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

Lipstick virtual try-on (VTO) experiences have become widespread across the e-commerce sector and assist users in eliminating the guesswork of shopping online. However, such experiences still lack in both realism and accuracy. In this work, we propose LipAT, a neural framework that blends the strengths of Physics-Based Rendering (PBR) and Neural Style Transfer (NST) approaches to directly apply lipstick onto face images given lipstick attributes (e.g., colour, finish type). LipAT consists of a physics aware neural lipstick application module (LAM) to apply lipstick on face images given its attributes and Lipstick Refiner Module (LRM) to improve the realism by refining the imperfections. Unlike the NST approaches, LipAT allows precise and controllable lipstick attribute preservation, without requiring crude approximations and inference of various intertwined environment factors (e.g., scene lighting, face structure etc) involved in image generation that is required for accurate PBR. We propose an experimental framework with quantitative metrics to evaluate different desirable aspects of the lipstick attribute driven try-on alongside user studies to further validate our findings. Our results show that LipAT considerably outperforms fully-automated PBR approaches in preserving realism and the NST approaches in preserving various lipstick attributes such as finish types.

1. Introduction

Throughout history, lipstick has been a go-to makeup that instantly changes the appearance of people to make them stand out and express their unique artistic style. With the increasing size and the diversity of the global lipstick market[1], it is challenging to identify the well-suited lipstick products for a particular face without physically trying it at a cosmetic outlet. Since manually trying lipstick products is not always possible (e.g., during online purchases), artificially applying lipstick to a face image has attracted tremendous research interest in computer vision.

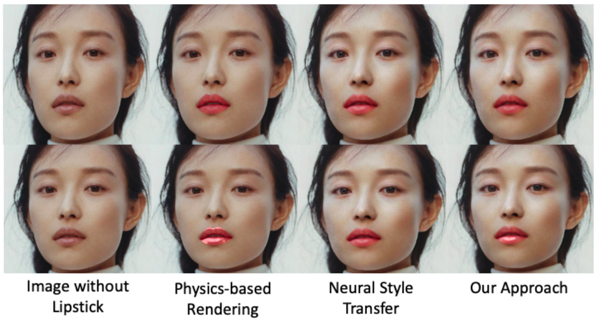

Figure 1. Results for lipstick virtual try-on with two lipstick products of the same colour (R=174, G=68, B=71) but different finish types – matte (top row) and glossy (bottom row). Each row shows the results by applying the lipstick attributes to the image in left using three techniques: PBR approach in [30]; Swatch-based variant of NST approach in [37]; and the proposed approach in this work (See Section 4.3 for more details about these techniques).

Our aim is to build a lipstick try-on framework that works in a fully automated manner, and can be used with face images captured on handheld devices, to accurately render lipstick of specified type to the desirable region while preserving its material properties (e.g. finish type, color). The system should accommodate variations in scene specific (e.g. lighting condition) and user specific (e.g. face structure, skin properties) factors while generating images, should be easy to use and scalable to be used for a large range of lipstick catalogues available on e-commerce sites.

Physics based rendering [13, 16, 17] presents itself as a viable option due to the greater flexibility in controlling/manipulating scene, user and material specific properties. However, accurate image generation with PBR [16, 17] requires full knowledge myriads of factors (e.g., scene lighting, face structure etc) that influence generation process. These factors are hard to control in the Virtual Try On (VTO) setup and difficult to recover from images taken in the wild. Thus, attempts to use PBR for try-ons rely on crude approximations of the actual image formation pro-

[1] https://www.alliedmarketresearch.com/lipstick-market

cess and are prone to render unrealistic images (see Fig. 1) due to errors in estimating the parameters influencing the rendering model via weak inference techniques [30].

Recent successes in neural network based implicit rendering and conditional image generation have led to neural style transfer (NST) as a promising approach for makeup VTO [2,7,25,31,37]. Approaches which adapt NST [9,18], focus solely on the problem of transferring the appearance of lipstick from a reference face image to a target face image. While NST's adversarial training procedure bypasses the requirement to explicitly reason about facial and scene specific rendering parameters, this approach has two major problems. First, such methods aim to learn (and preserve) lipstick appearance as *style* purely through unpaired images with and without lipstick. Due to the error in the decoupling lipstick as *style* from *content* describing rest of the image, they often end up transferring unwanted features like blemish, wrinkles and at times fake specularities while ignoring the target lighting condition during try-on (see Section 1 in Supplementary Material for supporting examples). Second, to facilitate VTO, these methods require large databases of face images consisting of at least one image of a person wearing each lipstick that we want to virtually try on – making the try-on unusable for a large portion of any e-commerce website's lipstick collection.

To address the limitations of the existing approaches (see Table 1), we propose LipAT, a novel neural framework for **Lip**stick **A**ttribute **T**ransfer. LipAT artificially applies a given lipstick product to a face image directly using lipstick attributes by blending the strengths of physics-based rendering approaches and neural lipstick transfer approaches.

The contributions of this work are to propose:

- A neural approach for virtually trying lipstick on natural face images by using lipstick material properties (e.g., colour, finish type, opacity) as the sole input. All neural modules in our method, except the discriminator in Section 3.2, are trained with a collection of real face images without lipstick and a set of lipstick attributes – which are readily available.
- A thorough experimental framework for evaluating different aspects of the rendered images with lipstick to address the lack of consistent evaluation scheme in previous works. Our framework includes a novel variant of FID for measuring the realism, called Patch-FID, which is empirically shown to be aligning with human perception of realism.

We verify the superiority of the proposed approach quantitatively and qualitatively in preserving realism and lipstick attributes over both PBR and NST approaches.

2. Related Work

In this section, we review recent approaches on lipstick simulation, which can be categorised into two: (1) PBR-based approaches; and (2) NST-based approaches.

	PBR	NST	LipAT
Robustness to wild images		✓	✓
No manual intervention		✓	✓
Lipstick attribute controllability	✓		✓
Does not require reference face image	✓		✓

Table 1. Comparison between PBR-based approaches, NST-based approaches and LipAT (our approach).

Physics-based Rendering Approaches. These techniques address the lipstick simulation problem from the computer graphics perspective by incorporating the knowledge from physics on the interaction of light with different cosmetic attributes and facial skin. Thus, these approaches typically involve many parameters to simulate the appearance of various optical properties (e.g., roughness score, reflection intensity and light intensity). The studies in [16,17] propose such a framework to quantify how various intrinsic layers (i.e., albedo, diffuse shading and specular highlights) are altered by applying a lipstick product using the physics-based models such as Kubelka-Munk [14] and Torrance-Sparrow [33] models. Despite the strong theoretical motivation behind this approach, it comes with the drawback of requiring a set of image pairs of the subjects with and without lipstick in order to tune the hyper-parameters. In [30], the face images are decomposed into different colour ranges – i.e., shadows; mid tones; and highlights. The appearance of lipstick attributes are simulated in each layer separately using various image filters (e.g., shadow/highlight filter, piecewise-linear intensity transformation filter). The works in [13,20] adopt 3D meshes of the lips predicted using facial key-points. The 3D meshes are used to incorporate texture of the lipstick (e.g., glossiness) with the help of the estimations of the environment reflections. Although these approaches yield high quality images, they involve parameters that require careful tuning for each image separately. Thus, it is challenging to fully automate these solutions. Also, these approaches heavily rely on off-the-shelf solutions for producing 3D meshes and for estimating environment lighting, thus, yield unrealistic results for some face images. In contrast, our framework draws insights from previous PBR approaches and relaxes their strong inductive biases in some of the operations in a data-driven manner.

Neural Makeup Transfer Approaches. These approaches aim to transfer the lipstick style of a reference face image to a different target image with the help of deep generative models. BeautyGlow [3] adopts Glow, deep generative model for transferring styles, to transfer makeup styles between images. BeautyGAN [19] addressed the makeup transfer and removal task by introducing a symmetric image to image architecture with a global cycle consistence loss. This work introduced local instance-level makeup loss terms to preserve the style of the reference lipstick, which is

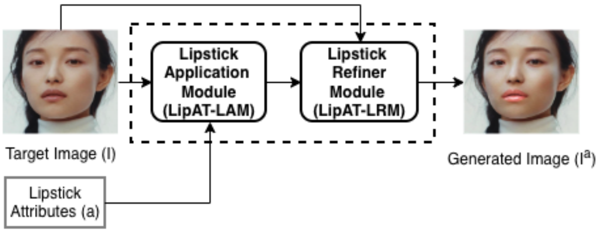
Target Image (I)
Lipstick Application Module (LipAT-LAM)
Lipstick Refiner Module (LipAT-LRM)
Generated Image (I^a)
Lipstick Attributes (a)

Figure 2. Overview of LipAT, which consists of physics-motivated lipstick application module and lipstick refiner module to simulate realistic lipstick application using lipstick attributes.

consistently used in most of the subsequent works. Earlier approaches are unable to simulate dramatic makeup styles. CPM [25] and LADN [5] were introduced to address such dramatic makeup styles. CPM adopts UV maps to align facial features of source and target faces. LADN proposes a pseudo labelling approach to provide distant supervision for transferring dramatic makeup styles with high-frequency details. Nevertheless, these approaches do not explicitly exploit correspondence at the semantic component level (e.g., eyes, lips), thus, yield inconsistent results for different poses and expressions. To address this limitation, most recent works [8, 31, 37] adopt segmentation masks or facial keypoints to identify corresponding regions between the reference and the target images, which are explicitly exploited when transferring makeup. Although these approaches can be easily extendable for other makeup products such as foundation and eye shadows, it has been found that these approaches typically transfer unwanted details of the reference image to the target image such as shadows, wrinkles and blemishes. In SOGAN [21], this issue was explored up to some extent by proposing a shadow and occlusion robust makeup transfer approach. Although this approach handles shadows and occlusion, the generated images from this approach are not realistic as the previous makeup transfer approaches. Overall, these neural approaches require the reference lipstick style on a face image – i.e., unable to transfer lipstick using their attributes directly. To address this research gap, the work in [12] proposed a technique to virtually apply makeups using their colour attribute. However, this approach does not focus on preserving other lipstick attributes such as finish types. Our framework is particularly motivated by this research gap, which proposes a neural framework to simulate face images with lipstick using lipstick attributes including complex attributes such as finish types.

3. Proposed Framework

This work proposes the first neural approach for applying a lipstick on a given face image using the attributes of the lipstick. We denote the output image of our approach as I^a by applying a lipstick product a to a target face image without lipstick I. As shown in Fig. 2, our framework consists of two modules. The first module is Lipstick Application Module (LipAT-LAM), a neural architecture to apply lipstick to a face image in a physics-aware manner. We denote the output of LipAT-LAM using a face-image without lipstick I and a lipstick a as $\hat{I^a}$. The architecture of LipAT-LAM is explicitly designed to exploit the knowledge motivated by geometry and reflectance models, which is particularly important to preserve lipstick attributes such as finish types. Unlike the existing PBR approaches, once trained, this module transfers lipstick attributes without requiring any manual tuning of parameters. However, this module heavily relies on the off-the-shelf pretrained neural blocks for face parsing [11] and specularity extraction [30]. This could make $\hat{I^a}$ unrealistic for some images. To address this limitation, LipAT-LAM is followed by Lipstick Refiner Module (LipAT-LRM). LipAT-LRM produces the final image I^a using the output from LipAT-LAM $\hat{I^a}$ and I as input. This module is learned to refine the potentially unrealistic outputs from LipAT-LAM due to the imperfection in off-the-shelf blocks in LipAT-LAM.

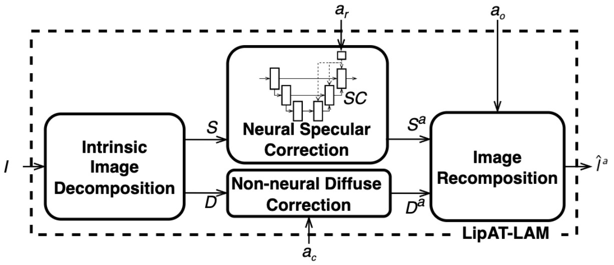
a_r
a_o
I
Intrinsic Image Decomposition
S
D
Neural Specular Correction
SC
Non-neural Diffuse Correction
S^a
D^a
Image Recomposition
$\hat{I}^a$
LipAT-LAM
a_c

Figure 3. Overview of LipAT-LAM, which first decomposes the given image I into its intrinsic components – diffuse D and specular S components, then updates S and D using the given roughness score a_r and base colour a_c of the lipstick, respectively. LipAT-LAM recomposes the output image $\hat{I^a}$ using the updated intrinsic components (S^a and D^a) and the lipstick opacity a_o.

3.1. Lipstick Application Module (LipAT-LAM)

This module is learned to apply lipstick to a target face image using lipstick attributes. This work characterises lipstick products using 3 attributes, namely: (1) a_c – RGB values of the base colour; (2) $a_r \in [0, 1]$ – gloss roughness, which controls the finish type of a lipstick product (e.g., $a_r = 0.7$ for a matte lipstick); and (3) $a_o \in [0, 1]$ – makeup opacity, which controls how strongly the lipstick is applied.

LipAT-LAM aims to keep the other parameters in conventional rendering engines (e.g., scene lighting parameters) consistent between I and $\hat{I^a}$, while simulating the selected parameters only on the lip area of I. To only update the lip of the face image I, LipAT-LAM guides the lipstick application process using a segmentation mask M^I of the lip region in I. This mask is produced using the off-the-

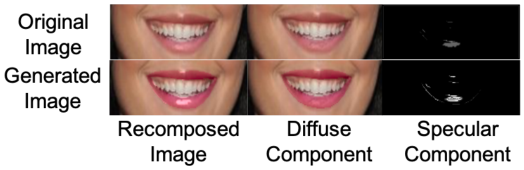

Figure 4. Intrinsic decomposition of LipAT-LAM. Top row shows an face image without lipstick I and the bottom row shows the same lip after applying a glossy lipstick using LipAT-LAM.

shelf neural face parser [11]. With the help of M^I, LipAT-LAM simulates lipstick attributes as described below.

Following the dichromatic reflection model in [28], LipAT-LAM first divides the target image I into two components: diffuse (D) and specular (S) components, which can be added together to reconstruct $I = D + S$. Here, S represents the highlights of I and the body reflection D represents I with highlights removed. LipAT-LAM updates the lip region of each component using the given attributes and combines them to reconstruct the target image with lipstick $\hat{I}^a$. Following [30], we adopt the shadow/highlight image filter in [30] to decompose the images as shown in Fig. 4.

3.1.1 Diffuse Component Update

Following [27, 30], we update the lip region of the diffuse component D using the base colour of the given lipstick a_c in the perceptually uniform LAB colour space (see Section 10 in Supplementary Material for more details about LAB):

$$\tilde{D}^a = \text{LAB}(D) + \text{LAB}(a_c) - \sum M^I \odot \text{LAB}(D) / \sum M^I$$
$$\hat{D}^a = M^I \odot \text{LAB}^{-1}(\tilde{D}^a) + (1 - M^I) \odot D \tag{1}$$

where $\text{LAB}(\cdot)$ and $\text{LAB}^{-1}(\cdot)$ are the colour conversion functions for RGB-LAB and LAB-RGB respectively and $\odot$ denotes elementwise multiplication.

3.1.2 Specular Highlight Update

To preserve the finish type of lipsticks – characterised by roughness, the specular highlights should be updated accordingly. Most previous works [13, 30] adopt either gamma correction on specular highlights or a PBR engine. The gamma correction-based techniques [30] can produce highlights aligning with the specular highlights of the original images, but requires manually tuning of γ for each image to produce realistic results. In contrast, rendering-based techniques [13] need to estimate the lighting profile of the original image to produce realistic outputs. To bridge these two research gaps, LipAT-LAM learns a conditional deep generative model $SC : (S, a_r) \rightarrow S^a$ parameterized using θ_{SC}, that accepts the specular highlight component of the face image without lipstick S and the roughness score of the lipstick a_r and returns the updated specular highlight component S^a as the output. We adopt the conditional U-Net architecture (see Section 2 in Supplementary Material) proposed in [24] as SC as shown in Figure 3.

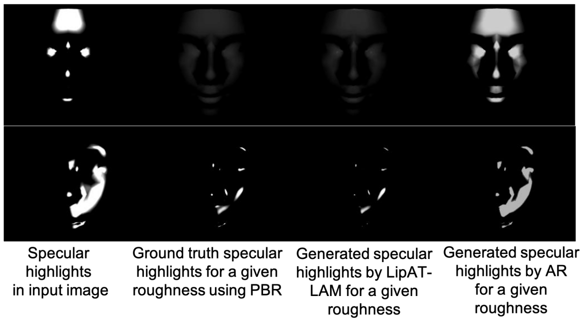
Specular highlights in input image
Ground truth specular highlights for a given roughness using PBR
Generated specular highlights by LipAT-LAM for a given roughness
Generated specular highlights by AR for a given roughness

Figure 5. Updated specular highlights for a given roughness score value (0.7 for top row and 0.1 for bottom row) and the specular highlights in the skin (left image in each row) using three different methods: using a PBR engine; LipAT-LAM; and AR [30].

To learn SC such that it updates specular highlights for a given roughness score without explicitly predicting scene lighting as in PBR, we adopt two loss functions: (1) PBR-based label reconstruction loss that aims to imitate the specular updates by a PBR engine in a synthetic rendering environment; and (2) gamma correction-based label reconstruction loss that aims to learn the gamma correction-based weak highlight updating operation [30] using real images.

$$L_{\text{LAM}} = L_{\text{pbr\_recon}} + L_{\gamma\text{\_recon}} \tag{2}$$

PBR-based label reconstruction loss. This loss term focuses on correcting specular highlights realistically. We adopt a synthetic dataset $\mathbb{D}^{\text{train}}_{\text{synthetic}}$ constructed using PyVista[2] physics-based rendering engine (see Section 4.1). $\mathbb{D}^{\text{train}}_{\text{synthetic}}$ consists of tuples $< S_{\text{pbr}}, a_r, S^a_{\text{pbr}} >$ where S_{pbr} is the specular highlight component of a face image without lipstick (assuming skin roughness as 0.3 [16]) with an arbitrary lighting profile; a_r is a roughness score; and S^a_{pbr} is the specular highlight component after updating S_{pbr} according to a_r. $\mathbb{D}^{\text{train}}_{\text{synthetic}}$ reflects accurate physics-aware specular updates as shown in the second column in Fig. 5. We formulate our PBR-based label reconstruction loss to force SC to update highlights in a physics-aware manner using the instances in $\mathbb{D}^{\text{train}}_{\text{synthetic}}$ as follows:

$$L_{\text{pbr\_recon}} = \sum\nolimits_{< S_{\text{pbr}}, a_r, S^a_{\text{pbr}} > \in \mathbb{D}^{\text{train}}_{\text{synthetic}}} ||S^a_{\text{pbr}} - SC(S_{\text{pbr}}, a_r)|| \tag{3}$$

Gamma correction-based label reconstruction loss. Since $\mathbb{D}^{\text{train}}_{\text{synthetic}}$ is generated under synthetic scene lighting, the images in $\mathbb{D}^{\text{train}}_{\text{synthetic}}$ may not reflect the lighting profile in real-world face images. We empirically observed that training SC only using $L_{\text{pbr\_recon}}$ reduced generalization to real-world face images. Thus, we introduce $L_{\gamma\text{\_recon}}$ based

[2] https://docs.pyvista.org/index.html

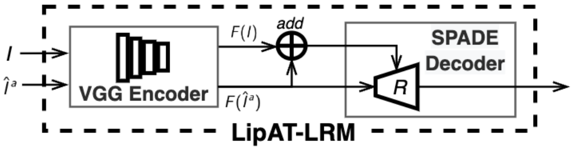

Figure 6. Overview of LipAT-LRM, which adopts SPADE-based neural architecture to refine the output from LipAT-LAM $\hat{I}^a$ with the help of the image features of I and $\hat{I}^a$ that are generated from pre-trained VGG-19 encoder.

on the specular components extracted from real face images in $\mathbb{D}^{\text{train}}_{\text{real}}$, which consists of tuples $< I, I_{\text{wl},a} >$ including unpaired face images without lipstick I, with lipstick I_{wl} and an attribute vector a. Following [30], we produce weak ground truth (denoted as S^a_γ) for S^a of I in $\mathbb{D}^{\text{train}}_{\text{real}}$ for a given a^r using gamma correction as follows:

$$S^a_\gamma = 1 - (1-S)^\gamma; \text{ where } \gamma = x \exp\left(-y \cdot a_r\right) \quad (4)$$

After regressing using the images in $\mathbb{D}^{\text{train}}_{\text{synthetic}}$, we set $x = 2.42$ and $y = 1.55$. Ideally, x and y should be manually tuned for each image individually as it depends on the lighting profile of each image, thus, S^a_γ could be unrealistic for some instances with fixed x and y values. However, the strength of S^a_γ is its ability to preserve the lighting profile of S^a in real images without explicitly estimating it. With these pseudo labels from gamma correction, we formulate our gamma-based label reconstruction loss as follows:

$$L_{\gamma\_\text{recon}} = \textstyle\sum_{<I,a>\in\mathbb{D}^{\text{train}}_{\text{real}}} ||S^a_\gamma - SC(S, a_r)|| \quad (5)$$

We learn θ_{SC} such that it minimizes $L_{\text{pbr\_recon}}$ for the images in $\mathbb{D}^{\text{train}}_{\text{synthetic}}$ and L_{γ_recon} for the images without lipstick in $\mathbb{D}^{\text{train}}_{\text{real}}$. As shown in Figure 5, the proposed neural specular highlight correction approach can realistically incorporate roughness score without manual intervention.

To construct the output of LipAT-LAM, D^a and S^a should be ideally blended using add blending ($D^a + S^a$). Following [30], we adopt screen blending instead, which is an operation similar to add blending, but softer.

$$\tilde{I}^a = 1 - (1 - D^a) \odot (1 - S^a) \quad (6)$$

Subsequently, $\hat{I}^a$ is produced by alpha blending the lip region of I and $\tilde{I}^a$ using opacity a_o as the blending strength:

$$\hat{I}^a = M^I \odot (a_o \cdot \tilde{I}^a + (1 - a_o) \cdot I) + (1 - M^I) \odot I \quad (7)$$

3.2. Lipstick Refiner Module (LipAT-LRM)

The output from LipAT-LAM $\hat{I}^a$ could be unrealistic for some examples due to two main reasons: (1) incorrect placement of lipstick due to the imperfections in binary lip masks M^I, particularly around the edge of the lip; and (2) unrealistic specularities in $\hat{I}^a$ that are not well-aligned with I. Since the differences of the multi-scale image features between I and $\hat{I}^a$ (e.g., edges) could be useful to identify the aforementioned imperfections, LipAT-LRM refines $\hat{I}^a$ with the help of the knowledge available in I (see Fig. 6).

LipAT-LRM first adopts pre-trained VGG-19 [29] to encode I and $\hat{I}^a$ into multi-scale feature representations. We denote the l^{th} level features from the pre-trained network for I and $\hat{I}^a$ as $F^l(I)$ and $F^l(\hat{I}^a)$ respectively, with $l \in [1, 2, 3, 4]$. Although LipAT-LRM may use any pre-trained image encoders [4, 32] capable of producing multi-resolution feature maps, we adopt VGG-19 due to its proven [25, 37] ability to provide informative features and to facilitate a fair comparison to other works [25, 37].

To refine $\hat{I}^a$ using $F^l(\hat{I}^a)$ while conditioning on $F^l(I)$, we adopt a neural architecture for refining images $R : (F(I), F(\hat{I}^a)) \rightarrow I^a$, which consists of multiple SPatially-Adaptive DE-normalization (SPADE) [26] blocks in cascade and parameterised using θ_R. Unlike other image generative neural blocks [24, 35], SPADE can effectively control pixel-level and semantic-level refinements via spatially adaptive normalization, which makes it ideal for region-specific image augmentation tasks as ours (see Section 3 in Supplementary Material for more architectural details).

To learn θ_R, LipAT-LRM adopts three loss functions: (1) cosmetic loss; (2) refine loss; and (3) adverserial loss.

$$L_{\text{LRM}} = \lambda_{\text{cos}} * L_{\text{cos}} + \lambda_{\text{ref}} * L_{\text{ref}} + \lambda_{\text{adv}} * L_{\text{adv}} \quad (8)$$

where $\lambda_{\text{cos}} (= 10)$, $\lambda_{\text{ref}} (= 10)$ and $\lambda_{\text{adv}} (= 1)$ control the weighting for each loss term, which were set by tuning.

Cosmetic Loss. This loss is proposed to preserve the simulated appearance of lipstick by LipAT-LAM, which was defined as L1 distance between the pixel values in the lip region of $\hat{I}^a$ and I^a:

$$L_{\text{cos}} = \textstyle\sum_{<I,a>\in\mathbb{D}^{\text{train}}_{\text{real}}} ||M^I \odot I^a - M^I \odot \hat{I}^a||_1 \quad (9)$$

Refine Loss. This loss is proposed to refine the imperfection of $\hat{I}^a$ using the knowledge in I. This is a edge-preserving perceptual loss function, which combines the mean-squared loss in the pre-trained feature space [37] and mean gradient error [10] in the pixel space between I and I^a to filter out unaligned edges and specularities.

$$L_{\text{ref}} = \textstyle\sum_{<I,a>\in\mathbb{D}^{\text{train}}_{\text{real}}} ||F^l(I^a) - F^l(I)||_1 + ||G(I^a) - G(I)||_2 \quad (10)$$

where G returns pixel-level gradients [10] for a given image using a Sobel operator.

Adversarial Loss. This loss function is proposed to generate natural looking images with high perceptual quality. L_{adv} adopts the least-square adversarial loss proposed in [22] by utilizing discriminator H parameterised with H_θ to classify real and fake images as follows:

$$L_{\text{adv}} = \textstyle\sum_{<I,I_{\text{wl}},a>\in\mathbb{D}^{\text{train}}_{\text{real}}} [H(I_{\text{wl}})^2] + [1 - H(I^a)^2] \quad (11)$$

We learn θ_R by minimizing L_{LRM} using the images with lipstick, without lipstick and attribute vectors in $\mathbb{D}^{\text{train}}_{\text{real}}$. Here, I_{wl} denotes the images with lipstick in $\mathbb{D}^{\text{train}}_{\text{real}}$, which are only used to optimize L_{adv}.

4. Experimental Setup

4.1. Datasets

Please refer to Section 4 in Supplementary Material for details about the construction of the following datasets.

Training Datasets. The training of LipAT involves two datasets: (1) $\mathbb{D}^{\text{train}}_{\text{real}}$ - This dataset consists of 10,000 unpaired face images without lipstick and with lipstick from CelebA-HQ [15], and 10,000 lipstick attribute vectors. The lipstick attribute vectors are sampled alternatively using a probability distribution fitted to the space of the lipstick products using public data and a uniform distribution. The images without lipstick and attribute vectors are used to learn the neural components in LipAT, and the images with lipstick are only used to optimize L_{adv}; and (2) $\mathbb{D}^{\text{train}}_{\text{synthetic}}$ - This dataset consists of 10,000 specular components that are rendered using 80 different face images under 25 different simulated scene lighting profiles and 5 different roughness scores with the help of PyVista PBR engine.

Test Datasets. The testing of LipAT involves two datasets: (1) $\mathbb{D}^{\text{test}}_{\text{up}}$ - Similar to $\mathbb{D}^{\text{train}}_{\text{real}}$, this dataset consists of 2048 unpaired face images with lipstick and without lipstick from CelebA-HQ and 2048 lipstick attribute vectors. There are no overlapping images between $\mathbb{D}^{\text{test}}_{\text{up}}$ and $\mathbb{D}^{\text{train}}_{\text{real}}$; and (2) $\mathbb{D}^{\text{test}}_{\text{wp}}$ - This dataset consists of 127 weakly paired images with and without lipstick and the corresponding lipstick attribute vectors that are inferred with the help of pre-trained neural models for makeup removal [31] and material attribute extraction [1, 23].

4.2. Evaluation Metrics

We quantitatively evaluate two aspects: (1) preservation of the realism; and (1) accuracy of the attribute preservation.

Patch-FID, a novel metric to evaluate realism. To evaluate the preservation of realism of the generated images, most previous works adopt Fréchet Inception Distance (FID) [6]. 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 [34] between the two Gaussian distributions $\mathcal{N}_{\text{real}}$ and $\mathcal{N}_{\text{gen}}$ estimated using the images in $\mathbb{D}_{\text{real}}$ and $\mathbb{D}_{\text{gen}}$ from an intermediate layer of the pre-trained Inception model [32]. If the selected intermediate layer outputs C number of channels, $\mathcal{N}_{\text{real}}$ and $\mathcal{N}_{\text{gen}}$ are modelled as C-variate Gaussian distribution after performing global average pooling to convert intermediate feature maps $\in \mathbb{R}^{C\times H\times W}$ of images to c-dimensional vectors. Here, H and W are the height and the width of a feature map. Due to this global average pooling operation, FID treats the features corresponding to all the pixels in face images equally. Thus, FID score becomes insensitive to the update in the targeted region when only a smaller region of real images are updated to generate $\mathbb{D}_{\text{gen}}$ – e.g., lip region of a full face image. We quantitatively verify this statement in Section 5 of Supplementary Material using an example.

To address this limitation, we propose Patch-FID, which estimates $\mathcal{N}_{\text{real}}$ and $\mathcal{N}_{\text{gen}}$ only using the activations from the Inception model corresponding to the updated regions in the images. To identify the intermediate features corresponding to lip region of an image, we masked the pixels of the image using black, except the pixels within a bounding box drawn around the lip region. Then, we identified the locations of the intermediate feature map of the masked image that have different values compared to the feature map of a completely black image. When estimating $\mathcal{N}_{\text{real}}$ and $\mathcal{N}_{\text{gen}}$, we only average values of the identified locations in each channel to represent the image as a C-dimensional vector, instead of performing global pooling. Our experiments show that Patch-FID is more suitable to evaluate the realism of our task and agreeing with human perception of realism. To compute the proposed Patch-FID we treat images with lipstick from $\mathbb{D}^{\text{test}}_{\text{up}}$ as $\mathbb{D}_{\text{real}}$ and applying lipstick to images without lipstick from $\mathbb{D}^{\text{test}}_{\text{up}}$ as $\mathbb{D}_{\text{gen}}$. We present formal derivation and more experiments with Patch-FID in Section 5 of Supplementary Material.

Other quantitative metrics. To quantitatively evaluate the second aspect – accuracy of the lipstick simulation, we adopt paired $\mathbb{D}^{\text{test}}_{\text{wp}}$ constructed in Section 4.1 and numerically evaluate how accurately we can reconstruct the images with lipstick in $\mathbb{D}^{\text{test}}_{\text{wp}}$ by applying lipstick to the corresponding images without lipstick. For this, we adopt Structural Similarity Index measure (SSIM) [36] and L1 distance. Since our work only focuses on lipstick simulation, only the pixels within the bounding box around the lip region are used to compute SSIM and L1 metrics.

4.3. Baselines

In this work, we compare the proposed approach with 8 baselines that are categorised as: (1) PBR-based approaches; (2) NST-based approaches; and (3) Hybrid approaches[3]. Under PBR-based approaches, we adopt three baselines: Colour-Transfer [27]; AR [30]; and LAM, which only consists of our lipstick application module.

Since almost all the existing NST-based approaches are unable to transfer lipstick using lipstick attributes, we modified SpMT [37], 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: (1) Swatch-SpMT, which transfer lip-

[3] See Section 6 in Supplementary Material for detailed descriptions about the baselines.

Method Type	Method	SSIM (↑)	L1 (↓)	Patch-FID (↓)	FID (↓)
PBR	Colour-Transfer [27]	0.7678	0.0328	24.3	**50.6**
	AR [30]	0.798	0.0311	25.4	50.7
	LAM	**0.799**	0.0310	23.2	50.8
NST	Swatch-SpMT	0.7592	0.0330	21.7	51.8
	Att-SpMT	0.6983	0.0371	24.8	52.3
Hybrid	LAM + CPM [25]	0.7437	0.0341	22.6	52.1
	LAM + SSAT [31]	0.7528	0.0336	22.1	51.6
	LAM + SpMT [37]	0.7692	0.0329	21.7	51.3
	Our Approach	0.797	**0.0309**	**20.2**	51.2

Table 2. Results for the lipstick attribute transfer task using different methods, categorised as Physics-Based Rendering (PBR); Neural Style Transfer (NST); and Hybrid (Hybrid).

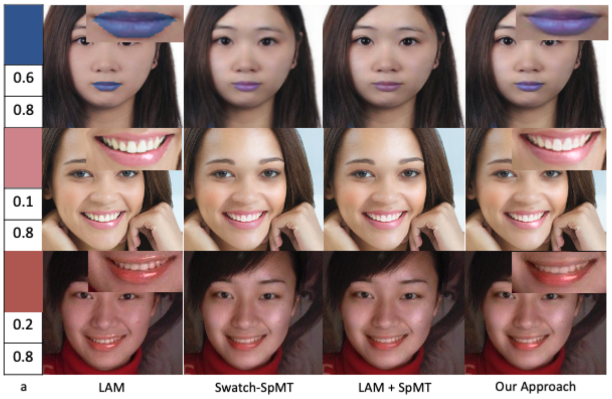

Figure 7. Generated images from different methods using the attributes vectors a in left of each row – each a vector gives the base color, roughness score (0.1 for glossy and 0.7 for matte), and opacity from top to bottom.

sticks via a swatch image; and (2) Att-SpMT, which transfer lipstick directly using lipstick attributes.

As the hybrid approaches, we combine LAM with three neural lipstick transfer approaches: (1) CPM [25]; (2) SSAT [31]; and (3) SpMT [37]. For each baseline, 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 neural lipstick transfer approach.

5. Results

5.1. Quantitative Evaluation

We quantitatively evaluate the lipstick application accuracy of different methods using SSIM and L1 metrics with D_{wp}^{test}. As shown in Table 2, LipAT outperforms all the compared baselines. The inferior performance of neural and hybrid approaches could be primarily due to the inability of their neural components to preserve finish types (see Figure 7). LipAT addresses these problems using its physics-aware LipAT-LAM module. This statement can be further verified as L1 (0.89 coefficient of determination with User Study 2) and SSIM (0.94 coefficient of determination with User Study 2) metrics report agreeing results with our user study on finish type preservation. In addition, we observed that most existing neural lipstick transfer modules are not generalising well to unseen lipstick attributes such as bluish colors as shown in the first row in Figure 7. Being able to train LipAT using lipstick attributes directly makes LipAT generalize well for such rarely seen and extreme cases.

Patch-FID vs FID. Table 2 also reports the evaluation with respect to the proposed Patch-FID and the conventional FID metrics. These metrics measure the realism of the generated images. However, we observed that Patch-FID yields highly correlated results with our user study on realism – i.e., the coefficient of determination between Patch-FID and User Study 1 is 0.99, while the same measure between FID and User Study 1 is 0.49 (see Section 5.2), indicating the ability of Patch-FID to effectively capture human perception of realism.

With the supporting results for patch-FID being a better measure of realism on this task, we can observe that PBR approaches are unable produce realistic images. In contrast, the approaches with deep generative models can preserve realism. Out of the neural and hybrid approaches, LipAT still outperforms other approaches by as much as 7.4% in Patch-FID. The high diversity of the generated images from LipAT due to finish type preservation could be a contributing factor for this performance gap. Overall, LipAT outperforms neural approaches in preserving the attributes such as finish type, and PBR approaches in preserving realism.

Ablation study. In Table 2, we compare LipAT with a weaker variant called as LAM, which only includes the LipAT-LAM module. With respect to SSIM and L1, LAM yields comparable results to the full model. This means that LipAT-LAM module largely contributes towards the accurate preservation of attribute in the final results from LipAT-LAM. Nevertheless, LAM yields unrealistic results for some examples as shown in the zoomed in images in Figure 7. We can quantitatively verify this as LipAT largely outperforms LAM by 14.9% in Patch-FID. These results verify the positive contribution of both modules in LipAT.

Controllability in LipAT. Almost all the existing neural lipstick transfer approaches do not allow controllability across lipstick attributes. In contrast, our approach allows controllable lipstick application as LipAT applies lipstick directly using lipstick attributes. Figure 8 shows how the results of LipAT vary when changing different lipstick attributes. As can be seen, LipAT effectively incorporates the attributes such as finish type, opacity and base color. To the best of our knowledge, this is the first neural approach that allows such controllable lipstick simulation.

LipAT's Failure Cases. There are two main limitations of LipAT: (1) LipAT's inability to apply lipstick when the neural lip parser used in LipAT fails to detect a lip region in a given face image, despite its ability to rectify misalign-

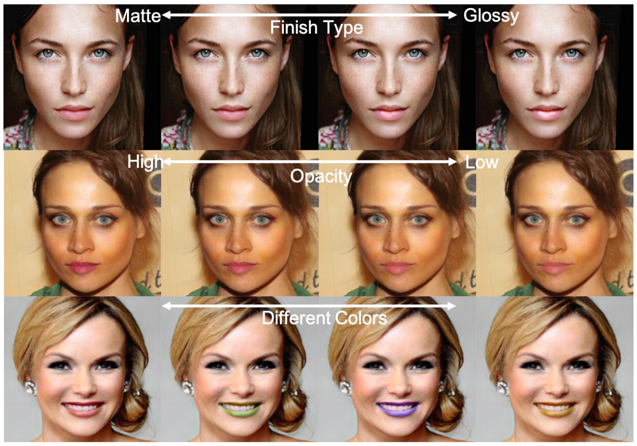

Figure 8. Controllability of LipAT with respect to different lipstick attributes such as finish type, opacity and base color.

Methods	User Study 1	User Study 2	User Study 3
Real Images	22.4%	62.3%	N/A
AR [30]	17.4%	**59.4%**	18.9%
Swatch-SpMT	19.7%	47.2%	22.8%
LAM + SpMT [37]	19.9%	48.4%	25.6%
Our Approach	**20.6%**	55.7%	**32.7%**

Table 3. Aggregated user study results – user studies 1 (preservation of realism) and 3 (accuracy of lipstick appearance) report the percentage of each method to be voted over the other methods; user study 2 (preservation of finish type) reports the accuracy of identifying the correct finish type of the images.

ments caused by the neural lip parser; and (2) LipAT's inability to accurately apply lipstick to face images that already have lipstick applied. Due to space limitation, we discuss these limitations in detail along with qualitative results in Section 10 of the supplementary material.

5.2. Qualitative Evaluation

For qualitative evaluation, we conducted 3 users studies (see Section 7 in supplementary material for more details) focusing on three aspects of the generated images.

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. We then asked participants to select the image with the most realistic lipstick application. As shown in Table 3, the neural-based approaches including our method yield the best results out of the artificial methods, largely outperforming the PBR-based baseline. This observation verifies the strength of deep generative models-based lipstick application techniques in preserving realism.

User study 2 - preservation of finish type. This study evaluates how accurately each method can incorporate finish types to 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 has same colour but different finish types (i.e., glossy and matte). The participants have been asked to select the image that simulates the appearance of a glossy lipstick. Our results (see Table 3) from this experiments show that the images from the existing neural-based solutions (e.g., Swatch-SpMT and LAM+SpMT) cannot incorporate finish types accurately as shown in Fig. 1. In contrast, LipAT yield superior results, verifying the potential of LipAT for incorporating finish types.

User study 3 - overall accuracy of the 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. We asked participants to select the generated image that gives the most accurate application of the lipstick in the reference image. Table 3 shows that our approach outperforms other baseline in this experiment. Also, we observed the results of this study well aligns with the combination of the results from Patch-FID and SSIM ($R^2 = 0.89$) instead of the results from each metric alone (see Fig. 22 in the supplementary material). Since Patch-FID and SSIM focus on the preservation of realism and material properties respectively, this observation further verifies the comprehensiveness of our quantitative evaluation framework and the agreement between the selected metrics and human perception.

6. Conclusion

We propose LipAT, a virtual lipstick try-on framework that takes a face image with lipstick attribute to try on as input. LipAT consists of two modules: LipAT-LAM, a physics-motivated neural module to apply lipstick attributes for a given face image using the attributes; LipAT-LRM - a neural image refining module to improve the realism of outputs from LipAT-LAM. Our framework requires no additional information such as a style image with a model face wearing the desired lipstick, facilitating a high level of scalability in addition to granular attribute controllability across complex lipstick attributes such as finish types. Our experiments show that LipAT yields visually realistic results compared to the state-of-the-art baselines while allowing the controllability across the complex lipstick attributes.

For future work, we intend to extend our approach to other makeup products such as foundation, eyebrow pencils etc. Since other makeup products have attributes different to lipstick (e.g., thickness in eyebrow pencils) and different effects on the intrinsic layers of faces from the PBR perspective, scaling LipAT to other makeup products requires additional research effort to address such challenges.

References

[1] Manuel Lagunas Arto, Sandra Malpica, Ana Serrano, Elena Garces, Diego Gutierrez, and Belen Masia. A similarity measure for material appearance. *Jornada de Jóvenes Investigadores del I3A*, 7, 2019. 6

[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. 2

[3] Hung-Jen Chen, Ka-Ming Hui, Szu-Yu Wang, Li-Wu Tsao, Hong-Han Shuai, and Wen-Huang Cheng. BeautyGlow: On-demand makeup transfer framework with reversible generative network. In *Proc. of CVPR*, pages 10042–10050, 2019. 2

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 5

[5] Qiao Gu, Guanzhi Wang, Mang Tik Chiu, Yu-Wing Tai, and Chi-Keung Tang. LADN: Local adversarial disentangling network for facial makeup and de-makeup. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10481–10490, 2019. 3

[6] 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. 6

[7] 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. 2

[8] Wentao Jiang, Si Liu, Chen Gao, Jie Cao, Ran He, Jiashi Feng, and Shuicheng Yan. PSGAN: Pose and Expression Robust Spatial-aware GAN for Customizable Makeup Transfer. In *Proc. of CVPR*, pages 5194–5202, 2020. 3

[9] 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. 2

[10] Nick Kanopoulos, Nagesh Vasanthavada, and Robert L Baker. Design of an image edge detection filter using the sobel operator. *IEEE Journal of solid-state circuits*, 23(2):358–367, 1988. 5

[11] 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. 3, 4

[12] Robin Kips, Pietro Gori, Matthieu Perrot, and Isabelle Bloch. Ca-gan: Weakly supervised color aware gan for controllable makeup transfer. In *Proc. of ECCV*, pages 280–296, 2020. 3

[13] 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, 2, 4

[14] Paul Kubelka. Ein beitrag zur optik der farbanstriche (contribution to the optic of paint). *Zeitschrift fur technische Physik*, 12:593–601, 1931. 2

[15] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. MaskGAN: Towards Diverse and Interactive Facial Image Manipulation. In *Proc. of CVPR*, 2020. 6

[16] 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, 2, 4

[17] Chen Li, Kun Zhou, Hsiang-Tao Wu, and Stephen Lin. Physically-based simulation of cosmetics via intrinsic image decomposition with facial priors. *IEEE transactions on pattern analysis and machine intelligence*, 41(6):1455–1469, 2018. 1, 2

[18] 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. 2

[19] 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

[20] TianXing Li, Zhi Yu, Edmund Phung, Brendan Duke, Irina Kezele, and Parham Aarabi. Lightweight Real-time Makeup Try-on in Mobile Browsers with Tiny CNN Models for Facial Tracking. *arXiv preprint arXiv:1906.02260*, 2019. 2

[21] Yueming Lyu, Jing Dong, Bo Peng, Wei Wang, and Tieniu Tan. Sogan: 3d-aware shadow and occlusion robust gan for makeup transfer. In *Proc. of ACM MM*, pages 3601–3609, 2021. 3

[22] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *Proc. of ICCV*, pages 2794–2802, 2017. 5

[23] MathWorks. Color-based segmentation using k-means clustering. 6

[24] 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. 4, 5

[25] 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. 2, 3, 5, 7

[26] 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. 5

[27] Erik Reinhard, Michael Adhikhmin, Bruce Gooch, and Peter Shirley. Color transfer between images. *IEEE Computer graphics and applications*, 21(5):34–41, 2001. 4, 6, 7

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

[29] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *Proc. of ICLR*, 2015. 5

[30] 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, 2, 3, 4, 5, 6, 7, 8

[31] 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. 2, 3, 6, 7

[32] 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. 5, 6

[33] Kenneth E Torrance and Ephraim M Sparrow. Theory for off-specular reflection from roughened surfaces. *Josa*, 57(9):1105–1114, 1967. 2

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

[35] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional GANs. In *Proc. of CVPR*, pages 8798–8807, 2018. 5

[36] Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multiscale Structural Similarity for Image Quality Assessment. In *Proc. of ACSSC*, pages 1398–1402, 2003. 6

[37] 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, 2, 3, 5, 6, 7, 8