LipAT: Beyond Style Transfer for Controllable Neural Simulation of Lipstick using Cosmetic Attributes

Amila Silva[1,2], Olga Moskvyak[2], Alexander Long[2], Ravi Garg[2], Stephen Gould[2,3],
Gil Avraham[2], Anton van den Hengel[2,4]
[1] The University of Melbourne, [2] Amazon,
[3] Australian National University, [4] The University of Adelaide

Abstract

This is the supplementary material for the paper titled "LipAT: Beyond Style Transfer for Controllable Neural Simulation of Lipstick using Cosmetic Attributes". The outline of this document is as follows. Section 1 motivates the solutions that allows lipstick virtual try-on using lipstick attributes directly. Sections 2 and 3 provide more details about the neural architectures in LipAT-LAM and LipAT-LRM. Then, we discuss the datasets used in this work in Section 4. Section 5 presents formal derivation and empirical justifications for the novel metric – i.e., Patch-FID, proposed in this work. We discuss more details about the selected baselines and user study in Sections 6 and 7. We provide more quantitative results in Sections 8 and 9 to show the superiority of LipAT, and discuss its limitations in Section 10. The final section in this manuscript recaps the preliminary on CIELAB colour space.

1. Beyond Lipstick Style Transfer

Neural Style Transfer (NST) [7, 12] aims to blend two images – a content image and a style reference image, such that the output image looks like the content image, but with the style of the reference image. Motivated by this literature, most recent neural lipstick application approaches [2, 6, 16, 22, 27] formulates the problem of lipstick application as a style transfer problem and exploits the strengths of the NST models to transfer the appearance of lipstick on a reference face image to a target face image. It has been found that NST-based approaches could produce realistic renderings without explicitly modelling facial and scene-specific parameters as in Physics-Based Rendering (PBR) techniques [9, 11, 21]. Nevertheless, there are several limitations in NST-based approaches for lipstick virtual try on (VTO).

First, due to the error in the decoupling lipstick as 'style', they often end up transferring unwanted features like blemish, wrinkles and at times fake specularities ignoring the target lighting condition. Fig. 1 presents a few examples produced using SSAT [22], a recently proposed NST-based

Figure 1. A few images produced using SSAT [22], a recently proposed lipstick transferring approach using reference face images with lipstick.

makeup transferring technique. As can be seen in the first two rows in Fig. 1, SSAT tends to transfer shadows and highlights in the reference image to the target image while completely ignoring the lighting profile of the target image. To address this issue, LipAT applies lipstick to a target image directly using the attributes of the given lipstick, instead of relying on the reference image with the given lipstick.

Second, NST-based approaches mostly preserve the colour of lipsticks – unable to preserve other attributes such as finish types, as they cannot accurately disentangle these other attributes from the colour of lipsticks by looking at the reference image. Also, it is critical to exploit the 3D geometry of the image to accurately preserve attributes such as finish type. LipAT addresses this challenge by having a physics-aware module to incorporate lipstick attributes such as finish type and opacity. LipAT is different from conventional PBR approaches as LipAT does not require scene-specific parameters (e.g., scene lighting) to be given, which are unavailable for images in wild.

Third, NST-based approaches require a big database of face images consisting of at least one image of a person wearing each lipstick that we want to virtually try on. Since such images are typically unavailable with most online lipstick products, NST-based approaches are unusable for a large portion of any e-commerce website's lipstick collection. Additionally, this limitation restricts the number of training instances available to train NST-based approaches. Most previous works adopt MT-dataset [13], which consists of 1115 non-makeup images and 2719 makeup images. Consequently, most NST-based approaches are not generalising well for unseen lipstick products and face images during training (see the last three rows in Fig. 1). Since LipAT applies lipstick using their attributes, LipAT is scalable and also it generalises to diverse lipstick products including extremely rare lipstick products (e.g., lipstick products with bluish colours). We present supporting examples for this statement in the main paper.

2. Implementation Details of LipAT-LAM

LipAT-LAM includes a neural block, denoted as SC, to correct specular highlights of a face image according to the given roughness scores. Motivated by the conditional U-Net architecture proposed in [15], we design SC as the architecture shown in Fig. 2. This architecture consists of multiple AdaIN blocks – adaptive instance normalization layer proposed in [5] – using which the specular highlight update process is conditioned on the given roughness score. Since the output of SC should be well-aligned with the specularities of the input image, UNet-based architecture has been selected to model SC as it includes skip connections to transfer multi-scale knowledge from the input image to the output image.

To train SC, we adopt $\mathbb{D}^{\text{train}}_{\text{real}}$ and $\mathbb{D}^{\text{train}}_{\text{synthetic}}$ (see Section 4 for more details about these datasets) and $L_{\text{pbr\_recon}}$ and L_{γ_recon} loss functions. To optimise, we use Adam optimiser and its hyper-parameters as learning rate $= 1e^{-4}$, decay factor$= 1e^{-5}$, $\beta_1 = 0, \beta_2 = 0.9$, epochs $= 100$, and batch size $= 32$. We train our network using 8 GPUs of size 48GB each.

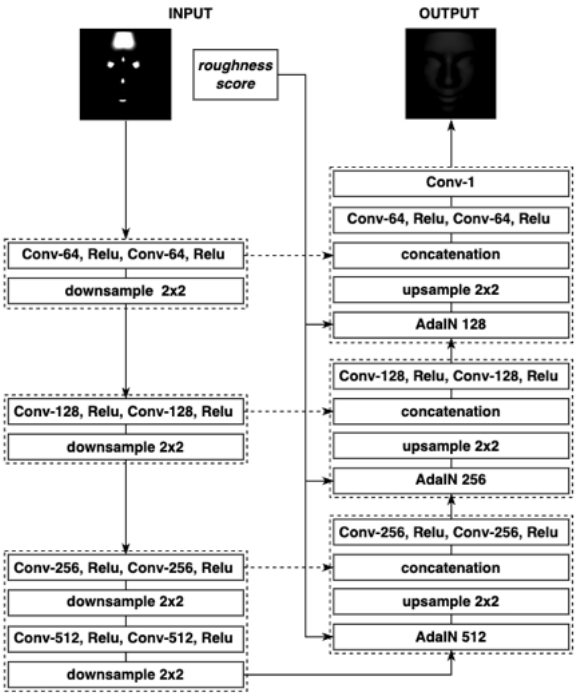

Figure 2. Design of the specular highlight correction (SC) module in LipAT-LAM. Each Conv layer adapts 3×3 kernel.

3. Implementation Details of LipAT-LRM

LipAT-LRM module initially extracts multi-scale features of I and $\hat{I}^a$ using pretrained VGG-19 encoder using ImageNet. Since most imperfections (e.g., unrealistic specularities, incorrect detection of lip) from LipAT-LRM can be identified by checking the differences of finer-level features (e.g., edges) in the lip regions of $\hat{I}^a$ and I, the intermediate features of VGG-19 encoder is a good choice due its ability to extract from granular to coarser features of the input image [20]. In particular, the features of the early layers ($l = 1, 2$) of VGG-19 for I could be useful to refine $\hat{I}^a$ as the early layers emphasis finer-level features such as edges.

To generate the refined image of $\hat{I}^a$, we adopt SPADE-based architecture [17] as shown in Fig. 3. Unlike other image generative neural blocks [15, 25], SPADE can effectively control pixel-level and semantic-level refinements via spatially adaptive normalisation, which makes it ideal for region-specific image augmentation tasks as ours. Unlike the conventional normalisation techniques, given a condition, SPADE produces the normalisation-related modula-

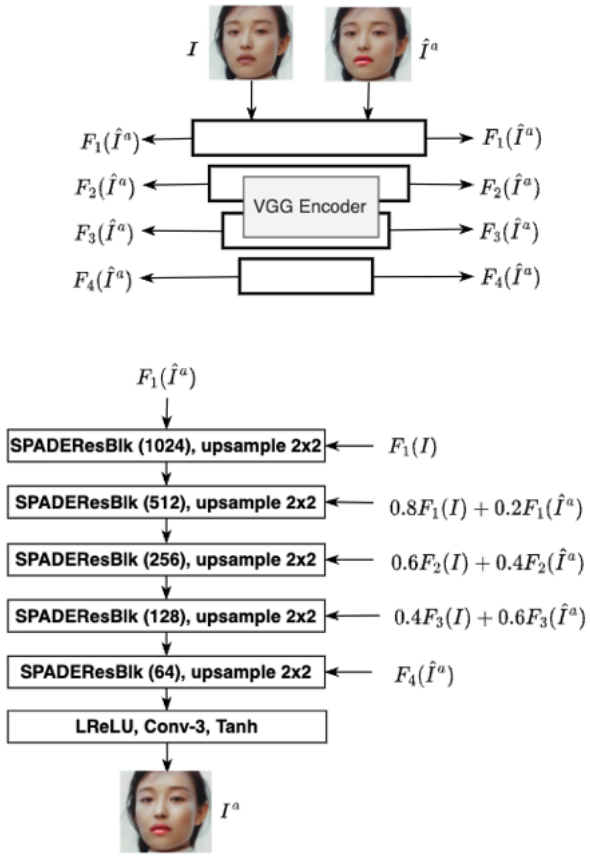

Figure 3. Design of LipAT-LRM

tion parameters as tensors with spatial dimensions to allow more flexibility for the granular-level image updates according to the given condition. SPADEResBlk in Fig. 3 denotes the same SPADE residual block proposed in [17], which consists of multiple SPADE blocks in cascade. Please refer to [17] for more details about the architectures of SPADEResBlk and SPADE.

LipAT-LRM consists of multiple SPADEResBlk in cascade as shown in Fig. 3. However, the condition for each block is different as they emphasise different levels of features. The conditions are computed as a weighted addition of the corresponding multiscale features of I and $\hat{I}^a$. As can be seen, LipAT emphasises the features of I from the early layers of VGG-19 more to exploit the finer-level knowledge in I more to refine I.

To train R, we adopt the images with lipstick and without lipstick in $\mathbb{D}^{\text{train}}_{\text{real}}$ (see Section 4 for more details about this dataset). To optimise, we adopt RMSprop optimiser and its hyper-parameters as learning rate $= 2e^{-4}$, $\beta_1 = 0, \beta_2 = 0.9$, epochs $= 200$, and batch size $= 32$. We train our network simultaneously using 8 GPUs of size 48GB each.

Figure 4. A few examples from D^{train}_{real}: (a) images without lipstick; and (b) image with lipstick

4. More Details on Datasets

4.1. Training Datasets

The training of LipAT involves two datasets: $\mathbb{D}^{\text{train}}_{\text{real}}$ and $\mathbb{D}^{\text{train}}_{\text{synthetic}}$.

$\mathbb{D}^{\textbf{train}}_{\textbf{real}}$ - This dataset consists of 10,000 face images without lipstick and 10,000 face images with lipstick from the CelebA-HQ dataset [10]. CelebA-HQ provides annotations under an attribute called *wearing lipstick*, which is 1 if the corresponding face image is wearing lipstick, 0 otherwise. The aforementioned two subsets of images are compiled with the help of this attribute.

Most previous lipstick simulation approaches [13, 16, 22, 27] are trained using MT-Dataset [13], which consists of 1115 non-makeup images and 2719 makeup images. Due to the small size of this dataset, most of the models trained using this dataset perform poorly for new face images as shown in Fig. 1. We aim to address this issue by having a large face-image dataset as the primary training dataset in LipAT, which gives better coverage of face images with respect to age, gender and scene lighting (see Fig. 4). During training, we adopt the images without lipstick in $\mathbb{D}^{\text{train}}_{\text{real}}$ to learn the neural components in both modules of LipAT. The images with lipstick are only used to serve real images to optimize L_{adv}.

Additionally, $\mathbb{D}^{\text{train}}_{\text{real}}$ consists of 10,000 lipstick attribute vectors as LipAT can be directly trained using attribute vectors, unlike other NST-based approaches. This distinctive trait allows unlimited control over the distribution of lipstick attributes used to train LipAT, which is leveraged in this work to improve the generalisability of LipAT. We fit a Gaussian kernel density estimation (KDE) function in the space of the lipstick attributes (RGB values of the base colors and roughness scores) using publicly available information related to lipstick products gathered from online sources. From these learned distribution, we observed that lipstick products tend to have reddish colours. Also, the matte lipstick products (roughness score around 0.7) are dominant in the market. Sampling lipstick attribute vectors for training according to such distribution allows LipAT to

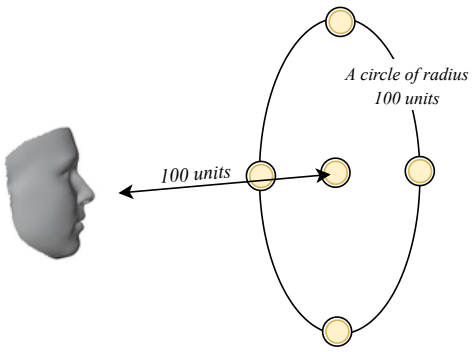

Figure 5. How the location of the spot light source in the PyVista rendering environment has been altered to replicate 5 different light directions for the images in $\mathbb{D}^{\text{train}}_{\text{synthetic}}$.

generalise well for real-world lipstick products. However, there are rare lipstick products (e.g., bluish colours) that are well covered from the learned distribution. We empirically observed that most existing NST-based lipstick VTO solutions perform poorly for such examples (see our main paper for a few examples). To address this problem, we sample 10,000 lipstick attribute vectors in $\mathbb{D}^{\text{train}}_{\text{real}}$ using the pre-trained KDE and a uniform distribution alternatively to produce a realistic set of lipstick attributes with extreme examples. We sampled the opacity attribute using a uniform distribution in the range $[0, 1]$.

$\mathbb{D}^{\textbf{train}}_{\textbf{synthetic}}$ - This dataset is used to optimize $L_{\text{pbr\_recon}}$ during the learning of the specular highlight correction module in LipAT. $\mathbb{D}^{\text{train}}_{\text{synthetic}}$ consists of 10,000 specular components corresponding to 80 unique face images that are rendered using PyVista physics-based rendering engine. We adopt the following steps to produce $\mathbb{D}^{\text{train}}_{\text{synthetic}}$.

For each face image, we produce its 3D mesh using media-pipe[1]. The produced 3D mesh is used to render the specular highlight components of the face image (i.e., reflection maps) using the metallic-roughness PBR workflow in PyVista under 5 different light directions (see Fig. 5), 5 light intensity values $\{1, 2, 5, 7.5, 10\}$ and 5 different roughness scores $\{0.1, 0.3, 0.5, 0.7, 0.9\}$. Figure 6 shows a few examples from $\mathbb{D}^{\text{train}}_{\text{synthetic}}$ under different variables of the rendering environments. Having such a diverse dataset makes the specular highlight correction module in LipAT-LAM robust against the unseen values from the space of the rendering parameters.

4.2. Test Datasets

The testing of LipAT involves following two datasets: $\mathbb{D}^{\text{test}}_{\text{up}}$ and $\mathbb{D}^{\text{test}}_{\text{wp}}$.

$\mathbb{D}^{\text{test}}_{\text{up}}$ - We adopt this dataset to measure the realism of the generated images using a novel variant of FID. Since

[1]https://google.github.io/mediapipe/solutions/face_mesh.html

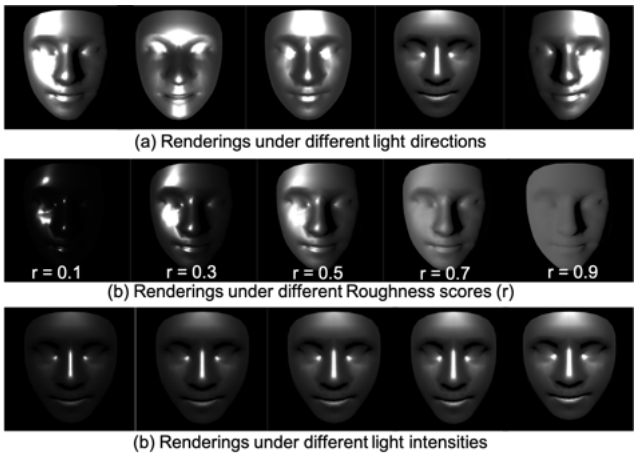

Figure 6. A few examples from $\mathbb{D}^{\text{train}}_{\text{synthetic}}$ under different variables of the rendering environment.

FID [4] requires at least 2048 images to be meaningful, this dataset consists of 2048 face images with lipstick and 2048 without lipstick from the CelebA-HQ dataset. There is no overlapping images between $\mathbb{D}^{\text{test}}_{\text{up}}$ and $\mathbb{D}^{\text{train}}_{\text{real}}$. Additionally, each image without lipstick in $\mathbb{D}^{\text{test}}_{\text{up}}$ associates with a lipstick attribute vectors – sampled as in $\mathbb{D}^{\text{train}}_{\text{real}}$. To compute FID, it requires a real dataset ($\mathbb{D}_{\text{real}}$) and a dataset consisting of generated images $\mathbb{D}_{\text{gen}}$[2]. We adopt the image with lipstick in $\mathbb{D}^{\text{test}}_{\text{up}}$ as $\mathbb{D}_{\text{real}}$, and the generated images by applying lipsticks to the images without lipstick in $\mathbb{D}^{\text{test}}_{\text{up}}$ as $\mathbb{D}_{\text{gen}}$.

$\mathbb{D}^{\text{test}}_{\text{wp}}$ - To quantitatively evaluate the accuracy of the lipstick simulation approaches, it ideally requires paired images of subjects with and without lipstick. Such publicly available datasets with enough images are unavailable. In this work, we construct a paired dataset using the face images with lipstick in CelebA-HQ using a weak labelling approach. Our approach adopts the makeup removal module proposed in [22], which can remove lipstick from an image given another reference face image without lipstick. This makeup removal module is unable to recover the lip attributes (e.g., colour) of the target face, thus, it simulates the lip attributes of the reference face image, which could be from a different person (see Fig. 7).

To address this limitation, we extract the image pairs in the CelebA-HQ dataset of the same person with and without lipstick using the annotations (i.e., *'user_id'* and *'wearing lipstick'* attributes) provided in CelebA-HQ. For each such pair, we adopt SSAT to remove the lipstick of the image with lipstick using an image without lipstick of the same person as the reference. Although the aforementioned approach considers the actual lip attributes when removing lipstick, the results could be unrealistic in some instances due to the differences in the lighting of the image pairs

[2]See section 5 for more details about this metric

Figure 7. Makeup removal results using SSAT [22]. The figure is taken from [22]. The first row is three nonmakeup reference images and the left column is target face image. The makeup removal results are displayed in the lower right corner

Figure 8. Each row shows an example from our weakly paired test dataset – the columns from left to rights represent the reference images without lipstick, target image with lipstick depicting the same person as the reference, and the generated image with the lipstick removed, respectively.

and imperfections in segmenting lips. By manually filtering out such unrealistic images, we constructed a weakly paired dataset consisting of 127 image pairs with and without lipstick. A few examples from this dataset are shown in Fig. 8.

To adapt $\mathbb{D}_{\text{wp}}^{\text{test}}$ for our quantitative evaluation framework, the images with lipstick in $\mathbb{D}_{\text{wp}}^{\text{test}}$ should be annotated with the corresponding lipstick attributes a. We adopt two data-driven approaches to extract the colour and roughness score attributes from the face images with lipstick.

To extract lipstick colour a^c from face images with lipstick, we adopt the approach proposed [14]. In this approach, we first extract the pixels of the lip of an image using face segmentation masks from media-pipe and convert the RGB values of the selected pixels to CIE-LAB colour space (see Section 11 for more details). Then, we cluster

Figure 9. Extracted lipstick attributes a for a few examples in $\mathbb{D}_{\text{wp}}^{\text{test}}$. a corresponding to each image is shown in the second row – each a vector gives the base color, roughness score (0.1 for glossy and 0.7 for matte), and opacity from left to right.

the selected pixels based on their a and b dimensions in the Lab values using K-Means into 5 clusters. Then, it returns RGB value corresponding to the center of the largest cluster as the colour of the lipstick product.

To extract roughness scores a^r, we first pre-trained a ResNet model using the self-supervised appearance-preserving contrastive loss proposed in [1] using the L dimension of the images in the MERL dataset. The MERL dataset consists of images of various objects (e.g., spheres, blobs) rendered using PBR under different material properties and different lighting environments. By pre-training only using L dimension of the images, we aim to make the image encoder robust against colour variations as colour information is not useful to predict the roughness of a surface. The adopted triplet loss function aims to learn similar representations for the images rendered using similar material properties. Subsequently, we trained a two-layer feed-forward neural network with ReLU activation to predict roughness using the features from the pre-trained image encoder. To train this neural network, we adopt the same dataset used to construct the lipstick attribute vectors in $\mathbb{D}_{\text{real}}^{\text{train}}$, which consists of public information related to the finish types of real-world lipstick products. We observed that this approach outperforms other conventional image encoders (e.g., EfficientNet, ResNet, VGG-19) trained using ImageNet and statistical heuristic-based approaches. This approach can predict the absolute roughness score with an accuracy of 0.14, which is enough to differentiate widely-known finish types (e.g., matte = 0.7, cream=0.3, glossy=0.1).

By adopting the aforementioned approaches, we weakly inferred a^c and a^r for the images with lipstick in $\mathbb{D}_{\text{wp}}^{\text{test}}$, and assumed that $a^o = 0.8$. Figure 9 shows the extracted attributes for a few examples in $\mathbb{D}_{\text{wp}}^{\text{test}}$.

5. Patch-FID

In this work, we introduce Patch-FID, a novel variant of FID that is suitable in cases where a small region of the images has been changed during the artificial augmentation – e.g., lip region in a full face image. We have empirically found that the conventional FID measure is insensitive for

such minor image augmentations. This section formally defines the proposed metric and compares it with conventional FID and a naive variant of FID using an example to show the superiority of Patch-FID.

Limitation of FID. To evaluate the preservation of realism of the generated images, most previous works adopt Fréchet Inception Distance (FID) [4]. For given two image datasets – the one consisting of real images $\mathbb{D}_{\text{real}}$ and the other consisting of generated images $\mathbb{D}_{\text{gen}}$, FID metric is formulated as the Wasserstein distance d_W [24] between the two Gaussian distributions estimated from $\mathbb{D}_{\text{real}}$ and $\mathbb{D}_{\text{gen}}$ as follows:

$$FID(\mathbb{D}_{\text{real}}, \mathbb{D}_{\text{gen}}) = d_W(\mathcal{N}_{\text{real}}, \mathcal{N}_{\text{gen}}) \tag{1}$$

In the conventional FID metric $\mathcal{N}_{\text{real}}$ and $\mathcal{N}_{\text{gen}}$ are estimated using the sets of image features – F_{real} and F_{gen} respectively, from an intermediate layer of the pre-trained Inception model [23] as follows:

$$F_{real} = \{\bigoplus_{\forall k} avg(f^l(I)[:,:,k]) | \forall I \in \mathbb{D}_{\text{real}}\} \tag{2}$$

$$F_{gen} = \{\bigoplus_{\forall k} avg(f^l(I)[:,:,k]) | \forall I \in \mathbb{D}_{\text{gen}}\} \tag{3}$$

where $f^l : I \rightarrow \mathbb{R}^{H \times W \times C}$ is the activations from the l^{th} layer of the pre-trained Inception model, which consists of C number of channels of size $H \times W$. $\bigoplus_{\forall k} avg(f^l(I)[:,:,k]) \in \mathbb{R}^C$ denotes the concatenation of the average of features in each channel. Since F_{real} and F_{gen} in Equations 2 and 3 treat the features corresponding to all the pixels in face images equally, FID score becomes insensitive when you are only updating a small region of real images to generate $\mathbb{D}_{\text{gen}}$ – e.g., lip region of a full face image.

Patch-FID. To address this limitation, we propose Patch-FID, which estimates F_{real} and F_{gen} only using the activations from the Inception model corresponding to the updated regions in the images. The set of locations of such activations P^I of an image I is identified with the help of a mask M^I for the lip region in I as follows:

$$P^I_k = \{(x,y) | f^l(M^I \cdot I)[x,y,k] \neq f^l(M^\psi \cdot I)[x,y,k]\})\} \tag{4}$$

where $x \in [0, H-1]$, $y \in [0, W-1]$ and M^ψ is a black image that masked out the whole face images. Our approach produces $F^I_\cdot$ for an given image I just using the locations captured using the Eq. 4 as follows:

$$F^I_\cdot = \bigoplus_{\forall k} \frac{1}{||P^I_k||} \sum_{(x,y) \in P^I_k} f^l(I)[x,y,k] \tag{5}$$

Then, Patch-FID computes F_{real} and F_{gen} by computing $F^I_\cdot$ from Equation 5 using the images in $\mathbb{D}_{\text{real}}$ and $\mathbb{D}_{\text{gen}}$ respectively. Then, the final score is computed using Eq. 1.

Figure 10. The selected example for the case study – (a) image without lipstick I; (b) image with lipstick I^a; and (c) the selected spatial locations of the intermediate feature maps of I to compute Patch-FID. The original feature maps are scaled to the scale of I to produce this figure to compare spatial locations. If a spatial location is frequently got selected across feature maps from different channels to compute Patch-FID, the colour of that location is pushed towards white in this figure.

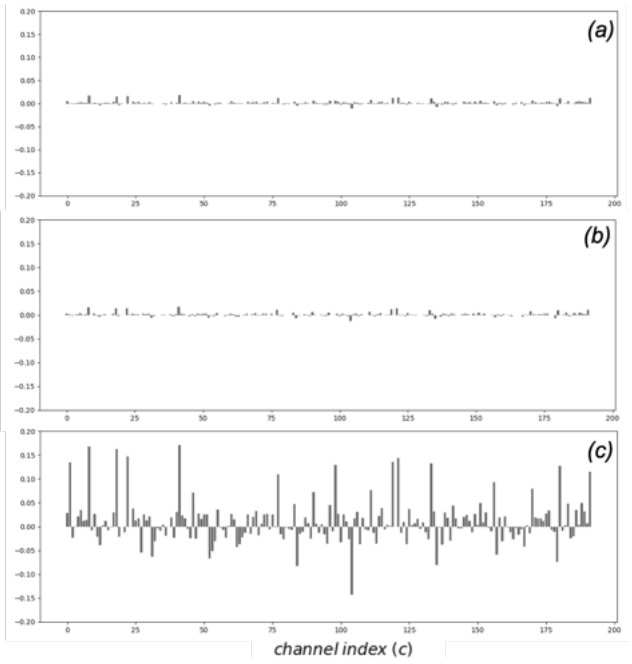

Figure 11. Difference of the inception feature values of I and I^a in Fig. 10 for each channel using different methods – (a) FID, (b) Masked-FID and (c) Patch-FID. These metrics are different due to their different strategies when pooling the spatial dimensions of each channel

Our experiments reported in the main paper showed that Patch-FID is more suitable and sensitive enough to evaluate our task.

Case Study. Here we compare Patch-FID with two other baselines using an example. As the baselines, we adopt conventional FID metric and a naive variant of FID, denoted as "*Masked-FID*". Masked-FID masks out the other pixels except for the lip region of the face images and computes FID using the masked images. As an example, we adopt the im-

Method	Wasserstein Distance
FID	0.0031
Masked-FID	0.0029
Patch-FID	**0.4087**

Table 1. Wasserstein Distance between I and I^a in Fig. 10 using their Inception features with different metrics.

age in Fig. 10 (a). As can be seen in Fig. 10 (c), Patch-FID only picks intermediate activations from Inception that are corresponding to the lip region. In contrast, Masked-FID and FID consider all the activations equally when computing the metrics. As a result, the intermediate channel values (after pooling spatial dimensions) for I and I^a in Figure 10 are similar when using Masked-FID and FID (see Figures 11 (a) and (b)). Thus, Masked-FID and FID become insensitive to the minor/ part-specific changes in the images. This can be quantitatively verified using Table 1, which reports a very small Inception-based Wasserstein distance values between I and I^a with Masked-FID and FID. These results further verify the importance of a metric like Patch-FID to measure realism in cases where a small region of the images has been changed during the artificial augmentation. Since Patch-FID is application-agnostic and FID is a special case of Patch-FID, it can be applied across different applications to evaluate the preservation of realism over FID.

6. Baselines

In this work, we compare the proposed with 8 baselines that are categorised as: (1) PBR-based baselines; (2) NST-based approaches; and (3) Hybrid approaches. Under PBR-based approaches, we adopt three baselines:

- Colour-Transfer [18] – this baseline adopts the colour transfer approach proposed in [18]. This baseline only considers the base colour attribute when applying lipstick, and performs an operation similar to the diffuse update in LipAT-LAM module. Thus, it does not preserve other lipstick attributes such as finish type as it is not explicitly updating the specular highlights of the images.
- AR [21] – this is our implementation of [21]. Since some of the parameters – e.g., γ values in gamma transformation, used in this work are manually tuned for each image in the paper, we set them to consistent values after tuning to make this baseline a fully automated solution for a fair comparison with our work and the other baselines.
- LAM – this baseline only consists of the lipstick application module (LAM) proposed in our framework. The comparison between this baseline and our full model allows identifying the contribution from LRM, the second module in our pipeline.

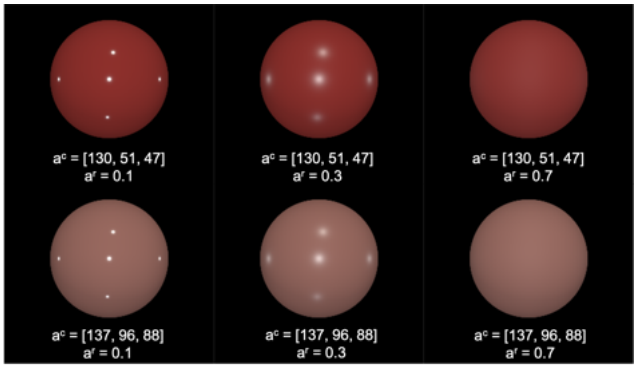

Figure 12. A few examples of swatch images rendered for different lipstick attributes using PyVista PBR engine.

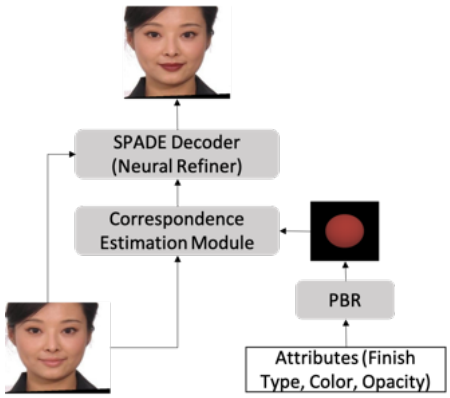

Figure 13. Overview of Swatch-SpMT neural baselines, which transfers lipstick attributes via a swatch image rendered according to the given lipstick attributes.

Since almost all the existing NST-based approaches are unable to transfer lipstick using lipstick attributes, we modified SpMT [27], a recently proposed neural makeup transfer approach, to create two neural-based baselines that can transfer lipstick without requiring a full face image with the reference lipstick:

- Swatch-SpMT - this baseline first represents the attributes of lipsticks using swatch images, that was constructed using PBR (see Fig. 12). By providing the lip area in the target image and the swatch image as the corresponding regions to the non-parametric correspondence estimation module proposed in SpMT [27], we modified SpMT to transfer lipsticks using swatch images that are rendered based on lipstick attributes. Overview of the modified framework is shown in Fig. 13. To train the model, we adopt the same architecture and the same set of objective functions proposed in SpMT [27] except its cosmetic loss. Since we do not have the reference face image with lipstick in

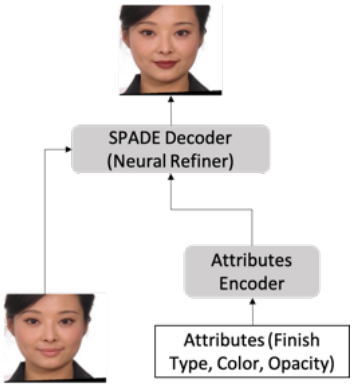

Figure 14. Overview of Att-SpMT neural baselines, which applies lipstick directly using lipstick attributes.

this setting, we adopt LipAT-LAM to generate pseudo labels to compute the cosmetic loss as $L_{\cos}$ in LipAT.

- Att-SpMT - this baseline trains the SPADE decoder in SpMT to transfer lipstick directly conditioning on the attributes. We adopt the architecture proposed for multimodal synthesis in [17], which maps the attributes into a latent space using which the SPADE decoder is conditioned on. The overview of this framework is shown in Fig. 14. The architecture of SPADE decoder is similar to the decoder in [27], and the attribute encoder is similar to the encoder proposed in [15].

Otherwise specified, we adopt the training hyperparameters proposed in the corresponding papers to train the NST-based baselines.

As the hybrid approaches, we combine LAM with three recently proposed neural lipstick transfer approaches: (1) CPM [16]; (2) SSAT [22]; and (3) SpMT [27]. For each of these baselines, we first adopt our LAM module to simulate the lipstick on the target image and then use it as the reference image to transfer lipstick using the selected neural lipstick transfer approaches. We adopt the pre-trained models of CPM, SSAT and SpMT available in their original repositories to collect results using hybrid approaches.

7. More Details on User Studies

This work conducts 3 user studies to qualitatively evaluate three aspects of the generated images – preservation of realism, preservation of finish types and overall accuracy. We conducted these user studies using Amazon Mechanical Turk. For each question in the user studies, we assigned 5 annotators and aggregated their results to get the reported results in the main paper. We adopt Fleiss' kappa [3] to measure the agreement between the annotators. We observed substantial agreement between the annotators for all three user studies – 0.71, 0.61, 0.65 Fleiss' kappa scores for User Studies 1, 2 and 3 respectively, which verifies the reliability of the conducted user studies.

Figure 15. Three examples from User Study 1 – Image 1 and Image 2 in each example could be be real images with lipstick or artificially altered images. The participants have to pick the image with the most realistic lipstick application.

7.1. User Study 1 - Preservation of Realism

In this study, participants have been shown two face images of the same person with the same lipstick for each round. The images could be real images with lipstick or artificially altered images by applying lipstick to real images using different methods as shown in Fig. 15. Then, the participants have been asked to pick the image that gives the most realistic application of lipstick. We also explicitly instructed participants to zoom in to the lip area of the images when judging similar images.

This user study consists of 1270 questions covering a balanced distribution between real images and the images rendered using four techniques – AR, Swatch-SpMT, LAM + SpMT and LipAT. Finally, the percentages of the images from different methods are picked as realistic are reported as the results of this study.

7.2. User Study 2 - Preservation of Finish Type

This study evaluates how accurately each method can incorporate finish types into the rendered images. For each round in this study, participants have been shown two artificially generated images of the same person with two lipstick products that have same colour but different finish types (i.e., glossy and matte) as shown in Fig. 16. Then, the participants have been asked to select the image that simulates the appearance of a glossy lipstick. From our preliminary experiments, we observed that humans struggle to predict the finish type of lipstick, even in real images, without hav-

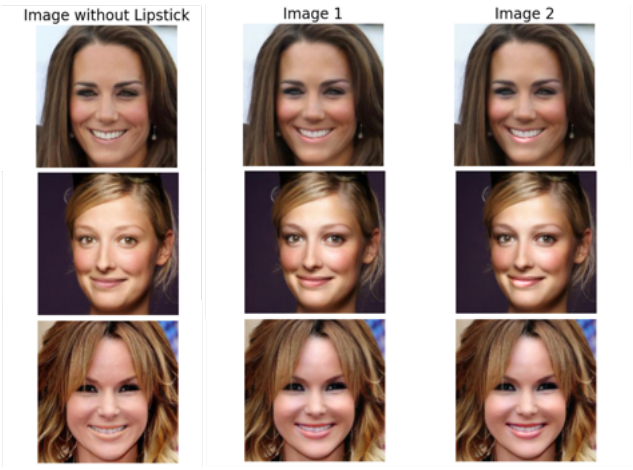

Figure 16. Three examples from User Study 2 – Image 1 and Image 2 are generated by applying two lipsticks with same color but different finish type to the image in left.

Figure 17. A set of lip images with glossy and matte lipsticks

ing additional clues. To make the judgements easier and accurate, we attached the corresponding face image without lipstick to each question (see Fig. 16) and also provided a set of examples for lips with glossy and matte lipsticks along with each question as shown in Fig. 17.

This user study consists of 508 questions covering a balanced distribution between the images rendered using four techniques – AR, Swatch-SpMT, LAM + SpMT and LipAT. Finally, the accuracy of picking the correct images with glossy lipstick from each method is reported as the result of this study.

7.3. User Study 3 - Overall Accuracy of Lipstick Appearance

This study evaluates the overall correctness of different methods. For each round, participants have been shown a reference image with lipstick and a sequence of generated images by artificially applying the lipstick on the reference image to a different face image using lipstick attributes (see Fig. 18).

This user study consists of 127 questions and the four generated images in each question are rendered using four techniques – AR, Swatch-SpMT, LAM + SpMT and LipAT.

Figure 18. Three examples from User Study 3. The participants have to pick the image out of Images 1, 2, 3 and 4 that simulates the appearence of the lipstick in the reference image.

Finally, the percentages of the images from different methods are picked as the most accurate images are reported as the results of this study.

8. More Quantitative Results

To further evaluate the ability of LipAT to incorporate finish types, Table 2 shows diversity between images generated by applying the same set of lipstick to the same set of image, but with different finish types. As can be seen, PBR-based solution and LipAT can produce significant diversity for different finish types compared, The same figures are negligible for all NST-based approaches. Although the diversity is not a good measure of the accuracy of lipstick application, this experiment shows that the inability of NST-based approaches to properly preserve finish types. We present a few examples in Fig. 20, which also shows aligning results with the aforementioned statements. In Fig. 21, we present a few LipAT's results for face images with dark skin tone to show that how LipAT generalises for different skin tones.

Quantitative Metrics vs User Studies. Here, we analyse the agreement of the results coming from user studies and from our quantitative evaluation. As can be seen in Figure 19, User Study 1 results align well with Patch-FID, which further verifies the suitability of the proposed novel variant of FID to measure realism, particularly compared to the conventional FID score. User study 2 focuses on preserving finish type. We can observe that SSIM is a suitable metric in this aspect to capture human perception of finish types.

Additionally, we can see that none of the quantitative metrics alone can produce aligning results with User Study 3. This could be because User Study 3 focuses on the overall correction of lipstick application, which should jointly consider both realism and attribute preservation. Thus, we

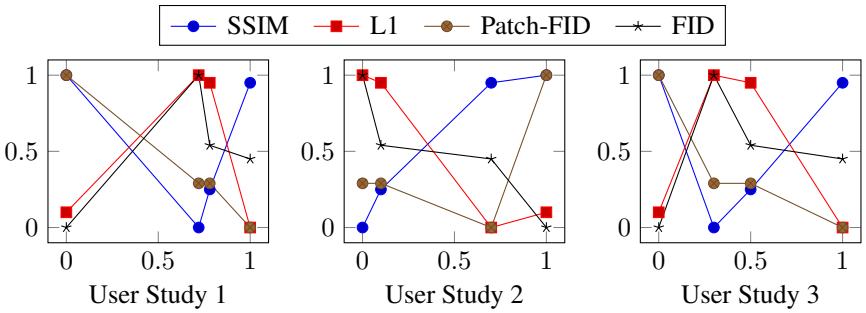

Figure 19. Analysis of the agreement between the user studies and different quantitative metrics. The standardised results from metrics and user studies are used for the plots. The results show that Patch-FID and SSIM metrics are well aligned with User Study 1 and User Study 2 results respectively.

Method Type	Method	SSIM (↓)	L1 (↑)	Patch-FID (↑)	FID (↑)
PBR	Colour-Transfer [18]	1	0.0000	0.0010	0.0002
	AR [21]	0.8766	0.0026	4.8238	0.0101
	LAM	**0.8651**	**0.0029**	**4.9184**	**0.0117**
NST	Swatch-SpMT	0.9978	0.0003	0.0527	0.0019
	Att-SpMT	0.9894	0.0009	0.0794	0.005
Hybrid	LAM + CPM [16]	0.9963	0.0004	0.0851	0.0024
	LAM + SSAT [22]	0.9951	0.0006	0.0910	0.0037
	LAM + SpMT [27]	0.9958	0.0005	0.0891	0.0032
	Our Approach	0.8801	0.0023	4.8380	0.0098

Table 2. Diversity analysis of the generated images using different finish types. For this analysis, we rendered two datasets using each method by applying the same color lipsticks for the images in $\mathbb{D}_{\text{wp}}^{\text{test}}$ with two different finish types – glossy and matte. The reported figures are computed using the corresponding images with glossy and matte finishes in the constructed datasets.

Method	SSIM (↑)	Patch-FID (↓)
LipAT-LAM	0.799	23.2
LipAT-LAM w/o $L_{\text{pbr-recon}}$	0.793	25.7
LipAT-LAM w/o $L_{\gamma\text{-recon}}$	0.770	23.1

Table 3. Ablation study of LipAT-LAM

Method	SSIM (↑)	Patch-FID (↓)
LipAT	0.797	20.2
LipAT w/o L_{cos}	0.678	21.3
LipAT w/o L_{ref}	0.791	22.9
LipAT w/o L_{adv}	0.794	21.8

Table 4. Ablation study of LipAT-LRM

produce a new metric as $\alpha_1 \cdot \text{SSIM} + \alpha_2 \cdot \text{Patch-FID}$ by combining the most suitable metric for measuring realism and attribute preservation. We then attempted to find α_1 and α_2 here such that this new metric maximises the agreement with the User Study 3 results. As shown in Fig. 22, we observed that this new metric is able to produce a better alignment – 0.89 R^2 value compared to the 0.86 and 0.74 R^2 values getting from SSIM and Patch-FID alone – with the User Study 3 results. These results further show the suitability of the proposed quantitative evaluation framework (i.e., the selected set of metrics and the weakly paired dataset) for evaluating different aspects of lipstick simulation approaches while agreeing to human perception.

9. Ablation Study

Since multiple loss functions are adopted to train the neural components in LipAT-LAM and LipAT-LRM, here we perform an ablation study to show the positive contribution of those loss functions.

Table 3 shows the importance of $L_{\text{pbr-recon}}$ and $L_{\gamma\text{-recon}}$ loss terms in the specular highlight correction module in LipAT-LAM. As can be seen, both loss terms are positively contributed towards the final performance. In particular, we observe that the model is not generalising well for wild images without $L_{\gamma\text{-recon}}$, and not accurately incorporate roughness without $L_{\text{pbr-recon}}$. Please see Fig. 5 in the main paper for examples.

Similarly, Table 4 shows that three loss terms used in LipAT-LRM have a positive impact. We empirically observed that each loss term plays a unique role: L_{cos} is important to preserve colour, L_{ref} is important to preserve finish types accurately, L_{adv} is important to preserve realism. Overall, this ablation study quantitatively verifies the importance of different components in LipAT, and the importance of how they are trained.

10. LipAT's Limitations

One of the limitations of LipAT is its reliance on the accuracy of the lip segmentation model. While LipAT can correct minor discrepancies between the predicted mask and the actual lip region, it fails to apply lipstick in scenarios where the lip segmentation technique is unable to detect the lip region in certain images (e.g., side face images and face

Figure 20. More results using a PBR technique (LAM), NST-based technique (LAM + SpMT) and LipAT. Each row shows the results generated by applying two lipsticks with the same colour, but different finish types.

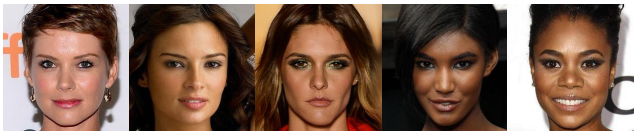

Figure 21. LipAT's results for face images with dark skin tones

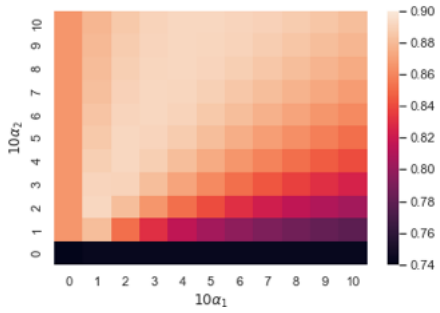

Figure 22. Coefficient of determination values for the regression task of predicting User Study 3 results using different combinations of SSIM and Patch-FID. In this analysis, independent variable is defined as $\alpha_1 * \text{SSIM} + \alpha_2 * \text{Patch-FID}$. All the variables – User Study 3 results, Patch-FID scores, SSIM scores, were standardised to the range $[0, 1]$ prior using for the regression task.

Figure 23. A few failure cases of the selected neural lip parser [8]

images with obscured lip region as shown in Fig. 23). To prevent these failures, users can be advised to avoid capturing such cases when testing LipAT in real-world settings. Additionally, handling these failure cases could be achieved by adopting a more dense and robust face landmark detection technique, considering that this research domain is active and continuously evolving [26].

Another limitation is that LipAT alone cannot accurately apply lipstick to face images that are already wearing lipstick. Previous works [2, 19] attempted to address this issue by assuming lipstick removal as another form of lipstick application, thus, adopt similar neural networks and features in the lip region for both lipstick application and removal. Although combining LipAT with an automated lipstick removal approach could enhance its scalability for face images with lipstick (see Fig. 24 for a few examples from such an approach), our efforts in constructing $\mathbb{D}_{wp}^{test}$ revealed that such existing makeup removal approaches can produce unrealistic results (refer to Section 4 for more details). This is primarily due to the complexity and challenges involved in lipstick removal as it lacks the necessary information to accurately determine the actual lip color of a person when looking at an image of their face with lipstick applied. Therefore, further research efforts are necessary to develop improved lipstick removal techniques, taking into account additional attributes of a face image such as the skin tone of other regions.

Figure 24. Lipstick simulation results on face images already wearing lipstick by combining LipAT with SSAT [22]. Please see how $\mathbb{D}_{wp}^{test}$ is constructed in Section 4 for more details how SSAT can be used for lipstick removal.

11. Preliminaries on CIELAB Colour Space

CIELAB is a device-independent color space that was designed to be a perceptually uniform space, where a given numerical change corresponds to a similar perceived change in color. Thus, CIELAB color space is useful for predicting small differences in color. CIELAB consists of three dimensions (L, a, b) to represent each color. L defines the lightness value, which is black at 0 and white at 100. The a and b dimensions represent the relative position of a color with respect to the green–red and blue-yellow opponent colors respectively, which range from -1 to 1. For example, having a negative number for an axis means the color is closer to green and a positive number means a color is closer to red. The conversions between RGB and CIELAB color spaces are done via CIEXYZ color space. Please refer to here for the detailed formulas for the conversions.

References

[1] Manuel Lagunas Arto, Sandra Malpica, Ana Serrano, Elena Garces, Diego Gutierrez, and Belen Masia. A similarity mea-

sure for material appearance. *Jornada de Jóvenes Investigadores del I3A*, 7, 2019. 5

[2] Huiwen Chang, Jingwan Lu, Fisher Yu, and Adam Finkelstein. PairedCycleGAN: Asymmetric Style Transfer for Applying and Removing Makeup. In *Proc. of CVPR*, pages 40–48, 2018. 1, 12

[3] Joseph L Fleiss. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378, 1971. 8

[4] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs Trained by a Two Time-scale Update Rule Converge to a Local Nash Equilibrium. *Proc. of NIPS*, 30, 2017. 4, 6

[5] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proc. of CVPR*, pages 1501–1510, 2017. 2

[6] Zhikun Huang, Zhedong Zheng, Chenggang Yan, Hongtao Xie, Yaoqi Sun, Jianzhong Wang, and Jiyong Zhang. Real-world Automatic Makeup via Identity Preservation Makeup Net. In *Proc. of IJCAI*, pages 652–658, 2021. 1

[7] Yongcheng Jing, Yezhou Yang, Zunlei Feng, Jingwen Ye, Yizhou Yu, and Mingli Song. Neural style transfer: A review. *IEEE transactions on visualization and computer graphics*, 26(11):3365–3385, 2019. 1

[8] Yury Kartynnik, Artsiom Ablavatski, Ivan Grishchenko, and Matthias Grundmann. Real-time facial surface geometry from monocular video on mobile gpus. *arXiv preprint arXiv:1907.06724*, 2019. 12

[9] Robin Kips, Ruowei Jiang, Sileye Ba, Edmund Phung, Parham Aarabi, Pietro Gori, Matthieu Perrot, and Isabelle Bloch. Deep Graphics Encoder for Real-Time Video Makeup Synthesis from Example. In *Proc. of CVPR*, pages 3889–3893, 2021. 1

[10] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. MaskGAN: Towards Diverse and Interactive Facial Image Manipulation. In *Proc. of CVPR*, 2020. 3

[11] Chen Li, Kun Zhou, and Stephen Lin. Simulating Makeup through Physics-based Manipulation of Intrinsic Image Layers. In *Proc. of CVPR*, pages 4621–4629, 2015. 1

[12] Jiayue Li, Qing Wang, Hong Chen, Jiahui An, and Shiji Li. A review on neural style transfer. In *Journal of Physics: Conference Series*, volume 1651, page 012156, 2020. 1

[13] Tingting Li, Ruihe Qian, Chao Dong, Si Liu, Qiong Yan, Wenwu Zhu, and Liang Lin. BeautyGAN: Instance-level facial makeup transfer with deep generative adversarial network. In *Proc. of ACM MM*, pages 645–653, 2018. 2, 3

[14] MathWorks. Color-based segmentation using k-means clustering. 5

[15] Gabriel Meseguer-Brocal and Geoffroy Peeters. Conditioned-u-net: Introducing a control mechanism in the u-net for multiple source separations. *arXiv preprint arXiv:1907.01277*, 2019. 2, 8

[16] Thao Nguyen, Anh Tuan Tran, and Minh Hoai. Lipstick ain't Enough: Beyond Color Matching for In-the-wild Makeup Transfer. In *Proc. of CVPR*, pages 13305–13314, 2021. 1, 3, 8, 10

[17] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic Image Synthesis with Spatially-adaptive Normalization. In *Proc. of CVPR*, pages 2337–2346, 2019. 2, 3, 8

[18] Erik Reinhard, Michael Adhikhmin, Bruce Gooch, and Peter Shirley. Color transfer between images. *IEEE Computer graphics and applications*, 21(5):34–41, 2001. 7, 10

[19] Steven A Shafer. Using color to separate reflection components. *Color Research & Application*, 10(4):210–218, 1985. 12

[20] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *Proc. of ICLR*, 2015. 2

[21] Kanstantsin Sokal, Siarhei Kazakou, Igor Kibalchich, and Matsvei Zhdanovich. High-quality AR Lipstick Simulation via Image Filtering Techniques. In *Proc. of CVPR Workshop on Computer Vision for Augmented and Virtual Reality*, 2019. 1, 7, 10

[22] Zhaoyang Sun, Yaxiong Chen, and Shengwu Xiong. SSAT: A Symmetric Semantic-aware Transformer Network for Makeup Transfer and Removal. In *Proc. of AAAI*, pages 2325–2334, 2022. 1, 3, 4, 5, 8, 10, 12

[23] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the Inception Architecture for Computer Vision. In *Proc. of CVPR*, pages 2818–2826, 2016. 6

[24] Leonid Nisonovich Vaserstein. Markov processes over denumerable products of spaces, describing large systems of automata. *Problemy Peredachi Informatsii*, 5(3):64–72, 1969. 6

[25] Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multiscale Structural Similarity for Image Quality Assessment. In *Proc. of ACSSC*, pages 1398–1402, 2003. 2

[26] Erroll Wood, Tadas Baltrušaitis, Charlie Hewitt, Matthew Johnson, Jingjing Shen, Nikola Milosavljević, Daniel Wilde, Stephan Garbin, Toby Sharp, Ivan Stojiljković, et al. 3d face reconstruction with dense landmarks. In *Proc. of ECCV*, 2022. 12

[27] Mingrui Zhu, Yun Yi, Nannan Wang, Xiaoyu Wang, and Xinbo Gao. Semi-parametric Makeup Transfer via Semantic-aware Correspondence. *arXiv preprint arXiv:2203.02286*, 2022. 1, 3, 7, 8, 10