future internet

MDPI

Article

Face Swapping Consistency Transfer with Neural Identity Carrier

Kunlin Liu [1], Ping Wang [1], Wenbo Zhou [1,*], Zhenyu Zhang [2], Yanhao Ge [2], Honggu Liu [1], Weiming Zhang [1] and Nenghai Yu [1]

[1] School of Information Science and Technology, University of Science and Technology of China, Hefei 230026, China; lkl6949@mail.ustc.edu.cn (K.L.); wp123@mail.ustc.edu.cn (P.W.); lhg9754@mail.ustc.edu.cn (H.L.); zhangwm@ustc.edu.cn (W.Z.); ynh@ustc.edu.cn (N.Y.)
[2] Tencent Youtu, Shanghai 200233, China; zhangjesse@foxmail.com (Z.Z.); halege@tencent.com (Y.G.)
* Correspondence: welbeckz@ustc.edu.cn

Abstract: Deepfake aims to swap a face of an image with someone else's likeness in a reasonable manner. Existing methods usually perform deepfake frame by frame, thus ignoring video consistency and producing incoherent results. To address such a problem, we propose a novel framework Neural Identity Carrier (NICe), which learns identity transformation from an arbitrary face-swapping proxy via a *U-Net*. By modeling the incoherence between frames as noise, NICe naturally suppresses its disturbance and preserves primary identity information. Concretely, NICe inputs the original frame and learns transformation supervised by swapped pseudo labels. As the temporal incoherence has an uncertain or stochastic pattern, NICe can filter out such outliers and well maintain the target content by uncertainty prediction. With the predicted temporally stable appearance, NICe enhances its details by constraining 3D geometry consistency, making NICe learn fine-grained facial structure across the poses. In this way, NICe guarantees the temporal stableness of deepfake approaches and predicts detailed results against over-smoothness. Extensive experiments on benchmarks demonstrate that NICe significantly improves the quality of existing deepfake methods on video-level. Besides, data generated by our methods can benefit video-level deepfake detection methods.

Keywords: deepfake generation; face swapping; consistency transfer

Citation: Kunlin, L.; Ping, W.; Wenbo, Z.; Zhenyu, Z.; Honggu, L.; Yangao, G.; Weiming, Z.; Nenghai, Y. Face Swapping Consistency Transfer with Neural Identity Carrier. *Future Internet* **2021**, *13*, 298. https://doi.org/10.3390/fi13110298

Academic Editor: Jari Jussila

Received: 8 November 2021
Accepted: 19 November 2021
Published: 22 November 2021

1. Introduction

Deepfake technique has ignited extensive interests in both academia and industry in recent years and inspires plenty of applications such as entertainment [1] and privacy applications [2]. It aims to swap a face of an image with someone else's likeness in a reasonable manner.

Recent studies have shown that high-fidelity face-swapping generation is achievable [3–5]. By disentangling the identity information and attribute information from images, they achieve excellent performance in frame-level face swapping [6,7]. These high-quality face-swapping results are spread in social media, which causes significant malicious influences. Researches about deepfake also attract tremendous attention in the academic community [8–10]. However, they swap faces by simply merging different features extracted from different person frame by frame, which may lead to unnatural results.

Generating continuous face-swapping sequences is a very challenging task. Directly generating face-swapping sequences might enhance the consistency, but it is computationally infeasible in the current environment. The main issue for the face-swapping task is how do we ensure continuity in final results. We try to find a way to inherit the continuity from the origin video directly. Inspired by prior work, we observe that the structure of a generator network is sufficient to capture the low-level statistics of a natural image or video [11,12]. Based on this observation, we conjecture that the flickering artifacts in a forged video are similar to the noise in the temporal domain. We can use a neural network to inherit the continuity from the origin video.

Until now, we have decided on the starting point of this task. However, the ending point is unreliable because of the proxy's instability, as shown in Figure 1. Directly using the previous face-swapping proxy as a reference will cause the results' artifacts because the artifacts in the face-swapping proxy will also be inherited. To address this issue, we introduce an aleatoric uncertainty loss that can tolerate the uncertainty in proxy data during our training. Furthermore, to get higher-quality results, we introduce static 3D detail supervision for fine-grained detail reconstruction.

In this paper, we propose a novel Neural Identity Carrier (NICe), which learns identity transformation from an arbitrary face-swapping proxy via a *U-Net*. To better model the inconsistency of face-swapping proxy, we introduce an aleatoric uncertainty loss that can tolerate the uncertainty in proxy data, and force our NICe to better learn the primary identity information in the meantime. Besides, we also introduce detail consistency transfer to guarantee the fine-grained detail information, i.e., moles and wrinkles. Extensive experiments on different types of face-swapping videos demonstrate the superiority of our method both qualitatively and quantitatively, including better retention of the attribute information from the target.

Figure 1. Previous methods suffers from two main problems in frame-level. First, they cannot inherit whole pose information from target image, i.e., gaze direction deviation. Besides, they cannot generate harmony results in complex environments, i.e., shadow areas.

The main contributions of this paper can be summarized as follows:

- We propose a novel Neural Identity Carrier (NICe), which learns identity transformation from an arbitrary face-swapping proxy via a *U-Net*.
- To better model the inconsistency of face-swapping proxy, we borrow the theory of aleatoric uncertainty. Moreover, we introduce aleatoric uncertainty loss to tolerate the uncertainty in proxy data and force our NICe to learn the primary identity information in the meantime.
- With the predicted temporally stable appearance, we further introduce static detail supervision to help NICe to generate results with more fine-grained details.
- We also verify that the refined forgery data can help to improve temporal-aware deepfake detection performance.

The rest of this paper is organized as follows. Related works of face swapping approach, uncertainty modeling, and 3D face reconstruction are presented in Section 2. A detailed description of the proposed method is explained in Section 3. Section 4 demonstrates the experimental results both quantitatively and qualitatively and provides ablation study results. Section 5 presents a discussion of the proposed work, including the advantage of the framework, limitations, and broader impact. Finally, Section 6 presents a conclusion of the whole work.

2. Related Work

In this section, we review the related work from three aspects: face-swapping approaches, uncertainty modeling, and 3D face reconstruction.

2.1. Face-Swapping Approaches

Face-swapping has a long history in vision and graphic research, going back nearly two decades. They are proposed due to privacy concerns first, while they are more used for entertainment [1]. The earliest swapping methods require manual adjustment [2]. Bitouk et al. propose an automatic face-swapping method [13]. However, these methods cannot produce satisfactory results. Recently, learning-based methods have achieved better performance. Deepfakes used auto-encoder to swap faces between identity and target [14]. Ivan et al. upgraded the structure and launched an open-source project, DeepFaceLab (DFL), which is the most popular one on the Internet [15]. Nirkin et al. used a fixed 3D face shape as the proxy to increase the controllability of face-swapping [5]. Nirkin et al. proposed subject-agnostic methods which can be applied to any pair of faces without training on them [4]. And Li et al. propose a two-stage method that can achieve high fidelity and occlusion aware face-swapping [3].

Previous methods suffer from their backbone heavily. For example, auto-encoder-based methods utilize an encoder to disentangle the target person's attribute and identity person's identity information and reconstruct them back by a decoder—a large amount of effective information lost in the encoder-decoder process [15]. GAN-based methods cannot deal with the problem of temporal consistency and produce abnormal results occasionally [4]. In this paper, we leverage a *U-Net* as neural identity carrier to carry the primary information from face-swapping proxy, significantly avoiding the loss of information and producing coherent results.

2.2. Uncertainty Modeling

There are two main types of uncertainty: epistemic (uncertainty of model) and aleatoric (uncertainty of data) in deep learning fileds [16]. Thus the predictive uncertainty should consist of two parts, epistemic uncertainty and aleatoric uncertainty. As the face-swapping proxy performs severe inconsistency, the main kind of uncertainty for this issue is the aleatoric uncertainty. Further, aleatoric uncertainty also has two sub-types: homoscedastic and heteroscedastic [17].

The homoscedastic regression assumes constant observation noise σ for all input point while the heteroscedastic regression, on the other hand, assumes that observation noise can vary with input [18,19]. Especially, the heteroscedastic models are helpful when parts of the observation space might have higher noise levels than others. In previous face-swapping work, the observation noise parameter σ is often fixed as part of the model's weight decay.

Previous methods point out that the observation noise parameter σ can be learned as a function when data is independent [17]. Given the output, we can perform MAP inference to find a single value for the model parameters θ:

$$\mathcal{L}(\theta) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{2\sigma(x_i)^2}||y_i - f(x_i)||^2 + \frac{1}{2}\log\sigma(x_i)^2 \quad (1)$$

where y_i is the ground truth of the output data, $f(\cdot)$ is the model's function, x_i is the input data point, N is the number of data points, σ is the model's observation noise parameter which captures how much noise we have in the outputs, and θ is the distribution's parameters to be optimized.

In our work, we realize the artifacts in face-swapping proxy always occur in facial outlines and local patches. The inconsistency of face-swapping results always performs like facial outline flicker, mouth area collapse, and eye shaking. We leverage aleatoric uncertainty to predict the output's difficult-to-generate area according to the input and reduce the weight of these areas.

2.3. 3D Face Reconstruction

3D face reconstruction has been a longstanding task in computer vision and computer graphics. It shows excellent potential in the face-swapping task. Previous face-swapping techniques tried to utilize 3DMM regression as auxiliary information to assist attribute disentanglement [20,21]. However, they only use coarse 3D reconstruction because they leverage 3D information to solve large-pose problems.

Recently, Chaudhuri et al. [22] learn the identity and expression corrective blend shapes with dynamic (expression-dependent) albedo maps. They model geometric details as part of the albedo map, and therefore, the shading of these details does not adapt to cases with varying lighting. Feng et al. propose to model facial details as geometric displacements and achieve significant improvement than previous methods [23].

Despite previous face-swapping methods utilizing 3D information to supervise their training, they only use the coarse information [4,7]. Motivated by these recent developments of 3D face reconstruction, NICe leverages the temporally stable information with static 3D detail information to build very realistic results while remedying the noise's affecting.

3. Methods

Existing face-swapping methods take identity and target image/video pairs as input. In this paper, we treat the face-swapping problem from a novel perspective. We focus on consistency inheritance in the whole process. Given an identity X_{id} and a target X_t, here X_{id} and X_t can be any portrait image or video, we first use existing face-swapping methods to generate a face-swapping proxy, denoted as X_{ref}.

Taking X_{ref} as references, we train a *U-Net* as a neural identity carrier to carry the primary information of the face-swapping proxy. During the training stage, we introduce a coarse encoder E_c and a detail encoder E_d to reconstruct a series of face parameters, including albedo coefficients α, separate linear identity shape β and detail δ, which will be used as constraints of the transfer learning to generate a photo-realistic result X_o.

3.1. Initial Face Swapping

As shown in the left of Figure 2a, current face-swapping methods can be regarded as a facial attributes disentanglement and re-combination process of identity and target portraits, in which X_{id} provides identity information of the identity and X_t provide attribute information of the target. We use existing face-swapping methods to generate face-swapping proxies X_{ref}. By fusing the identity and attribute embeddings, the swapped results X_{ref} will inherit X_{id}'s identity traits and have X_t's other information. Due to the limitation of existing methods, the X_{ref} can suffer from the problems of inconsistency and visual artifacts.

3.2. Consistency Transfer

After obtaining X_{ref} as a reference, we focus on the consistency transfer. The consistency transfer consists of two parts, coherence consistency transfer—inheriting the coherence from input video and detail consistency transfer—inheriting the static detail information from identity image.

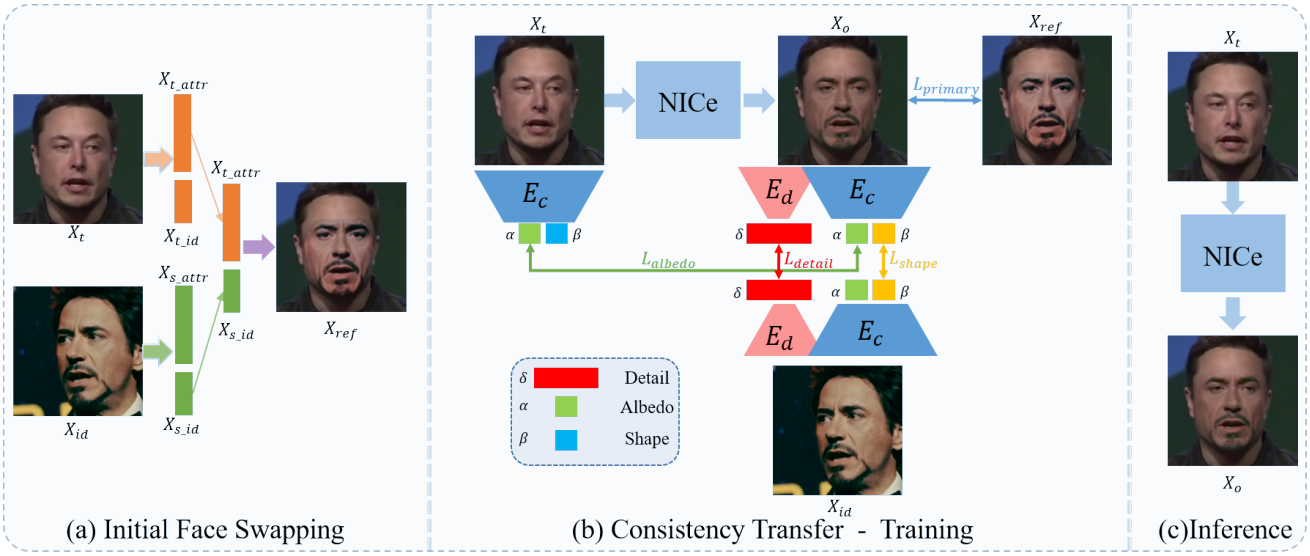

Figure 2. The pipeline of our proposed framework. In the initial face-swapping stage, the face-swapping proxy X_{ref} is obtained by swapping the identity face X_{id} to the target face X_t. We utilize the NICe to extract the face-swapping proxy's information and train the NICe under 3D supervision in the consistency transfer stage. We can directly input a target image/video for inference. This framework is efficient in producing coherent and realistic swapped results.

3.2.1. Coherence Consistency Transfer

As mentioned before, applying swapping algorithms independently to each frame often leads to temporal inconsistency in the generated video due to the discrete input distribution. Inspired by the DVP [12], utilizing CNN to simulate unstable processing algorithms is an efficient way to improve the temporal consistency of video produced by image algorithms. The flickering artifacts in an imperfect swapped video are similar to the noise in the temporal domain, while convolutional networks can reconstruct noise-free content before the noise. Thus we believe the temporal noise of the initial swapped video can be corrected by the re-expression of the neural identity carrier. As shown in Figure 2b, we take *U-Net* as a NICe to remove the flickering artifacts based on face-swapping proxy X_{ref}. During the training stage, the neural identity carrier takes X_t as input, and generate the re-expression result X_o.

3.2.2. Detail Consistency Transfer

Prior face-swapping methods rely on heavy training on input data to synthesize realistic and abundant details, such as wrinkles and moles. But the excessive training will cause the carrier's degradation that the *U-Net* will no longer learn the noise-free contents but noises themselves. Thus, the over-trained *U-Net*'s results are inevitably direct to the flickers, and visual artifacts appear. On the contrary, the basic facial information can not be preserved well if we train the model insufficiently. To address the problem, we propose to introduce a novel 3D representation manner to help enhance the detail information of X_o without suffering the issues brought by the excessive or insufficient training process.

According to the observation, one individual will show different details when taking different expressions and poses. The detail information of a subject is not all static. To address this issue, we suppose that the detail information should be separated into two parts, dynamic detail, which represents expression-related detail information, and static detail, which represents resident detail information. In this paper, we utilize a detail UV displacement map D to represent the details (both dynamic and static). By extracting the static detail information from identity images, NICs can learn fine-grained facial structure.

3.3. Static Detail Extractor

Getting a static detail extractor is not easy. First we adopt a pre-trained state-of-the-art 3D reconstruction model [23] as a coarse encoder. This coarse encoder E_c enables 3D disentanglements in FLAME's model space [24] and regress a series of FLAME parameters, geometry parameters β, ψ and θ, albedo parameters α, camera parameters c and lighting parameters l. Among geometry parameters, β describes the shape information, ψ is the expression parameters, θ represents other coarse geometry information, such as the angle of jaw, nose, and eyeballs.

We conjecture that the dynamic detail information can be represented by the expression parameters ψ and the pose-related parameter θ. To gain an efficient static detail representation, we propose to train an extractor E_d, with the same architecture as E_c, to extract the static detail information, i.e., moles and wrinkles, from input images.

As shown in Figure 3, the extractor E_d encodes input image I_j into a latent code δ which represents static detail of I_j. Subsequently, we concatenate the latent code δ with expression parameters ψ and pose parameters θ. Such a combination is finally decoded by displacement decoder F_d to displacement D. The process of decode detail feature can be formulated as,

$$D = F_d(\delta, \theta, \psi) \tag{2}$$

where δ controls the static detail, θ and ψ both control the dynamic detail. Then, we convert D to a normal map. And By converting original geometry M and its surface normal N to UV space, denoted as M_{un} and N_{un}, we can calculate the detail geometry M_d from them. We formulate this process as

$$M_d = M_{uv} + D \odot N_{uv} \tag{3}$$

Once the detail geometry M_d obtained, the detail normal N_d can be derived easily. Then we obtain the detail rendering result I'_r through rendering M_d with detail normal N_d as

$$I'_r = \mathcal{R}(M_d, B(\alpha, l, N_d), c) \tag{4}$$

where $\mathcal{R}$ is a differentiable mesh renderer [25] and B is the shaded texture, represented in UV coordinates. The obtained detail parameters are then used to constrain the network for more realistic results.

3.4. Training Losses

In the first stage, the initial face-swapping method can be any existing method. We primarily introduce the training process of consistency transfer and static detail extractor in this section. There are two trainable parts in our framework: static detail extractor E_d and neural identity carrier *U-Net*. To train a high-quality carrier network, we need to train a good extractor first.

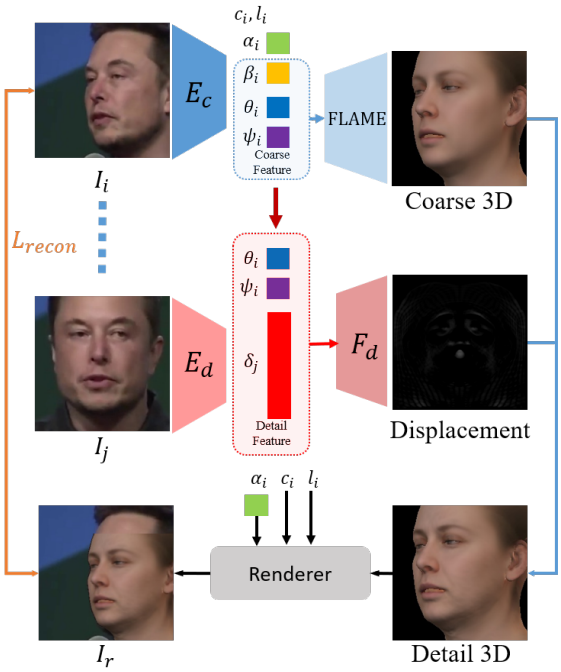

Figure 3. Illustration of our 3D detail extractor's training process. E_c is the state-of-the-art 3D reconstruction model which disentangles the input face. The disentangled face parameters are then recombined into coarse feature and detail feature respectively.

3.4.1. Static Detail Extractor Training

In Section 3.3, we introduce the pipeline of detail reconstruction. Given a set of images from one individual, the detail reconstruction is trained by minimizing $\mathcal{L}_{recon}$, formally as

$$\mathcal{L}_{recon} = \mathcal{L}_{pho} + \mathcal{L}_{mrf} + \mathcal{L}_{sym} + \mathcal{L}_{chr} + \mathcal{L}_{reg}, \tag{5}$$

with photometric loss $\mathcal{L}_{pho}$, ID-MRF loss $\mathcal{L}_{mrf}$, soft symmetry loss $\mathcal{L}_{sym}$, coherence loss $\mathcal{L}_{chr}$ and regularization loss $\mathcal{L}_{reg}$.

The photometric loss L_{pho} computes the distance of the input image I and the rendering I_r as $L_{pho} = \|V_I \odot (I - I_r)\|$. Here, V_I is a binary mask generated by a face segmentation method [5] which represents the facial region, and $\odot$ denotes the Hadamard product. The photometric loss L_{pho} can enforce more attention focused on the facial region and awareness of occlusions with the help of mask V_I.

Besides, we adopt the Implicit Diversified Markov Random Fields (ID-MRF) loss for geometric details reconstruction [26]. Given two images of the same subject, the ID-MRF loss minimizes the distance between these two images on VGG19's feature level. As the same setting as previous work [26], we compute the ID-MRF loss on layers *conv*3_2 and *conv*4_2 of VGG19 as

$$\mathcal{L}_{mrf} = 2\mathcal{L}_M(conv4\_2) + \mathcal{L}_M(conv3\_2), \tag{6}$$

where $\mathcal{L}_M(layer)$ denotes the VGG19's feature-level distance between I'_r and I on layer *layer* of VGG19.

In consideration of occlusions, we also add a soft symmetry loss to regularize the non-visible face parts. The soft symmetry loss can be formulated as

$$\mathcal{L}_{sym} = V_{uv} \odot (D - Flip(D)), \tag{7}$$

where V_{uv} denotes the facial mask in UV space, and $Flip$ denotes the horizontal flip operation.

As mentioned in Section 3.2, detail information is divided into two parts, dynamic and static. We believe that replacing the static detail codes of another image of the same subject should have no effect on the final rendered image, which conforms to the logical evidence that one specific person should have his own consistent static detail code. Formally, given two images I_i and I_j of the same subject, the loss is defined as

$$\begin{aligned}\mathcal{L}_{chr} = \| I_i - \mathcal{R}(M(\beta_i, \theta_i, \psi_i), A(\alpha_i), \\ F_d(\delta_j, \psi_i, \theta_i), l_i, c_i)) \|^2\end{aligned} \tag{8}$$

where β_i, θ_i, ψ_i, α_i, l_i, and c_i are the parameters of I_i, while δ_j is the detail code of I_j.

Finally, the detail displacements D are regularized by $\mathcal{L}_{reg} = \|D\|_{1,1}$ to reduce noise.

3.4.2. Neural Identity Carrier Training

Given target and reference image/video X_t, X_{ref} and an identity identity image X_{id}, the transfer network is trained by minimizing

$$\mathcal{L}_{transfer} = \mathcal{L}_{primary} + \mathcal{L}_{3D} \tag{9}$$

To learn the process of identity transformation, a primary loss is essential. In consideration of the artifacts in face-swapping proxy, we model the uncertainty at the same time. We adjust the aleatoric uncertainty loss to fit our scenario. The primary loss is formulated as

$$\begin{aligned}\mathcal{L}_{primary} = & \frac{1}{2\sigma(X_i)^2} ||VGG(X_o) - VGG(X_{ref})||^2 \\ & + \frac{1}{2} \log \sigma(X_i)^2\end{aligned} \tag{10}$$

where $VGG(\cdot)$ denotes the VGG features which consist of features from layers $conv1\_2$, $conv2\_2$, $conv3\_2$, $conv4\_2$ and $conv5\_2$. The σ denotes the model's noise parameter—predicting how much noise we have in the outputs. It is noteworthy that we learn the noise parameter σ implicitly from the loss function. $\mathcal{L}_{primary}$ is a basic perceptual error between X_o and X_{ref}. This loss can basically guarantee that the NICe can learn identity transformation from arbitrary face-swapping proxy.

To enhance the quality of simulation, we adopt 3D losses with trained static detail extractor E_d and coarse encoder E_c. 3D losses consist of three components, albedo loss $\mathcal{L}_{albedo}$, shape loss $\mathcal{L}_{shape}$ and detail loss $\mathcal{L}_{detail}$, formulated as

$$\mathcal{L}_{3D} = \mathcal{L}_{albedo} + \mathcal{L}_{shape} + \mathcal{L}_{detail} \tag{11}$$

In consideration of the swapping area, the skin consistency between the face and the neck can be perceived by the human vision system easily. We utilize albedo loss to improve albedo consistency between X_o and X_t. The albedo loss is defined as

$$\mathcal{L}_{albedo} = \|\alpha_{X_o} - \alpha_{X_t}\| \tag{12}$$

where α_{X_o} and α_{X_t} are the albedo coefficients of X_o and X_t, encoded by E_c respectively.

The shape loss $\mathcal{L}_{shape}$ focuses on identity preserving. Formally, we minimize

$$\mathcal{L}_{shape} = \|\beta_{X_o} - \beta_{X_{id}}\| \tag{13}$$

where β_{X_o} and X_{id} are the shape parameters of X_o and X_{id} encoded by E_c respectively.

The detail loss $\mathcal{L}_{detail}$ can greatly enhance detail information. We define it as

$$\mathcal{L}_{detail} = \|\delta_{X_o} - \delta_{X_{id}}\| \tag{14}$$

where δ_{X_o} and $\delta_{X_{id}}$ are detail information's latent code of X_o and X_{id} encoded by E_d respectively.

4. Experiments

In this part, we compare our framework with several state-of-the-art face-swapping methods by taking them as face-swapping proxies, including FaceSwap [11], DeepFakes [14], FSGAN [4] and FaceShifter [3]. The initial swapped face videos of FSGAN are built by ourselves, while others are collected from the FF++ dataset [27].

4.1. Quantitative Evaluation

For the quantitative evaluation, we compare the temporal consistency and attribute differences among the results of ours and others. We use the stability error e_{stab} to measure the temporal consistency:

$$e_{stab}(O_t, O_{t-1}) = M^f \odot ||O_t - \mathcal{W}_{t-1}^t(O_{t-1})||^2, \tag{15}$$

where $e_{stab}(O_t, O_{t-1})$ measures the coherence between two adjacent output O_t and O_{t-1}, M^f is the facial area mask, $\mathcal{W}_{t-1}^t(\cdot)$ is the function to warp O_{t-1} to time step t using the ground truth backward flow as defined in [28], O_t and O_{t-1} are the results of frame t and $t-1$. Here, we only evaluate stability in facial regions. Lower stability error indicates more stable results. For the entire video, we use average errors instead. As shown in Table 1, Our method outperforms all mentioned methods which means that our method produces more steady results.

Table 1. Temporal coherence e_{stab} comparison of different face-swapping methods. DF denotes Deepfake, FS denotes FaceSwap, FSGAN denotes FSGAN, and Fshift denotes FaceShifter. Our framework can reduce the stability error of swapped results which represent better temporal coherence.

Methods	DF	FS	FSGAN	FShift
e_{stab}	1.471	1.518	1.498	1.214
Ours-e_{stab}	0.944	1.026	0.928	0.930

We also evaluate the attribute differences, including gaze direction, pose, 2d landmark, and 3d landmark with Openface [29]. A lower difference indicates better inheritance. As shown in Table 2, our method can inherit more attribute information than previous methods.

Table 2. Quantitative comparisons among different face-swapping methods of gaze direction, pose, 2D landmarks, and 3D landmarks. Our method apparently reduces the attribute differences which represents that our method can better inherit the attributes from the target video.

Methods	Gaze	Pose	2D lmk	3D lmk
DF	2.360	2.827	3.302	3.500
Ours-DF	**2.038**	**2.611**	**3.055**	**3.270**
FS	3.555	0.864	1.639	1.581
Ours-FS	**2.665**	**0.729**	**1.379**	**1.340**
FSGAN	2.803	1.469	1.768	1.801
Ours-FSGAN	**2.226**	**1.290**	**1.560**	**1.609**
FShift	2.471	1.085	1.750	1.801
Ours-FShift	**2.201**	**0.945**	**1.650**	**1.647**

4.2. Qualitative Evaluation

For visually demonstrating the superiority of our framework in temporal consistency, we select nine continuous frames in Figure 4 for comparison. It can be observed that the results of FaceSwap are volatile due to the independent deformation for face alignment in each frame, which our framework can significantly solve. FSGAN also suffers a serious consistency problem; adjacent frames' brightness can not maintain stability. This is mainly because that its blending network cannot capture consistent information. Therefore, the facial region becomes brighter and brighter from left to right, while our method can still get very stable results.

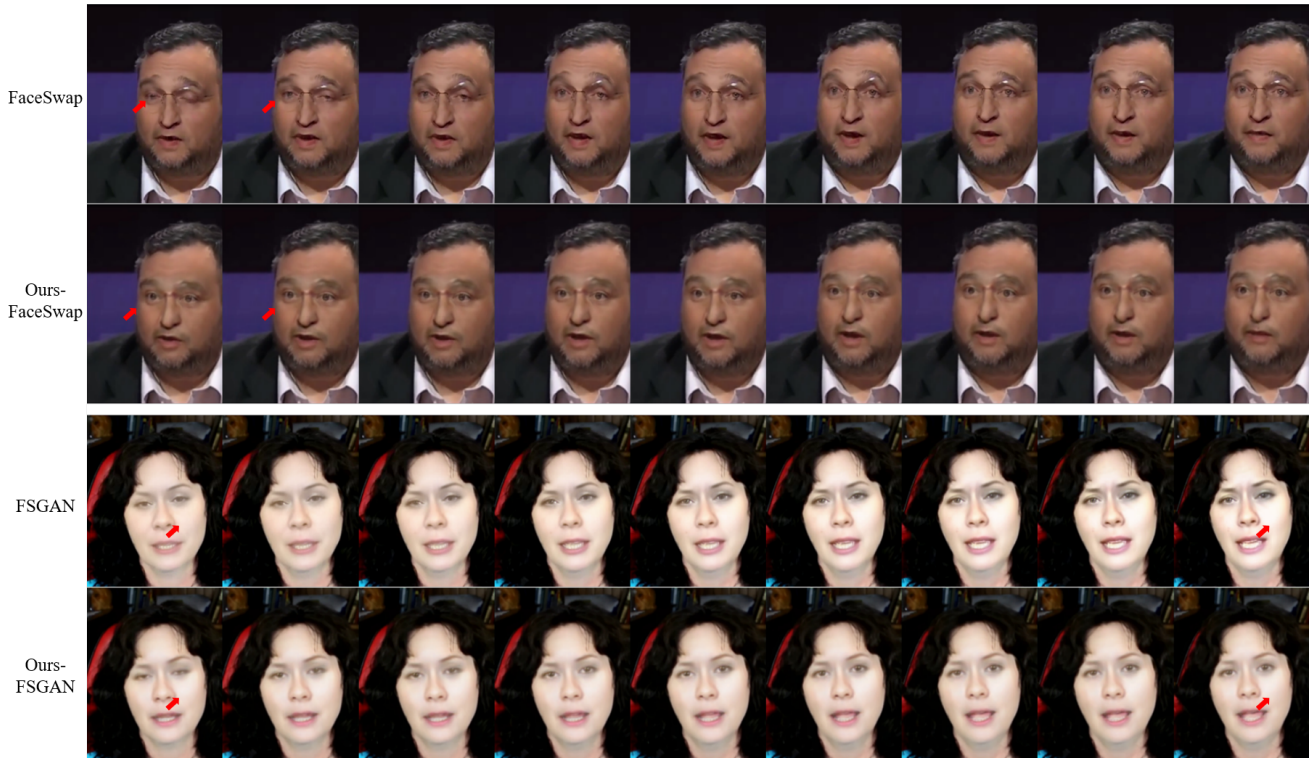

Figure 4. The qualitative evaluation results of our method. The results of FaceSwap are unstable and full of traces of deformation. FSGAN cannot deal with brightness well, which causes bad coherence in the temporal domain. Our method can significantly eliminate the inconsistency in the temporal domain and produce satisfactory results.

4.3. Ablation Study

In this part, we investigate the efficiency of the proposed 3D loss and visualize the corresponding results. We use FSGAN as a basic face-swapping method in this experiment. The results in Figure 5 demonstrate that adopting detail losses can significantly enhance the re-generation quality. Details become richer after adopting detail losses. More specifically, detail information, such as the eyeglasses' shading in row 1 and wrinkles in row 2, are more abundant, which makes the results more realistic.

Figure 5. Ablation study on 3D loss. Under the constraint of 3D loss, the generated result can obtain more detail information and make results more realistic.

4.4. Ability to Improve Forgery Detection

We conduct additional experiments to verify that data synthesized by our framework can help to enhance current forgery detection. We take I3D [30] as the baseline, which is the most efficient video-level forgery detection method and has a recognized generalization ability. We train the baseline on 100 videos from FF++ [27] datasets and evaluate the cross-dataset performance on CelebDF-v2 [31]. Then we utilize our framework to refine the previous 100 videos from FF++ and train I3D on them. Finally, we merge the refined videos with initial videos and train I3D on them. As shown in Figure 6, a model trained on our data can achieve better performances, which indicates that our framework has great value to enhance the current deepfake datasets.

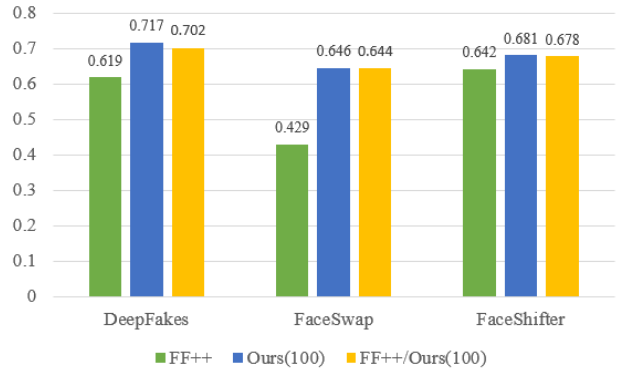

Figure 6. The testing accuracy comparison results on CelebDF-v2 of detection models trained on different datasets. The training data generated by our method provides better temporal coherence and quality which is challenging for detection and is able to help promote the ability of detection models.

5. Discussion

In this section, we discuss the advantages and limitations of our work. Besides, we discuss the broader impact of our work which may bring severe ethical problems.

5.1. Advanced Framework

As shown in Figure 2a, most previous face-swapping methods can be regarded as the facial attributes disentanglement and re-combination between identity and target. It is noteworthy that the face reconstruction models in such methods do not play a fixed role in training and inference. They use attributes from a natural portrait image for training while using edited attributes for inference. Apparently, switching the latent codes between different subjects must have a bad effect on the final result. As shown in Figure 2b,c, unlike

previous methods, our framework only takes X_t as input in both the training and inference stage. The identity information of X_{id} is already learned by NICe and keeps constant in the inference. Thus the final output results X_o can significantly retain more attributes of X_t, such as gaze direction.

Figure 7 gives examples of attributes preservation, here we use DeepFaceLab [15] for comparison. Although DeepFaceLab can produce high-quality swapped results with plenty of post-processing operations, it still suffers from detail inconsistency, such as gaze direction and motion blur. But our framework perfectly inherits the gaze direction and motion blur from target X_t.

Figure 7. Examples of attributes preservation. The first row shows that our method can inherit gaze direction from the target. The second row shows that our method can preserve the same motion blur as the target.

5.2. Limitations

Our framework must use the existing face-swapping method's result as a proxy, which also brings a limitation. The face-swapping proxy limits the quality of generated results. Specifically, if the face-swapping proxy cannot provide satisfying facial content as a reference, our method cannot produce a high-fidelity face even though we introduce detail consistency as supervision.

5.3. Broader Impact

Face-swapping algorithms always face severe ethical problems. We sincerely notice the ethical problem.

Conquering the harmful effects of face-swapping algorithms needs the research of detection algorithms and the investigation of manipulation methods. However, the detection ability always depends on the generation ability. It is challenging to detect high-quality face-swapping videos because attackers can set off a public opinion storm by producing a high-quality video regardless of the costs.

As to deepfake detection, the detectors always need enormous spoofing data to build a robust detection model. Although several datasets have been proposed, there is always a lack of high-quality data. Our method can be leveraged to enhance significantly previous face-swapping methods and build more extensive datasets with coherent and high-quality results.

In the future, we'll expand the current Deepfake dataset (synthesized by our framework) to advance state of the art in Deepfake detection algorithms. With the help of our methods, the high-quality deepfake dataset could be established with high temporal consistency deepfake content.

6. Conclusions

In this paper, we propose a novel neural identity carrier (NICe), which learns identity transformation from an arbitrary face-swapping proxy via a *U-Net*. By neural identity carrier's re-expression and aleatoric uncertainty model, we can eliminate the flickers in the face-swapping proxy. We further introduce static detail supervision to improve the final

results' detail. With the help of NICe, we can revive previous face-swapping methods and strengthen any face-swapping methods.

Author Contributions: Conceptualization, K.L.; Data curation, K.L. and P.W.; Formal analysis, K.L.; Funding acquisition, W.Z. (Weiming Zhang); Investigation, K.L., P.W. and H.L.; Methodology, K.L., W.Z. (Wenbo Zhou) and Z.Z.; Project administration, W.Z. (Wenbo Zhou) and Weiming Zhang; Resources, W.Z. (Wenbo Zhou), Y.G. and Weiming Zhang; Software, K.L. and P.W.; Supervision, W.Z. (Wenbo Zhou), Z.Z., W.Z. (Weiming Zhang) and N.Y.; Validation, K.L., W.Z. (Wenbo Zhou) and H.L.; Visualization, K.L.; Writing—original draft, K.L.; Writing—review & editing, K.L., P.W., W.Z. (Wenbo Zhou), Z.Z., Y.G., H.L., W.Z. (Weiming Zhang) and N.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported in part by the Natural Science Foundation of China under Grant 62002334 and U20B2047, by Anhui Science Foundation of China under Grant 2008085QF296, by Exploration Fund Project of University of Science and Technology of China under Grant YD3480002001, and by Fundamental Research Funds for Central Universities under Grant WK2100000011.

Data Availability Statement: Not Applicable, the study does not report any data.

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

References

1. Alexander, O.; Rogers, M.; Lambeth, W.; Chiang, M.; Debevec, P. Creating a photoreal digital actor: The digital emily project. In Proceedings of the 2009 Conference for Visual Media Production, London, UK, 12–13 November 2009; pp. 176–187.
2. Blanz, V.; Scherbaum, K.; Vetter, T.; Seidel, H.P. Exchanging faces in images. *CGF* **2004**, *23*, 669–676. [CrossRef]
3. Li, L.; Bao, J.; Yang, H.; Chen, D.; Wen, F. Advancing High Fidelity Identity Swapping for Forgery Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 14–19 June 2020.
4. Nirkin, Y.; Keller, Y.; Hassner, T. FSGAN: Subject agnostic face swapping and reenactment. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Korea, 27–28 October 2019; pp. 7184–7193.
5. Nirkin, Y.; Masi, I.; Tuan, A.T.; Hassner, T.; Medioni, G. On face segmentation, face swapping, and face perception. In 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), Xi'an, China, 15–19 May 2018; pp. 98–105.
6. Chen, R.; Chen, X.; Ni, B.; Ge, Y. SimSwap: An Efficient Framework For High Fidelity Face Swapping. In Proceedings of the MM '20: The 28th ACM International Conference on Multimedia, New York, NY, USA, 12 October 2020; pp. 2003–2011. [CrossRef]
7. 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. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21, Montréal, QC, Canada, 21 August 2021*; Zhou, Z.H., Ed.; International Joint Conferences on Artificial Intelligence Organization: 2021; pp. 1136–1142. Available online: https://arxiv.org/pdf/2106.09965.pdf (accessed on 18 November 2021).
8. Li Fan, W.L.; Cui, X. Deepfake-Image Anti-Forensics with Adversarial Examples Attacks. *Future Internet* **2021**, *13*, 288. [CrossRef]
9. Hewage, C.; Ekmekcioglu, E. Multimedia Quality of Experience (QoE): Current Status and Future Direction. *Future Internet* **2020**, *12*, 121. [CrossRef]
10. Khalil, S.S.; Youssef, S.M.; Saleh, S.N. iCaps-Dfake: An Integrated Capsule-Based Model for Deepfake Image and Video Detection. *Future Internet* **2021**, *13*, 93. [CrossRef]
11. Ulyanov, D.; Vedaldi, A.; Lempitsky, V. Deep image prior. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018.
12. Lei, C.; Xing, Y.; Chen, Q. Blind Video Temporal Consistency via Deep Video Prior. Advances in Neural Information Processing Systems. 2020. Available online: https://proceedings.neurips.cc//paper/2020/hash/0c0a7566915f4f24853fc4192689aa7e-Abstract.html (accessed on 18 November 2021).
13. Bitouk, D.; Kumar, N.; Dhillon, S.; Belhumeur, P.; Nayar, S.K. Face Swapping: Automatically replacing faces in photographs. *ACM SIGGRAPH* **2008**, *27*, 39. Available online: https://dl.acm.org/doi/abs/10.1145/1399504.1360638 (accessed on 18 November 2021). [CrossRef]
14. DeepFakes. FaceSwap. 2017. Available online: https://github.com/deepfakes/faceswap (accessed on 6 February 2019).
15. Perov, I.; Gao, D.; Chervoniy, N.; Liu, K.; Marangonda, S.; Um'e, C.; Dpfks, M.; Luis, R.; Jiang, J.; Zhang, S.; et al. DeepFaceLab: A simple, flexible and extensible face swapping framework. *arXiv* **2020**, arXiv:2005.05535.
16. Kiureghian, A.D.; Ditlevsen, O. Aleatory or epistemic? Does it matter? *Struct. Saf.* **2009**, *31*, 105–112. [CrossRef]
17. Kendall, A.; Gal, Y. What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision? *NIPS* 15 March 2017; pp. 5580–5590. Available online: https://arxiv.org/abs/1703.04977 (accessed on 18 November 2021).

18. Nix, D.; Weigend, A. Estimating the mean and variance of the target probability distribution. In Proceedings of the 1994 IEEE International Conference on Neural Networks (ICNN'94), Orlando, FL, USA, 28 June–2 July 1994; Volume 1, pp. 55–60. Available online: https://ieeexplore.ieee.org/abstract/document/374138 (accessed on 18 November 2021). [CrossRef]
19. Le, Q.V.; Smola, A.J.; Canu, S. Heteroscedastic Gaussian Process Regression. In Proceedings of the 22nd International Conference on Machine Learning, Bonn, Germany, 7–11 August 2005; Association for Computing Machinery: New York, NY, USA; pp. 489–496. Available online: https://dl.acm.org/doi/abs/10.1145/1102351.1102413 (accessed on 18 November 2021). [CrossRef]
20. Nagano, K.; Seo, J.; Xing, J.; Wei, L.; Li, Z.; Saito, S.; Agarwal, A.; Fursund, J.; Li, H. PaGAN: Real-Time Avatars Using Dynamic Textures. *ACM Trans. Graph.* **2018**, *37*. [CrossRef]
21. Thies, J.; Zollhöfer, M.; Nießner, M. Deferred Neural Rendering: Image Synthesis Using Neural Textures. *ACM Trans. Graph.* **2019**, *38*. [CrossRef]
22. Chaudhuri, B.; Vesdapunt, N.; Shapiro, L.; Wang, B. Personalized face modeling for improved face reconstruction and motion retargeting. *arXiv* **2020**, arXiv:2007.06759.
23. Feng, Y.; Feng, H.; Black, M.J.; Bolkart, T. Learning an Animatable Detailed 3D Face Model from In-The-Wild Images. 2020. Available online: http://xxx.lanl.gov/abs/2012.04012 (accessed on 18 November 2021).
24. Li, T.; Bolkart, T.; Black, M.J.; Li, H.; Romero, J. Learning a model of facial shape and expression from 4D scans. *ACM Trans. Graph. (Proc. SIGGRAPH Asia)* **2017**, *36*, 194–201. [CrossRef]
25. Ravi, N.; Reizenstein, J.; Novotny, D.; Gordon, T.; Lo, W.Y.; Johnson, J.; Gkioxari, G. Accelerating 3D Deep Learning with PyTorch3D. *arXiv* **2020**, arXiv:2007.08501.
26. Wang, Y.; Tao, X.; Qi, X.; Shen, X.; Jia, J. Image Inpainting via Generative Multi-column Convolutional Neural Networks. In *Advances in Neural Information Processing Systems*; 2018; pp. 331–340. Available online: https://arxiv.org/abs/1810.08771 (accessed on 18 November 2021).
27. Rössler, A.; Cozzolino, D.; Verdoliva, L.; Riess, C.; Thies, J.; Nießner, M. FaceForensics++: Learning to Detect Manipulated Facial Images. International Conference on Computer Vision (ICCV). 2019. Available online: https://openaccess.thecvf.com/content_ICCV_2019/html/Rossler_FaceForensics_Learning_to_Detect_Manipulated_Facial_Images_ICCV_2019_paper.html (accessed on 18 November 2021).
28. Chen, D.; Liao, J.; Yuan, L.; Yu, N.; Hua, G. Coherent Online Video Style Transfer. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017. Available online: https://openaccess.thecvf.com/content_iccv_2017/html/Chen_Coherent_Online_Video_ICCV_2017_paper.html (accessed on 18 November 2021).
29. Baltrusaitis, T.; Zadeh, A.; Lim, Y.C.; Morency, L. OpenFace 2.0: Facial Behavior Analysis Toolkit. In Proceedings of the 2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018), Xi'an, China, 15–19 May 2018; pp. 59–66. Available online: https://ieeexplore.ieee.org/abstract/document/8373812 (accessed on 18 November 2021). [CrossRef]
30. Carreira, J.; Zisserman, A. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017. Available online: https://openaccess.thecvf.com/content_cvpr_2017/html/Carreira_Quo_Vadis_Action_CVPR_2017_paper.html (accessed on 18 November 2021).
31. Li, Y.; Yang, X.; Sun, P.; Qi, H.; Lyu, S. Celeb-DF: A Large-scale Challenging Dataset for DeepFake Forensics. In Proceedings of the IEEE Conference on Computer Vision and Patten Recognition (CVPR), Seattle, WA, USA, 14–19 June 2020. Available online: https://openaccess.thecvf.com/content_CVPR_2020/html/Li_Celeb-DF_A_LargeScale_Challenging_Dataset_for_DeepFake_Forensics_CVPR_2020_paper.html (accessed on 18 November 2021).