cycle Consistency Loss: To verify the effectiveness of the proposed Cycle Consistency Loss, we train the model without it (w/o Cycle). And we use the same Pose Error metric (the $\mathcal{L}_2$ distance between pose vectors extracted by HRNet (Sun et al., 2019)) as (Li et al., 2019) to evaluate the effectiveness of the cycle consistency loss. It can be observed from Table 4 and Figure 7 that the Pose Error rises a lot and some occlusions are missed (marked in green boxes) without the Cycle Consistency Loss. This proves that the Cycle Consistency Loss is very useful for preserving attributes of the target.

Detail Recover Block: To verify the effectiveness of the Detail Recover Block in our framework, we remove it from the full pipeline (w/o DRB). As shown in Table 4, all the attribute-related evaluation metrics (Pose Error and Expression Error) are worse than the full model. Besides, a lot of details (e.g., background details) are lost in the swapped image, as can be seen in the fifth column of Fig 7. This indicates that the multi-level attribute features through the skip connections transfer detailed information validating that the Detail Recover Block is beneficial.

4.5. Robustness

In this subsection, we will evaluate our proposed method's robustness. We conduct several experiments on more wild images, video sequences, and even unnatural images (facial landmark images). Furthermore, an identity interpolation experiment is also performed.

Wild Images: We first perform face swapping on more wild images. As shown in Figure 10 (a) and Figure 10 (b), these target faces come from different categories, including game character faces, and human faces, or art paintings, and all of these images are unseen during training. The results show that our method works well on all kinds of facial images and can preserve the target expression well when transferring identity, even if the expression of the source image and the target image are very different. Our method can also handle pose changes, different face skin colors, and different lighting.

Table 4. Quantitative ablation study.

Method	ID Retrieval Accuracy (%) ↑			Pose	Expression Error ↓		
	CosFace	ArcFace2	SphereFace	Error ↓	DFER	FECNet	DLN
w/o Exp	98.81	99.59	99.13	3.04	4.24	0.32	0.41
w/o Cycle	**98.95**	99.68	99.28	3.41	4.15	0.25	0.26
w/o DRB	98.85	**99.93**	**99.34**	3.54	4.00	0.26	0.35
Single ArcFace	95.70	99.45	97.59	2.68	3.63	0.20	0.23
Single FaceNet	76.29	64.60	74.88	**2.46**	3.68	0.21	0.25
Ours	98.76	99.71	99.15	2.63	**3.52**	**0.20**	**0.23**

Table 5. Ablation results for the finetune strategy.

Method	w/o FT	Ours
FID	43.56	**28.95**

Video Sequences: We then perform our face swapping method on video sequences. The swapped results are shown in Figure 11. It can be seen that the swapped faces not only preserve the identity and the expression well but also have a natural facial movement along with the time sequence. This shows that our face swapping method demonstrates promising performance when applied to videos.

Identity Interpolation: We further perform identity interpolation to explore how identity embeddings influence the face swapping results. Specially, we perform Spherical Linear Interpolation (Kremer, 2008) (Slerp) between two different identity embeddings:

$$f_{id} = Slerp(f_{id}^A, f_{id}^B; \alpha) = \frac{\sin((1-\alpha)\times\theta)}{\sin(\theta)} f_{id}^A + \frac{\sin(\alpha\times\theta)}{\sin(\theta)} f_{id}^B \quad (13)$$

where f_{id}^* are identity vectors and θ is the angle between the two vectors. It can be observed from Figure 12 that as α increases, the identity of the swapped face gradually transitions from A to B. This proves that our model can adapt well to the distribution of identity features.

Unnatural images: To further study our method, we swap source human faces to target unnatural faces (facial landmark images). As can be observed from Figure 13, the swapped faces properly reflect the identities of source faces and the attributes (expression and pose) of target faces. This reveals that our facial image encoder can extract correct attribute representations, and our decoder also synthesizes faces correctly without texture being consistent with target images.

4.6. Limitation

Finally, we evaluate our method on images in extreme situations. As shown in Figure 14, failure cases are produced by our method when the face is severely occluded or under extreme angles. The reasons for these cases mainly come from two aspects, 1) There are fewer such extreme faces in the training set, so the encoder cannot learn how to extract accurate attribute information from them. 2) Heavy occlusions or extreme poses can lead to inaccurate identity embeddings and expression embeddings, so the loss function based on these embeddings can not provide adequate supervision during training. 3) The temporal information is not taken into account to enhance the consistency across subsequent frames. In the future, we will further investigate the model of temporal information to enhance the video face-swapping quality.

5. Conclusion

This work proposes a new face swapping method for game character face swapping, allowing the game user to generate customized identity-specific game CG videos. We collect a game face dataset and propose a fine-tuning strategy to improve the image quality of cross-domain face swapping. Besides, we propose a novel expression loss to guide the expression of the swapped face image close to that of the target image. Moreover, the compound identity embeddings are proposed to alleviate individual face recognition models' unexpected bias (e.g. expression variance). Qualitative, quantitative experiments on both human data and game data show that the proposed method is well adapted to the problem of cross-domain face swapping and outperforms the state-of-the-art methods.

6. Ackonwledgement

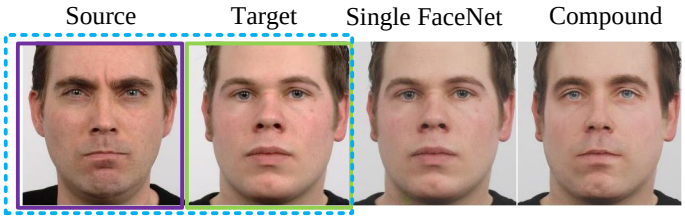
We thank all anonymous reviewers and ACs for their constructive suggestions. This work is supported by the 2022 Hangzhou Key Science and Technology Innovation Program (No. 2022AIZD0054), the Key Research and Development Program of Zhejiang Province (No. 2022C01011), and the National Key R&D Program of China (No. 2022YFF09022303). This research is funded in part by ARC (Australian Research Council)-Discovery grant (DP220100800 to XY) and ARC-DECRA (Discovery Early Career Researcher Award) grant (DE230100477 to XY). We thank all anonymous reviewers and ACs for their constructive suggestions.

References

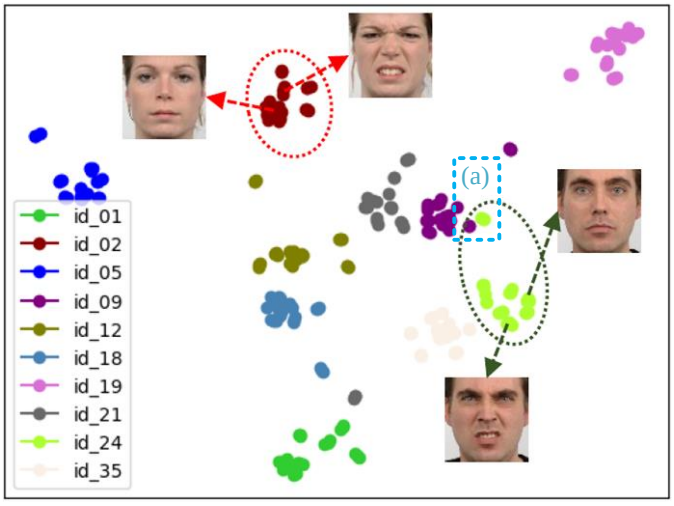
An, R., Zhang, W., Zeng, H., Chen, W., Deng, Z., Ding, Y., 2022. Global-to-local expression-aware embeddings for facial action unit detection. arXiv:2210.15160.

Bitouk, D., Kumar, N., Dhillon, S., Belhumeur, P., Nayar, S.K., 2008. Face swapping: automatically replacing faces in photographs, in: ACM SIGGRAPH, pp. 1--8.

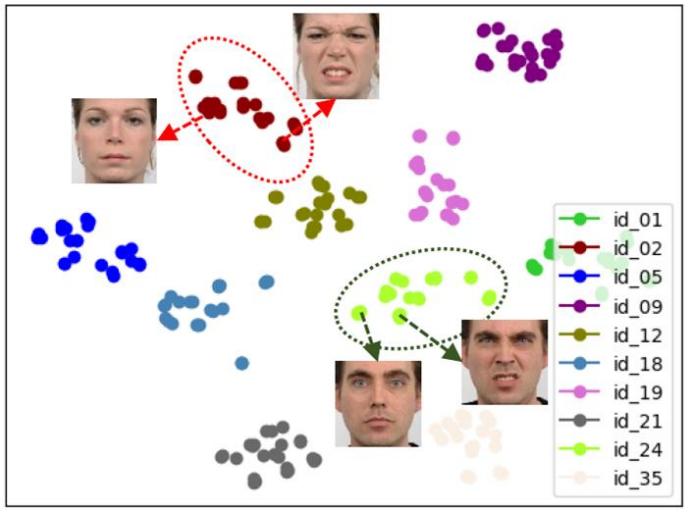
Blanz, V., Scherbaum, K., Vetter, T., Seidel, H.P., 2004. Exchanging faces in images, in: Computer Graphics Forum, Wiley Online Library. pp. 669--676.



(a) Compound identity embeddings avoid the failure case. The ID embeddings of the source and target face images are indicated in the below figure.



(b) Identity distribution of FaceNet.



(c) Identity distribution of Arcface.

Fig. 9. Visualization of the influence of facial expressions on identity embedding. Ten identities and eight emotions for each identity are sampled from RaFD (Langner et al., 2010).

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

Chen, D., Chen, Q., Wu, J., Yu, X., Jia, T., 2019a. Face swapping: realistic image synthesis based on facial landmarks alignment. Math. Probl. Eng. 2019.

Chen, R., Chen, X., Ni, B., Ge, Y., 2020. Simswap: An efficient framework for high fidelity face swapping, in: Proceedings of the 28th ACM International Conference on Multimedia, pp. 2003--2011.

Chen, Y., Wang, J., Chen, S., Shi, Z., Cai, J., 2019b. Facial motion prior networks for facial expression recognition, in: 2019 IEEE Visual Communications and Image Processing (VCIP), IEEE. pp. 1--4.

Choi, Y., Choi, M., Kim, M., Ha, J.W., Kim, S., Choo, J., 2018. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 8789--8797.

DeepFakes, 2019. Deepfakes. https://github.com/deepfakes/faceswap. Online; Accessed March 1, 2021.

Deng, J., Guo, J., Xue, N., Zafeiriou, S., 2019. Arcface: Additive angular margin loss for deep face recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4690--4699.

Ekman, P., Friesen, W.V., 1971. Constants across cultures in the face and emotion. J Pers Soc Psychol 17, 124.

Gao, G., Huang, H., Fu, C., Li, Z., He, R., 2021. Information bottleneck disentanglement for identity swapping, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3404--3413.

Goodfellow, I.J., Erhan, D., Carrier, P.L., Courville, A., Mirza, M., Hamner, B., Cukierski, W., Tang, Y., Thaler, D., Lee, D.H., et al., 2013. Challenges in representation learning: A report on three machine learning contests, in: International conference on neural information processing, Springer. pp. 117--124.

Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative adversarial networks. arXiv preprint arXiv:1406.2661 .

Happy, S., Routray, A., 2014. Automatic facial expression recognition using features of salient facial patches. IEEE Trans. Affective Comput. 6, 1--12.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770--778.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S., 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems 30.

Ioffe, S., Szegedy, C., 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift, in: International conference on machine learning, PMLR. pp. 448--456.

Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A., 2017. Image-to-image translation with conditional adversarial networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1125--1134.

Jiang, D., Song, D., Tong, R., Tang, M., 2023. Styleipsb: Identity-preserving semantic basis of stylegan for high fidelity face swapping, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.

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

Karras, T., Laine, S., Aila, T., 2019. A style-based generator architecture for generative adversarial networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4401--4410.

Kervadec, C., Vielzeuf, V., Pateux, S., Lechervy, A., Jurie, F., 2018. Cake: Compact and accurate k-dimensional representation of emotion. arXiv preprint arXiv:1807.11215 .

Khan, S., Chen, L., Yan, H., 2017. Co-clustering to reveal salient facial features for expression recognition. IEEE Trans. Affective Comput. 11, 348--360.

Kim, D.H., Baddar, W.J., Jang, J., Ro, Y.M., 2017. Multi-objective based spatio-temporal feature

(a) Results of frontal face images.

(b) Results of posed facial images.

Fig. 10. Face swapping results. We present the frontal face and posed face results. Please zoom in for more details.

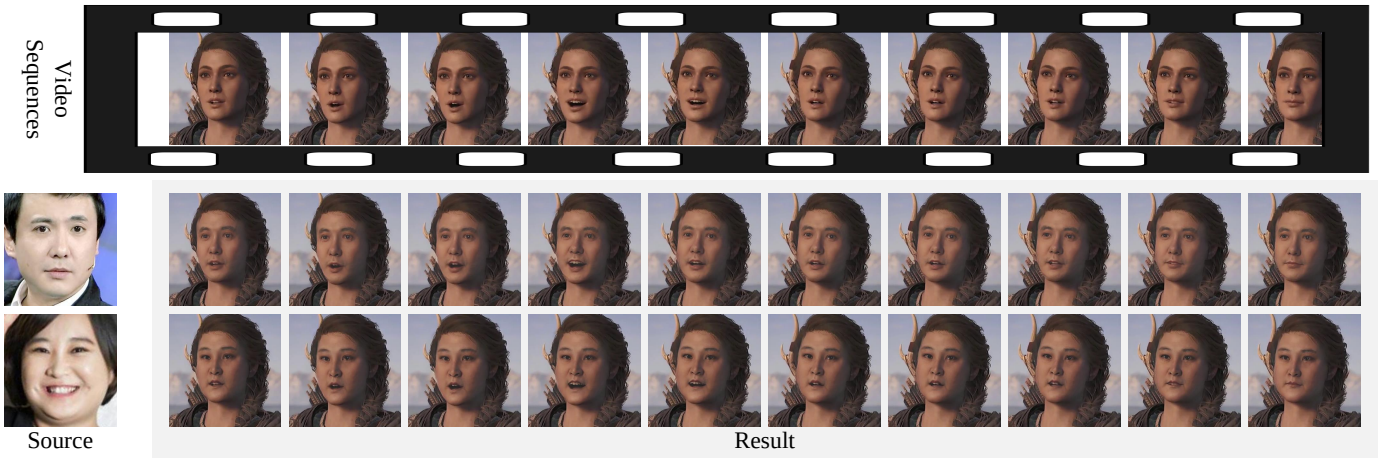


Fig. 11. Face swapping on video sequences. Please zoom in for more details.

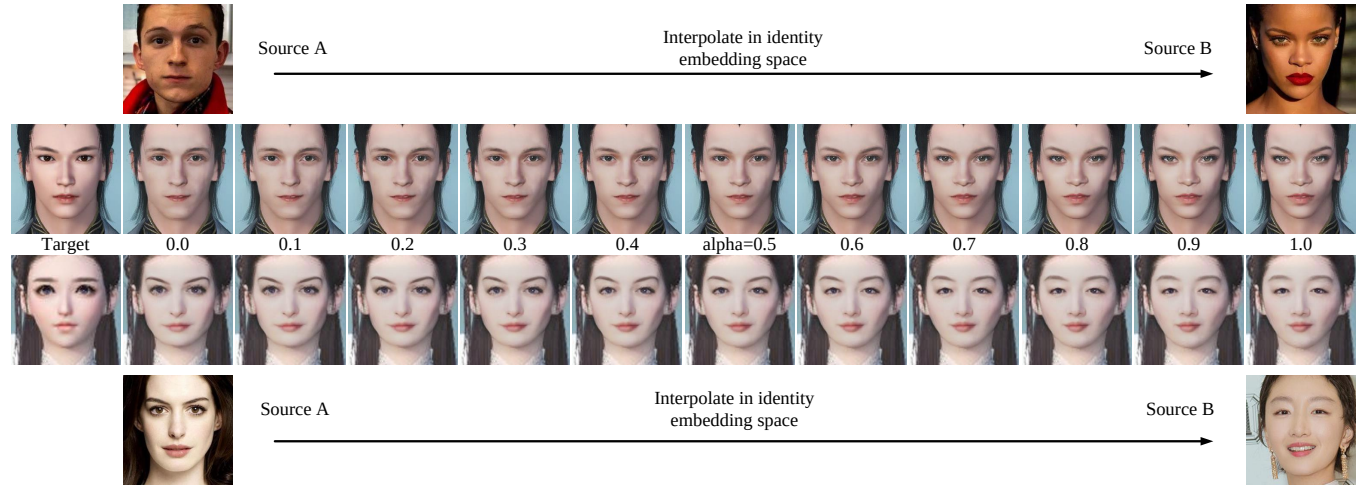


Fig. 12. The results of identity embedding interpolation results. During the interpolation process, our generated faces have semantically meaningful identity transitions. Please zoom in for more details.

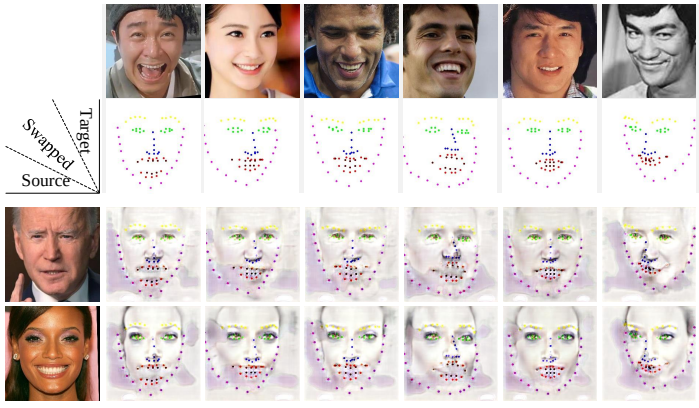


Fig. 13. Face swapping on facial landmark images. We utilize the facial landmarks as the target images, and the results indicate that our face swapping only changes the identity information without introducing other attributes.

representation learning robust to expression intensity variations for facial expression recognition. IEEE Trans. Affective Comput. 10, 223--236.

Kingma, D.P., Ba, J., 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 .

Kremer, V.E., 2008. Quaternions and slerp. Embots. dfki. de/doc/seminar_ca/Kremer_Quaternions. pdf .

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems 25, 1097--1105.

Langner, O., Dotsch, R., Bijlstra, G., Wigboldus, D.H., Hawk, S.T., Van Knippenberg, A., 2010. Presentation and validation of the radboud faces database. Cogn Emot 24, 1377--1388.

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

Li, M., Xu, H., Huang, X., Song, Z., Liu, X., Li, X., 2018. Facial expression recognition with identity and emotion joint learning. IEEE Trans. Affective Comput. .

Li, S., Deng, W., 2020. Deep facial expression recognition: A survey. IEEE Trans. Affective Comput. .

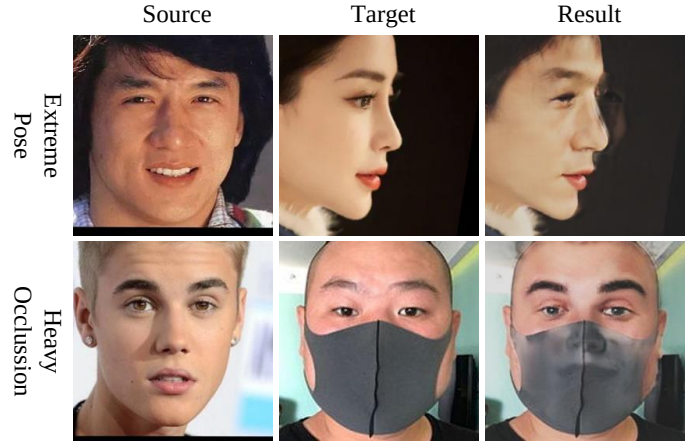


Fig. 14. The examples of failure cases. Heavy occlusions or extreme poses lead to inaccurate identity and expression embeddings and ultimately influence the swapped images.

Lim, J.H., Ye, J.C., 2017. Geometric gan. arXiv preprint arXiv:1705.02894 .

Lin, Y., Lin, Q., Tang, F., Wang, S., 2012. Face replacement with large-pose differences, in: Proceedings of the 20th ACM international conference on Multimedia, pp. 1249--1250.

Liu, W., 2018. Sphereface: Deep hypersphere embedding for face recognition. https://github.com/wy1iu/sphereface. Accessed March 1, 2021.

Liu, W., Wen, Y., Yu, Z., Li, M., Raj, B., Song, L., 2017a. Sphereface: Deep hypersphere embedding for face recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 212--220.

Liu, X., Vijaya Kumar, B., You, J., Jia, P., 2017b. Adaptive deep metric learning for identity-aware facial expression recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 20--29.

Liu, Z., Li, M., Zhang, Y., Wang, C., Zhang, Q., Wang, J., Nie, Y., 2023. Fine-grained face swapping via regional gan inversion.

Lucey, P., Cohn, J.F., Kanade, T., Saragih, J., Ambadar, Z., Matthews, I., 2010. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression, in: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops, pp. 94--101. doi:10.1109/CVPRW.2010.5543262.

MarekKowalski, M., 2021. Faceswap. [EB/OL]. https://github.com/MarekKowalski/FaceSwap Accessed March 1, 2021.

Mollahosseini, A., Hasani, B., Mahoor, M.H., 2017. Affectnet: A database for facial expression, valence, and arousal computing in the wild. IEEE Trans. Affective Comput. 10, 18--31.

Natsume, R., Yatagawa, T., Morishima, S., 2018. Fsnet: An identity-aware generative model for image-based face swapping, in: Asian Conference on Computer Vision, Springer. pp. 117--132.

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

Nguyen, D.H., Kim, S., Lee, G.S., Yang, H.J., Na, I.S., Kim, S.H., 2019. Facial expression recognition using a temporal ensemble of multi-level convolutional neural networks. IEEE Trans. Affective Comput. .

Nirkin, Y., Keller, Y., Hassner, T., 2019. Fsgan: Subject agnostic face swapping and reenactment, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 7184--7193.

Nirkin, Y., Masi, I., Tuan, A.T., Hassner, T., Medioni, G., 2018. On face segmentation, face swapping, and face perception, in: IEEE International Conference on Automatic Face & Gesture Recognition, IEEE. pp. 98--105.

Park, T., Liu, M.Y., Wang, T.C., Zhu, J.Y., 2019. Semantic image synthesis with spatially-adaptive normalization, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2337--2346.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S., 2019. Pytorch: An imperative style, high-performance deep learning library, in: Advances in Neural Information Processing Systems 32. Curran Associates, Inc., pp. 8024--8035.

Petrov, I., Gao, D., Chervoniy, N., Liu, K., Marangonda, S., Umé, C., Jiang, J., RP, L., Zhang, S., Wu, P., et al., 2020. Deepfacelab: A simple, flexible and extensible face swapping framework. arXiv preprint arXiv:2005.05535 .

Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., Nießner, M., 2019. FaceForensics++: Learning to detect manipulated facial images, in: International Conference on Computer Vision (ICCV).

Sun, K., Xiao, B., Liu, D., Wang, J., 2019. Deep high-resolution representation learning for human pose estimation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5693--5703.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z., 2016. Rethinking the inception architecture for computer vision, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2818--2826.

Thies, J., Zollhofer, M., Stamminger, M., Theobalt, C., Nießner, M., 2016. Face2face: Real-time face capture and reenactment of rgb videos, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2387--2395.

TreB1eN, 2018. Insightface pytorch. https://github.com/TreB1eN/InsightFace_Pytorch. Online; Accessed March 1, 2021.

Ulyanov, D., Vedaldi, A., Lempitsky, V., 2016. Instance normalization: The missing ingredient for fast stylization. arXiv preprint arXiv:1607.08022 .

Unity, 2021. The leading platform for creating interactive, real-time content. https://unity.com/. Online; Accessed July 16, 2021.

Vemulapalli, R., Agarwala, A., 2019. A compact embedding for facial expression similarity, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5683--5692.

Wang, H., Wang, Y., Zhou, Z., Ji, X., Gong, D., Zhou, J., Li, Z., Liu, W., 2018. Cosface: Large margin cosine loss for deep face recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 5265--5274.

Wang, K., Peng, X., Yang, J., Meng, D., Qiao, Y., 2020. Region attention networks for pose and occlusion robust facial expression recognition. IEEE Trans. Image Process. 29, 4057--4069.

Wang, M., 2018. Cosface pytorch. https://github.com/MuggleWang/CosFace_pytorch. Accessed March 1, 2021.

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

Woo, S., Park, J., Lee, J.Y., Kweon, I.S., 2018. Cbam:

Convolutional block attention module, in: Proceedings of the European conference on computer vision (ECCV), pp. 3--19.

WuJie, 2020. Facial-expression-recognition.pytorch. https://github.com/WuJie1010/Facial-Expression-Recognition.Pytorch. Online; Accessed March 1, 2021.

Xu, Z., Yu, X., Hong, Z., Zhu, Z., Han, J., Liu, J., Ding, E., Bai, X., 2021. Facecontroller: Controllable attribute editing for face in the wild. arXiv preprint arXiv:2102.11464 .

Zeng, H., Zhang, W., Chen, K., Zhang, Z., Li, L., Ding, Y., 2022. Paste you into game: Towards expression and identity consistency face swapping, in: 2022 IEEE Conference on Games (CoG), IEEE Press. p. 1{8.

Zeng, H., Zhang, W., Fan, C., Lv, T., Wang, S., Zhang, Z., Ma, B., Li, L., qiong Ding, Y., Yu, X., 2023. Flowface: Semantic flow-guided shape-aware face swapping .

Zhang, K., Zhang, Z., Li, Z., Qiao, Y., 2016. Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Process Lett. 23, 1499--1503.

Zhang, S., Zhu, X., Lei, Z., Shi, H., Wang, X., Li, S.Z., 2017. S3fd: Single shot scale-invariant face detector, in: Proceedings of the IEEE international conference on computer vision, pp. 192--201.

Zhang, W., Guo, Z., Chen, K., Li, L., Zhang, Z., Ding, Y., 2021a. Prior aided streaming network for multi-task affective recognitionat the 2nd abaw2 competition. arXiv preprint arXiv:2107.03708 .

Zhang, W., Ji, X., Chen, K., Ding, Y., Fan, C., 2021b. Learning a facial expression embedding disentangled from identity, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6759--6768.

Zhang, W., Qiu, F., Wang, S., Zeng, H., Zhang, Z., An, R., Ma, B., Ding, Y., 2022. Transformer-based multimodal information fusion for facial expression analysis, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2428--2437.

Zhang, Y., Zeng, H., Ma, B., Zhang, W., Zhang, Z., Ding, Y., Lv, T., Fan, C., 2023. Flowface++: Explicit semantic flow-supervised end-to-end face swapping. arXiv:2306.12686.

Zhu, H., Fu, C., Wu, Q., Wu, W., Qian, C., He, R., 2020. Aot: Appearance optimal transport based identity swapping for forgery detection, in: Neural Information Processing Systems (NeurIPS).

Zhu, Y., Li, Q., Wang, J., Xu, C.Z., Sun, Z., 2021. One shot face swapping on megapixels, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4834--4844.