IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 7 July 2022, accepted 21 July 2022, date of publication 5 August 2022, date of current version 12 August 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3196668

RESEARCH ARTICLE

GHOST—A New Face Swap Approach for Image and Video Domains

ALEXANDER GROSHEV, ANASTASIA MALTSEVA, DANIIL CHESAKOV, ANDREY KUZNETSOV, AND DENIS DIMITROV
Artificial Intelligence Research Institute (AIRI), 121170 Moscow, Russia
Corresponding author: Andrey Kuznetsov (kuznetsoff.andrey@gmail.com)

ABSTRACT Deep fake stands for a face swapping algorithm where the source and target can be an image or a video. Researchers have investigated sophisticated generative adversarial networks (GAN), autoencoders, and other approaches to establish precise and robust algorithms for face swapping. However the achieved results are far from perfect in terms of human and visual evaluation. In this study, we propose a new one-shot pipeline for image-to-image and image-to-video face swap solutions - GHOST (Generative High-fidelity One Shot Transfer). We take the FaceShifter (image-to-image) architecture as a baseline approach and propose several major architecture improvements which include a new eye-based loss function, face mask smooth algorithm, a new face swap pipeline for image-to-video face transfer, a new stabilization technique to decrease face jittering on adjacent frames and a super-resolution stage. In the experimental stage, we show that our solution outperforms SoTA face swap architectures in terms of ID retrieval (+1.5% improvement), shape (the second best value) and eye gaze preserving (+1% improvement) metrics. We also established an ablation study for our solution to estimate the contribution of pipeline stages to the overall accuracy, which showed that the eye loss leads to 2% improvement in the ID retrieval and 45% improvement in the eye gaze preserving.

INDEX TERMS Deep fake, face swap, GHOST, AEI-Net, eye loss, face mask smooth, stabilization, super resolution.

I. INTRODUCTION

Deep fake [1] is a technique of swapping an original face (target) with another one (source) in an image or video. Different deep fake synthesis approaches exist, when source and target data can be presented as image or video data. Different combinations of source and target data type become a starting point for various pipelines like entire face swap [2], attribute manipulation [3], identity swap [4] and expression swap [5]. High-quality synthesis methods can be applied in the movie industry, in extending face datasets used to train detection and recognition models, anti-spoofing attacks modeling, etc. One can also imagine many design applications, such as makeup manipulation, hair styling, and specific attribute modeling for different types of faces.

However there are a lot of face swap methods developed, a number of problems still exist which lead to visual artifacts. It also disturbs a feeling of face liveness when watching the swap results. Moreover moving from static to dynamic domain leads to new artifacts and problems in swap technology. It means that if a perfect image-to-image face swap algorithm is developed, it does not mean that it could be easily applied to the video domain. If one needs to create a perfect swap solution an additional fine tuning stage is needed [3]. Starting from this point we set the tasks and problems to solve during research. First, the proposed new solution should be single-shot, which means that we use a single source image to swap the target face in an image or a video without an additional training stage. The next point is to outperform the quality metrics of existing SoTA solutions and to maintain target face liveness when the swap process is finished. Finally, the deep fake generation performance should improve with our solution.

In this paper we propose a new face swap approach for image-to-video and image-to-image tasks - GHOST (Generative High fidelity One Shot Transfer). We take the FaceShifter

The associate editor coordinating the review of this manuscript and approving it for publication was Gianluigi Ciocca.

(image-to-image) architecture as a baseline and suggest several major changes. To improve the quality of identity transfer we add an eye loss function, which allows to preserve eye direction as in the target person. Moreover, we propose adaptive blending to transfer the shape of the source face more effectively. To decrease face jittering on adjacent frames in the image-to-video pipeline we develop a new stabilization technique. Besides, we apply super resolution post-processing to get a high quality transfer. As a result we have developed image-to-image and single-shot image-to-video pipelines with high quality of identity transfer.

The proposed paper structure is organized as follows. In the next section we provide a brief overview of two types of deep fake synthesis approaches and corresponding papers: identity-oriented and pretrained. Section III is devoted to the general description of the GHOST approach and includes the architecture details and loss function parts. In the next section we describe the general image-to-image swap pipeline, dive into details of image-to-video swap process, super resolution post processing step and the training stage details. In the Experiments section we describe the quality metrics, the details and results of the AEI-Net architecture research. We explored Identity and Attribute Encoders and AAD Generator. We also did the loss function research, compared with SoTA models which showed that our method outperformed existing solutions and evaluated the overall quality in terms of ablation study when we add eye gaze loss function. In order to estimate the visual quality of our solution several face swap results are also presented in this section. The final section is devoted to results and conclusions including future plans of our research.

Our contributions can be summarized as follows. First, we added several modifications to the loss function: eye gaze loss and reconstruction loss parts. Second, a new blending approach was designed and embedded in the face swap pipeline. Third, we developed a new stabilization technique to decrease face jittering on adjacent frames. Overall we proposed one-shot image-to-image and image-to-video face swap approaches which significantly improve quality and performance.

II. RELATED WORK

In recent years, many attempts have been made to solve high-quality face swaps. Beginning with some classic approaches [2], [6], which were good for that time in terms of quality and performance, but provided unstable results, they were significantly improved with the development of convolutional neural networks. All the methods can be divided into two main categories: identity-oriented, which can only transfer the identity of people, whose faces were used to train the model, and pretrained, which only needs to be pretrained once and after that could be applied to any pair of people.

A. IDENTITY-ORIENTED APPROACHES

One of the first well-known methods was Korshunova *et al.*'s [4] paper, where they used a multilevel convolutional network to create a CageNet, which allowed the transfer of the identity of Nicolas Cage to different images. The model achieved a higher quality level than the previous models, but the output of the model was still unstable in uncommon poses or harsh lighting conditions. Nirkin *et al.*'s method [7] showed good results using 3D segmentation masks of images for transfer. This method allowed us to obtain more photorealistic results using adversarial loss; however, the texture of the results was blurry, and uncommon angles of the face remained a key obstacle for stable generation. DeepFakes *et al.* [8] proposed a method that created an entire section of new approaches. The idea was based on the encoder–decoder architecture, where they trained two autoencoders with a common encoder for different identities and then used them to acquire a face swap. One of the current state of the art approaches, based on the idea of autoencoders is DeepFaceLab (DFL) [3], which developed this idea a lot and achieved a very good quality of the transfer. However, both encoder–decoder methods are restricted by the pair of people on which they are trained. Although the quality of the DFL faceswap is almost unlimited, it is a time-consuming operation to train a new model for each pair of persons, and the result strictly depends on the quality of the videos for these two persons. Another approach, based on the autoencoder concept, was introduced by Disney [9]. Their method was developed for 1024 × 1024 images, and also allowed the transfer of the identities of five different people to one another, but the old problem remained – the method required FullHD videos for both the source and target person of transfer for training.

B. PRETRAINED APPROACHES

Although identity-oriented approaches may produce qualitative results, it may be important to transfer new identities without additional training. Siarohin *et al.* [5] used affine transformations of segmentation masks to transfer one image to another and arrange a face swap. The main restrictions of this type of swap are large head turns and different face shapes of the source and target, which cause many artefacts. With the rise of GANs, many GAN-based approaches have been put into face swap practice. Zakharov *et al.* [10] introduced a system for generating a new face image of a person based on a few images of him or her and landmarks for the new pose. However, the key limitation of this approach is that if the landmarks of the two persons differ significantly, the output quality may decrease. MegaFS [11] learns to find the corresponding latent code in the pretrained StyleGAN2 [12] latent space to generate a face swap image, which allows the generation of a high-resolution result but is restricted by the capabilities of StyleGAN2.

Based on state-of-the-art results, some core and good-quality face swap approaches include FaceShifter [13] and SimSwap [14]. In both architectures, the authors extracted identity information from the source image using the pretrained ArcFace model [15]. In FaceShifter they use source and target features from different layers to generate a

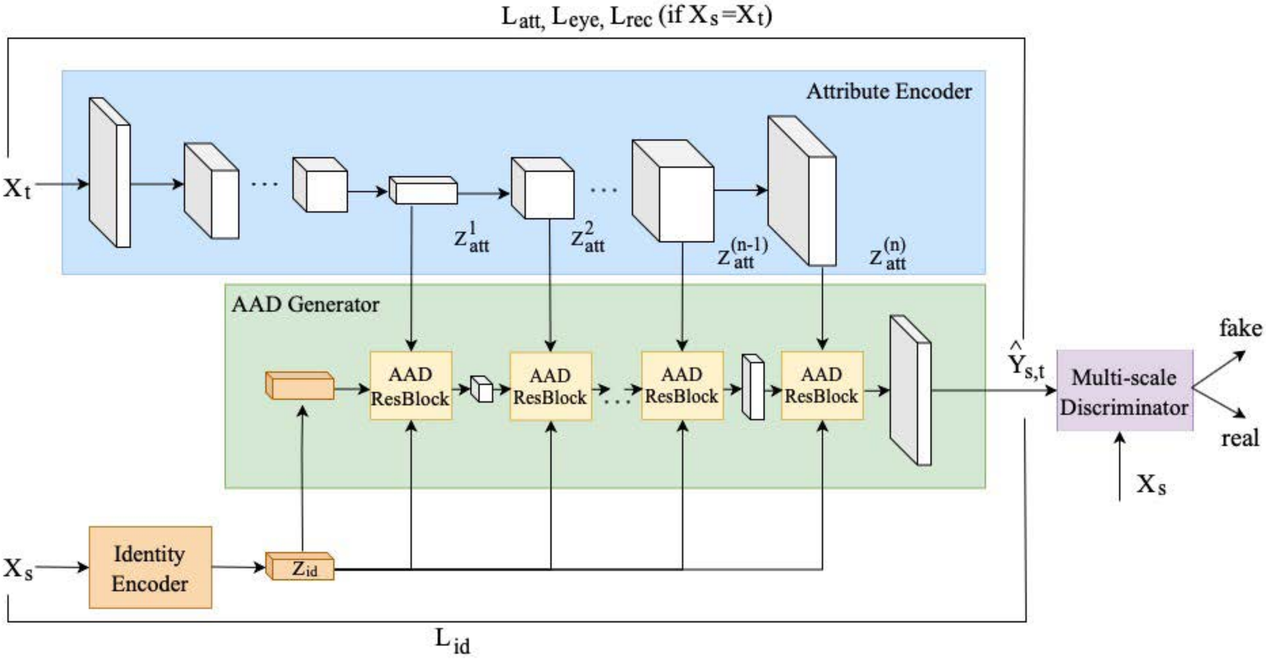

FIGURE 1. The general architecture of the GHOST approach based on the faceshifter model.

new image, while in the SimSwap the authors encode target image, blend this features with identity information and then decode it to get the final image. The HifiFace [16] solution uses a similar idea, introducing a 3D shape-aware identity extractor, which allows them to improve metrics of identity transfer, but in return to decrease the perception of the realism of the transfer.

III. THE GHOST APPROACH

Most of the state-of-the-art (SoTA) architectures have advantages and disadvantages. The disadvantages usually include different face edge errors, eye gaze inconsistency, and low quality, especially when swapping a face from a single image to a video. Therefore, we took the approach from the FaceShifter [13] model as a baseline, specifically the AEI-Net part, and implemented several new steps to improve the output results in terms of quality and stability.

Let X_s and X_t be the cropped faces from the source and target images, respectively. Let $\hat{Y}_{s,t}$ be the new generated face. X_s, X_t and $\hat{Y}_{s,t}$ share the same dimensions $256 \times 256 \times 3$. The architecture of the proposed approach presented in Fig. 1 consists of the following main parts:

1) The identity encoder is a pretrained ArcFace model that extracts the vector z_{id} with the size 1×512 from the source image X_s, which keeps information about a source person identity.
2) The attribute encoder is a model with a U-Net [17] architecture that extracts attribute features from the target image: X_t - $z_{att}^1, z_{att}^2, \ldots, z_{att}^n$.
3) The AAD generator is a model that sequentially mixes the attribute vector evaluated from X_t and the identity vector evaluated from X_s using AAD ResBlocks and generates a new face $\hat{Y}_{s,t}$ with source identity and target attribute features.
4) A multiscale discriminator [18] is a model that is utilised to improve the output synthesis quality by comparing real and fake images.

To outperform the baseline approach, we developed a loss function that contains several additional features which lead to better stability and higher swap accuracy. These improvements allow our model to achieve better performance in terms of the output quality. It should be noted that we carried out various experiments with the architecture itself and loss improvements, which are described further in the experiments section.

The general loss of the AEI-Net architecture part contains the following parts:

1) L_{rec} represents the reconstruction loss. We randomly use $X_s = X_t$ as the model input and require that the output value to be $\hat{Y}_{s,t} = X_t$:

$$L_{rec} = \begin{cases} \|\hat{Y}_{s,t} - X_t\|_2^2, & \text{if } X_s = X_t \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

2) L_{att} represents attribute loss. We require that $z_{att}^1, z_{att}^2, \ldots, z_{att}^n$ features for $\hat{Y}_{s,t}$ and X_t are close:

$$L_{att} = \sum_{k=1}^{n} \|z_{att}^k\left(\hat{Y}_{s,t}\right) - z_{att}^k\left(X_t\right)\|_2^2 \tag{2}$$

3) L_{id} represents identity loss. We assume that the identity encoder outputs for $\hat{Y}_{s,t}$ and X_s are similar in terms of

cosine similarity:

$$L_{id} = 1 - cos\left(z_{id}\left(\hat{Y}_{s,t}\right), z_{id}\left(X_s\right)\right) \tag{3}$$

4) L_{adv} represents the GAN loss based on multi-scale discriminator (adversarial loss).

Let us now proceed with our loss modifications. Firstly, we have modified the reconstruction loss using the idea from the SimSwap model [14]. In the original architecture of FaceShifter the idea of reconstruction loss is that if the model takes two identical images of a person as source and target, the output should be the same photo. However, $X_s = X_t$ does not require X_s and X_t belong to the same person $P^i \in P$, where P is the set of persons. Due to the fact that we used the datasets with several photos per person, it was possible to implement this loss modification:

$$L_{rec} = \begin{cases} \|\hat{Y}_{s,t} - X_t\|_2^2, & \text{if } X_s \in P^i \text{ and } X_t \in P^i \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

Another important component in the visual perception of face swap results is the transfer of gaze direction, especially when we use an image-to-video swap. In this case every single frame should represent the same direction of sight in order to establish a realistic perception. Therefore, we decided to add a special eye loss to correct the gaze direction. L_{eye} compared the eye heatmaps $hm(\cdot)$ between X_t and $\hat{Y}_{s,t}$ obtained using the face keypoint model [19] as follows:

$$L_{eye} = \|hm(\hat{Y}_{s,t}) - hm(X_t)\|_2^2 \tag{5}$$

The overall loss is written as:

$$L = \lambda_{adv}L_{adv} + \lambda_{att}L_{att} + \lambda_{rec}L_{rec} + \lambda_{id}L_{id} + \lambda_{eye}L_{eye} \tag{6}$$

IV. PROCESSING PIPELINE

A. GENERAL PIPELINE

When a model is trained, it can be used it to swap the face of a person from one image to another. However, as the model is trained on the cropped faces – we cannot just apply it to any images, we must first crop faces from both source and target images first. After we applied our model to the cropped images, we inserted the result of the swap back into the initial target image. The main problem here is that although we save the attributes of X_t on the $\hat{Y}_{s,t}$ – they are not exactly the same. Therefore, if we would insert $\hat{Y}_{s,t}$ directly back into the target image, we can clearly see the edges of such an operation. The solution we propose is quite simple: we first blend the output of the model $\hat{Y}_{s,t}$ with X_t, and then insert the result back into the target image. Sometimes we also apply post-processing techniques to the output of the model (e.g., super resolution) to improve the quality of the swap result. Therefore the general pipeline of the GHOST solution (shown in Fig. 2) is as follows:

1) Detect and crop faces for the source and target images using a pretrained face detector. Let X_s and X_t be the cropped faces.
2) Apply our model to X_s and X_t and obtain the result – $\hat{Y}_{s,t}$.
3) Apply post-processing steps to $\hat{Y}_{s,t}$ (e.g., super resolution).
4) Perform $\hat{Y}_{s,t}$ and X_t blending to get the final face $\hat{Y}^X_{s,t}$.
5) Insert the result $\hat{Y}^X_{s,t}$ back into the target image.

More details on how we process the images with several faces and how we implement $\hat{Y}_{s,t}$ and X_t blending with further result $\hat{Y}^X_{s,t}$ embedding back into the target image are described in the next section.

B. IMAGE-TO-VIDEO FACE SWAP

To establish a high-quality face swap from the source image to the target video, we proposed some steps that are described in this section. Face extraction was the first stage of our image-to-video pipeline. For each original frame, we must detect and identify crop-found faces. We used the SCRFD algorithm [20] as the default face detector. The proposed model could perform multiple face swaps. The user can select a specific person or set of persons on the video to which they want to transfer. In this case, the identity vectors are calculated for each target and detected face, and the cosine similarity between the vectors is considered to determine the correct person in the video. ArcFace [15] was used as a face recognition model. In the proposed pipeline, multiple-source face selection is supported for further processing.

Moreover, during the extraction step, we saved the transformation matrix for each face in the frame. This information is used to insert the face swap result back into the original frame; we call this process *blending*. However, if we insert the entire image obtained by our model, visual artefacts usually appear on the edge of the inserted area in the original frame and are visible. This effect occurs because of the incomplete correspondence of the brightness of the source image and target frame and because of the possible blurring of the image synthesized by our model. Therefore, it was necessary to ensure a smooth transition from the source image to the generated face. Therefore, we used segmentation masks.

A face mask is a binary image that determines which pixels belong to the face and which do not. Thus, we can determine the exact location of the face and perform precise contour cropping. To make the face insert seamless, we added Gaussian blurring to the edges. The results of these modifications are presented in Fig. 3.

When all faces are detected, we provide an opportunity to smooth the bounding boxes using information from adjacent frames. This function improves the stability of detection and removes the effect of face jittering in the resulting video. The pipeline stage scheme is shown in Fig. 4.

Moreover, we found that the shapes of the target and generated faces may not match, as the model attempts to maintain the shape of the source face X_s. To address this problem, we tracked the landmarks for the generated and target faces on the video. In the case of a significant difference in the coordinates of landmarks, we modified the face mask. If the face obtained by the model completely covered the face in the

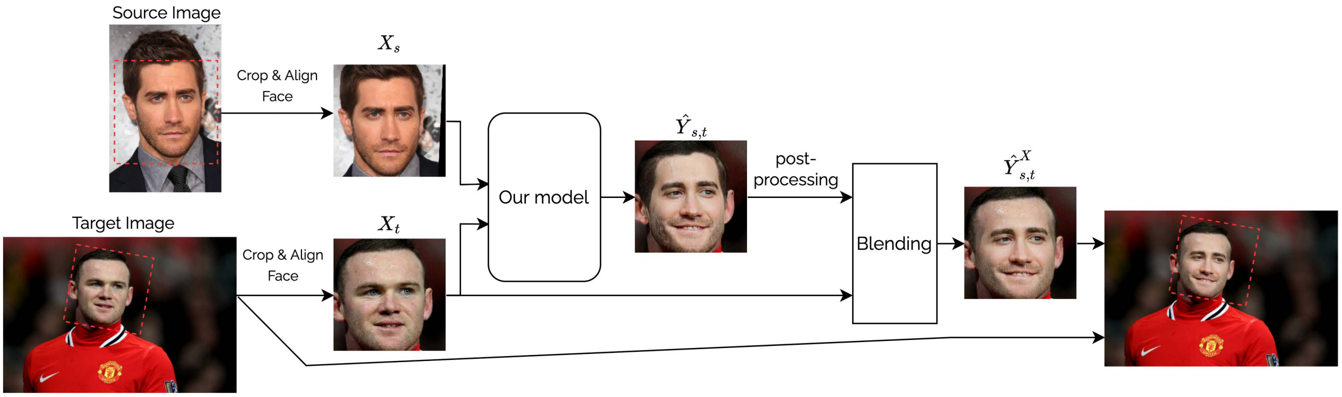

FIGURE 2. The general GHOST pipeline for image-to-image task.

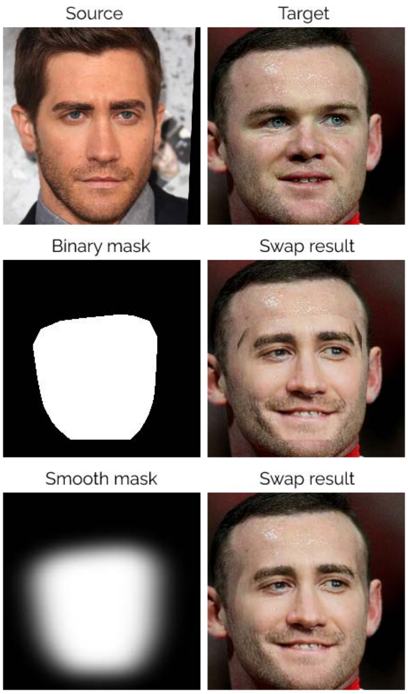

FIGURE 3. The initial source and target images (top row), the general binary mask and the swap result (middle row) and the result of swap with a smoothed binary face mask (bottom row).

video, we increased the mask, thereby creating the effect of transferring not only the face but also the shape of the head. Otherwise, we reduce the mask and increase the blurring effect to transfer only to the central parts of the face, such as the nose, mouth, and eyes.

In the next section, we describe the super resolution post-processing step that leads to significant quality improvement in comparison with existing SoTA approaches.

C. SUPER RESOLUTION POST-PROCESSING

We found that the generated image could look blurry when compared to the original video, so we decided to add a face enhancement module. This makes the facial features stronger and the composed image more natural. Therefore, we followed the face renovation approach proposed in [21]. We trained a neural network to restore the original quality of the degraded images. The experiments were performed on the FFHQ [22], which was resized to 256 × 256 pixels.

Our degradation function includes the following items:

- blurring the image;
- JPEG compression;
- random image downsampling.

The face enhancement module results are shown in Fig. 5.

D. TRAINING DETAILS

We trained our model using the VGGFace2 [23] dataset. To increase the training set quality, we removed all images with the sizes smaller than 250. For other images, we cropped and aligned the images to the resolution of 256.

The training stage contains 2 steps. As for the first step we disabled the eye loss and set other loss weights as follows: $\lambda_{adv} = 1$, $\lambda_{att} = 10$, $\lambda_{id} = 15$, $\lambda_{rec} = 10$. For the second training part, we changed λ_{id} to 70 and set the eye loss as $\lambda_{eye} = 1200$. We trained our model for 12 epochs within a batch size of 16. Training experiments were carried out on NVidia Tesla V100 32 GB GPU.

We set $X_s = X_t$ ratio to 0.2. The Adam optimizer was used with $\beta_1 = 0$, $\beta_2 = 0.999$ – these parameters are used to compute moving averages of gradient and its square. The learning rate was set to $4 \cdot 10^{-4}$ and the weight decay was set to 10^{-4} for both generator and discriminator.

V. EXPERIMENTS

In this section we first describe the metrics used and then proceed to the architecture and loss function research. Finally, we compare the GHOST model with SoTA models and conduct ablation study experiments.

A. METRICS

We used CosFace [24] as a face recognition model to evaluate the identity preservation measure for ID retrieval. We then calculated the cosine similarity and selected the nearest values for each pair of generated and original faces. The

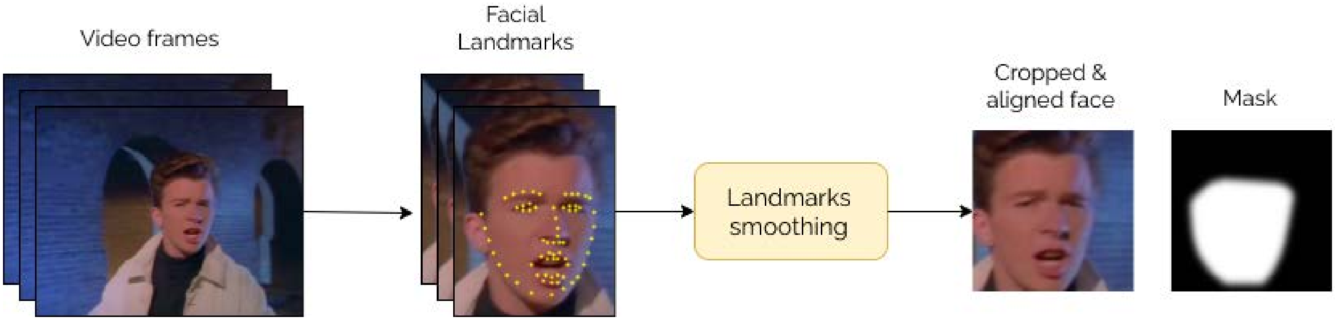

FIGURE 4. Source and target images (top row), the result of swapping using a binary face mask (middle row) and the result of our approach with smoothing the face mask (bottom row).

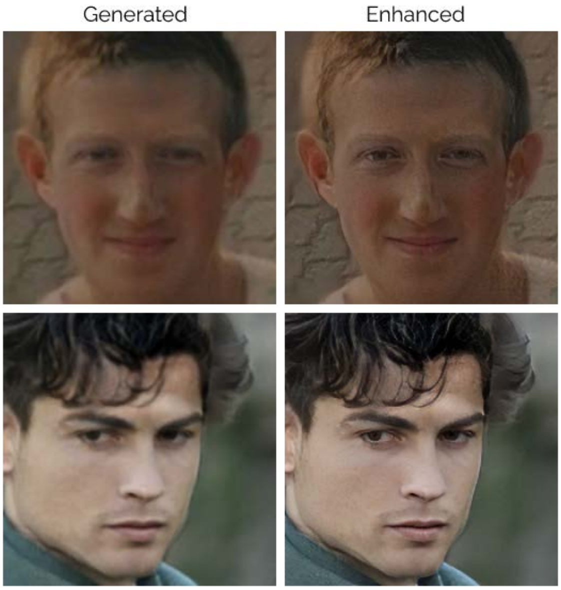

FIGURE 5. Examples of super resolution step execution: input low resolution faces (left column) and resolution enhancement result (right column).

ID retrieval measure evaluates whether a source is correctly preserved.

Further, we used the 3D face reconstruction model Ringnet [25] and pose estimator [26] to evaluate the shape, expression and pose metrics. The shape metric was used to compare the shapes and identities of X_s and $\hat{Y}_{s,t}$ face areas. The expression metric is responsible for preserving the facial expressions and emotions between X_t and $\hat{Y}_{s,t}$. The pose and poseHN metrics evaluate the preservation of the head pose between X_t and $\hat{Y}_{s,t}$. All of these metrics are evaluated using the L2 distance.

The effectiveness of the gaze direction transfer is computed using the L2 distance between the eye landmarks evaluated for swapped and original faces.

B. ARCHITECTURE RESEARCH

We conducted experiments by modifying each component of the AEI-Net architecture, while retaining the entire model. Furthermore, we describe the obtained results.

1) IDENTITY ENCODER RESEARCH

The ArcFace [15] model is a SoTA model used to solve the task of obtaining an identity vector from a person image, so we decided to retain it. However, there are several different implementations of this model and we used the following settings:

1) Use different versions of ArcFace as the identity encoder.
2) Use the average of the identity vectors that are derived from the output of different versions of the ArcFace model.

Overall, we decided to use a single version of ArcFace [15], because using multiple encoders to produce vectors with further averaging operations makes the transfer process more time-consuming and has a low quality impact.

2) ATTRIBUTE ENCODER RESEARCH

The attribute encoder allows us to obtain face attribute features (such as pose, expression, and image colour) with different layers, providing the generator with detailed information about the target image. The following approaches were attempted:

1) Using U-Net to reduce the number of channels in the extracted feature maps allows to accelerate the model. Adding ResNet blocks to the U-Net encoder part did not improve the generation quality.
2) Using Linknet as an attribute encoder leads to a reduction in the feature dimension and increases the inference rate of the model because feature maps in the decoder are summarised and not concatenated. Reducing the dimensions of the attributive features improves the identity of the swap result images. However, in the video, we obtained slightly unstable generation.
3) Using ResNet as an attribute encoder allows a lightweight architecture to be obtained. In this experiment, we rejected the idea of an encoder-decoder architecture but took feature maps at different layers of ResNet. As a result, ResNet provides more information about attributes from the target image compared to encoder-decoder architectures. During the face swap process, this encoder allows the preservation of various facial details, such as a quiff and ears.

TABLE 1. Comparison of different attribute encoders after post-processing (blending).

Method	ID retrieval	Shape (ringnet)	Exp (ringnet)	Pose (ringnet)	PoseHN	Eye ldn
GHOST (ResNet)	89.9	**0.62**	0.44	**0.045**	2.41	1.92
GHOST (LinkNet)	90.2	0.63	0.51	0.057	3.09	**1.91**
GHOST (U-Net)	**90.61**	0.64	**0.436**	0.047	**2.26**	2.02

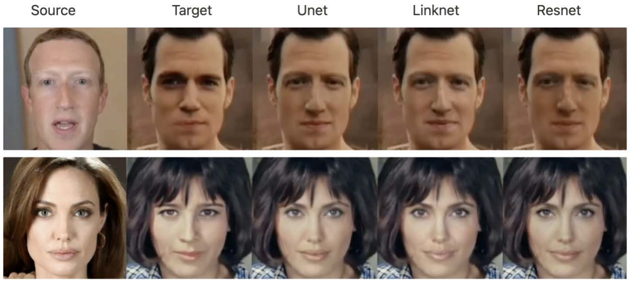

FIGURE 6. The GHOST model results with different attribute encoders after post-processing (blending).

TABLE 2. Comparison of different number of AAD U-Net blocks after post-processing (blending).

Method	ID retrieval	Shape (ringnet)	Exp (ringnet)	Pose (ringnet)	PoseHN	Eye ldn
GHOST (1 block)	89.92	0.64	0.48	0.048	**2.23**	2.17
GHOST (2 blocks)	90.61	0.64	**0.436**	**0.047**	2.26	**2.02**
GHOST (3 blocks)	**91.74**	**0.61**	0.55	0.057	2.69	2.45

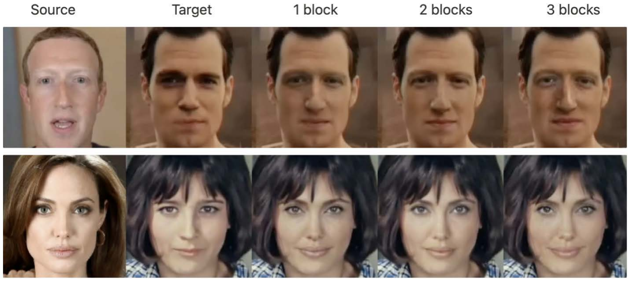

FIGURE 7. The GHOST model output results achieved when using various number of AAD blocks in the generator's AAD ResBlock after blending.

Table 1 and Fig. 6 show the GHOST model results with the use of different attribute encoders after post-processing (blending). U-Net based attribute encoder achieved the highest values for three metrics in comparison with the other two encoders. According to this we selected U-Net for our solution. It should be mentioned that increasing the number of AAD blocks leads to the preservation of more features of X_t which makes the face swap result far from X_s. However, we want to solve the task of higher result similarity to the source face.

3) AAD GENERATOR RESEARCH

The AAD generator mixes attribute features from X_t, the identity vector of X_s and the results of previous generation steps using AAD ResBlocks. Every AAD ResBlock consists of several AAD blocks, in which the AdaIN and Spade blocks extract information from the identity and attribute features, respectively. During the experiments, we tested several hypotheses and obtained the following results:

1) Varying the number of AAD blocks in the AAD ResBlock improved the overall model quality. The use of fewer blocks speeds up the model significantly but results in a small reduction in the model quality. To facilitate the training of a lightweight version of the model, knowledge distillation was applied. L2 and perceptual loss require that the output of the lighter version of the model be close to the output of the original model. When using more blocks, we could improve the generation quality and identity transfer; however, the model performance decreased.
2) The use of AdaConv [27] and Attention [28] in the AAD block instead of AdaIN did not lead to any quality improvements; therefore, we settled on the same AdaIN and Spade blocks as originally.

Table 2 shows metrics for models with different number of AAD blocks in the generator's AAD ResBlock after blending. Fig. 7 shows examples of all three versions of the model. It can be seen that 2 AAD U-Net blocks lead to the best values for three metrics in comparison with the other number of blocks.

C. LOSS FUNCTION RESEARCH

To correct the eye gaze we tried several approaches:

1) Compare the eye heatmaps between X_t and $\hat{Y}_{s,t}$ obtained by the face keypoint model using the L2 measure – L_{eye} loss.
2) Compare the eye gaze direction between X_t and $\hat{Y}_{s,t}$ using L2 measure.
3) The eye areas obtained using masks are compared between X_t and $\hat{Y}_{s,t}$ using the L1 measure.
4) The eye areas are compared between X_t and $\hat{Y}_{s,t}$ using the eye discriminator to improve the quality of eye generation.

During the experiments, we settled on the first approach, as it provides significant improvements in terms of realism and keeping the gaze direction stable as it is in X_t.

In addition, we tried to add losses to improve the identity transfer:

1) We pretrained the classifier on VggFace2, which determines the person class from the image. Then, the classifier is used to predict the class labels for X_s and $\hat{Y}_{s,t}$, which are compared using cross-entropy loss.
2) We trained a classifier for image pairs $(X_{s1}, X_{s2}) \rightarrow 1$ and $(X_{s1}, X_t) \rightarrow 0$. It should be learned to distinguish whether there is one person in the two pictures

TABLE 3. Quantitative comparison with SoTA on FaceForensics++ dataset.

Method	ID retrieval	Shape (ringnet)	Exp (ringnet)	Pose (ringnet)	PoseHN	Eye ldmk
FaceSwap (2016) [30]	58.82	0.75	**0.305**	0.045	1.94	4.22
DeepFakes (2018) [8]	72.42	0.65	0.696	0.110	7.35	11.6
FaceShifter (2019) [13]	82.02	0.67	0.420	0.040	**1.93**	2.48
SimSwap (2021) [14]	87.42	0.72	0.340	**0.035**	2.13	2.91
HifiFace (2021) [16]	89.17	**0.64**	0.510	0.048	2.13	2.04
GHOST	**90.61**	**0.64**	0.436	0.047	2.26	**2.02**

TABLE 4. Quantitative comparison with SoTA models.

Method	ID retrieval	Pose
FaceSwap (2016) [30]	54.19	2.51
DeepFakes (2018) [8]	77.65	4.59
FaceShifter (2019) [13]	97.38	2.96
SimSwap (2021) [14]	92.83	**1.53**
HifiFace (2021) [16]	98.48	2.63
GHOST	**98.67**	3.00

or not. Then, using the pretrained classifier, we wanted to obtain $(X_{s1}, \hat{Y}_{s,t}) \rightarrow 1$ and $(X_t, \hat{Y}_{s,t}) \rightarrow 0$ for the generator.

3) We also attempted to train an identity discriminator. This idea is similar to the previous one, but here the discriminator is trained in the process.

Unfortunately, the identity losses yielded little or no improvement in quality.

D. COMPARISON WITH SoTA MODELS

As a result of the proposed GHOST architecture and pipeline research, we performed a comparison with the existing SoTA face swap models. In order to evaluate our model quality, we reproduced the experiment proposed by the FaceShifter model authors in [13]. Taking into account every face swap method presented in FaceForensics++ [29] dataset, we collected a test set by randomly picking 10 frames from each video. When comparing the SimSwap model with ours, we generated face swap datasets based on the same pairs of source and target images as it was done by the SimSwap authors. Both models were compared using the following metrics: ID retrieval, pose, shape and expression errors, and gaze direction. The comparison results are presented in Tables 3 and 4 (for each method, the publication year or the GitHub model upload year is given).

It can be seen that our solution achieved the highest scores for ID retrieval, shape, and eye gaze metrics. In terms of expression and pose values, our model is very close to the best scores. However, there are ways to improve it in the future. Overall, the output image results achieved by the proposed model look very natural and do not discourage an independent viewer.

TABLE 5. Average inference time on a 10 second video.

Video resolution	Frames per second	Inference time
1280 × 720	30	15.8 s
1920 × 1080	30	23.7 s

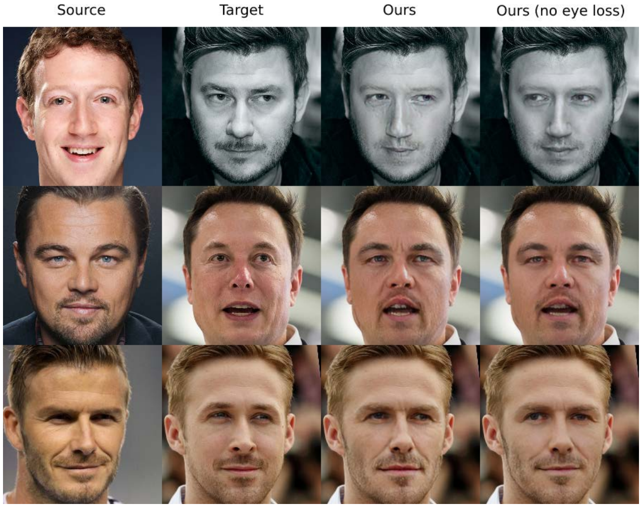

FIGURE 8. Comparison between our models with and without eye loss.

TABLE 6. Quantitative experiments on FaceForensics++ dataset.

Method	ID retrieval	Eye ldn
GHOST	90.61	2.02
GHOST (no eye loss)	88.79	3.70

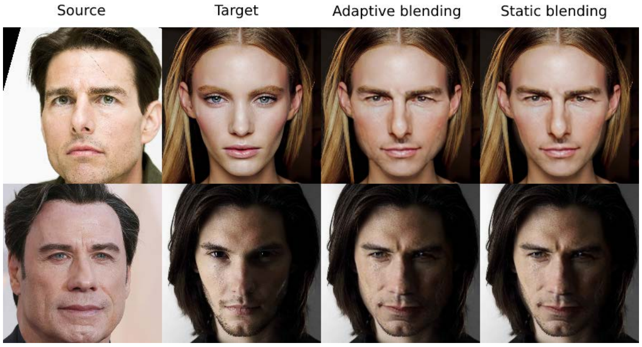

FIGURE 9. Comparison between blending methods.

E. INFERENCE TIME

In Table 5 we show the average GHOST model inference time for different video resolutions. The inference was performed on a single NVidia Tesla V100 32 GB GPU for a 10 second video. Since our approach works image-to-image, the inference time will change in direct proportion to the number of frames per second. The table shows the time of the full model run, taking into account the source and target preprocessing and the compilation of the final video.

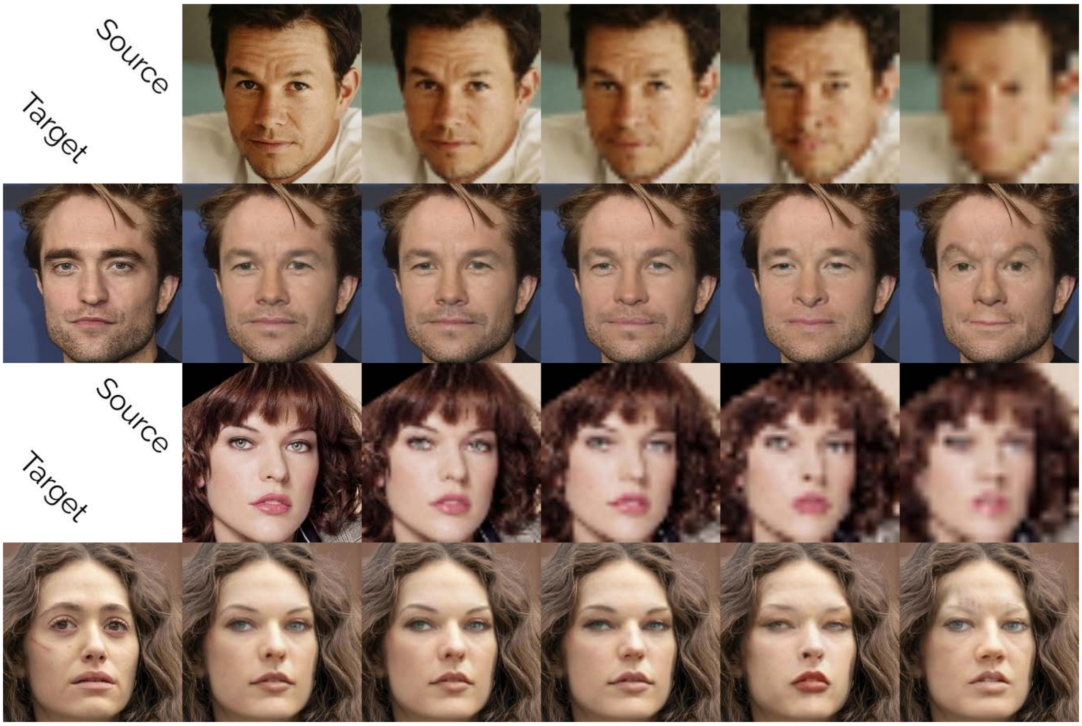

FIGURE 10. GHOST results on images with different resolutions. Source image was taken in original size and 3x, 5x, 7x and 9x downscaled versions.

TABLE 7. Comparison of blending methods on FaceForensics++ dataset.

Method	ID retrieval	Shape
GHOST (adaptive blending)	90.61	0.64
GHOST (static blending)	87.61	0.65

F. ABLATION STUDY

In this section, we analyze the proposed features of the GHOST pipeline. We attempted to estimate the efficiency of the L_{eye} function and blending using face masks.

We noticed that eye gaze direction has a significant impact on the visual quality of the model; therefore, we decided to add a loss function based on the distance between the pupils of the original and generated images. To evaluate the effectiveness of the proposed loss function in the gaze direction, we trained the model with and without this loss function (see Fig. 8). Table 6 presents a quantitative comparison.

We have introduced an adaptive blending algorithm that attempts to transfer the shape of a source's face. This is an improvement of the static algorithm, when the face mask is built without considering the difference in shape of the source and target faces. We calculated the metrics for the adaptive and static blending methods (Fig. 9), the results are presented in Table 7.

G. LIMITATIONS

We also analyzed the GHOST model for image resolution limitations. In the case of a low quality or blurry image, the visual identity of the person becomes weaker. The source identity information is contained as a vector encoded by the ArcFace model [15]. In case of a low quality image, the vector distorts and retains the identity information less accurately. As a result, the generated face will be less similar to the source, as shown in Fig. 10.

VI. CONCLUSION

In this paper, we proposed a new pipeline for image-to-image and image-to-video face swap – GHOST. The FaceShifter AEI-Net part is used as the baseline. We further implemented several upgrades and proposed a final solution that includes stabilization improvements for deep fake video synthesis. The proposed new steps and the final new deep fake synthesis pipeline provide high quality deep fake images and video, which is verified by the experimental results. We provided the following upgrades that lead to the quality improvement.

First, we added several modifications to the loss function: eye gaze loss and reconstruction loss parts. Second, a new blending approach was evaluated and embedded in a face swap pipeline. Third, we developed a new stabilization technique to decrease face jittering on adjacent frames. As a result, we developed a new image-to-image and a single-shot image to video face swap pipelines with several post-processing steps. In the experimental stage, we compared our solution with existing SoTA architectures and obtained the highest scores in terms of ID retrieval (90.61), shape (0.64), and eye gaze (2.02) metrics. In order to evaluate the impact of the eye gaze loss we did the ablation study for this specific architecture upgrade. The study showed that the eye gaze loss leads to the quality improvement in terms of ID retrieval metric by 2%. The inference performance of our

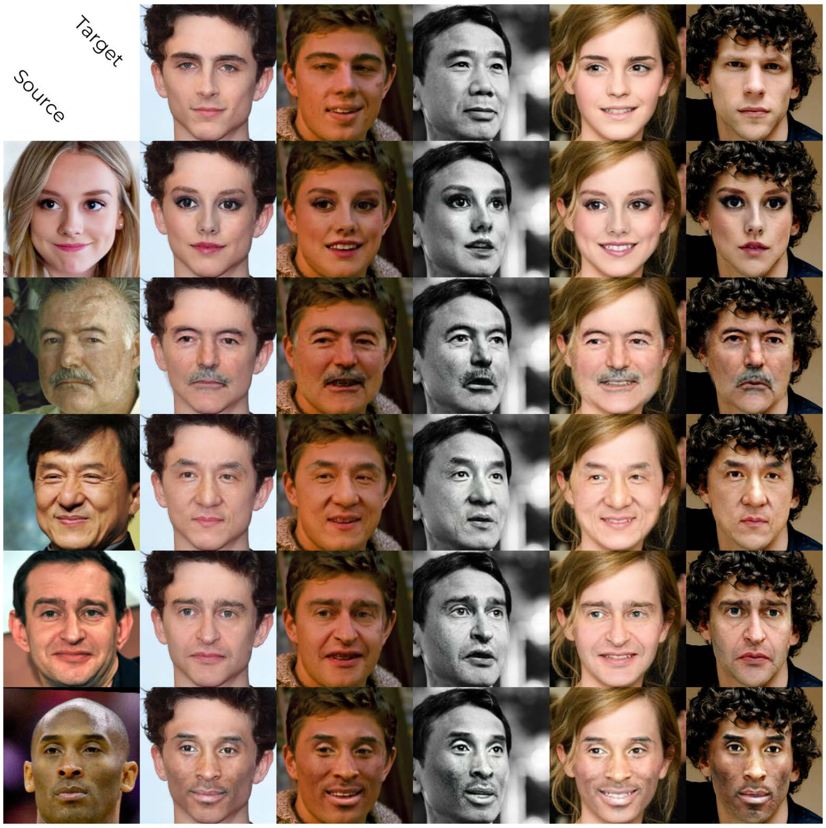

FIGURE 11. The output results matrix for a list of source and target images.

solution for a FullHD video (1920 × 1080) with a single face to be replaced is 12 FPS in average.

Finally, we did some visual evaluation of the generated swap results. Several examples of the proposed new face swap pipeline output are shown in Fig. 11.

In terms of future research we consider the fine tuning process of our solution in a GAN pipeline, where the proposed architecture will be used as a generator and SoTA deep fake detection models will be used as a discriminator. We expect to obtain the quality impact for general and difficult cases (lighting conditions, extreme head position, etc.). Another future tasks concern adding specific attributes functionality to the face swap process and implementing CLIP-like (Contrastive Language-Image Pretraining) functionality to our pipeline to control the face swap process by some text prompt (e.g., using the text prompt "make lips red" we can obtain the blending result with red lips even if they were not red in the source image).

The results of our research are presented as open-source by the codebase and the trained model which are available at our GitHub repo and can be used for research purposes.

REFERENCES

[1] T. T. Nguyen, Q. V. H. Nguyen, D. T. Nguyen, D. T. Nguyen, T. Huynh-The, S. Nahavandi, T. T. Nguyen, Q.-V. Pham, and C. M. Nguyen, "Deep learning for deepfakes creation and detection: A survey," 2019, *arXiv:1909.11573*.

[2] D. Bitouk, N. Kumar, S. Dhillon, P. Belhumeur, and S. K. Nayar, "Face swapping: Automatically replacing faces in photographs," *ACM Trans. Graph.*, vol. 27, no. 3, pp. 1–8, Aug. 2008.

[3] I. Perov, D. Gao, N. Chervoniy, K. Liu, S. Marangonda, C. Umé, M. Dpfks, C. S. Facenheim, R. P. Luis, J. Jiang, S. Zhang, P. Wu, B. Zhou, and W. Zhang, "DeepFaceLab: Integrated, flexible and extensible face-swapping framework," 2020, *arXiv:2005.05535*.

[4] I. Korshunova, W. Shi, J. Dambre, and L. Theis, "Fast face-swap using convolutional neural networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3677–3685.

[5] A. Siarohin, S. Roy, S. Lathuilière, S. Tulyakov, E. Ricci, and N. Sebe, "Motion-supervised co-part segmentation," 2020, *arXiv:2004.03234*.

[6] V. Blanz, K. Scherbaum, T. Vetter, and H.-P. Seidel, "Exchanging faces in images," in *Computer Graphics Forum*, vol. 23. Hoboken, NJ, USA: Wiley, 2004, pp. 669–676.

[7] Y. Nirkin, Y. Keller, and T. Hassner, "FSGAN: Subject agnostic face swapping and reenactment," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 7184–7193.

[8] *DeepFakes*. Accessed: Jun. 21, 2022. [Online]. Available: https://github.com/deepfakes/faceswap

[9] J. Naruniec, L. Helminger, C. Schroers, and R. M. Weber, "High-resolution neural face swapping for visual effects," *Comput. Graph. Forum*, vol. 39, no. 4, pp. 173–184, 2020.

[10] E. Zakharov, A. Shysheya, E. Burkov, and V. Lempitsky, "Few-shot adversarial learning of realistic neural talking head models," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9459–9468.

[11] Y. Zhu, Q. Li, J. Wang, C. Xu, and Z. Sun, "One shot face swapping on megapixels," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 4834–4844.

[12] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of StyleGAN," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8110–8119.

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

[14] R. Chen, X. Chen, B. Ni, and Y. Ge, "Simswap: An efficient framework for high fidelity face swapping," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 2003–2011.

[15] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4690–4699.

[16] 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," 2021, *arXiv:2106.09965*.

[17] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," 2015, *arXiv:1505.04597*.

[18] T. Park, M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu, "Semantic image synthesis with spatially-adaptive normalization," 2019, *arXiv:1903.07291*.

[19] X. Wang, L. Bo, and L. Fuxin, "Adaptive wing loss for robust face alignment via heatmap regression," 2019, *arXiv:1904.07399*.

[20] J. Guo, J. Deng, A. Lattas, and S. Zafeiriou, "Sample and computation redistribution for efficient face detection," 2021, *arXiv:2105.04714*.

[21] L. Yang, S. Wang, S. Ma, W. Gao, C. Liu, P. Wang, and P. Ren, "HiFaceGAN: Face renovation via collaborative suppression and replenishment," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 1551–1560.

[22] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," 2018, *arXiv:1812.04948*.

[23] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "VGGFace2: A dataset for recognising faces across pose and age," 2017, *arXiv:1710.08092*.

[24] 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," 2018, *arXiv:1801.09414*.

[25] S. Sanyal, T. Bolkart, H. Feng, and M. J. Black, "Learning to regress 3D face shape and expression from an image without 3D supervision," 2019, *arXiv:1905.06817*.

[26] N. Ruiz, E. Chong, and J. M. Rehg, "Fine-grained head pose estimation without keypoints," 2017, *arXiv:1710.00925*.

[27] P. Chandran, G. Zoss, P. Gotardo, M. Gross, and D. Bradley, "Adaptive convolutions for structure-aware style transfer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 7972–7981.

[28] S. Liu, T. Lin, D. He, F. Li, M. Wang, X. Li, Z. Sun, Q. Li, and E. Ding, "AdaAttN: Revisit attention mechanism in arbitrary neural style transfer," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 6649–6658.

[29] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Niessner, "FaceForensics++: Learning to detect manipulated facial images," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1–11.

[30] *FaceSwap*. Accessed: Jun. 21, 2022. [Online]. Available: https://github.com/ondyari/FaceForensics/tree/master/dataset/FaceSwapKowalski

ALEXANDER GROSHEV works as a Research Fellow at AIRI. His research interests include machine learning, in particular computer vision tasks. His main investigation topics include generative adversarial networks, deep fake synthesis, and image forgery detection.

ANASTASIA MALTSEVA received the master's degree from the Higher School of Electronics and Computer Science, South Ural State University, in 2020. She works as a Research Fellow at AIRI. She has three publications indexed by Scopus. Her research interests include machine learning, in particular computer vision and multi-modal neural networks.

DANIIL CHESAKOV works as a Research Fellow at AIRI. His research interests include machine learning, in particular computer vision tasks and simulation tasks. His main investigation topics include generative adversarial networks, deep fake synthesis, and deep fake detection.

ANDREY KUZNETSOV received the Ph.D. degree, in 2013. He works as a CV Lead at AIRI and an Associate Professor with the Department of Geoinformatics and Information Security, Samara National Research University. He has 56 publications indexed in Scopus, including publications in Q1 journals and on such conferences as ICPR and ICIAR. His research interests include image processing and machine learning algorithms development in remote sensing data analysis, digital image forgery detection, and algorithm theory application in comprehensive computer vision multi-domain methods design.

DENIS DIMITROV works as a CV & Multimodal Research Lead at AIRI and a Researcher with the Department of Probability Theory, Faculty of Mechanics and Mathematics, Lomonosov Moscow State University. He has a number of publications indexed in Scopus, including publications in Q1 journals such as *Mathematics*, *Acta Mathematica Sinica*, and on such conferences as ICML, IJCAI, and ICDAR. His research interests include both strictly mathematical issues concerning statistical estimation of the f-divergences and applications, such as multivariate inhomogeneities detection, feature selection, handwritten text recognition, generative computer vision models, and multimodal models.