electronics

MDPI

Article

AmazingFS: A High-Fidelity and Occlusion-Resistant Video Face-Swapping Framework

Zhiqiang Zeng, Wenhua Shao *, Dingli Tong and Li Liu

School of Computer and Information and Engineering, Xiamen University of Technology, Xiamen 361024, China; 2222031149@stu.xmut.edu.cn (D.T.)
* Correspondence: 2222031146@stu.xmut.edu.cn

Abstract: Current video face-swapping technologies face challenges such as poor facial fitting and the inability to handle obstructions. This paper introduces Amazing FaceSwap (AmazingFS), a novel framework for producing cinematic quality and realistic face swaps. Key innovations include the development of a Source-Target Attention Mechanism (STAM) to improve face-swap quality while preserving target face expressions and poses. We also enhanced the AdaIN style transfer module to better retain the identity features of the source face. To address obstructions like hair and glasses during face-swap synthesis, we created the AmazingSeg network and a small dataset AST. Extensive qualitative and quantitative experiments demonstrate that AmazingFS significantly outperforms other SOTA networks, achieving amazing face swap results.

Keywords: face swapping; attention mechanism; style transfer; face segmentation

check for updates

Citation: Zeng, Z.; Shao, W.; Tong, D.; Liu, L. AmazingFS: A High-Fidelity and Occlusion-Resistant Video Face-Swapping Framework. *Electronics* **2024**, *13*, 2986. https://doi.org/10.3390/electronics13152986

Academic Editor: George A. Tsihrintzis

Received: 19 June 2024
Revised: 19 July 2024
Accepted: 22 July 2024
Published: 29 July 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

The purpose of face-swapping technology is to combine the identity information of the source face, such as skin texture and facial contours, with the attribute information of the target face, such as posture, shape, expressions, and lighting, to generate a facial image that encapsulates these details [1]. The face-swapping result needs to retain the facial attributes of the target face while transferring the identity of the source face to the target face. In recent years, automatic and realistic video face-swapping technology has garnered increasing interest due to its wide applications in fields such as movies, games, and the entertainment industry. However, accurately and realistically extracting and integrating the identity information of the source face with the attribute characteristics of the target face to achieve high-fidelity face swapping remains a challenging issue.

Video face-swapping technology is distinct from traditional face swapping as it is always conducted on a one-to-one basis using state-of-the-art (SOTA) methods such as Deepfake, FaceSwap, DeepFaceLab, Simswap, Inswapper, and BlendSwap [2–7]. Even with 300,000 iterations of training using these methods, achieving high-fidelity face swapping remains challenging. In scenes with obstructions, problems like poor edge fitting of the face, artifacts, and distortions can still arise. The goal of face swapping technology is not only to achieve high realism in the swapped faces, but also to ensure that (1) the identity of the resulting face closely matches that of the source face, and (2) the result possesses a photorealistic quality, true to the expressions and posture of the target face, and aligns with the details of the target image such as lighting, background, and obstructions.

According to these requirements, we propose AmazingFS, a novel face swapping framework. AmazingFS leverages extensive data from both the target face and the source face to perform video face swapping and achieve cinematic-level face-swap results with high fidelity, as illustrated in Figure 1.

AmazingFS incorporates multiple strategies including attention mechanisms, style transfer modules, and facial segmentation networks. To address unclear facial features, we propose a novel Source-Target Attention Mechanism (STAM). STAM enables AmazingFS to

discern which conditional features, such as identity information, need to be discarded and which non-conditional features, such as background information, need to be retained in the target face. Our experiments show that STAM significantly enhances identity transfer by improving the model's focus on relevant features, thereby enhancing face-swapping results. STAM combines Channel Attention Modules (CAM) [8] and Spatial Attention Modules (SAM) [9], adaptively adjusting the importance within the feature maps, allowing for the model to more effectively capture and process critical features. In face-swapping tasks, this attention mechanism helps the model better recognize and retain the key features of the source face while accurately blending the target face's features into the source face.

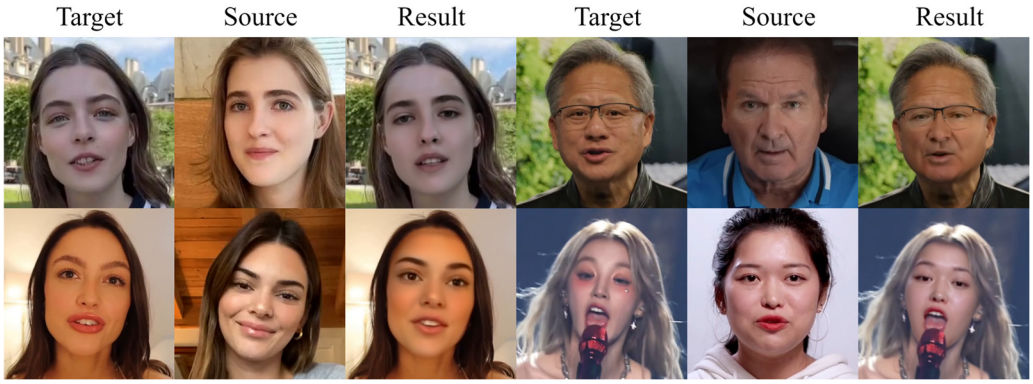


Figure 1. The face-swap results generated by AmazingFS involve replacing the face from the source image with the face in the target image.

To better adapt to the stylistic features of the target image while preserving the identity features of the source face, we have improved the AdaIN [10] module. Originally used for image style transfer and generative tasks, AdaIN effectively adjusts the style attributes of feature maps. We introduced a dynamic parameter adjustment mechanism using a neural network to predict the mean and standard deviation of style features instead of directly using the statistics of input features. This allows for adaptive parameter adjustment based on input features, producing more natural and realistic face-swapping effects.

To address the occlusion problem, we designed the AmazingSeg network simplified from TernausNet [11] and created a small dataset AST selected from 5000 obscured face images from the large AFLW [12] dataset with obstructions like hair, glasses, and hands. We manually annotated these 5000 obscured face images to distinguish facial areas from obstructed regions. The trained AmazingSeg model can be applied to any video face-swap involving obstructions, allowing for the model to automatically learn and address obstructions during the face synthesis stage.

The main contributions of this paper are as follows:

- We propose AmazingFS, a groundbreaking face-swapping framework that delivers cinematic-quality realistic face swaps with superior performance shown in extensive qualitative and quantitative experiments compared to other SOTA networks.
- We designed the STAM attention mechanism to improve the alignment of face edges and the detailed depiction of facial features.
- We improved the AdaIN style transfer module to better enhance the identity features of the source face while preserving the expressions and posture of the target face.
- We created the AmazingSeg dataset AST and trained the AmazingSeg network to handle obstructions such as hair and glasses during the face-swapping synthesis process.

2. Related Work

In recent years, advancements in deep learning have propelled the rapid development of face-swapping technology. This technology has evolved from simple image editing based on traditional image processing techniques to significant popularization through

deep learning methods. Key face-swapping technologies and tools have made substantial impacts in both academia and industry.

Deepfake [2], based on Generative Adversarial Networks (GAN) proposed by Ian Goodfellow et al., uses GANs to generate lifelike images and videos, creating facial images of target individuals and replacing faces in existing videos. FaceSwap [3] is an open-source tool that combines traditional image processing and deep learning methods to achieve high-quality face swapping, with strong community support and a broad user base. DeepFaceLab [4] employs GANs and autoencoders for face swapping, while offering customizable options for user-specific adjustments and optimizations. SimSwap [5] emphasizes maintaining the characteristics of the source face while achieving high-fidelity replacement of the target's facial features, utilizing a multi-scale feature fusion technique to enhance naturalness. Inswapper [6] uses deep learning techniques to seamlessly blend facial features between different images, preserving original expressions and lighting conditions for realistic results. BlendSwap [7] uses advanced neural networks to merge multiple facial images into a single composite face, maintaining high fidelity in skin texture, color, and detail. FaceDancer [13] focuses on real-time face-swapping technology for live broadcasting and video calling, achieving low-latency swaps through an optimized neural network architecture. HiFiFace [14] emphasizes maintaining high resolution and detail during the swap by using advanced generative models and detail enhancement techniques. DiffFace [15] employs advanced image generation algorithms that produce highly realistic results and excel in complex scenes and details. FaceShifter [16] uses a multi-stage generative model that gradually refines the swap for highly natural results.

Attention mechanisms in AI-driven face-swapping technology have significantly improved the quality and fidelity of generated images. Initially, GAN models handled global features, but attention mechanisms now enable a finer focus on local features, resulting in more lifelike swaps. Mnih et al. [17] introduced a pioneering visual attention model that enhances image processing efficiency and accuracy by selectively focusing on specific regions. Jaderberg et al. [18] developed Spatial Transformer Networks that dynamically manipulate images, providing robustness against input variations. Dai et al. [19] and Zhu et al. [20] proposed deformable convolutional networks that adapt to geometric variations, improving performance in detection and segmentation tasks. Hu et al. [21] introduced Squeeze-and-Excitation Networks that dynamically adjust channel weights to improve model performance. Woo et al. [22] presented the Convolutional Block Attention Module (CBAM) which enhances focus on crucial features and improves performance. Wang et al. [23] developed Non-local Neural Networks to model global relationships between features facilitating long-range dependency modeling. Dosovitskiy et al. [24] showcased the Vision Transformer model, applying the Transformer architecture to image recognition tasks with superior image processing capabilities.

Style transfer plays a crucial role in AI face-swapping technology by transferring style features from a target image onto a source image to generate realistic effects. Early methods based on neural style transfer by Gatys et al. [25] used convolutional neural networks to extract and apply style features, but were computationally intensive. Johnson et al. [26] introduced a fast-style transfer method by training a feedforward network to generate target images directly, improving efficiency, but limiting flexibility, as each style required a separate model. Huang et al. proposed Adaptive Instance Normalization (AdaIN), a further improved style transfer method that combines instance normalization with style encoding to apply the statistical features of the target image to the source image. AdaIN significantly enhances flexibility by handling multiple styles within a single model, eliminating the need to train a separate model for each style. We have improved the AdaIN mechanism and applied it to our model, providing an efficient and flexible solution for face swapping, ensuring real-time performance and realistic results.

Face segmentation technology is essential for enhancing the quality of face swapping. FaceXFormer [27] is a transformer-based encoder-decoder architecture that can handle multiple facial analysis tasks within a single framework. STN-iCNN [28] is an end-to-end face parsing framework that improves overall segmentation performance by using a localization network to accurately crop facial features. BiSeNet V2 [29] processes spatial information and semantic information through two separate paths and then combines them, resulting in higher segmentation accuracy and faster processing speed. TernausNet is a convolutional neural network (CNN) based on the U-Net architecture, designed specifically for image segmentation tasks. A key feature of TernausNet is its use of a pre-trained VGG 11 model in the encoder which enhances the efficiency of feature extraction. This approach enables TernausNet to achieve higher accuracy and stability in image segmentation. In the context of face-swapping technology, TernausNet is utilized for the precise segmentation of facial regions and extraction of key features, thereby improving the naturalness and consistency of face-swapping results. Its superior performance in handling complex scenes and intricate details provides a robust foundation for achieving high-quality facial replacements. We have implemented AmazingSeg occlusion handling technology, an improved version of TernausNet in the face-swapping domain. This enhancement significantly boosts the authenticity and visual appeal of the swapped images.

3. Methods

We propose a framework, AmazingFS, that successfully transfers the identity from the source face to the target face while preserving the attributes of the target face, thereby enhancing the fitting accuracy and realism of the face swap (Section 3.1). We introduce a novel attention mechanism, STAM, enabling the model to capture critical information about the face (Section 3.2). Subsequently, we explain the principles behind the improvements to AdaIN. AdaIN+ significantly enhances the identity features of the source face in the face swap results (Section 3.3). Additionally, the AmazingSeg facial segmentation network addresses occlusion issues in face swapping (Section 3.4). Finally, we describe the joint loss function used in our approach (Section 3.5).

3.1. Network Architecture

AmazingFS is a GAN-based framework designed for face-swapping, with its overall architecture shown in Figure 2. The generator includes an encoder, an intermediate layer (Inter), and a decoder. The target video face and the source video face are processed through encoders and decoders with shared weights and the intermediate layer. Unlike traditional methods, the AmazingFS algorithm does not require strict alignment of facial expressions between the source and target faces. This effective paradigm addresses the problem of unpaired faces while maintaining high fidelity and perceptual quality of the generated facial images. By utilizing a shared encoder and intermediate layer, AmazingFS solves the unpaired problem, with only one decoder needed to accomplish the face-swapping task. The result is the replacement of the source face with the target face rather than the entire target person.

The encoder is composed of multiple Conv2DBlocks, with the output dimensions of each Conv2DBlock gradually increasing from dim to dim $\times$ 2 to dim $\times$ 4 and then to dim $\times$ 8. After passing through these four convolutional layers, the image is flattened through a Flatten operation to enter the intermediate layer. The intermediate layer contains two fully connected layers and a Reshape operation to convert the flattened feature vector back into a shape suitable for convolution operations. The UpscaleBlock is used for upsampling, which restores low-resolution feature maps to higher resolution.

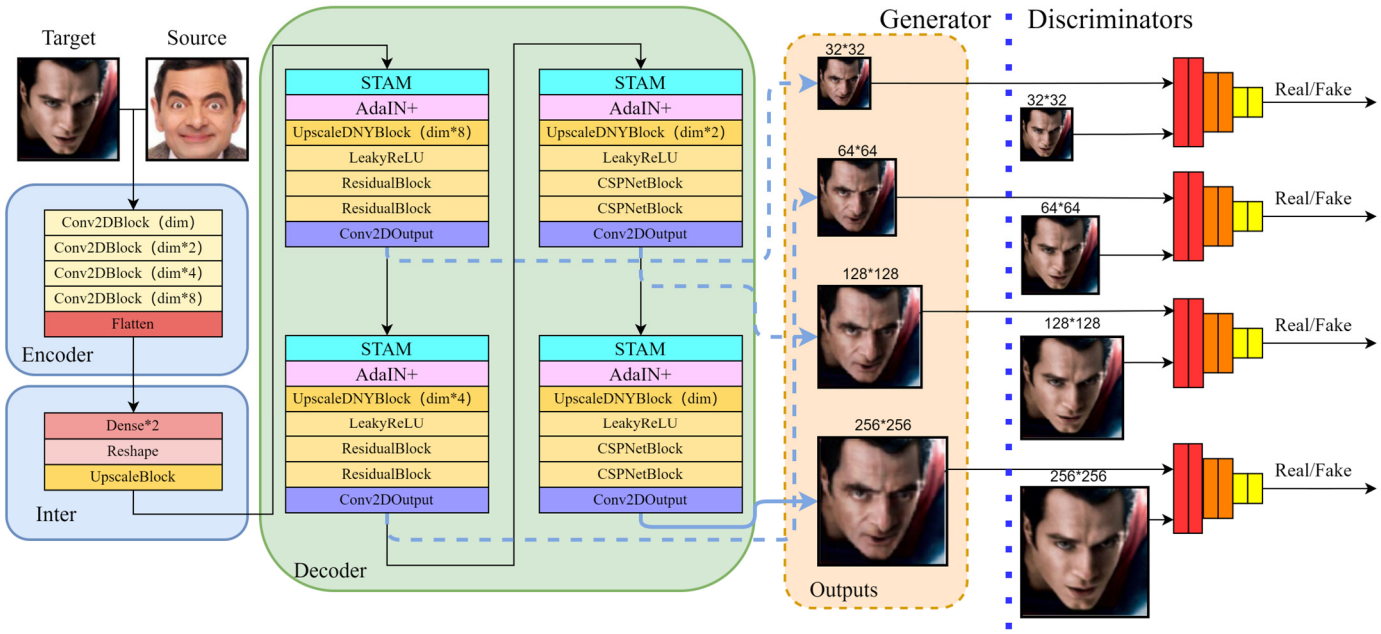


Figure 2. The overall network architecture of AmazingFS.

The decoder includes several modules, each serving a specific function to enhance overall image quality. The Source-Target Attention Mechanism (STAM) provides adaptive attention mechanisms in both space and time to improve the detail and dynamic performance of generated images. AdaIN+ enables style and content disentanglement by adjusting normalization parameters. The UpscaleDNYBlock employs Disney upsampling technology to efficiently upscale low-resolution images to high-resolution images. The LeakyReLU activation function, with a small slope, prevents the "dying" phenomenon of ReLU and enhances the model's non-linear representation capabilities. ResidualBlock adds short connections to mitigate gradient vanishing issues in deep networks. CSPNetBlock boosts the learning capability and efficiency of the network through partial feature separation and fusion. Finally, Conv2DOutput serves as the convolutional output layer. The input dimensions of each module progressively decrease from dim $\times$ 8 to dim, with the decoder outputting images at resolutions of 32 $\times$ 32, 64 $\times$ 64, 128 $\times$ 128, and 256 $\times$ 256. These multi-scale fusion operations restore complete image features.

The encoder and intermediate layer convert input facial images into latent feature vectors through mapping and transformation in the feature space. The encoder extracts and compresses facial features to convert high-dimensional image data into low-dimensional feature vectors. The intermediate layer exchanges and fuses facial features in the feature space to achieve feature transfer and face-swapping effects between different faces. The decoder converts the low-dimensional feature vectors processed by the encoder back into high-dimensional images. By training this neural network, it reconstructs facial images resembling the original from the input feature vectors. The decoder employs STAM, AdaIN+, Disney upsampling blocks, and activation functions. It adds convolutional layers before doubling the feature map size to generate four different scales of RGB images for multi-scale fusion operations. This process restores the complete image feature vectors while retaining the identity features of the source face and recovering the facial expressions and poses to achieve a high-fidelity reconstruction of facial images.

After the generator outputs images of multiple resolutions, these images are input into corresponding resolution discriminators for real or fake discrimination. The discriminator consists of multiple convolutional layers and activation functions to distinguish between generated and real images. Each discriminator has a similar structure with multiple Conv2D layers and LeakyReLU activation functions. Through layer-by-layer downsampling and feature extraction, a real or fake label is finally output via a fully connected layer. Discrimi-

nators are trained at different resolutions to ensure the generator can produce high-quality images at various scales. This structure enables AmazingFS to achieve a high-fidelity reconstruction of facial images.

3.2. STAM

In the field of face-swapping technology, facial images are crucial due to their complex structures and diversity. The face-swapping process involves numerous detailed changes and simply adding attention modules does not adequately address issues related to facial details in feature selection. To enhance the model's ability to select features and capture details, we optimized the attention mechanism algorithm and introduced the Source-Target Attention Mechanism (STAM) into the face-swapping system. This enhancement aims to improve the quality and stability of swapped facial images.

STAM combines the advantages of the Channel Attention Module (CAM+) and the Spatial Attention Module (SAM) through a parallel design and a feature recalibration module, significantly improving the performance and stability of the face-swapping model. This approach allows for the better handling of complex facial images, resulting in high-quality face-swapping outcomes. As illustrated in Figure 3, STAM integrates CAM+ and SAM. CAM+ performs average pooling and max pooling on the input feature map, compressing it to a size of $1 \times 1 \times C$. The compressed feature map is then fed into two MLP networks for computation, and spatial attention weights are generated through a Sigmoid activation function. Given the minimal background noise and complex texture details of facial images, the CAM+ module adopts an expansion perceptron instead of a reduction perceptron. This convolutional attention module effectively captures facial details, enhancing model stability. Specifically, the CAM+ module first elevates the image features from $1 \times 1 \times C$ to $1 \times 1 \times rC$, where r is the expansion coefficient set to 4 based on experimental comparisons, and then reduces it back to $1 \times 1 \times C$. This process enables the model to capture facial details more finely during feature selection, resulting in more natural and seamless facial integration in the face-swapping results, thereby maximizing the advantages of the attention mechanism.

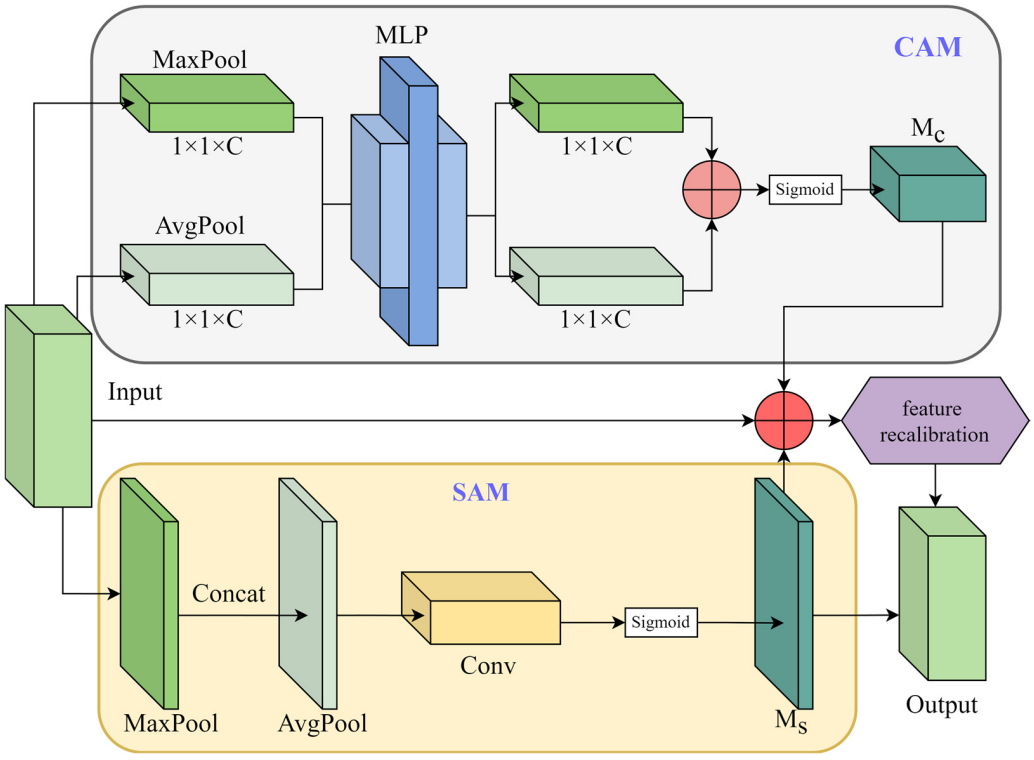


Figure 3. Internal Structure of STAM.

In designing the STAM module, various factors were considered. Firstly, choosing a parallel approach for CAM+ and SAM instead of a cascading approach ensures that the computation of one attention mechanism does not interfere with the other, thereby improving model stability. Secondly, CAM+ effectively extracts global features through average pooling and max pooling, while SAM complements these global features by focusing on local spatial information, making feature extraction more comprehensive. In the CAM+ module, the input features are processed through max pooling and average pooling, resulting in two $1 \times 1 \times C$ feature maps. These maps are then processed through an expansion perceptron (MLP) to generate channel attention weights. In the SAM module, the input features are similarly processed through max pooling and average pooling, and the results are concatenated and then processed through a convolutional layer to generate spatial attention weights. Thus, STAM extracts both channel and spatial information through the parallel CAM+ and SAM modules, and finally recalibrates these features through the feature recalibration module. The feature recalibration module better matches the facial details of the source and target images, accurately capturing the texture, lighting, and structural features of the face. This enhances the model's robustness, reducing potential flaws and inconsistencies during the face-swapping process, resulting in more stable and reliable final swapped images.

The parallel formula is shown as Equation (1):

$$\mathrm{F1} = \mathrm{W}((\mathrm{Mc}(\mathrm{F})\otimes\mathrm{F})\oplus(\mathrm{Ms}(\mathrm{F})\otimes\mathrm{F}))\otimes\mathrm{F} \tag{1}$$

In the proposed structure, F represents the original input feature map, Mc(F) represents the output of the feature map after channel attention computation, and Ms(F) denotes the output of the feature map after spatial attention computation. The operation $\otimes$ indicates a multiplication operation for weighted feature maps, while $\oplus$ indicates an addition operation for weighted feature maps. W is the feature recalibration weight matrix, which is generated through an additional learning layer. Furthermore, the STAM is structured to first expand and then reduce dimensionality, and the CAM and SAM are arranged in parallel. The overall architecture is illustrated in Figure 3. Equation (2) provides the formula for computing channel attention. Equation (3) details the formula for spatial attention.

$$\mathrm{Mc}(\mathrm{F}) = \sigma(\mathrm{MLP}(\mathrm{AvgPool}(\mathrm{F})) + \mathrm{MLP}(\mathrm{MaxPool}(\mathrm{F})) = \sigma(\mathrm{W1}(\mathrm{W2}(F^{c}_{Avg})) + \mathrm{W1}(\mathrm{W2}(F^{c}_{Max}))) \tag{2}$$

$$\mathrm{Ms}(\mathrm{F}) = \sigma(f^{7\times7}[\mathrm{AvgPool}(\mathrm{F}); \mathrm{MaxPool}(\mathrm{F})]) = \sigma(f^{7\times7}(\left[F^{s}_{Avg}; F^{s}_{Max}\right])) \tag{3}$$

In this context, σ represents the Sigmoid activation function, W1 and W2 denote the weights of the upsampling Multi-Layer Perceptron (MLP), $\mathrm{F}^{c}_{\mathrm{Avg}}$ and $\mathrm{F}^{s}_{\mathrm{Avg}}$ represent the average pooling features, and $\mathrm{F}^{c}_{\mathrm{Max}}$ and $\mathrm{F}^{s}_{\mathrm{Max}}$ denote the maximum pooling features. The notation $\mathrm{f}^{7\times7}$ indicates a 7×7 convolution operation with filters.

3.3. AdaIN+

To better adapt to the target image's style characteristics while maintaining the source facial identity features, we improved the AdaIN module. The traditional AdaIN module normalizes the features by simply computing the mean and standard deviation of the content and style features, which may lead to suboptimal image quality. To address this issue, we propose a dynamic parameter adjustment mechanism and integrate a multi-scale fusion approach. Additionally, we introduce a regularization strategy to ensure the naturalness and consistency of the generated images. The traditional AdaIN operation is implemented by the following formula:

$$\mathrm{AdaIN}(\mathrm{x}, \mathrm{y}) = \sigma(\mathrm{y})(\frac{x - \mu(\mathrm{x})}{\sigma(\mathrm{x})}) + \mu(\mathrm{y}) \tag{4}$$

x denotes the facial content features, y represents the facial style features, and μ and σ signify the mean and standard deviation calculations, respectively. To enhance the quality of the generated images, we introduce a dynamic parameter adjustment mechanism. Specifically, we employ a convolutional neural network to predict the mean and standard deviation of the style features, rather than directly using the statistics of the input features. The improved approach is as follows:

$$\mathrm{AdaIN}(\mathrm{x},\mathrm{y}) = \sigma(\mathrm{f}(\mathrm{y}))\left(\frac{x-\mu(\mathrm{x})}{\sigma(\mathrm{x})}\right) + \mu(\mathrm{f}(\mathrm{y})) \tag{5}$$

f represents the neural network used to predict the mean and standard deviation based on the input style features y. Additionally, we introduce a multi-scale fusion method in the improved AdaIN module. This involves decomposing the input content features into multiple scales, calculating the mean and standard deviation for each scale separately, and then performing a weighted sum of these scaled results. The specific implementation is as follows:

$$\mathrm{MultiAdaIN}(\mathrm{x},\mathrm{y}) = \sum_{i=1}^{n} \omega_i \ \mathrm{AdaIN}(x_i, y_i) \tag{6}$$

n is the number of scales, which we set to 4. ω_i represents the weighting parameters for each scale, while x_i and y_i denote the content and style features at the i-th scale, respectively. To prevent the generated images from being excessively smooth or distorted, we use weighting parameters in the improved AdaIN module to perform a weighted sum of the outputs from different scales. By learning these weighting parameters, the model automatically adjusts the contribution of each scale during training, resulting in a more natural face-swapping effect. The final improved approach is as follows:

$$\mathrm{RegMultiAdaIN}(\mathrm{x},\mathrm{y}) = \sum_{i=1}^{n} \alpha_i \cdot \mathrm{AdaIN}(x_i, y_i) \tag{7}$$

α_i are the weighting parameters obtained through a regularization strategy, used to balance the information contribution from different scales. With these improvements, our model can capture both detailed and global information, thereby generating images that are more consistent with the identity information of the source face. This weighted summation method not only balances the information from different scales, but also prevents the information from any single scale from excessively influencing the final generated result.

3.4. AmazingSeg

In this study, we developed AmazingSeg, a technique for precise facial region segmentation to enhance face-swapping tasks. AmazingSeg uses a trained neural network model to automatically identify and segment facial features in images. Its core advantage is its efficiency and accuracy, significantly improving face replacement outcomes. The workflow of AmazingSeg is as follows: Initially, we created a small dataset called AST that includes faces with occlusions. This dataset is used to train the facial segmentation model. During the data preparation phase, we annotated the images by explicitly marking the facial regions in each image. Accurate annotations are crucial, as they directly impact segmentation quality. We then trained a deep learning segmentation model using these annotated data and employed cross-validation methods to ensure the model's generalization capability. The training process relies on substantial computing resources and was accelerated using high-performance GPUs.

AmazingSeg utilizes a joint loss function to optimize the model's learning process. This function combines pixel-level cross-entropy loss, Dice loss, and boundary loss to ensure segmentation accuracy, particularly in handling occlusions. After training, the model underwent multiple tests and validations to ensure robustness under various lighting conditions, facial expressions, and occlusion scenarios. The resulting model efficiently and accurately segments facial regions, automatically processing input images to generate precise segmentation results. In practical applications, we used AmazingSeg for face replacement tasks. The model processes input images to generate segmentation masks for

facial areas. These masks guide the replacement of target facial features onto the source image, achieving high-quality face replacement effects while maintaining facial detail integrity.

The architecture of the AmazingSeg network shown in Figure 4 produces high-precision segmentation masks through convolutional operations, multi-scale feature extraction, pooling, upsampling, and skip connections. This architecture effectively addresses occlusion issues, significantly enhancing the robustness and accuracy of segmentation tasks. The comprehensive application of these techniques enables AmazingSeg to perform exceptionally well in complex scenes with severe occlusions, ensuring its reliability and practicality in real-world applications.

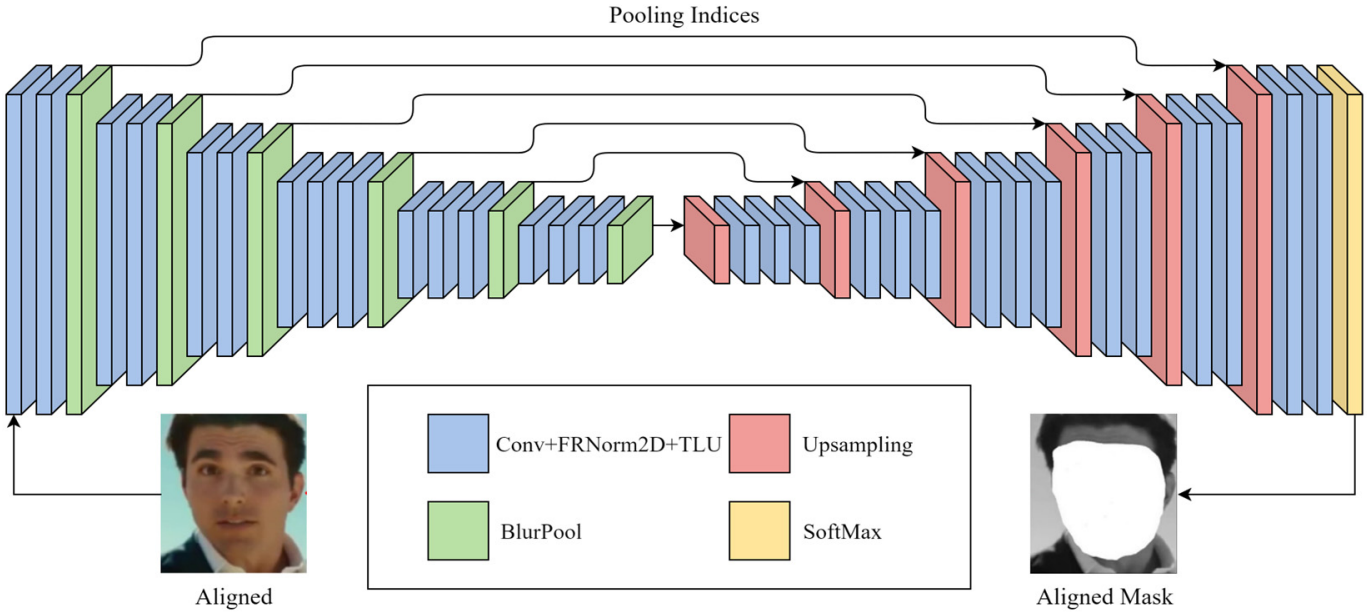


Figure 4. The overall network architecture of AmazingSeg.

3.5. Loss Function

In AmazingFS, various loss functions are primarily used to train the model to achieve high-quality face swapping. The following are the five main loss functions utilized in AmazingFS and their formulas:

- Reconstruction Loss

$$L_{Rcon} = \| I_R - I_T \| \tag{8}$$

If the source face and the target face belong to the same identity, the generated result should resemble the target face. We use reconstruction loss as a regularization term to ensure that the reconstructed face closely approximates the target face.

- Adversarial Loss

$$L_{adv} = -E[\log D(G(z))] \tag{9}$$

The adversarial loss is provided by the discriminator in the Generative Adversarial Network (GAN) and is used to train the generator to produce realistic images, making the generator's output indistinguishable from real images. Here, E represents the mathematical expectation, D denotes the discriminator, and G denotes the generator. z is the random noise vector, and D(G(z)) represents the discriminator's output for the image generated by the generator.

- Perceptual Loss

$$L_{per} = \sum_{t=0}^{n} \cdot \| \varnothing_1(I_R) - \varnothing_1(I_T) \| \tag{10}$$

Perceptual loss uses a pre-trained neural network (e.g., VGG) to extract features, ensuring that the generated image is perceptually akin to the original image. $\varnothing_i$ represents the features extracted by the pre-trained neural network (e.g., VGG) at the i-th layer.

- Style Loss

$$L_{style} = \sum_{i=0}^{n} \cdot \| G_{\varnothing i}(I_R) - G_{\varnothing i}(I_T) \| \tag{11}$$

Style loss ensures that the style features of the generated image are consistent with those of the target image. It is commonly used in style transfer tasks and calculates the correlation of feature maps using the Gram matrix.

- Id Loss

$$L_{id} = 1 - \left(\frac{V_R \cdot V_S}{\|V_R\|_2 \|V_S\|_2} \right) + \mu(y) \tag{12}$$

Identity loss is used to constrain the distance between the vectors V_S and V_R. We use cosine similarity to compute this distance. The overall loss function L_{total} can be expressed as:

$$L_{total} = \alpha L_{Rcon} + \beta L_{adv} + \gamma L_{per} + \delta L_{style} + \eta L_{id} \tag{13}$$

where the parameters are defined as follows: $\alpha = 1.0$, $\beta = 0.1$, $\gamma = 0.01$, $\delta = 0.01$, and $\eta = 1$.

In the training of AmazingSeg within AmazingFS, the loss functions are primarily used to optimize the generated masks to accurately segment the facial regions. The loss functions can be categorized as follows:

- Cross-Entropy Loss

Cross-entropy loss is primarily used to measure the difference between the generated masks and the ground truth masks. In the face-swapping task, its role is to ensure that the generated masks accurately match the actual facial and background regions, providing an initial segmentation. The formula is as follows:

$$L_{CE} = -\frac{1}{N} \sum_{i=1}^{n} [y_i log(p_i) + (1 - y_i) log(1 - p_i)] \tag{14}$$

Here, y_i is the ground truth label indicating whether a pixel belongs to the target facial region (1 for belonging, 0 for not belonging). p_i represents the probability predicted by the model that a pixel belongs to the target facial region. N is the total number of pixels in the image.

- Dice Loss

Dice loss is used to measure the overlap between the predicted mask and the ground truth mask. In face-swapping tasks, its role is to ensure that the generated mask highly overlaps with the actual facial region, thereby improving the accuracy of facial region segmentation. The formula is as follows:

$$L_{Dice} = 1 - \frac{2 \sum_{i=1}^{N} p_i y_i}{\sum_{i=1}^{N} p_i + \sum_{i=1}^{N} y_i} \tag{15}$$

- Boundary Loss

Boundary loss is minimized to generate masks with clearer boundaries, making the face-swapping effect appear more natural. This loss function helps to avoid abrupt transitions between the facial region and the background, resulting in a smoother and more realistic face-swapping effect. The formula is as follows:

$$L_{\text{Boundary}} = \frac{\sum_{i=1}^{n} |\nabla p_i - \nabla y_i|}{\text{N}} \tag{16}$$

Here, ∇ denotes the gradient operation used to calculate the image edges. We combine these loss functions to form a comprehensive loss function for training the AmazingSeg model. The overall loss function L_{All} can be expressed as:

$$L_{All} = \lambda_1 L_{CE} + \lambda_2 L_{Dice} + \lambda_3 L_{\text{Boundary}} \tag{17}$$

Here, λ_1 = 0.5, λ_2 = 0.5 and λ_3 = 0.1.

4. Datasets and Training

4.1. Datasets and Processing

The AmazingFS dataset is divided into a base dataset and a one-to-one dataset. The source dataset for training the AmazingFS base dataset is the publicly available CelebA dataset [30]. First, we use S3FD [31] as the face detector to locate the target face in the given data. To ensure the stability of the facial landmarks, we use the 2DFan [32] facial landmark extraction algorithm and the classic point pattern alignment and transformation method proposed by Umeyama [33] to calculate the similarity transformation matrix for face alignment. To improve the quality of face swapping, we align and crop the images to a standard size of 256 × 256 pixels. The one-to-one model training dataset of AmazingFS consists of the source face dataset from one video and the target face dataset from another video. Both the source and target faces undergo preprocessing steps including face detection, face alignment, and cropping to a standard size.

The AST dataset used for training the AmazingSeg network is derived from the AFLW face database. It includes 5000 facial images with occlusions, as shown in Figure 5. The AFLW face database is a large-scale collection featuring a wide range of poses and viewpoints, with each face annotated with 21 key points. It encompasses images influenced by various factors such as pose, expression, lighting, and ethnicity. From the approximately 25,000 images in the AFLW database, we selected 5000, comprising 53% female and 47% male subjects. Given the importance of addressing occlusion in face-swapping technology, we specifically chose images with occlusions. The images in the AFLW database are affected by various factors including pose, expression, lighting, and ethnicity, adding to their complexity and diversity. To enhance the model's ability to handle occlusions, we selected representative occluded images for re-annotation. This re-annotation process was carried out manually. Specifically, we used solid lines to annotate the facial regions of each image and dashed lines to differentiate occlusions in some images, as shown in Figure 6. This meticulous annotation process ensured accuracy and detail. The precise annotation of each facial feature laid the groundwork for subsequent mask synthesis and enabled the face-swapping model to better identify and handle occluded facial areas. These steps ultimately led to the construction of the AST dataset for training the AmazingSeg network. This dataset includes a wide variety of facial images, covering different genders, poses, expressions, and occlusion scenarios. High-quality manual annotations enhance the dataset's accuracy, providing a solid foundation for training the face-swapping model.



Figure 5. Our self-created occluded face dataset AST.



Figure 6. AST preprocessing.

4.2. Training Details

We trained the AmazingFS generator from scratch with weights randomly initialized using a normal distribution and the Adam optimizer (β_1 = 0.5; β_2 = 0.999) with a learning rate of 0.0002. The learning rate decayed exponentially by 0.97 every 100K steps. All our

networks were trained on an NVIDIA GeForce RTX 4070 GPU (NVIDIA Corporation, Santa Clara, CA, USA) and an Intel Core i7-13700HX processor (Intel Corporation, Santa Clara, CA, USA) with a batch size of 4. We used a progressive multi-scale approach starting from a resolution of 32 × 32 and ending at 256 × 256, with a total of 500 K training steps. The target and source images were randomly adjusted for brightness, contrast, and saturation. For ablation experiments, each configuration was trained for 300 K steps. For comparative experiments with other recent methods, we controlled the batch size and number of training steps, similarly.

The face segmentation network AmazingSeg also needed to be trained from scratch, using the same hardware configuration of an NVIDIA GeForce RTX 4070 GPU and an Intel Core i7-13700HX processor. During training, the model parameters were first initialized, and the standard Adam optimizer was used for parameter updates. The learning rate was set to 0.0001 and dynamically adjusted during training to ensure effective convergence of the model. The training data were standardized to improve the stability and performance of the model. The batch size was set to 4, and training was stopped after 300 K steps. We used the cross-entropy loss function to evaluate the performance of the AmazingSeg network. Additionally, the model was regularly saved during training for subsequent analysis and adjustments.

5. Experiment

5.1. Qualitative Results

We used videos from the publicly available FaceForensics++ dataset as our source images. This dataset includes 977 face videos downloaded from YouTube and 977 face-swapped videos generated by DeepFakes. We demonstrated the effects of 12 sets of videos: 12 target face videos, and 12 source face videos. Six sets served as our unoccluded comparative experiments, and the other six as our occluded comparative experiments. All these images were excluded from our training set. We performed face swapping on all source and target pairs.

In unoccluded scenarios, we compared our method with mainstream SOTA methods: FaceSwap, DeepFaceLab, SimSwap, Inswapper, and BlendSwap, as shown in Figure 7. FaceSwap is a popular open-source project, but when there are significant differences in resolution, lighting conditions, or angles between the source and the target images, the generated face-swapped images may appear blurred or unnatural, with unclear edges or noticeable stitching marks. Additionally, for images with rich facial details such as beards or eyebrows, these features may be inaccurately fitted or distorted. DeepFaceLab can maintain the same facial shape as the target face, but still suffers from poor fitting or blurred details when dealing with images with significant differences in lighting, angles, and skin tones. SimSwap performs well in lighting processing but cannot maintain the target face's pose well. The model fails to correctly reconstruct textures in some areas, resulting in artifacts in some face-swapped results. Inswapper and BlendSwap perform well in detail fusion and handling facial occlusions, but still have shortcomings in lighting and skin tone consistency. Our AmazingFS method addresses these issues better by generating believable face-swapping results while achieving better performance in retaining target attributes. Additionally, it demonstrates strong identity preservation capabilities and presents a framework for better image restoration. STAM is a specially designed attention mechanism to overcome the loss of facial details by better focusing on details such as the eyes and facial edges of the target person. AdaIN+ reduces the adverse effects of style mismatches, enhancing the identity similarity of the resulting faces. Overall, our AmazingFS perfectly preserves lighting and facial styles while capturing the target face's pose well, generating high-quality face-swapped results and retaining the source face's identity.



Figure 7. *Cont.*



Figure 7. AmazingFS vs. Other SOTA Methods: Unoccluded Face Comparison Experiments.

In occlusion scenarios, we compared AmazingFS with FaceSwap, DeepFaceLab, SimSwap, Inswapper and BlendSwap. The comparison results in Figure 8 show that each SOTA's method has shortcomings in face-swapping images. FaceSwap produces blurred images and unnatural facial blending. The facial details and contours are noticeably blurred, and the alignment and fusion of facial features lack harmony. DeepFaceLab suffers from loss of facial details and significant skin tone differences. The details of facial features, such as eyes and mouths, are not well restored, leading to a lack of vivid facial expressions. When dealing with noticeable skin tone differences, DeepFaceLab tends to produce unnatural transitions, causing obvious image artifacts. SimSwap generates unnatural facial features and loses detail. The facial features, especially expressions and contours, are not natural enough. SimSwap's performance is unsatisfactory in handling rich facial details, leading to detail loss and uneven edges between the face and the background. Inswapper has some occlusion resistance but performs poorly in processing facial details and expressions. The generated images tend to have poor fusion in complex backgrounds. BlendSwap, although having good occlusion resistance, still exhibits unnatural facial detail restoration and skin tone transitions. It tends to lose details when handling complex facial expressions, resulting in less clear outcomes. In contrast, AmazingFS demonstrates superior performance. It excels in aligning and fusing facial features, generating natural and realistic images. AmazingFS accurately restores facial details and handles expressions and features well. It achieves natural skin tone transitions with a harmonious overall effect. Most importantly, AmazingFS exhibits excellent occlusion resistance. When faced with occlusions like hair, glasses, and microphones, the face-swapping effect remains natural and comfortable, with hardly any noticeable flaws.

In the experimental results of AmazingFS, an "incomplete face swap" phenomenon appears. This happens because AmazingFS retains the source face features too strongly, making certain features too prominent and overshadowing the target face features. When AmazingFS keeps distinct features like jawlines, eye shapes, or cheekbones, these elements can dominate the swapped image and make the swap seem less complete. AmazingFS excels in handling lighting and skin tone consistency. This ensures that the lighting and skin tone of the swapped face blend seamlessly with the rest of the image and reduce visual discrepancies. However, this strength can also make the swapped face resemble the source face too closely and create the impression that the faces have not been completely swapped. For example, if the source face has a particular skin texture or lighting condition that is well preserved, the target face may inherit these characteristics too faithfully, making the distinction between the two faces less apparent. The attention mechanism designed to focus on key features might overly emphasize prominent aspects of the source face like unique eye shapes or mouth curvature. Similarly, the style fusion module, which integrates stylistic elements of both faces, might blend the features in a way that favors the source face's distinctive traits. As a result, the final generated image can seem "incomplete" in terms of the swap, appearing to retain more of the source face than intended. This illusion of incomplete face swapping is particularly noticeable when swapping faces within the same ethnicity. For instance, swapping Asian faces with other Asian faces or Caucasian faces with other Caucasian faces might result in subtler differences in facial features due to inherent similarities within the same ethnic group. The minimal variation in features like skin tone, facial structure, and eye shape means that even slight retention of source features can make the swap seem less significant. Consequently, AmazingFS's ability to maintain lighting and skin tone consistency, while generally advantageous, can exacerbate the perception of an incomplete swap in these scenarios.

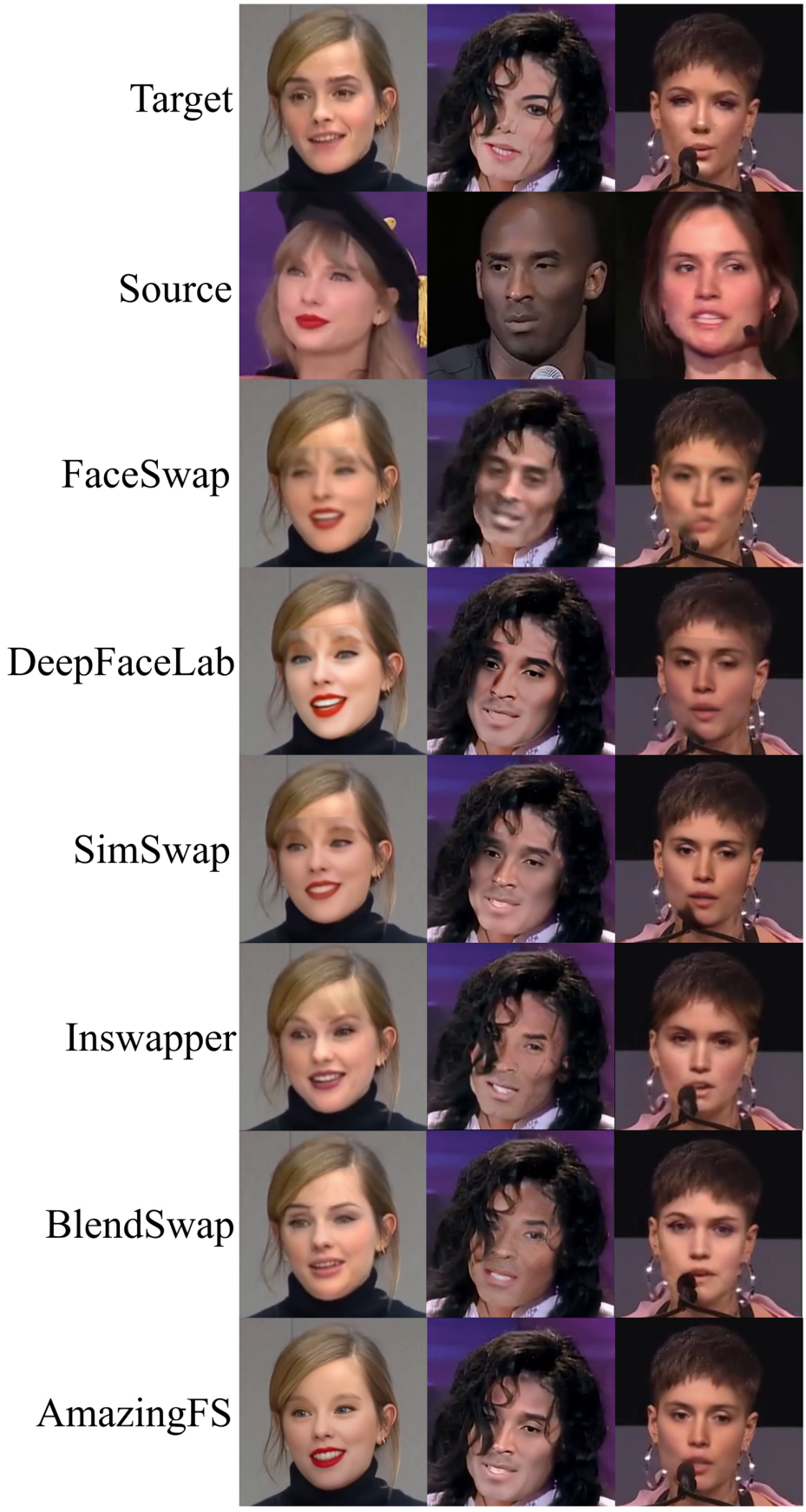


Figure 8. *Cont.*



Figure 8. AmazingFS vs. Other SOTA Methods: Occluded Face Comparison Experiments.

We downloaded face videos from YouTube to perform frame-by-frame face swapping and demonstrated five sets of results, extracting the swapped face results from the 1st, 50th, 100th, 150th and 200th frames. AmazingFS exhibits significant advantages in face swapping across different frames, primarily due to its use of attention mechanisms, AdaIN+ and AmazingSeg technologies. The attention mechanism enables AmazingFS to more accurately capture and process facial features and details. This ensures the generated face-swapped images maintain high quality and a natural, realistic appearance under various complex backgrounds and lighting conditions. As shown in Figure 9, with an increasing number of frames, facial feature alignment and fusion remain precise, and detail processing remains on point. AdaIN+ allows for AmazingFS to better adapt and integrate different styles and textures of source images. This plays a crucial role in maintaining facial consistency and detail restoration. Despite significant style differences between the source and target images, AdaIN+ helps achieve a natural face-swapping effect. AmazingSeg segmentation technology improves the segmentation accuracy of facial regions. This makes the face-swapping effect more natural and delicate. When dealing with hair, accessories, glasses, and microphones of different sizes, positions, and angles, AmazingSeg can accurately segment facial regions, avoiding common issues of unsmooth edges or improper occlusion handling, ensuring that the face-swapping effect remains natural and comfortable. By combining attention mechanisms, AdaIN+, and AmazingSeg technologies, AmazingFS significantly enhances the quality of face-swapped images. Its performance is stable and excellent across different frames, capable of handling various complex facial features and backgrounds, resulting in natural, realistic, and detailed face-swapping effects, demonstrating movie-level face-swapping quality.

5.2. Quantitative Results

We perform a quantitative comparison on the FaceForensics++ video dataset using SSIM, ID preservation, pose error, expression error, and face shape error to further demonstrate the effectiveness of our AmazingFS. For FaceSwap, DeepFaceLab, and SimSwap, we uniformly sample ten frames from each video to form a 10K test set.

- SSIM

The Structural Similarity Index (SSIM) measures the similarity between two images by comparing their luminance, contrast, and structural information. Specifically, SSIM divides the images into small blocks and calculates the mean, variance, and covariance of these blocks. These statistics are then used to measure the luminance similarity, contrast similarity, and structural similarity between the images, resulting in an SSIM value between 0 and 1. A value closer to 1 indicates higher similarity and better quality of the face-swapped image.

- ID Preservation

The ID preservation metric evaluates identity consistency by comparing the original face image with the face-swapped image. We extract feature vectors from both images and calculate the similarity between these feature vectors. A higher similarity indicates that the face-swapped image retains the identity features of the original face, verifying the authenticity and quality of the face swap. We use the pre-trained face recognition model ArcFace [34] to ensure accurate feature extraction.

- Pose Error

The pose error metric determines the accuracy of the facial pose by analyzing the positions and angles of facial keypoints. The algorithm identifies keypoints (e.g., eyes, nose, mouth) on both the source and target faces and solves the perspective-n-point problem. This involves estimating the camera's (viewpoint's) pose using a set of 3D model points and their 2D image correspondences. The resulting rotation matrix can be converted into Euler angles (pitch, yaw, roll). The differences in these Euler angles are calculated using cosine similarity and angle differences to determine the 3D facial pose. The pose metric evaluates the naturalness and realism of the face swap by comparing the facial pose in the

swapped image with the target face's pose. Smaller differences indicate better alignment with the original image's head pose, resulting in a higher evaluation.



Figure 9. AmazingFS Face Swapping Results Across Different Frames.

- Expression Error

Expression error evaluation is based on analyzing and comparing facial expressions in the images. By comparing the differences in facial expressions between two images, we measure their similarity. We use facial feature extraction techniques such as facial keypoint detection and facial expression classification, and then calculate the Euclidean distance

between the 2D landmarks of the target face and the swapped face. Smaller distances indicate better expression retention.

- Shape Error

The core idea of the shape error metric is to measure the geometric similarity between the result face and the target face. By detecting keypoints on both the source and target faces, we obtain the positions of these keypoints and then calculate the differences between the two sets of keypoints. We use the Euclidean distance to quantify these differences, with smaller values indicating higher matching degrees in shape.

Table 1 presents a quantitative comparison of five face-swapping algorithms (FaceSwap, DeepFaceLab, SimSwap, Inswapper, and BlendSwap) across several key metrics, including SSIM, ID preservation, pose error, expression error, and shape error. The results indicate that AmazingFS outperforms the others in all metrics. Specifically, AmazingFS achieved the highest SSIM score of 0.82, indicating superior image quality. In terms of ID preservation, AmazingFS leads with a score of 95.12, demonstrating its excellence in maintaining identity consistency. In contrast, other face-swapping methods fall short on these key metrics. For pose and expression, AmazingFS attained relatively small errors of 2.35 and 23.45, respectively, highlighting its superior performance in preserving pose and expression. Finally, in face shape, AmazingFS again scored the best with 44.61, indicating high efficiency in geometric shape matching. Therefore, AmazingFS shows significant advantages in the realism and quality of face-swapping effects.

Table 1. Quantitative Experiments of Different SOTA Methods.

Method	SSIM ↑	ID ↑	Pose ↓	Expression ↓	Shape ↓
FaceSwap	0.68	75.61	4.85	40.54	49.52
DeepFaceLab	0.76	89.56	2.65	30.26	45.29
SimSwap	0.79	92.25	3.56	27.18	46.16
Inswapper	0.80	93.52	2.68	23.73	45.95
BlendSwap	0.77	91.79	2.81	25.82	47.27
AmazingFS (ours)	**0.82**	**95.12**	**2.35**	**23.45**	**44.61**

At the same time, we designed an evaluation system comprising five subjective evaluation metrics to assess the effectiveness of face-swapping algorithms. These metrics are identity consistency, attribute retention, anti-occlusion capability, detail preservation, and overall fidelity. Each metric is rated on a scale from 0 to 5, with higher scores indicating better performance.

- Identity Consistency

Identity primarily measures whether the face-swapped image retains the original identity features. Higher scores indicate a greater similarity between the face-swapped image and the original face in terms of identity features.

- Attribute Retention

Attribute evaluates whether the face-swapped image retains the original face's attribute features, such as age, gender, and emotional expression. Higher scores indicate better attribute retention.

- Anti-Occlusion Capability

Anti-occlusion assesses the performance of the face-swapping algorithm in handling partial occlusions, such as wearing glasses or hats. Higher scores indicate that the face-swapped image maintains good recognition performance, even under occlusion.

- Detail Preservation

Details primarily measures whether the face-swapped image retains the original face's detailed features, such as skin texture and hair. Higher scores indicate better detail preservation.

- Overall Fidelity

Fidelity evaluates the overall visual realism of the face-swapped image. Higher scores indicate more realistic face-swapping effects.

To ensure the objectivity and reliability of the evaluation results, we invited 50 volunteers to subjectively score the face-swapping results. Each volunteer viewed the face-swapped images and rated them based on the five metrics mentioned above. Finally, as shown in Table 2, we performed statistical analysis on the scores for each metric to evaluate the performance of the face-swapping algorithms across different aspects.

Table 2. Subjective Evaluation Scores.

Method	Identity ↑	Attribute ↑	Anti-Occlusion ↑	Details ↑	Fidelity ↑
FaceSwap	2.95	2.98	2.14	3.12	3.58
DeepFaceLab	3.86	3.36	2.65	4.15	3.95
SimSwap	3.75	3.84	2.35	4.27	4.31
Inswapper	3.99	4.12	3.54	3.89	4.67
BlendSwap	3.82	4.01	3.72	4.42	4.38
AmazingFS (ours)	**4.38**	**4.25**	**4.75**	**4.66**	**4.83**

5.3. Ablation Studies

We conducted four different types of ablation experiments to demonstrate the effectiveness of our designed STAM, AdaIN+, and AmazingSeg modules.

5.3.1. CAM+ Upscaling Study

We performed an in-depth exploration of the optimal dimensions for the channel attention mechanism (CAM) upscaling module in the STAM module of AmazingFS. By employing various upscaling strategies and using the structural similarity index (SSIM), pose, and expression as evaluation metrics, we systematically analyzed the impact of different upscaling strategies on model performance. Table 3 details the results of six experimental groups.

Table 3. Performance of Different Upscaling Dimensions for CAM in STAM.

Group	Upscaling Strategy	Feature Map Size	SSIM ↑	Pose ↓	Expression ↓
A	$1 \times 1 \times 1/4$	$1 \times 1 \times C/16$	0.74	6.29	28.15
B	$1 \times 1 \times 2$	$1 \times 1 \times 2C$	0.75	5.26	26.56
C	$1 \times 1 \times 3$	$1 \times 1 \times 3C$	0.79	3.54	24.78
D	$1 \times 1 \times 4$	$1 \times 1 \times 4C$	**0.82**	**2.35**	**23.45**
E	$1 \times 1 \times 5$	$1 \times 1 \times 5C$	0.81	3.92	25.29
F	$1 \times 1 \times 10$	$1 \times 1 \times 10C$	0.76	4.18	26.11

First, Group A adopted the strategy of directly reducing the channel dimensions to $1 \times 1 \times C/16$. The results showed that this method performed the worst, with an SSIM of 0.74, a pose error of 6.29, and an expression error of 28.15. This indicates that simple dimensionality reduction leads to significant information loss, greatly diminishing the model's performance in facial feature extraction and reconstruction. Specifically, this strategy failed to effectively preserve the details of facial landmarks and contours, resulting in poor overall quality of the face-swapped images. Next, Groups B and C increased the channel dimensions to $1 \times 1 \times 2C$ and $1 \times 1 \times 3C$, respectively, and the results showed improved model performance. Group B achieved an SSIM of 0.75, a pose error of 5.26, and an expression error of 26.56, while Group C further improved to an SSIM of 0.79, a pose error of 3.54, and an expression error of 24.78. This demonstrates that increasing the channel dimensions can better capture fine-grained facial details, such as the eyes, nose, and

mouth, thereby reducing information loss and enhancing the model's performance in facial feature generation and alignment. Group D adopted the strategy of increasing the channel dimensions to $1 \times 1 \times 4C$, and the results showed the best performance. This suggests that moderately increasing the channel dimensions can significantly enhance the model's feature representation capability, making it more effective in handling facial expression details and complex facial poses. However, when the channel dimensions were further increased to $1 \times 1 \times 5C$ (Group E) and $1 \times 1 \times 10C$ (Group F), the model's performance did not improve further, and even declined in some metrics. Group E achieved an SSIM of 0.81, a pose error of 3.92, and an expression error of 25.29, while Group F had an SSIM of 0.76, a pose error of 4.18, and an expression error of 26.11. This indicates that excessive increase in channel dimensions can lead to feature redundancy, negatively impacting the model's performance. Specifically, these strategies may fail to effectively focus on critical facial keypoints, resulting in redundant and unstable feature representations. Ultimately, we set the channel dimension upscaling factor to 4 in the channel attention module. This strategy effectively preserves the original facial features and reduces information loss while significantly enhancing the network's feature extraction capabilities, thereby improving the authenticity and visual consistency of the face-swapped images.

5.3.2. Attention Mechanisms in Face-Swapping

In this ablation study, we analyzed the impact of different attention mechanisms and feature processing methods adopted in the STAM module on the model's performance. Table 4 presents the performance of various experimental groups when applying different attention mechanisms and feature processing methods. The results indicate that Group G, which employs an improved channel attention mechanism (CAM+), spatial attention mechanism (SAM), and feature recalibration method, performed the best with an SSIM of 0.82, a pose error of 2.35, and an expression error of 23.45. In contrast, Group A, which did not use any attention mechanisms or feature processing methods, performed the worst, with an SSIM of 0.72, a pose error of 7.99, and an expression error of 34.02. This demonstrates that attention mechanisms and feature processing methods play a crucial role in enhancing the model's performance. Although Group F performed well in terms of expression retention, it was behind Group G in pose retention and overall image quality. A comparison between Groups D and E revealed that the parallel use of attention modules outperformed their serial use. This is likely because the parallel structure can more effectively capture and process both local facial details and global information, thereby improving the model's performance in facial feature extraction and reconstruction. Additionally, Groups D and E, which did not use the feature recalibration module, performed worse than Groups F and G, which did. This further underscores the importance of feature recalibration in enhancing the model's sensitivity to facial details and improving image reconstruction quality. The feature recalibration module adjusts the weights of feature maps, allowing the model to focus more on key facial features, thus enhancing the realism and visual consistency of the final generated images.

Table 4. Ablation Study of the Internal Modules of STAM.

Group	CAM+	SAM	Cascade	Parallel	Feature Recalibration	SSIM ↑	Pose ↓	Expression ↓
a	×	×	×	×	×	0.72	7.99	34.02
b	√	×	×	×	×	0.74	7.15	32.21
c	×	√	×	×	×	0.74	6.93	33.99
d	√	√	√	×	×	0.75	4.52	29.54
e	√	√	×	√	×	0.78	2.91	24.18
f	√	√	√	×	√	0.80	3.54	**22.95**
g	√	√	×	√	√	**0.82**	**2.35**	23.45

This ablation study demonstrates that employing multiple attention mechanisms and feature processing methods can significantly enhance the performance of face-swapping models, particularly in facial feature generation, alignment, and reconstruction. Specifically, CAM+ and SAM improve the model's sensitivity and accuracy to facial features by capturing both facial details and global information. Meanwhile, the feature recalibration module further enhances the model's feature expression capability and focus on critical features, thereby improving the overall quality of face-swapped images. This finding is significant for the further optimization and development of advanced face-swapping technology. By deeply understanding and applying these attention mechanisms and feature processing methods, we can achieve better results in facial feature generation, alignment, and reconstruction, resulting in more natural and realistic face-swapped images.

5.3.3. Enhanced AdaIN Strategies for Face-Swapping

We improved the AdaIN module to enhance the performance of the face-swapping model, with results shown in Table 5. The results indicate that Group 5, which employed AdaIN, a dynamic parameter adjustment mechanism, multi-scale fusion, and regularization strategies, performed the best. These strategies collectively optimized the model's performance. Specifically, AdaIN normalizes the source image features and then recalibrates them using the mean and standard deviation of the target image, resulting in face-swapped images that retain the identity features of the source image while adopting the style features of the target image. The dynamic parameter adjustment mechanism enhanced the model's adaptability, allowing it to better handle different facial features and expression changes. Multi-scale fusion captured details at various scales, ensuring the delicacy and fineness of the image. The regularization strategy maintained the naturalness and consistency of the image, preventing overfitting.

Table 5. Performance of AdaIN using different strategies.

Group	AdaIN	Dynamic Parameter Adjustment	Multi-Scale Fusion	Regularization	ID ↑	Pose ↓	Expression ↓
1	×	×	×	×	82.18	6.06	30.26
2	√	×	×	×	89.76	5.12	29.12
3	√	√	×	×	92.54	4.25	27.54
4	√	√	√	×	94.08	2.81	24.66
5	√	√	√	√	**95.12**	**2.35**	**23.45**

The combination of these strategies resulted in the best performance for Group 5 in terms of quality, naturalness, and consistency. Specifically, Group 5 achieved an identity preservation (ID) score of 95.12, a pose error of 2.35, and an expression error of 23.45. This demonstrates that the comprehensive use of these advanced feature processing and adjustment strategies can significantly enhance the realism and visual consistency of face-swapped images. In contrast, Group 1, which did not use any improvement strategies, performed the worst, with an identity preservation score of 82.18, a pose error of 6.06, and an expression error of 30.26. This further proves the critical role of AdaIN and its improvement strategies in enhancing model performance. Group 2, which used only AdaIN, showed significant improvement, with an identity preservation score increasing to 89.76 and reductions in pose and expression errors. However, it was only after the introduction of the dynamic parameter adjustment mechanism that Group 3's performance further improved, with an identity preservation score of 92.54 and pose and expression errors reduced to 4.25 and 27.54, respectively. When the multi-scale fusion strategy was added, Group 4's performance significantly increased, with an identity preservation score reaching 94.08 and further reductions in pose and expression errors to 2.81 and 24.66, respectively. Finally, with the addition of the regularization strategy in Group 5, the model performance reached its optimal level.

This study demonstrates that the improved AdaIN module, combined with dynamic parameter adjustment, multi-scale fusion, and regularization strategies, can significantly enhance the performance of face-swapping models, particularly in facial feature generation and style transfer. By working synergistically, these strategies improve the model's adaptability and stability in handling complex facial features, resulting in more natural and realistic face-swapped images.

5.3.4. Comprehensive Strategies for Enhanced Face-Swapping

In this research, we conducted a detailed analysis of the impact of different combinations of STAM, AdaIN+, AmazingSeg, and multi-scale strategies on the performance of face-swapping models. The results, shown in Table 6, revealed that Group 9, which employed all of these strategies comprehensively, performed the best, achieving an SSIM of 0.82, an identity preservation (ID) score of 95.12, and a shape error reduced to 44.61. This indicates that AmazingFS, through the combined use of these improved strategies, significantly enhanced the model's structural similarity, identity preservation, and shape fidelity.

Table 6. The Impact of STAM, AdaIN+, AmazingSeg, and Multi-Scale Strategies on the Performance of AmazingFS.

Group	STAM	AdaIN+	AmazingSeg	Multi-Scale	SSIM ↑	ID ↑	Shape ↓
1	×	×	×	×	0.63	75.26	60.25
2	√	×	×	×	0.68	78.26	58.05
3	×	√	×	×	0.71	84.87	54.82
4	×	×	√	×	0.69	80.03	56.31
5	×	×	×	√	0.66	76.26	56.28
6	√	√	×	√	0.72	86.19	50.28
7	√	×	√	√	0.76	82.18	46.25
8	×	√	√	√	0.72	89.92	47.29
9	√	√	√	√	**0.82**	**95.12**	**44.61**

The STAM module improves the quality and naturalness of face-swapped images by enhancing the model's ability to capture both details and global information. AdaIN+ effectively transfers style features by dynamically adjusting feature parameters, retaining the identity features of the source image while incorporating the style features of the target image. The AmazingSeg module excels in handling facial occlusions and detail restoration, ensuring consistency and authenticity in generated images even in complex scenarios. The multi-scale strategies capture and fuse features at different scales, enhancing the model's adaptability and stability when processing images with varying resolutions and detail levels.

The experimental results showed that while the use of a single strategy offered some improvement, the effects were limited. For instance, using STAM, AdaIN+, AmazingSeg, or the multi-scale strategy alone resulted in certain improvements in SSIM and ID, but not to the optimal level. Significant performance enhancement was only achieved when multiple strategies were combined. For example, the group that combined STAM, AdaIN+, and the multi-scale strategy performed well, but not as well as Group 9, which utilized all strategies. These results indicate that the comprehensive application of STAM, AdaIN+, AmazingSeg, and multi-scale strategies maximizes the overall performance of face-swapping models. This not only significantly enhances structural similarity, identity preservation, and shape fidelity, but also improves the naturalness and consistency of the generated images. By deeply understanding and applying these advanced feature processing and adjustment strategies, we can achieve better results in facial feature generation, alignment, and reconstruction, resulting in more natural and realistic face-swapped images.

5.3.5. Analyzing the Role of AmazingSeg

As shown in Figure 10, the comparison results of face-swapping technology with and without AmazingSeg demonstrate its effectiveness in segmenting the face area more accurately and handling occlusions such as hair, glasses, and microphones. AmazingSeg focuses on learning the segmentation of the face area, making the face-swapping effect more natural and realistic. In the first row of comparison results, without AmazingSeg, there is a noticeable color difference and contour inconsistency between the target face and the source face, especially at the edges of the face and hair area, making the face-swapping effect look unnatural. After using AmazingSeg, the fusion effect of the face significantly improves, with more natural skin tone transitions and more harmonious contours, particularly with finer handling of the edges between the face and hair. The second row shows similar improvements. Without AmazingSeg, the integration of the target face's eyeglasses and facial features is inconsistent, leading to distortions in the source face's features and an unrealistic overall effect. After using AmazingSeg, the fusion of the glasses and face appears more natural, and the features of the source face are better preserved, making the face-swapping effect more realistic and natural. The third and fourth rows further illustrate the advantages of AmazingSeg. Without AmazingSeg, there are noticeable traces at the edges of the fusion between the source face and the target face. After using AmazingSeg, the edge transitions are more natural, and the overall effect looks more harmonious. The comparison results in the last row also significantly demonstrate the role of AmazingSeg in improving the face-swapping effect. Without AmazingSeg, the features of the source face are distorted during the face-swapping process, especially in the facial contours and makeup, making the face-swapping effect appear unrealistic. After using AmazingSeg, the features of the source face are well preserved, with natural integration of facial contours and makeup, resulting in a more realistic overall effect.



Figure 10. Comparison Experiment of Face-Swapping with and without AmazingSeg.

5.3.6. Analyzing the Role of GAN

In Figure 11, a comparative analysis of face-swapping effects with and without the use of Generative Adversarial Networks (GAN) is presented. GAN, an advanced deep learning technology, utilizes two neural networks—a generator and a discriminator—that work antagonistically to produce highly realistic images. This technique significantly enhances realism and detail retention in face-swapping tasks. Results without GAN exhibit unnatural facial textures and lighting, along with poor integration. Faces with glasses show poorly fused eyes and facial features, appearing very unnatural. Faces with beards display poor integration of the beard with facial features, and the overall appearance looks mismatched and unrealistic. In contrast, using GAN markedly improves integration, with more harmonious lighting and better fusion of glasses and facial features, resulting in more natural textures overall. GAN-enhanced results show well-aligned facial features and present a more realistic effect even in complex scenarios involving glasses and beards. This experimental comparison demonstrates that GAN methods provide better facial feature fidelity and adaptability in face-swapping tasks. In AmazingFS, employing a GAN framework for face-swapping achieves higher quality and more natural effects.



Figure 11. Comparison Experiment of Face-Swapping with and without GAN.

5.4. Discussion and Prospects

We propose the AmazingFS method, which integrates STAM, AdaIN+, AmazingSeg, and multi-scale strategies to achieve significant performance improvements. AmazingFS offers several key advantages. It excels in image quality, with the combination of the STAM module and multi-scale strategies ensuring that the generated face-swapped images capture both fine details and global information, thereby enhancing the overall naturalness of the images. The AdaIN+ module dynamically adjusts feature parameters, effectively transferring identity features from the source image while incorporating style features from the target image, significantly improving the identity preservation and style consistency of the face-swapped images. Additionally, the AmazingSeg module demonstrates exceptional

performance in handling facial occlusions and detail restoration, ensuring consistency and authenticity in generated images even in complex scenarios, and enhancing the model's adaptability and stability when dealing with complex facial features. As shown in Figure 12, AmazingFS effectively manages occlusions and meticulously integrates fine details such as the eyes and mouth, demonstrating the method's robustness and precision. This detailed comparison further validates the superior capability of AmazingFS in maintaining realistic and consistent facial features.



Figure 12. AmazingFS Detail Display.

Despite the superior performance of AmazingFS in generating realistic and consistent images, its high computational complexity currently limits real-time face-swapping. Producing face-swapping videos requires several preprocessing and post-processing steps, including face alignment, face detection, face cropping, face blending, and sharpening. These steps are computationally intensive and time-consuming. In the preprocessing phase, face alignment and detection require precise localization and correction of facial features to ensure consistent key points between the source and target images, while face cropping requires accurately extracting the facial region from the image. In the post-processing phase, face blending ensures the seamless integration of the face-swapped image with the background, and sharpening enhances the clarity and detail of the final image. The substantial computational load of these processes limits the feasibility of real-time face-swapping. Therefore, future research should focus on optimizing computational efficiency to achieve faster and real-time face-swapping capabilities.

6. Conclusions

We propose AmazingFS, a novel and efficient face-swapping framework designed for high-fidelity video face-swapping with occlusion resistance. Our proposed STAM attention mechanism effectively focuses on key features by integrating spatial and channel attention mechanisms, enhancing the realism and consistency of the face-swapping effect. To effectively adjust the style attributes of the target face feature map, we improved the AdaIN module, enabling the generated face-swapping results to better retain the identity features of the source face. Additionally, we designed AmazingSeg for face segmentation and created the occluded face dataset AST to address occlusion issues in various scenarios during video face-swapping. Extensive experiments show that our method produces higher fidelity results both quantitatively and qualitatively compared to previous SOTA face-swapping methods, demonstrating strong competitiveness. AmazingFS can produce movie-level face-swapping effects, but AmazingFS has room for improvement in real-time applications. Future research will focus on optimizing the framework to enhance its performance in scenarios that require real-time face-swapping effects.

Author Contributions: Conceptualization, W.S. and Z.Z.; Methodology, W.S. and D.T.; Writing—original draft, W.S. and D.T.; Writing—review and editing, Z.Z. and W.S.; Supervision, Z.Z., W.S., D.T. and L.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The original data presented in the study are openly available in [CelebA-HQ] at [https://www.kaggle.com/datasets/lamsimon/celebahq].

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Kohli, A.; Gupta, A. Detecting deepfake, faceswap and face2face facial forgeries using frequency cnn. *Multimed. Tools Appl.* **2021**, *80*, 18461–18478. [CrossRef]
2. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y.; Bengio, Y. Generative Adversarial Networks. *Commun. ACM* **2014**, *63*, 139–144. [CrossRef]
3. Korshunova, I.; Shi, W.; Dambre, J.; Theis, L. Fast face-swap using convolutional neural networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 3677–3685.
4. Liu, K.; Perov, I.; Gao, D.; Chervoniy, N.; Zhou, W.; Zhang, W. Deepfacelab: Integrated, flexible and extensible face-swapping framework. *Pattern Recognit.* **2023**, *141*, 109628. [CrossRef]
5. Chen, R.; Chen, X.; Ni, B.; Ge, Y. Simswap: An efficient framework for high fidelity face swapping. In Proceedings of the 28th ACM International Conference on Multimedia, Seattle, WA, USA, 12–16 October 2020; pp. 2003–2011.
6. Nguyen, T.T.; Nguyen, Q.V.H.; Nguyen, D.T.; Nguyen, D.T.; Huynh-The, T.; Nahavandi, S.; Nguyen, T.T.; Pham, Q.-V.; Nguyen, C.M. Deep learning for deepfakes creation and detection: A survey. *Comput. Vis. Image Underst.* **2022**, *223*, 103525. [CrossRef]
7. Shiohara, K.; Yang, X.; Taketomi, T. Blendface: Re-designing identity encoders for face-swapping. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2–6 October 2023; pp. 7634–7644.
8. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.

9. Zhu, X.; Cheng, D.; Zhang, Z.; Lin, S.; Dai, J. An empirical study of spatial attention mechanisms in deep networks. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 6688–6697.
10. Huang, X.; Belongie, S. Arbitrary style transfer in real-time with adaptive instance normalization. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 1501–1510.
11. Iglovikov, V.; Shvets, A. Ternausnet: U-net with vgg11 encoder pre-trained on imagenet for image segmentation. *arXiv* **2018**, arXiv:1801.05746.
12. Koestinger, M.; Wohlhart, P.; Roth, P.M.; Bischof, H. Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In Proceedings of the 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops), Barcelona, Spain, 6-13 November 2011; pp. 2144–2151.
13. Rosberg, F.; Aksoy, E.E.; Alonso-Fernandez, F.; Englund, C. Facedancer: Pose-and occlusion-aware high fidelity face swapping. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 2–7 January 2023; pp. 3454–3463.
14. Wang, Y.; Chen, X.; Zhu, J.; Chu, W.; Tai, Y.; Wang, C.; Li, J.; Wu, Y.; Huang, F.; Ji, R. Hififace: 3d shape and semantic prior guided high fidelity face swapping. *arXiv* **2021**, arXiv:2106.09965.
15. Kim, K.; Kim, Y.; Cho, S.; Seo, J.; Nam, J.; Lee, K.; Kim, S.; Lee, K. Diffface: Diffusion-based face swapping with facial guidance. *arXiv* **2022**, arXiv:2212.13344.
16. Li, L.; Bao, J.; Yang, H.; Chen, D.; Wen, F. Faceshifter: Towards high fidelity and occlusion aware face swapping. *arXiv* **2019**, arXiv:1912.13457.
17. Mnih, V.; Heess, N.; Graves, A. Recurrent Models of Visual Attention. *Adv. Neural Inf. Process. Syst.* **2014**, 27.
18. Jaderberg, M.; Simonyan, K.; Zisserman, A. Spatial transformer networks. *Adv. Neural Inf. Process. Syst.* **2015**, 28.
19. Dai, J.; Qi, H.; Xiong, Y.; Li, Y.; Zhang, G.; Hu, H.; Wei, Y. Deformable convolutional networks. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 764–773.
20. Zhu, X.; Hu, H.; Lin, S.; Dai, J. Deformable Convnets V2: More Deformable, Better Results. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 9300–9308.
21. Hu, J.; Shen, L.; Sun, G. Squeeze-and Excitation Networks. *EEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 2011–2023. [CrossRef] [PubMed]
22. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. CBAM: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018.
23. Wang, X.; Girshick, R.; Gupta, A.; He, K. Nonlocal Neural Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018.
24. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv* **2020**, arXiv:2010.11929.
25. Gatys, L.A.; Ecker, A.S.; Bethge, M. Image style transfer using convolutional neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2414–2423.
26. Johnson, J.; Alahi, A.; Fei-Fei, L. Perceptual losses for real-time style transfer and super-resolution. In Proceedings of the Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; Proceedings, Part II 14. Springer International Publishing: Berlin/Heidelberg, Germany, 2016; pp. 694–711.
27. Narayan, K.; VS, V.; Chellappa, R.; Patel, V.M. FaceXFormer: A Unified Transformer for Facial Analysis. *arXiv* **2024**, arXiv:2403.12960.
28. Yin, Z.; Yiu, V.; Hu, X.; Tang, L. End-to-end face parsing via interlinked convolutional neural networks. *Cogn. Neurodyn.* **2021**, *15*, 169–179. [CrossRef] [PubMed]
29. Yu, C.; Gao, C.; Wang, J.; Yu, G.; Shen, C.; Sang, N. Bisenet v2: Bilateral network with guided aggregation for real-time semantic segmentation. *Int. J. Comput. Vis.* **2021**, *129*, 3051–3068. [CrossRef]
30. Zhang, Y.; Yin, Z.; Li, Y.; Yin, G.; Yan, J.; Shao, J.; Liu, Z. Celeba-spoof: Large-scale face anti-spoofing dataset with rich annotations. In Proceedings of the Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Proceedings, Part XII 16. Springer International Publishing: Berlin/Heidelberg, Germany, 2020; pp. 70–85.
31. Zhang, S.; Zhu, X.; Lei, Z.; Shi, H.; Wang, X.; Li, S.Z. S3fd: Single shot scale-invariant face detector. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 192–201.
32. Bulat, A.; Tzimiropoulos, G. How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 1021–1030.
33. Umeyama, S. Least-squares estimation of transformation parameters between two point patterns. *IEEE Trans. Pattern Anal. Mach. Intell.* **1991**, *13*, 376–380. [CrossRef]
34. Deng, J.; Guo, J.; Xue, N.; Zafeiriou, S. Arcface: Additive angular margin loss for deep face recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 4690–4699.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.