MuNeRF: Robust Makeup Transfer in Neural Radiance Fields

Yu-Jie Yuan†, Xinyang Han†, Yue He, Fang-Lue Zhang, and Lin Gao*

***Abstract*—There has been a high demand for facial makeup transfer tools in fashion e-commerce and virtual avatar generation. Most of the existing makeup transfer methods are based on the generative adversarial networks. Despite their success in makeup transfer for a single image, they struggle to maintain the consistency of makeup under different poses and expressions of the same person. In this paper, we propose a robust makeup transfer method which consistently transfers the makeup style of a reference image to facial images in any poses and expressions. Our method introduces the implicit 3D representation, neural radiance fields (NeRFs), to ensure the geometric and appearance consistency. It has two separate stages, including one basic NeRF module to reconstruct the geometry from the input facial image sequence, and a makeup module to learn how to transfer the reference makeup style consistently. We propose a novel hybrid makeup loss which is specially designed based on the makeup characteristics to supervise the training of the makeup module. The proposed loss significantly improves the visual quality and faithfulness of the makeup transfer effects. To better align the distribution between the transferred makeup and the reference makeup, a patch-based discriminator that works in the pose-independent UV texture space is proposed to provide more accurate control of the synthesized makeup. Extensive experiments and a user study demonstrate the superiority of our network for a variety of different makeup styles.**

***Index Terms*—Makeup Transfer, Neural Radiance Field, Patch GAN.**

I. Introduction

With the explosive development of the metaverse and digital human, there has been a high demand for innovative solutions for facial image generation. Deep neural networks have significantly advanced face synthesis and enabled intelligent face editing tools [1], [2]. Recently, facial makeup transfer has attracted a good amount of research interest, which has broad application prospects, such as virtual makeup try-on in fashion e-commerce and VR/AR games. Existing methods [3]–[5] have enabled users to see themselves in different makeup styles, even when there are certain differences between the user's photo and the reference makeup image. Nevertheless, they mainly focus on 2D makeup transfer without resorting to complex facial geometry modeling, while 3D-consistent makeup transfer with different poses and expressions provides greater value to users. The users of virtual makeup applications usually need to view the generated face under different poses and expressions continuously. For example, when virtually trying a new cosmetic product, users often rotate their heads and make facial movements to check their look. Therefore, a robust makeup transfer method capable of preserving visual consistency across different 3D facial poses and movements would greatly increase the accessibility and applications of virtual makeup technology.

Reference — Makeup results with different poses and expressions

Fig. 1. Our method is able to transfer makeup styles (left column) to facial images of different poses and expressions while preserving the geometry and appearance consistency.

There remain two challenges for transferring makeup with arbitrary poses and expressions. First, most 2D-based approaches struggle to handle dramatically different poses and expressions since they usually align facial features on frontal faces or limit the head rotation angle in the data-processing stage. Although some works such as [4] can cope with large pose and expression differences, their results on facial videos suffer from asymmetrical makeup flaws and flickering. Second, the convolution-based approaches cannot preserve the consistency of makeup details. The reason is that the convolution operation tends to fuse pixels in the receptive field when learning and interpreting features, leading to considerable appearance differences between different views when transferring style features.

In this work, we explore the challenges of accurately and consistently transferring the makeup style of a single reference image to facial images in any pose and expression. To achieve this goal, we need a 3D-aware representation that can robustly

† Authors contributed equally
* Corresponding Author is Lin Gao (gaolin@ict.ac.cn).
Yu-Jie Yuan, Xinyang Han, Yue He, and Lin Gao are with the Beijing Key Laboratory of Mobile Computing and Pervasive Device, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, and also with the University of Chinese Academy of Sciences, Beijing, China. E-Mail: {yuanyujie, heyue19s, gaolin}@ict.ac.cn, hanxinyang20a@mails.ucas.ac.cn
Fang-Lue Zhang is with Victoria University of Wellington, New Zealand. E-mail: fanglue.zhang@ecs.vuw.ac.nz

disentangle the inherited facial geometry and appearance information. Recently, Neural Radiance Fields (NeRFs) have shown competence in representing 3D face information implicitly for view synthesis with geometry consistency [6], [7]. We thus build a novel framework based on dynamic NeRFs for transferring the makeup style from a reference image to the facial images of the same person. Our method has two-stages. In the first stage, a dynamic NeRF is trained to reconstruct a face model from the given video. The trained density module is then reused in the next stage as the representation of the inherent 3D geometry of the target face. In the second stage, we cascade the fixed density module and a makeup module to render facial images with makeups. It is the key for maintaining the consistency of the facial geometry in the final results. We also propose a novel hybrid makeup loss, which considers the characteristics of the makeup applied on different facial parts, to improve the makeup details in the generated images. To further reduce the effect of confounding factors caused by such as poses, we employ a patch-based discriminator working on UV maps to enhance the appearance consistency between the generated image with any pose and expression and the makeup reference.

The contributions of this paper are:

- We propose a novel NeRF-based makeup transfer method, *MuNeRF,* which is capable of automatically applying reference makeup on facial images under different poses and expressions consistently.
- We propose a novel hybrid makeup loss which considers the makeup characteristics of different facial parts. It works well for both light and extreme makeup styles.
- We introduce a patch-based discriminator working on UV maps which aligns the distribution between the transferred makeup and the reference makeup to improve the consistency across different poses and expressions.
- Extensive experiments and a user study are conducted to demonstrate the superiority of our method over other state-of-the-art methods in terms of visual authenticity and the consistency across the generated facial images.

II. Related Works

A. 2D Facial Makeup Transfer

Facial makeup is an interesting and popular topic in computer graphics and vision. Compared with style transfer, makeup transfer demands more accuracy of color distribution and more delicate details. Given a face image with a target makeup style, makeup transfer aims at perfectly imitating the style on this face image. Before the introduction of neural network, makeup transfer mainly relies on image warping and blending. Guo et al. [8] propose to decompose the face image and the reference makeup image into three layers: face structure layer, skin detail layer, and color layer. Then they mix the layers of the two images to achieve makeup transfer. Since CycleGAN [9] proposes a classical solution to unpaired image-to-image translation, it is widely used as a base structure in the makeup transfer task. For example, PairedCycleGAN [10] incorporates a makeup transfer network with a makeup removal network to achieve cycle-consistent training. To better transfer color distribution, BeautyGAN [3] introduces a local histogram matching, which achieves realistic frontal makeup results with light makeup style. LADN [11] is the first to handle dramatic makeup styles with multiple overlapping local discriminators. However, it may generate noticeable artifacts. Thao et al. [5] decompose extreme makeup and view it as a combination of colors and patterns. The method extends warped faces in UV space to transfer color and learns masks for patterns. PSGAN [12] focuses on dealing with new poses and expressions varying from source face with learned attentive matrices. SOGAN [13] introduced a flip attention module that utilizes facial symmetry to overcome issues with shadows and occlusions. Sun et al. [14] decompose the face image into four independent parts, which enables local control of the transferred makeup. [15] is the first to introduce Transformer [16] into this task to learn better shape transformation. The Glow model [17] is also used in makeup transfer [18]. To better align corresponding facial parts of the source and target faces, [19] proposes a novel symmetric semantic-aware network, working well on light makeup styles but failing on extreme ones. It also extends to video makeup transfer. However, it suffers from the inconsistency of makeup transfer details and lighting effects.

B. Neural Radiance Fields

Recently, neural rendering [20], especially Neural Radiance Field (NeRF) [21] has received substantial attention. NeRF implicitly encodes the geometry and appearance of a scene using a multi-layer perceptron (MLP) network. Many works have presented diverse methods to improve NeRF by providing better synthesis effects [22], [23] and faster training or inference speed [24]–[27], or adapt it to dynamic scenes [28], [29] and re-lighting tasks [30], performing geometry or appearance editing [31]–[33] etc. Incorporated with the prior models [34], [35], dynamic human face and body modeled by NeRF [6], [7], [36], [37] have been fully studied. However, the deformation field formulated by MLP-based networks cannot handle topological changes such as expressions. HyperNeRF [38] proposes to use an additional MLP to model different topological states of dynamic scenes as different hyper-planes in high-dimensional space. To cope with the diversity of facial image datasets, 2D image generation networks [2] are further extended to the generation of 3D-aware face image [39], [40]. The implicit 3D face can be edited by modifying the pose or expression parameters [41], the attribute values specified by the user [42], or semantic masks [43]–[45]. Some works [46], [47] also explore how to generate dynamic face videos from speech. HeadNeRF [48] proposes a parameterized human head model based on NeRF, which can generate realistic face images under different views with various parameters, including identity, expression, appearance and pose. NeP [49] decouples the geometry and appearance of the face and enables facial appearance editing by operating on the UV map. Although they can also apply makeup edits, other professional image editing software and some tedious labouring work are needed to generate a desirable makeup style. Our method is complementary to existing NeRF-based facial image generation

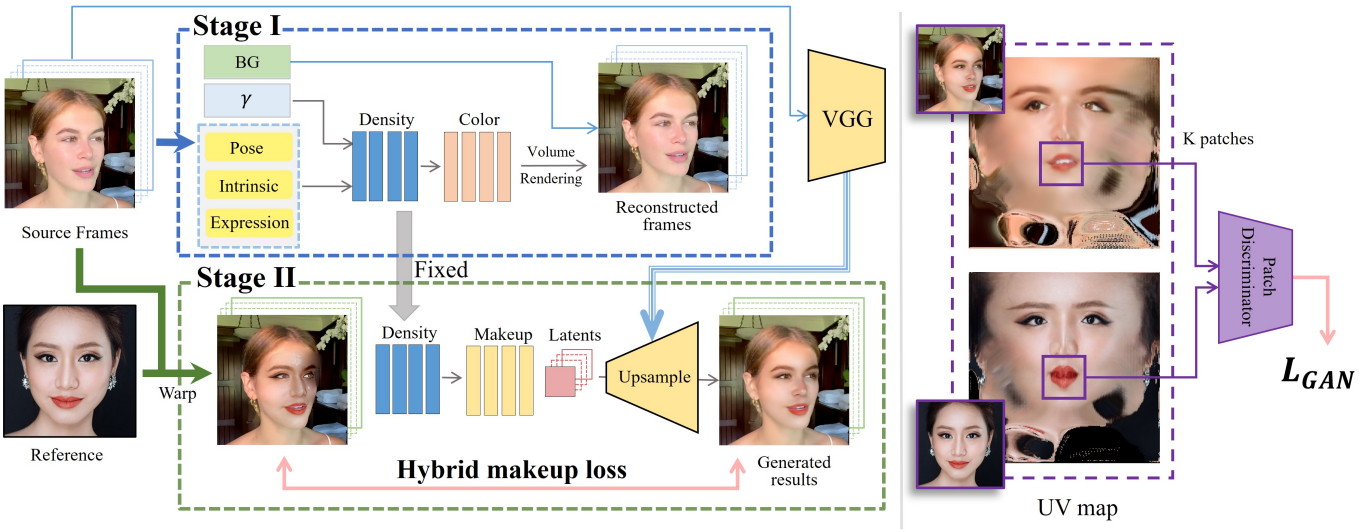


Fig. 2. Our framework. We employ a two-stage scheme. Firstly, we reconstruct the 3D NeRF representation of the input monocular video. Subsequently, we transfer the reference makeup style to different facial poses and expressions through a dedicated makeup module. A VGG network is introduced to extract global convolutional features which are fed into the upsampling module in the second stage to maintain 3D consistency of the makeup transfer results. The patch-based discriminator working on the facial UV texture maps is illustrated on the right. 'BG' denotes the background image and γ is a per-frame trainable vector used to eliminate estimation errors.

techniques, by providing an automated approach of applying makeup styles on the predicted appearances.

III. Method

Our goal is to generate facial makeup images consistently with given pose and expression parameters. Our network is built upon a recent dynamic facial NeRF model, NeRFace [7]. Given an input face video, we first train a dynamic NeRF model with facial poses and expressions to obtain the inherent face geometry represented in the density module (Sec. III-A). We then cascade the fixed density module and a MLP-based makeup module to learn how to apply the makeup features extracted from the reference image on the inherent face geometry (Sec. III-B). To imitate more detailed makeup features, we propose a patch-based discriminator on UV maps to better control the synthesized regions for important facial parts. What's more, a hybrid makeup loss is specially designed based on makeup characteristics. It works with other losses to supervise the training of the makeup module (Sec. III-C). Our framework preserves the consistency among the makeup transfer results of all the input frames and is able to generate new makeup images for unseen poses and expressions.

A. NeRF and Dynamic Face Reconstruction

A Neural Radiance Field (NeRF) represents a 3D scene as a continuous volumetric space and uses a multi-layer perceptron (MLP) to model both its geometry and appearance. This representation is encapsulated in a function F, which predicts color $\mathbf{c}$ and density σ based on the 3D position $\mathbf{p} = (x, y, z)$ and view direction $\mathbf{d} = (\phi, \theta)$ as inputs. To capture high-frequency scene details effectively, NeRF incorporates positional encoding $\zeta(\cdot)$, which transforms each input $(\mathbf{p}, \mathbf{d})$ into a higher-dimensional space. The function F is expressed as follows:

$$F_\Theta : (\zeta(\mathbf{p}), \zeta(\mathbf{d})) \rightarrow (\mathbf{c}, \sigma), \tag{1}$$

where Θ represents the parameters of the network. Since we need to model dynamic faces from the input monocular video and generate different poses and expressions, we adopt the dynamic NeRF face reconstruction work, NeRFace [7]. It extends NeRF to represent 3D faces with dynamic expressions with an extra input δ describing facial expressions and can be expressed as:

$$F_\Theta : (\zeta(\mathbf{p}), \zeta(\mathbf{d}), \delta, \gamma) \rightarrow (\mathbf{c}, \sigma), \tag{2}$$

where γ is a learnable latent vector to compensate for errors resulting from parameter estimation. Specifically, given the input face video Y comprising N frames y_i, we extract the camera intrinsic parameters, pose parameters P_i, and facial expression parameters δ_i for each frame using preprocessing tools provided by [50]. Utilizing the estimated pose parameters and camera intrinsics, the dynamic NeRF model maps the ray corresponding to each pixel into camera space and samples a specified number of points along the ray. Then the position $\mathbf{p}$ of a sampled point, the view direction $\mathbf{d}$ of the corresponding ray, and the estimated facial expression parameter δ are fed into two multi-layer perceptrons (MLPs): the density prediction module F_Θ^d and the color prediction module F_Θ^c. These modules predict the density value σ and the color value $\mathbf{c}$ of the sampled point, respectively. Additionally, a per-frame trainable code γ is introduced into the network to compensate for estimation errors. Finally, the density and color predictions are aggregated using volume rendering [51] to compute the pixel color corresponding to the ray $\mathbf{r}(t) = \mathbf{c} + t\mathbf{d}$:

$$\mathcal{C}(\mathbf{r}; \Theta, P, \delta, \gamma) : \int_{z_{near}}^{z_{far}} F_\Theta^d(\mathbf{r}(t)) \cdot F_\Theta^c(\mathbf{r}(t), \mathbf{d}) \cdot T(t) dt, \tag{3}$$

Here, z_{near} and z_{far} denote the closest and farthest depths, respectively. $T(t)$ represents the accumulated transmittance along the ray from z_{near} to t. Notably, our approach incorporates a two-stage volume rendering scheme inspired by [21], which involves the concurrent training of a coarse NeRF and a fine NeRF. Utilizing the dynamic NeRF model, we can derive the corresponding reconstructed frame y_i^Y.

Our approach is fundamentally grounded in the dynamic NeRF representation and leverages the 4D field to enable consistent makeup transfer across various facial poses and expressions. Our network architecture comprises two distinct components, each trained in separate stages. The complete network structure is depicted in Fig. 2. The first part is the dynamic 3D face reconstruction from a monocular face video. Then, we stabilize the density module in NeRF and train a dedicated makeup module to achieve consistent makeup effects.

B. Consistent Makeup Transfer

After the first stage of training is completed, we obtain an implicit reconstruction of face geometry. Then the goal of the second stage is to faithfully transfer the makeup distribution of the reference makeup image X to the dynamic face reconstructed in the first stage. The second part of the network reuses the density module F_Θ^d trained in the first stage and employs a new color module, F_Θ^m, which functions as the makeup module, to apply the reference makeup style. Unlike the color module in the first stage that predicts a 3-channel color for each pixel, the makeup module predicts a feature vector for each pixel and a low-resolution feature map by volume rendering instead, the same as [52], [53]. The feature map will be decoded into the final makeup image afterwards. This strategy reduces network training overhead, allowing us to generate the entire image during training which is convenient for the subsequent patch-based discrimination and other supervisions. We use an upsampling module for decoding the feature map, which consists of several downsampling layers, ResNet blocks and upsampling layers. Although the fixed density module can maintain the geometry consistency for the image generation with different poses and expressions, the convolutional operations after the makeup module may introduce appearance differences that break the 3D consistency. To alleviate this issue, we add the global convolutional features (VGG features) of the pre-makeup image when translating the feature map to the final makeup image. The global features are fused during deconvolution and upsampling with the corresponding dimensions, thus obtaining a higher resolution and a better preservation of geometry. During training, the pre-makeup image comes from the input video, and for the novel pose or expression to synthesize, it comes from the synthetic result of the dynamic NeRF model from the first stage. As shown later in our experiments, the consistency of the makeup transfer results is comparable to that of the dynamic NeRF trained in the first stage.

To supervise the training of our makeup module, we create a pseudo ground truth y_i^W for each source image y_i with the target makeup distribution. We adopt the creation method in LADN [11], where the pseudo ground truth y_i^W is generated by image warping based on landmark-based face matching and blending the face of the reference image X to the source image y_i by Poisson image editing [54] within OpenCV. We fix the density module trained in the first stage during the second training stage. This training scheme ensures the accuracy of reconstructed face geometry in different poses and expressions. It also speeds up the training process and reduces the GPU memory cost for training. The training process in the second stage is supervised by our novel hybrid makeup loss and a dense-landmark color loss that measure the differences between the transfer result and the pseudo ground truth, which are introduced in Sec. III-C. It should be noted that the pseudo ground truth is not perfect and the density module provides correspondence across different views to help eliminate the artifacts through the complementary information from different views. For better supervision of makeup consistency, we also build a patch-based discriminator working on facial UV texture maps to lift the makeup quality on important facial parts and ensure no confounding factors are included. The patch-based discriminator is shown on the right side of Fig. 2. After training the makeup module to convergence, the final makeup result y_i^X will be obtained.

Patch-based Discriminator on UV maps. The pseudo ground truth made by face morphing has some artifacts which may affect the makeup transfer results of the makeup module. To mitigate the effect, we further introduce a patch-based discriminator to correct the makeup distribution error of the pseudo ground truth. The original patch-based discriminator [9] directly judges whether sampled patches satisfy generative objectives. However, the underlying geometry information of the patches from the reference and the synthesized image could be considerably different, causing the ineffectiveness of the discriminator on identifying patches with properly transferred makeup features. Therefore, we choose to convert facial images to the UV-map domain to remove the confounding factors, such as poses and expressions, for assessing the synthesized makeup effects. We use PRNet [55] to map each face pixel to a fixed semantic facial point on the UV plane. We then get the UV textures using a texture mapping method [55] to represent the appearance information that is invariant to poses and expressions. The corresponding UV texture map of the synthesized image y_i^X and the reference makeup image X are denoted as ${y_i^X}_{uv}$ and X_{uv}, respectively. Moreover, instead of randomly sampling patches from UV maps, we select patches from fixed facial landmark positions on the UV map, such as lips, nose, eyes, and eyebrows. This encourages the discriminator to focus only on those facial parts related to makeup. Although the UV map also contains identity information that could be affected after modification, our fixed density module and the global feature of the original facial image can help to keep the geometry consistency in the generated images. Thus, the discriminator will mainly take effect on the makeup styles at patch-level. We also presented the ablation experiments of conducting discrimination on the facial image domain in Sec. IV-D. With the help of patch-based discriminator on UV maps, we can avoid obvious facial defects caused by mixing color distributions and provide more accurate control of the transferred makeup style.

C. Loss Functions

In this section, we introduce the losses to supervise our two-stage training. The first stage reconstructs the dynamic face

model before makeup from the source video. The dynamic NeRF F_Θ is trained by the color reconstruction loss:

$$L_{RGB} = \sum_{i=1}^{N} \sum_{j \in pixels} \|\mathcal{C}(\mathbf{r}_j; \Theta, P_i, \delta_i, \gamma_i) - I_i[j]\|^2, \quad (4)$$

where $I_i[j]$ is the ground truth color of the j-th pixel on the i-th input frame, Θ is the optimized weights.

In the second stage, we reuse the density module and keep it fixed. Then we replace the color module with a makeup module and train it from scratch. Due to the possible artifacts in the pseudo ground truth and the uneven distribution of the significant makeup-related features, the original color loss L_{RGB}, which treats all pixels equally, is not suitable for the second stage. Therefore, we propose a novel hybrid makeup loss to supervise the color prediction. The aforementioned patch-based discriminator is also introduced to align makeup style among generated frames. We also introduce a dense-landmark color loss to refine the details in makeup distribution.

Hybrid Makeup Loss. Our motivation for designing the hybrid makeup loss is our observation that certain aspects of the makeup effect occupy relatively small areas, such as the makeup around the eyes. These intricate details can be overlooked during optimization. Therefore, we design the hybrid makeup loss to specify varying weights to different makeup regions. This approach strengthens the constraints on detailed regions, ensuring the fidelity of the final makeup effect. The hybrid makeup loss consists of two parts. First, we crop K ($K = 7$) patches, containing key facial parts from both the pseudo ground truth y_i^W and the generated result y_i^X. We find the corresponding patch in the source image for a patch in the pseudo ground truth using histogram matching [3]. Then a L_2 loss is calculated between the matched patches and the generated result patches. Compared to the static reference image X, the pseudo ground truth has the same expression and pose as the original frame, making the makeup loss calculation easier and more accurate. To further enhance the fine-grained makeup details on the eye and lip region, we add an additional loss for some extreme eye shadows and lip makeups that directly calculates the L_1 loss on the eye and lip patch pairs. Second, we obtain the skin region using a facial mask [56] and crop M skin patches surrounding the key facial parts on y_i^W and y_i^X. Then the L_2 loss of each skin patch pair is added to the hybrid makeup loss with different weights. Specially, the patches around lips and eyes have twice the weights of other skin patches. In all, the hybrid makeup loss can be formulated as (for brevity, we omit the notation i):

$$L_{hybrid} = L_{key} * \lambda^{key} + L_{skin} * \lambda^{skin} \quad (5)$$

where,

$$L_{key} = \sum_{k=1}^{K} \|\mathcal{H}(p_k^W, p_k), p_k^X\|_2 + \|p_{eye,lip}^W, p_{eye,lip}^X\|_1$$

$$L_{skin} = \sum_{m=1}^{M} \|s_m^W, s_m^X\|_2 * \lambda_m^{part}$$

where p_k^W, p_k and p_k^X represent each key part patches of the pseudo ground truth y^W, non-makeup source image y, and our synthesized result y^X. $\mathcal{H}$ represents the histogram mapping. $p_{eye,lip}^W$ and $p_{eye,lip}^X$ represent the patches on the eye and lip regions of y^W and y^X. s_m^W and s_m^X denote each skin part of y^W and y^X, respectively. λ^{key} is the weight for key feature parts loss, λ^{skin} is for skin parts loss and λ_m^{part} represents the individual weight for each skin part according to the makeup characteristics, where the weights for the skin patches surrounding the key facial parts should be higher to obtain better effects.

PatchGAN Loss. We use a patch-based discriminator on UV maps, which involves a GAN loss L_{GAN} formulated as:

$$\begin{aligned} L_{GAN} = \max_G \min_D (&E_{p \sim P_{uv}^y}[\log(D(p))] \\ &+ E_{\hat{p} \sim P_{uv}^X}[\log(1 - D(\hat{p}))]), \end{aligned} \quad (6)$$

where G represents the generator, that is, our makeup NeRF that generates the makeup images, including the density module and the makeup module. D represents our patch-based discriminator on UV maps. P_{uv}^y is the set of image patches in ${y_i^X}_{uv}$ and P_{uv}^X is the set of image patches in X_{uv}.

Dense-landmark Color Loss. We use [55] to reconstruct an explicit 3D mesh with vertex color from the face image, where the mesh vertices are viewed as dense landmarks. Unlike [4] that only considers the nose region, we consider the dense landmarks from more appointed facial regions such as lip, eye, nose and cheek. Then the L_2 loss between the colors of dense landmarks obtained from y_i^W and y_i^X is calculated point-by-point:

$$L_{dense} = \sum_{d=1}^{D} \|\mathcal{D}_d(y_i^X) - \mathcal{D}_d(y_i^W))\|^2. \quad (7)$$

Here, $\mathcal{D}_d$ denotes the function to obtain the d-th dense face landmark color from the image and D represents the number of dense landmarks we select.

Finally, the total loss function of the second stage is:

$$L_{total} = \lambda_{hybrid} L_{hybrid} + \lambda_{GAN} L_{GAN} + \lambda_{dense} L_{dense}, \quad (8)$$

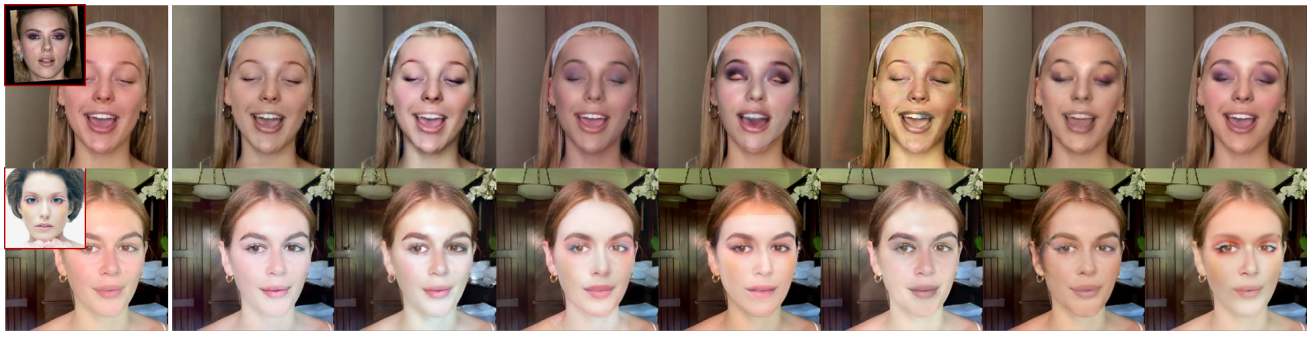
where λ_{hybrid}, λ_{GAN} and λ_{dense} are the adjusting weights.

Please refer to our supplementary file for more training and implementation details of our method.

IV. Results and Experiments

A. Comparisons with Other Methods

Some existing NeRF-based methods [57], [58] are able to apply style transfer on facial images. However, they pay more attention to the transfer of overall style and fail to generate accurate details, which can not satisfy the high demand for details in makeup transfer task. NeP [49] provides an interactive method to apply makeup on face images. But it needs tedious manual editing work on UV maps, which requires a high-level professional skills to achieve satisfactory results. Therefore, we compare our method with the following state-of-the-art automated 2D makeup transfer methods, including BeautyGAN [3], PSGAN [12], SCGAN [59], SSAT [19], CPM [5] and LADN [11]. Based on the open-source models provided by these methods, we fine-tune the models on our



(a) Inputs (b) BeautyGAN (c) SCGAN (d) SSAT (e) CPM (f) PSGAN (g) LADN (h) Ours

Fig. 3. We compare the makeup transfer effect on a single frame with six existing methods. It can be seen that our results have more reasonable color distributions while maintaining the geometric structure of the source face.

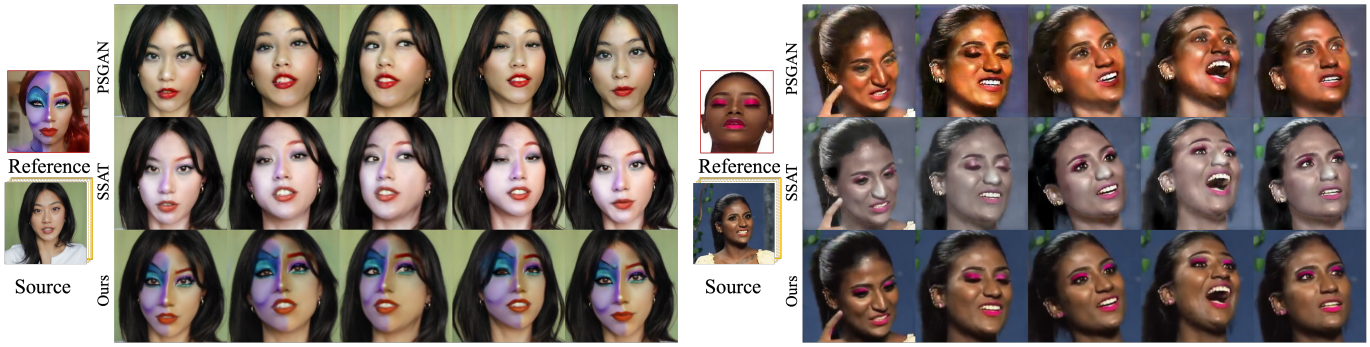


Fig. 4. Comparisons on video makeup transfer with PSGAN [12] and SSAT [19]

training set for fairness. We also show more results in the supplementary file and dynamic results in the video.

We first compare our method with other makeup transfer methods in terms of makeup transfer quality on a single frame. We show the comparison results in Fig. 3. We can see that our results have more accurate and reasonable color distributions on facial features while maintaining the geometric structure of the source face. In contrast, PSGAN, SCGAN, and SSAT fail to transfer accurate makeup details in the eye region, as shown in the first row. The results of CPM have visible pasting artifacts when the skin colors are different between the source and reference faces.

We then show our great advantages of keeping consistency when transfer makeup to facial videos. It should be noted that most of the existing methods only consider makeup transfer on a single input image. Even though a few researches [12], [19] claim that they could perform continuous makeup transfer for facial videos, we found the consistency achieved by those methods is poor, especially under some exaggerated poses and expressions. We compare our method with PSGAN [12] and SSAT [19] on the video makeup transfer task, since they claim that their method can handle such cases. The results are shown in Fig. 4. Notably, the makeup effects of our results are more visually pleasant and consistent, while we can see obvious visual defects in the results of PSGAN and SSAT. This demonstrates that the consistency of our makeup transfer under different poses and expressions is superior to the existing methods. We also show the comparisons with pseudo ground truth in the supplementary file.

B. More Makeup Transfer Results

Partial Makeup Transfer. Our method can be applied to perform partial makeup transfer and multi-reference makeup transfer since our hybrid makeup loss and discriminator are both patch-based. We first generate pseudo ground truth for each makeup reference and use the facial mask to select the desired makeup region from these pseudo ground truths. We then fuse these desired regions to form the entire target pseudo ground truth image to train our model. Fig. 5 provides examples where we transfer makeup for single facial parts and their combinations.

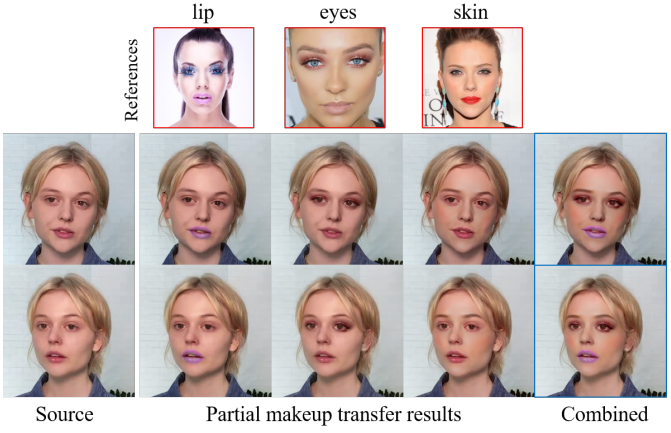


Fig. 5. Partial makeup transfer results. We show the transferred makeup on the lips, eyes, and skin, and the combination of the three.

De-makeup. Our method can also utilize makeup transfer to achieve de-makeup effects. For a face video with an arbitrary makeup, we select an arbitrary face image without makeup and

	Video 1	Video 2	Video 3	Video 4	Video 5	Average
w/o conv.	0.862	0.858	0.852	0.876	0.873	0.8642
Ours	0.821	0.854	0.848	0.867	0.826	0.8432

TABLE I
THE SIMILARITIES BETWEEN THE EXPRESSION VECTORS OF THE RECONSTRUCTED AND THE ORIGINAL FRAMES FOR CONSISTENCY EVALUATION.

use our method to transfer this unpasteurized effect to the 3D face with makeup. The results are shown in Fig. 6. It can be seen that our method can also achieve a consistent de-makeup effect for makeup faces.

Pose and Expression Interpolation. Since our network takes poses and expressions as explicit controls of dynamic neural radiance fields, interpolation between different poses and expressions can be achieved. Our method is good at transferring makeup style to unseen poses and expressions while keeping makeup distribution consistent across all the frames. Some results are shown in Fig. 7.

More Cases. We show more makeup transfer examples in Fig. 8 and the supplementary.

C. Evaluations

Consistency. Although our makeup module has convolution operations, our method still guarantees strong 3D consistency by incorporating the features of pre-makeup images during upsampling and employing a PatchGAN working on the UV map for supervision. Because we use NeRFace as the 3D representation, different views are actually different facial poses. In order to quantitatively evaluate the consistency under different poses, inspired by EG3D [40], we compare the 3DMM [34] expression vectors extracted from the original video frames, the corresponding reconstructed images using our model without convolution operations (w/o conv.) and our full model. We then calculate the average cosine similarity between the expression vectors of the 'w/o conv.' model or the full model and the original frames. It can be seen from Tab. I that these two similarities are comparable, indicating the convolution operations has a nearly neglectable impact on the 3D consistency.

Generalization. In order to illustrate the generalizability of our method, we manually split the images into two groups according to the head position and pose. We use one group as our training set and the other as the test set to see whether our model can cope with large pose differences. In Fig. 9, we show the makeup transfer results of some representative test frames in the bottom two rows, and the corresponding most similar training images in the first row. It can be seen that our method exhibits good generalizability on unseen images with obvious differences from the training set.

Time Statistics. We present the time statistics for each stage in Table II. Due to the limitations of the NeRF representation, our method requires training a NeRF network for each individual and also requires some time to train makeup modules for different reference makeups. While other methods can be directly applied to novel persons without the need for retraining, it's essential to emphasize that there already exist acceleration technologies capable of expediting NeRF training and rendering, such as iNGP [27] and NeRFAcc [60]. On a practical note, each user can initially establish their facial NeRF representation, which is a one-time process. Subsequent makeup transfers from different makeup styles won't consume excessive time. We've also compared the inference time required to generate a makeup transfer image with other methods, revealing that our method falls within a moderate range in terms of inference time. Incorporating NeRF acceleration technology is expected to further reduce this time, potentially enabling real-time performance.

D. Ablation Study

We demonstrate the effect of the components of our method on the performance of the makeup transfer with several ablation experiments.

Effectiveness of Losses. To show the effectiveness of the losses used in the second training stage, we test several variants of the training scheme. In Fig. 10(a), we show the result of the model trained with only the RGB color loss between the pseudo ground truth and the generated image. Because the RGB loss treats all pixels equally, it fails to generate some important details, such as the lip region in the first row, and the eyebrow region in the bottom example. The results in (b) show that our proposed hybrid makeup loss can provide better-aligned makeup features for key facial parts. However, it uses histogram matching to generate target color distribution for lips and nose, where large deformation may happen, so it may still miss some detailed makeup features. By inspecting (b) and (c), we can find that the patch discriminator and the PatchGAN loss enable a direct comparison between the generated results and the reference makeup image regardless of their geometry difference, which can further help capture the makeup details. Compared with the results in (c), the results of our full model (d) show that the dense landmark loss can further improve the details by providing more accurate colors (the top row) and sharpening the makeup (the bottom row).

Patch-based Discriminator. The effectiveness of our patch-based discriminator working on the UV map has been shown in the aforementioned experiments. In addition to working on the UV map, we also test other alternatives of applying discrimination. We train two other discriminators for directly distinguishing the patch pairs from the final output facial image and the corresponding pesudo ground truth or the original reference makeup image. Fig. 11 shows the comparison results. It can be seen that applying the adversarial loss on the pose-invariant UV texture maps can provide better control of the generated makeup texture, saving the training process from huge and non-convergence network parameters of continuously changing input variables. Using the patches on the pseudo ground truth image and the original reference image cannot guide the network effectively due to the significant different poses and expressions.

E. Comparison with Warping

Our method uses a warped makeup face as the pseudo ground truth for color supervision, offering color distribution information to guide makeup transfer. This supervision manner

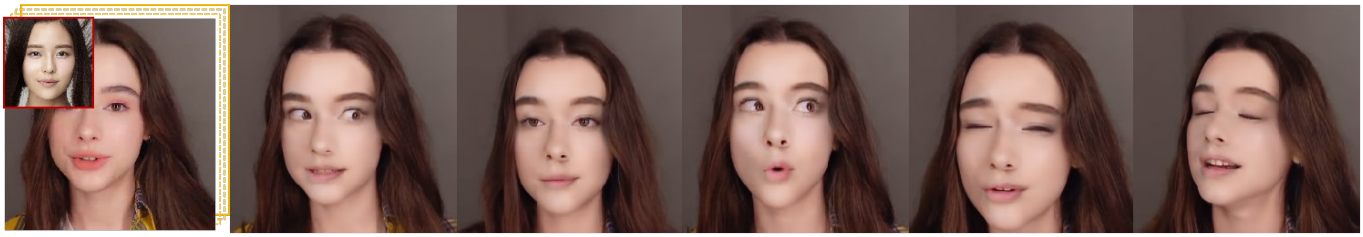


Fig. 6. De-makeup results. Our method can also achieve a consistent de-make effect for makeup faces.

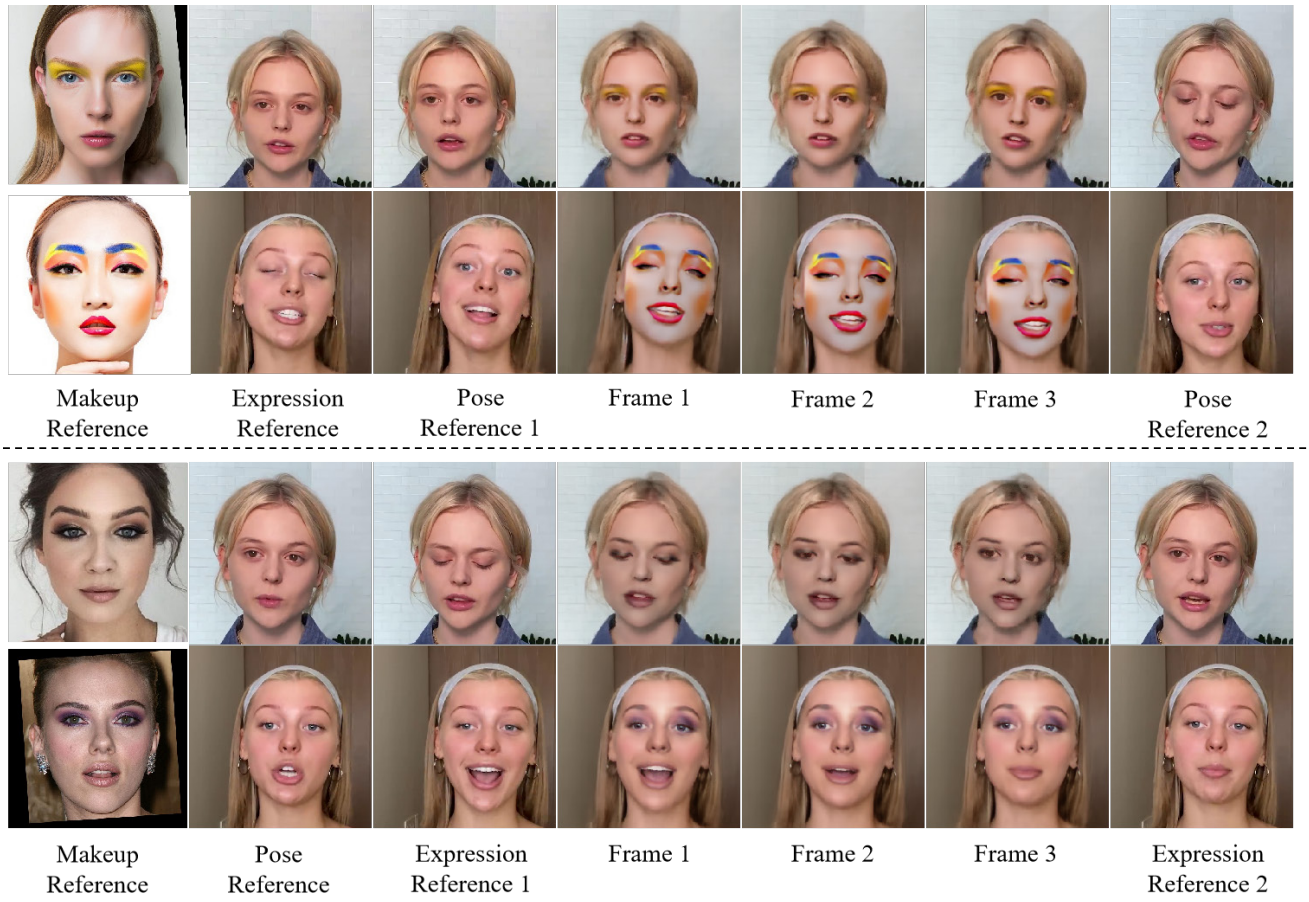


Fig. 7. Makeup transfer results during pose and expression interpolation.

	BeautyGAN	PSGAN	SSAT	SCGAN	CPM	Ours - 1st stage	Ours - 2nd stage	Ours - inference
Time	5.25s	1.96s	0.30s	0.57s	9.63s	∼4h	∼1h	1.75s

TABLE II

WE SHOW THE TRAINING TIME OF EACH STAGE AND ALSO COMPARE THE INFERENCE TIME WITH OTHER METHODS. IT CAN BE SEEN THAT OUR METHOD IS AT A MODERATE LEVEL IN INFERENCE TIME.

is also used in some other face makeup transfer works [11], [19]. However, the pseudo ground truth images may exhibit geometric distortions and visual defects, as depicted in Fig. 12 (b). By utilizing NeRF as the 3D implicit representation, our method can effectively mitigate these artifacts by leveraging complementary information obtained through correspondences across different views. Additionally, our PatchGAN, which operates on the UV map, further enhances appearance consistency. As illustrated in Fig. 12 (c), the results of our method successfully rectify the visual artifacts present in the pseudo ground truths.

F. User Study

For the subjective evaluation of the results, we conduct a user study. We prepare makeup transfer results of 20 images and 10 videos and compare our results with five state-of-art methods [3], [5], [12], [19], [59]. 30 participants are asked to rank the results from best to worst in terms of realism, faithfulness, and overall quality. The user study result is reported in Tab. III, where our method has shown apparent advantages, especially for the results of videos. It should be noted that the realism of the video results also evaluates the consistency, as inconsistency will affect the realism. Please refer to our

Fig. 8. More makeup transfer examples.

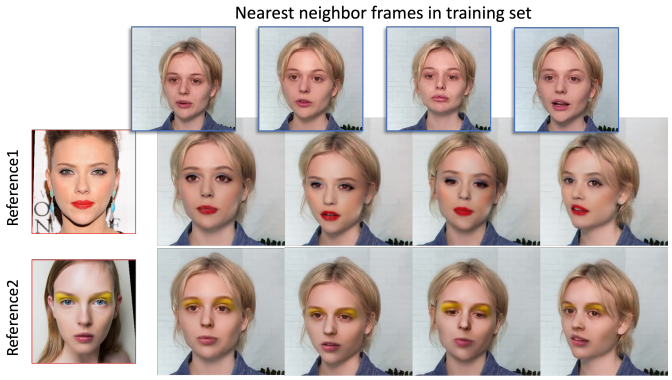


Fig. 9. Generalizability for various poses and expressions. The first row shows the training images with the nearest poses to the makeup transfer results in the bottom two rows.

supplementary file for more results and information of our user study.

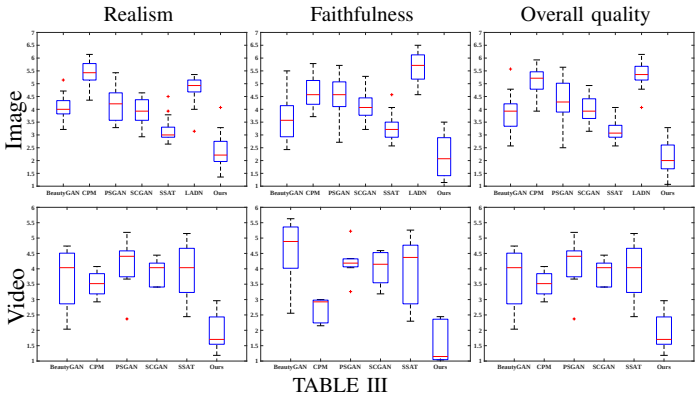


TABLE III

BOX PLOTS OF THE AVERAGE REALISM, FAITHFULNESS AND OVERALL QUALITY PERCEPTION SCORES FOR EACH OF THE COMPARED APPROACHES.

G. Limitations

We show some failure cases in Fig. 13. First, our method may fail to transfer some high-frequency makeup details. As shown in the first row, the white spots on both sides of the face in the reference image are neglected. The reason lies in the trade-off made by the NeRF while ensuring consistency across all the training frames. This can be mitigated by blending the

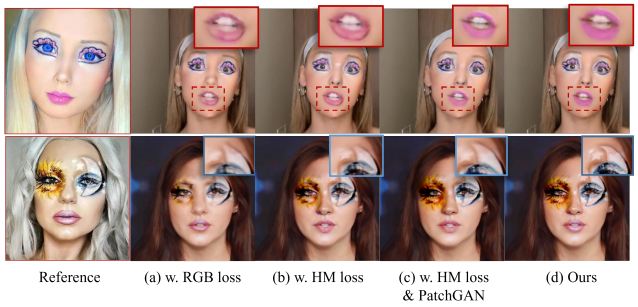


Fig. 10. Ablation study on the losses. (a-c) show the results of the model trained with the RGB loss on the pseudo ground truth (RGB loss), hybrid makeup loss (HM loss), and the combination of PatchGAN loss and HM loss, respectively. (d) uses all the proposed losses.

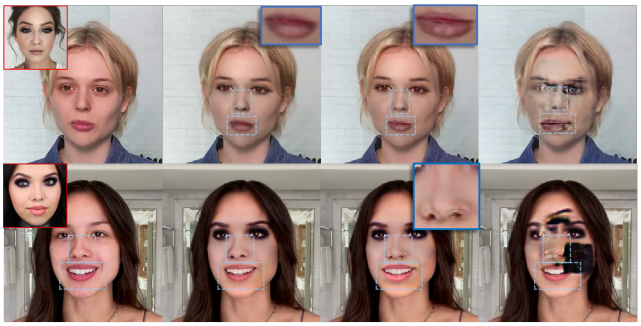
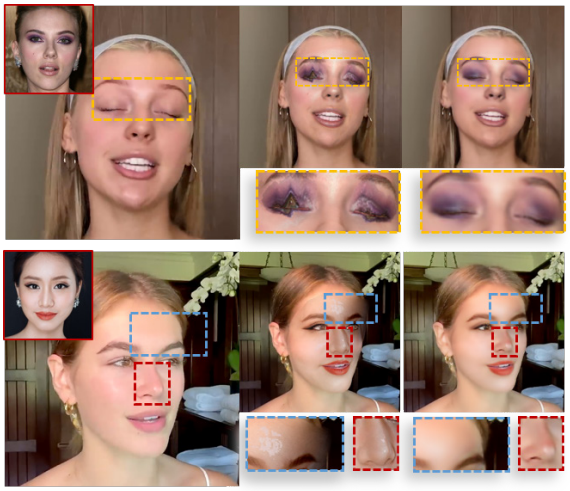


Fig. 11. Ablation study on our patch-based discriminator. We perform discrimination on patches from the pseudo ground truth (on Pseudo GT) and the original reference image (on Reference). The UV-map-based discriminator (on UV map) provides the best performance.

warped results and our transfer results with delicate pattern masks. Second, as in the bottom row, our method may generate false highlights when the lighting conditions are significantly different, e.g. darker environment. Our method assumes that the different shades on a makeup face come from the use of cosmetics of different color and brightness, so that the shades are transferred as part of makeup effects. It can be resolved by incorporating user interaction to specify where to apply makeup. Last, due to the inability of our NeRF framework to render in real time, our method is unable to achieve real-time makeup transfer. Also, the training takes several hours. However, these can be solved by incorporating current NeRF acceleration technologies, such as iNGP [27].

V. CONCLUSIONS

We propose a novel framework based on dynamic neural radiance fields for consistently transferring makeup styles to facial images with any facial pose and expression. Our method maintains the geometry and appearance consistency among all the synthesized facial images with dramatically different poses and expressions. In our two-stage training scheme, we first obtain the implicit 3D geometry representation and then apply reference makeup styles through a makeup module. Specifically, to ensure the quality and consistency of the synthesized results, a novel hybrid makeup loss that considers the characteristics of the makeup applied on different facial parts, and a UV-map-based patch discriminator that works in the pose-independent space are proposed. Extensive experiments



(a) Reference (b) Pseudo GT (c) Ours

Fig. 12. Comparison with pseudo ground truth. Although there are artifacts on the pseudo ground truth, our method can eliminate them with the help of the complementary information learned from different views and the patch discriminator on the UV map.

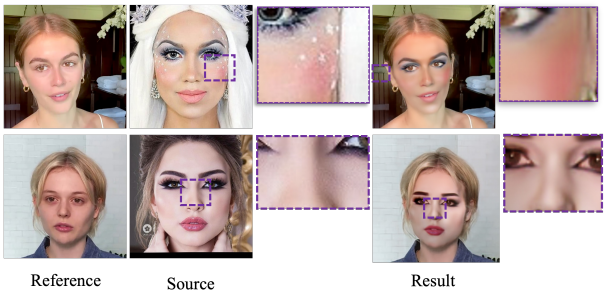


Fig. 13. Failure cases.

and a user study demonstrate the superiority of our method, which achieves the best performance of visual quality and consistency in transferring makeup to multiple facial images with different poses and expressions.

Acknowledgments

This work was supported by National Natural Science Foundation of China (No. 62322210), Beijing Municipal Natural Science Foundation for Distinguished Young Scholars (No. JQ21013) and Beijing Municipal Science and Technology Commission (No. Z231100005923031).

References

[1] H. Kim, P. Garrido, A. Tewari, W. Xu, J. Thies, M. Niessner, P. Pérez, C. Richardt, M. Zollhöfer, and C. Theobalt, "Deep video portraits," *ACM Transactions on Graphics (TOG)*, vol. 37, no. 4, pp. 1–14, 2018.

[2] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of StyleGAN," in *Proc. CVPR*, 2020.

[3] T. Li, R. Qian, C. Dong, S. Liu, Q. Yan, W. Zhu, and L. Lin, "Beautygan: Instance-level facial makeup transfer with deep generative adversarial network," in *Proceedings of the 26th ACM international conference on Multimedia*, 2018, pp. 645–653.

[4] S. Liu, W. Jiang, C. Gao, R. He, J. Feng, B. Li, and S. Yan, "Psgan++: Robust detail-preserving makeup transfer and removal," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[5] T. Nguyen, A. T. Tran, and M. Hoai, "Lipstick ain't enough: Beyond color matching for in-the-wild makeup transfer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 13 305–13 314.

[6] K. Park, U. Sinha, J. T. Barron, S. Bouaziz, D. B. Goldman, S. M. Seitz, and R. Martin-Brualla, "Nerfies: Deformable neural radiance fields," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 5865–5874.

[7] G. Gafni, J. Thies, M. Zollhofer, and M. Nießner, "Dynamic neural radiance fields for monocular 4d facial avatar reconstruction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 8649–8658.

[8] D. Guo and T. Sim, "Digital face makeup by example," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2009, pp. 73–79.

[9] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232.

[10] H. Chang, J. Lu, F. Yu, and A. Finkelstein, "Pairedcyclegan: Asymmetric style transfer for applying and removing makeup," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 40–48.

[11] Q. Gu, G. Wang, M. T. Chiu, Y.-W. Tai, and C.-K. Tang, "Ladn: Local adversarial disentangling network for facial makeup and de-makeup," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 10 481–10 490.

[12] W. Jiang, S. Liu, C. Gao, J. Cao, R. He, J. Feng, and S. Yan, "Psgan: Pose and expression robust spatial-aware gan for customizable makeup transfer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5194–5202.

[13] Y. Lyu, J. Dong, B. Peng, W. Wang, and T. Tan, "Sogan: 3d-aware shadow and occlusion robust gan for makeup transfer," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 3601–3609.

[14] Z. Sun, F. Liu, W. Liu, S. Xiong, and W. Liu, "Local facial makeup transfer via disentangled representation," in *Proceedings of the Asian Conference on Computer Vision*, 2020.

[15] Z. Wan, H. Chen, J. An, W. Jiang, C. Yao, and J. Luo, "Facial attribute transformers for precise and robust makeup transfer," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 1717–1726.

[16] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[17] D. P. Kingma and P. Dhariwal, "Glow: Generative flow with invertible 1x1 convolutions," *Advances in neural information processing systems*, vol. 31, 2018.

[18] H.-J. Chen, K.-M. Hui, S.-Y. Wang, L.-W. Tsao, H.-H. Shuai, and W.-H. Cheng, "Beautyglow: On-demand makeup transfer framework with reversible generative network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10 042–10 050.

[19] Z. Sun, Y. Chen, and S. Xiong, "Ssat: A symmetric semantic-aware transformer network for makeup transfer and removal," *arXiv preprint arXiv:2112.03631*, 2021.

[20] A. Tewari, O. Fried, J. Thies, V. Sitzmann, S. Lombardi, K. Sunkavalli, R. Martin-Brualla, T. Simon, J. M. Saragih, M. Nießner, R. Pandey, S. Fanello, G. Wetzstein, J.-Y. Zhu, C. Theobalt, M. Agrawala, E. Shechtman, D. B. Goldman, and M. Zollhofer, "State of the art on neural rendering," *Computer Graphics Forum*, vol. 39, 2020.

[21] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," in *European conference on computer vision*. Springer, 2020, pp. 405–421.

[22] L. Liu, J. Gu, K. Zaw Lin, T.-S. Chua, and C. Theobalt, "Neural sparse voxel fields," *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[23] S. Wizadwongsa, P. Phongthawee, J. Yenphraphai, and S. Suwajanakorn, "Nex: Real-time view synthesis with neural basis expansion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 8534–8543.

[24] A. Yu, R. Li, M. Tancik, H. Li, R. Ng, and A. Kanazawa, "PlenOctrees for real-time rendering of neural radiance fields," in *ICCV*, 2021.

[25] S. J. Garbin, M. Kowalski, M. Johnson, J. Shotton, and J. Valentin, "Fastnerf: High-fidelity neural rendering at 200fps," *arXiv preprint arXiv:2103.10380*, 2021.

[26] C. Reiser, S. Peng, Y. Liao, and A. Geiger, "Kilonerf: Speeding up neural radiance fields with thousands of tiny mlps," 2021.

[27] T. Müller, A. Evans, C. Schied, and A. Keller, "Instant neural graphics primitives with a multiresolution hash encoding," *ACM Trans. Graph.*, vol. 41, no. 4, pp. 102:1–102:15, Jul. 2022. [Online]. Available: https://doi.org/10.1145/3528223.3530127

[28] A. Pumarola, E. Corona, G. Pons-Moll, and F. Moreno-Noguer, "D-nerf: Neural radiance fields for dynamic scenes," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 10 318–10 327.

[29] E. Tretschk, A. Tewari, V. Golyanik, M. Zollhofer, C. Lassner, and C. Theobalt, "Non-rigid neural radiance fields: Reconstruction and novel view synthesis of a dynamic scene from monocular video," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 12 959–12 970.

[30] X. Zhang, P. P. Srinivasan, B. Deng, P. Debevec, W. T. Freeman, and J. T. Barron, "Nerfactor: Neural factorization of shape and reflectance under an unknown illumination," *arXiv preprint arXiv:2106.01970*, 2021.

[31] Y.-J. Yuan, Y.-T. Sun, Y.-K. Lai, Y. Ma, R. Jia, and L. Gao, "Nerf-editing: geometry editing of neural radiance fields," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18 353–18 364.

[32] Y.-H. Huang, Y. He, Y.-J. Yuan, Y.-K. Lai, and L. Gao, "Stylizednerf: consistent 3d scene stylization as stylized nerf via 2d-3d mutual learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18 342–18 352.

[33] Chong Bao and Bangbang Yang, Z. Junyi, B. Hujun, Z. Yinda, C. Zhaopeng, and Z. Guofeng, "Neumesh: Learning disentangled neural mesh-based implicit field for geometry and texture editing," in *European Conference on Computer Vision (ECCV)*, 2022.

[34] V. Blanz and T. Vetter, "A morphable model for the synthesis of 3d faces," in *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, 1999, pp. 187–194.

[35] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "SMPL: A skinned multi-person linear model," *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, vol. 34, no. 6, pp. 248:1–248:16, Oct. 2015.

[36] S. Peng, Y. Zhang, Y. Xu, Q. Wang, Q. Shuai, H. Bao, and X. Zhou, "Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans," in *CVPR*, 2021.

[37] L. Liu, M. Habermann, V. Rudnev, K. Sarkar, J. Gu, and C. Theobalt, "Neural actor: Neural free-view synthesis of human actors with pose control," *ACM Trans. Graph.(ACM SIGGRAPH Asia)*, 2021.

[38] K. Park, U. Sinha, P. Hedman, J. T. Barron, S. Bouaziz, D. B. Goldman, R. Martin-Brualla, and S. M. Seitz, "Hypernerf: A higher-dimensional representation for topologically varying neural radiance fields," *arXiv preprint arXiv:2106.13228*, 2021.

[39] E. R. Chan, M. Monteiro, P. Kellnhofer, J. Wu, and G. Wetzstein, "pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 5799–5809.

[40] E. R. Chan, C. Z. Lin, M. A. Chan, K. Nagano, B. Pan, S. De Mello, O. Gallo, L. J. Guibas, J. Tremblay, S. Khamis *et al.*, "Efficient geometry-aware 3d generative adversarial networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 16 123–16 133.

[41] S. Athar, Z. Xu, K. Sunkavalli, E. Shechtman, and Z. Shu, "Rignerf: Fully controllable neural 3d portraits," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*. IEEE, 2022, pp. 20 332–20 341.

[42] K. Kania, K. M. Yi, M. Kowalski, T. Trzciński, and A. Tagliasacchi, "Conerf: Controllable neural radiance fields," *arXiv preprint arXiv:2112.01983*, 2021.

[43] K. Jiang, S.-Y. Chen, F.-L. Liu, H. Fu, and L. Gao, "Nerffaceediting: Disentangled face editing in neural radiance fields," in *ACM SIGGRAPH Asia 2022 Conference Proceedings*, ser. SIGGRAPH Asia'22. New York, NY, USA: Association for Computing Machinery, 2022.

[44] J. Sun, X. Wang, Y. Shi, L. Wang, J. Wang, and Y. Liu, "Ide-3d: Interactive disentangled editing for high-resolution 3d-aware portrait synthesis," *arXiv preprint arXiv:2205.15517*, 2022.

[45] J. Sun, X. Wang, Y. Zhang, X. Li, Q. Zhang, Y. Liu, and J. Wang, "Fenerf: Face editing in neural radiance fields," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 7672–7682.

[46] Y. Guo, K. Chen, S. Liang, Y. Liu, H. Bao, and J. Zhang, "Ad-nerf: Audio driven neural radiance fields for talking head synthesis," in *IEEE/CVF International Conference on Computer Vision, ICCV*. IEEE, 2021, pp. 5764–5774.

[47] S. Yao, R. Zhong, Y. Yan, G. Zhai, and X. Yang, "Dfa-nerf: Personalized talking head generation via disentangled face attributes neural rendering," *CoRR*, vol. abs/2201.00791, 2022.

[48] Y. Hong, B. Peng, H. Xiao, L. Liu, and J. Zhang, "Headnerf: A real-time nerf-based parametric head model," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 20 374–20 384.

[49] L. Ma, X. Li, J. Liao, X. Wang, Q. Zhang, J. Wang, and P. Sander, "Neural parameterization for dynamic human head editing," *arXiv preprint arXiv:2207.00210*, 2022.

[50] Y. Guo, K. Chen, S. Liang, Y. Liu, H. Bao, and J. Zhang, "Ad-nerf: Audio driven neural radiance fields for talking head synthesis," *arXiv preprint arXiv:2103.11078*, 2021.

[51] J. T. Kajiya and B. P. Von Herzen, "Ray tracing volume densities," *ACM SIGGRAPH computer graphics*, vol. 18, no. 3, pp. 165–174, 1984.

[52] M. Niemeyer and A. Geiger, "Giraffe: Representing scenes as compositional generative neural feature fields," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 11 453–11 464.

[53] J. Gu, L. Liu, P. Wang, and C. Theobalt, "Stylenerf: A style-based 3d-aware generator for high-resolution image synthesis," *arXiv preprint arXiv:2110.08985*, 2021.

[54] P. Pérez, M. Gangnet, and A. Blake, "Poisson image editing," in *ACM SIGGRAPH 2003 Papers*, 2003, pp. 313–318.

[55] Y. Feng, F. Wu, X. Shao, Y. Wang, and X. Zhou, "Joint 3d face reconstruction and dense alignment with position map regression network," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 534–551.

[56] C. Yu, C. Gao, J. Wang, G. Yu, C. Shen, and N. Sang, "Bisenet v2: Bilateral network with guided aggregation for real-time semantic segmentation," *International Journal of Computer Vision*, vol. 129, no. 11, pp. 3051–3068, 2021.

[57] J. Sun, X. Wang, Y. Zhang, X. Li, Q. Zhang, Y. Liu, and J. Wang, "Fenerf: Face editing in neural radiance fields," *arXiv preprint arXiv:2111.15490*, 2021.

[58] P. Zhou, L. Xie, B. Ni, and Q. Tian, "Cips-3d: A 3d-aware generator of gans based on conditionally-independent pixel synthesis," *arXiv preprint arXiv:2110.09788*, 2021.

[59] H. Deng, C. Han, H. Cai, G. Han, and S. He, "Spatially-invariant style-codes controlled makeup transfer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 6549–6557.

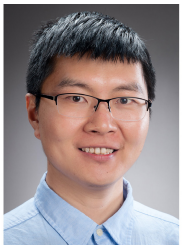
[60] R. Li, H. Gao, M. Tancik, and A. Kanazawa, "Nerfacc: Efficient sampling accelerates nerfs." *arXiv preprint arXiv:2305.04966*, 2023.

VI. Biography Section

Yu-Jie Yuan received the bachelor's degree in mathematics from Xi'an Jiaotong University in 2018. He is currently a Ph.D. candidate in the Institute of Computing Technology, Chinese Academy of Sciences. His research interests include computer graphics and neural rendering.

Xin-Yang Han is an undergraduate at the School of Computer Science and Technology, University of Chinese Academy of Sciences. His research interests include computer vision and graphics.

Yue He received the Bachelor's and Master's degrees from University of Chinese Academy of Sciences. She is currently working at the Ant Group. Her research interests include computer vision and graphics.



Fang-Lue Zhang is currently a senior lecturer with Victoria University of Wellington, New Zealand. He received the Doctoral degree from Tsinghua University in 2015. His research interests include image and video editing, computer vision, and computer graphics. He received Victoria Early-Career Research Excellence Award in 2019 and Fast-Start Marsden Grant from New Zealand Royal Society in 2020. He is on the editorial board of Computer & Graphics. He is a committee member of IEEE Central New Zealand Section.

Lin Gao received the bachelor's degree in mathematics from Sichuan University and the PhD degree in computer science from Tsinghua University. He is currently a Professor at the University of Chinese Academy of Sciences. He has been awarded the Newton Advanced Fellowship from the Royal Society and the Asia Graphics Association Young Researcher Award. His research interests include computer graphics and geometric processing.