Toward Tiny and High-quality Facial Makeup with Data Amplify Learning

Qiaoqiao Jin[1]*, Xuanhong Chen[1,3], Meiguang Jin[2], Ying Chen[2], Rui Shi[1],
Yucheng Zheng[1], Yupeng Zhu[1], and Bingbing Ni[1]**

[1]Shanghai Jiao Tong University, Shanghai 200240, China
[2]Alibaba Group [3]Institute of Cultural and Creative Industry, USC-SJTU
{jinqiaoqiao, nibingbing}@sjtu.edu.cn

Abstract. Contemporary makeup approaches primarily hinge on unpaired learning paradigms, yet they grapple with the challenges of inaccurate supervision (e.g., face misalignment) and sophisticated facial prompts (including face parsing, and landmark detection). These challenges prohibit low-cost deployment of facial makeup models, especially on mobile devices. To solve above problems, we propose a brand-new learning paradigm, termed "Data Amplify Learning (DAL)," alongside a compact makeup model named "TinyBeauty." The core idea of DAL lies in employing a Diffusion-based Data Amplifier (DDA) to "amplify" limited images for the model training, thereby enabling accurate pixel-to-pixel supervision with merely a handful of annotations. Two pivotal innovations in DDA facilitate the above training approach: (1) A Residual Diffusion Model (RDM) is designed to generate high-fidelity detail and circumvent the detail vanishing problem in the vanilla diffusion models; (2) A Fine-Grained Makeup Module (FGMM) is proposed to achieve precise makeup control and combination while retaining face identity. Coupled with DAL, TinyBeauty necessitates merely **80K** parameters to achieve a state-of-the-art performance without intricate face prompts. Meanwhile, TinyBeauty achieves a remarkable inference speed of up to **460 fps** on the iPhone 13. Extensive experiments show that DAL can produce highly competitive makeup models using only **5** image pairs. Please visit `https://github.com/TinyBeauty` for code and demos.

Keywords: Facial Makeup · Image Synthesis · Stable Diffusion

1 Introduction

Facial makeup aims to enhance facial appearance by applying cosmetic components on facial images, such as vibrant lipstick, intense eyeshadow, and eyeliner. Its primary application scenarios revolve around mobile devices, including short videos and live broadcasts, which makes makeup a time- and resource-sensitive task that demands efficient solutions. Specifically, mobile deployment

* This work was completed during the author's internship at Alibaba.

** Corresponding author: Bingbing Ni.

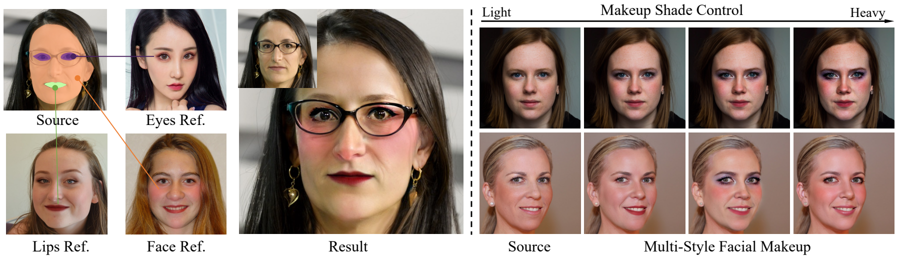
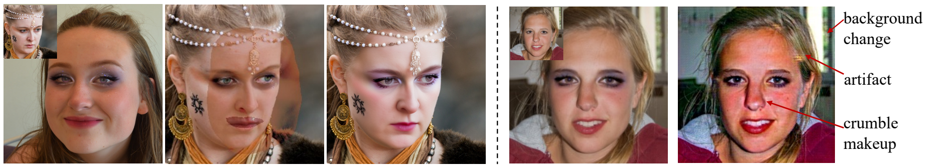


Fig. 1: Our **Diffusion-based Data Amplifier** is capable of learning from just several images to generate high-quality makeup visuals in diverse styles, offering flexible control such as makeup shade control and customized local editing.

faces unique challenges due to stringent restrictions on computational resources, which require extremely small model sizes (e.g., <100K) and minimized inference latency (typically <30ms) to work within such constrained resources. However, the current advanced makeup methods [6, 16, 17, 21, 23, 39, 40] suffer from large model sizes, generally exceeding 1M, and heavily rely on complex face prompts pipelines to ensure accurate makeup transfer. These factors hinder the practical application of existing approaches.

The root cause behind the excessive model sizes and complex inference pipelines in current mainstream makeup models can be attributed to the underlying learning framework, which is inherently flawed. The annotation of paired makeup data is expensive and challenging to achieve style scalability. Consequently, unpaired data has become the mainstream data protocol, which requires makeup models to carefully employ adversarial training [6, 17, 21] for achieving stable unpaired learning. Moreover, unpaired data often suffers from significant facial misalignment issues, which necessitates the inclusion of facial spatial prompts and other operations (e.g., warping) to assist in the unpaired learning process. For example, explicit incorporation of facial landmarks is used to guide the network in identifying the position of lipstick, eyelashes, and eyeshadow. However, these additional prompts further contribute to the complexity of the model. Simultaneously, the data misalignment poses a challenge for unpaired learning models to employ stable and straightforward reconstruction loss (e.g., L1, L2). Instead, they rely on inaccurate supervision, such as color histograms matching [21], earth mover's distance, due to the lack of alignment information. As well known, these inaccurate supervision methods often rely on theoretical approximations (e.g., approximate upper and lower bounds), leading to poor model robustness. For instance, as shown in Fig. 2, even with a model size of 10M, EleGANt [40] still frequently produces unsatisfactory results. The above challenges become more pronounced as the model size continues to decrease. The reduced model scale makes the model more sensitive to the accuracy of supervision while also requiring the abandonment of various auxiliary alignment techniques.

To address the above challenges, we propose a novel learning framework named **Data Amplify Learning** enabling stable and robust training supervi-



(a) Source/GT (b) PGT by EleGANt (c) Our Data Amplifier (d) Makeup by TinyBeauty (e) Makeup by EleGANt

Fig. 2: Left: Data evolution in facial makeup: unpaired data→pseudo-paired data→paired data (generated by data amplifier). **Right**: Comparison between TinyBeauty trained on paired data and EleGANt [40] trained on pseudo-paired data.

sion even with a limited number of image annotations. In the heart of DAL is a meticulously designed **Diffusion-based Data Amplifier**, responsible for amplifying the limited labeled image pairs (typically 5 pairs) into a larger dataset suitable for model training. This process allows DAL to transform the inaccurate supervision into pixel-to-pixel learning (e.g., L1), reducing the optimization difficulty and enabling accurate gradient propagation. Naively using mainstream diffusion-based control methods [35,41] as the Data Amplifier can lead to subpar outcomes, including identity mismatching, over-smoothing details, and inaccurate makeup results. To alleviate above challenges, two ingredients of the DDA are proposed: (1) Instead of directly generating makeup results, we design a Residual Diffusion Model (RDM) for the Data Amplifier. RDM consists of two branches: one is responsible for face reconstruction, and the other is dedicated to applying makeup. By utilizing the residual between the reconstructed face and the input face, we can transfer this residual to the makeup branch, resulting in high-fidelity and sharpened makeup results. (2) Unlike direct image injection [41], we propose a Fine-Grained Makeup Module (FGMM) that couples fine-grained makeup with facial identity representation to control makeup process, which effectively avoids the loss of facial identity. Additionally, we introduce semantic region labels for makeup, further refining the control of makeup application, rendering a precise makeup control generation. DAL greatly relaxes the optimization methods, allowing us to abandon face prompts and over-parameterization methods, thereby focusing on network design for mobile computing. Benefiting from this, a single **TinyBeauty** model is designed without reliance on external face prompts. TinyBeauty exceeds prior makeup transfer performance using only 14 convolution layers, enabling efficient deployment on mobile devices. Furthermore, to produce more complete makeup components compared to prior works, we utilize an edge operator to constrain the learning of eyeliner, enhancing the realism and fidelity of generated makeup styles.

Quantitatively, TinyBeauty outperforms EleGANt [40] with a significant +5.21dB PSNR increase (17.3% improvement) on the FFHQ [18] dataset and +1.49dB (4.55% improvement) on the MT [21] dataset. It also operates 13× faster than BeautyGAN [21] and delivers a swift 2.18ms latency, ensuring smooth smartphone integration on an iPhone13. Extensive experimental results demonstrate that DAL achieves impressive performance in training makeup models, even with a minimal dataset of only 5 image pairs.

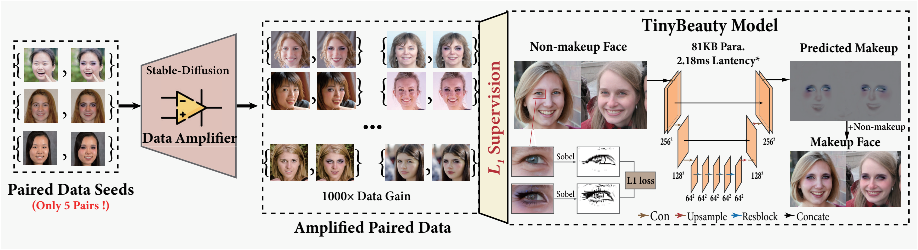


Fig. 3: Overview of the Data Amplify Learning framework. The Data Amplify Learning process contains two components: **(1)A data amplifier** which utilizes a pretrained diffusion model to amplify a small set of seed data into a larger synthesized dataset. **(2) A lightweight model** which is trained on the amplified data to accurately learn the makeup styles while retaining identity features of the original images.
*The latency is the inference time on an iPhone 13 device.

2 Related Work

2.1 Diffusion-based Image-to-Image Translation

Compared to previously widely used GAN-based [1–4, 12, 26] models , recent diffusion models [9, 31, 34] have demonstrated superior performance in high-fidelity image generation tasks, especially when few-shot images are available. Typically, conditions are introduced to steer the image generation process, generally in two forms: text-based and image-based guidance. DALL·E2 [24] and Imagen [32] apply text encoders from pre-trained language models [8, 27] for image generation guidance. Textual Inversion [11] and DreamBooth [30] learn special tokens from example images. However, solely using text for conditional guidance is indirect and inaccurate, making it difficult to ensure stable, consistently generated images. To strengthen conditional constraints, SD Image Variations2 [33] and Stable unCLIP3 [36] directly fine-tune text-conditioned diffusion models on image embeddings. IP-Adapter [41] and InstantID [38] use a combined image and text prompt for data generation conditions to keep the identity of the condition image. While these methods yield commendable outcomes in image stylization, they fail to preserve the facial detail like wrinkles of the original image, thus inadequate in applying facial makeup.

2.2 Facial Makeup

Traditional facial makeup methods [13, 20] rely on facial landmarks to warp predetermined beauty materials on the facial images for makeup application, which is efficient but unrealistic. BeautyGAN [21] introduces a dual GAN approach for makeup transfer with a color histogram loss using the proposed dataset. PSGAN [17] and FAT [37] propose attention mechanisms to address pose/expression changes, significantly increasing model size. SCGAN [6] encodes styles into component-wise codes while EleGANt [40] simplifies this complex optimization problem as L1 loss by generating pseudo ground truth. However, EleGANt still consumes significant computational overhead due to its use of multiscale attention modules. By abandoning CycleGAN's structure, BeautyREC [39] achieves a lighter model compared to previous work. Unlike makeup transfer methods, the face beautification proposed in [22] targets many-to-many translation by integrating beauty score prediction.Typically, all these methods incorporate parsing maps and/or landmarks during pre-processing on the unpaired MT dataset, forming cumbersome pipelines.

3 Method

We propose a novel learning scheme called **Data Amplify Learning (DAL)** to replace the unstable and inaccurate unpaired learning. DAL leverages a diffusion model to generate previously inaccessible paired data, which is then employed as training material for a compact tiny model. The core of DAL is a **Diffusion-based Data Amplifier (DDA)** to generate large amounts of paired data using only 5 makeup images as data seeds. (Sec. 3.1). Benefit from the DDA-amplified data, a single **TinyBeauty Model** is designed to replace the previous cumbersome makeup pipeline, facilitating the integration of the makeup model into mobile devices. (Sec. 3.2).

3.1 Diffusion-based Data Amplifier

Diffusion models [14, 28] are probabilistic generative models trained to learn a data distribution by performing an iterative denoising process. To project non-makeup image domain X into makeup domain Y, we finetune a pre-trained Stable Diffusion (SD) model [35] $\mathcal{F}_{init}$ using low-rank adaption [15]. We aim to guide the model to apply makeup style on input non-makeup data under conditional constraints. The constraint of the finetuning process of the SD model ϵ_θ is defined as:

$$L_{simple} = \mathbb{E}_{\mathbf{z}_0, \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \mathbf{c}, t} || \epsilon - \epsilon_\theta(\mathbf{z}_t, \mathbf{c}, t) ||^2, \tag{1}$$

where $\mathbf{z}_0$ represents the latent feature of the manually annotated makeup image with condition $\mathbf{c}$, ϵ denotes the noise added to $\mathbf{z}_0$, $t \in [0, T]$ denotes the time step of diffusion process, $\mathbf{z}_t = \alpha_t \mathbf{z}_0 + \delta_t \epsilon$ is the noisy data at t step. To generate makeup images, we incorporate conditions to guide the inference process in the fine-tuned SD model $\mathcal{F}_{fine}$.

$$y = \mathcal{F}_{fine}(x, \mathbf{c}), \tag{2}$$

where x is the non-makeup image, and y is the paired makeup image.

To generate high-quality paired makeup data, our DDA is required to contain the subject's original facial features, skin texture details, such as wrinkles and spots, and consistent, precise makeup styles across various portraits. However,

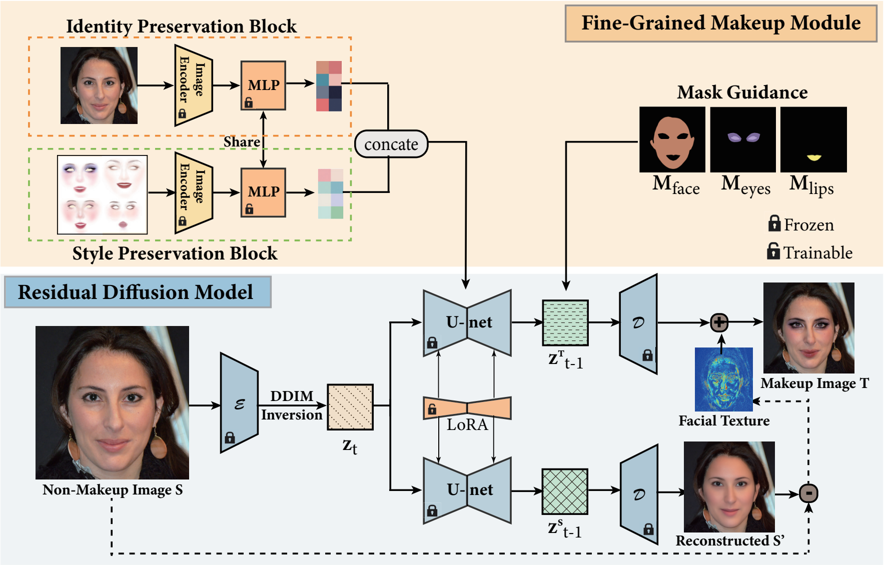


Fig. 4: Overview of the Diffusion-based Data Amplifier (DDA). Our DDA leverages a Residual Diffusion Model for high-fidelity texture preservation, minimizing distortion and avoiding unnatural mask-like appearances. It also employs a Fine-Grained Makeup Module including Identity Preservation Block (IPB) to maintain the original facial features, Style Preservation Block (SPB) to guarantee consistent makeup style application, and facial masks to specify makeup region.

as shown by Gal et al. [11], stable diffusion models face a trade-off between image reconstruction and editability, making it difficult to apply makeup while preserving the original face. To overcome these obstacles, we introduce a Residual Diffusion Model that preserves texture and detail, reducing distortion and mask-like effects. Moreover, we propose a Fine-Grained Makeup Module to ensure the precise application of makeup to the appropriate facial areas and generate visually consistent makeup styles, as shown in Fig. 4.

Residual Diffusion Models To address the issue of facial features like wrinkles being smoothed out during the denoising step of diffusion-based portrait image generation, we introduce a novel framework that utilizes parallel diffusion branches during the inference phase—the conditioned and unconditioned branches. The conditioned branch operates with content condition $\mathbf{c}_{con}$ and style condition $\mathbf{c}_{sty}$ to produce a smooth makeup image, while unconditioned branch proceeds without any conditions to output a smooth non-makeup image.

We define the concept of residual detail R_d by considering the difference between the original image x and its smooth counterpart produced by the unconditioned branch, $\mathcal{F}_{fine}(x)$. This difference, $R_d = x - \mathcal{F}_{fine}(x)$, represents the facial detail residuals, which are the essential features we aim to preserve to maintain realism. Furthermore, the makeup residual R_m is captured by the difference between the outputs of the two branches, $R_m = \mathcal{F}_{fine}(x, (\mathbf{c}_{sty} + \mathbf{c}_{con})) - \mathcal{F}_{fine}(x)$. To synthesize the enhanced portrait, we apply the following equation:

$$y_{detail} = \mathcal{F}_{fine}(x) + \lambda_m R_m + \lambda_d R_d, \tag{3}$$

where controlling the coefficients of the makeup λ_m and detail residuals λ_d allows us to adjust the makeup intensity and detail sharpness, respectively. Usually, we set λ_m to 1 and set λ_d to 0.8. This control mechanism is akin to an annealing concept, where increasing λ_d to 1.0 can lead to an oversaturation effect due to the addition of makeup-related details. This framework not only improves the quality of makeup in generated images but also has broader applications in other high-fidelity image generation tasks involving diffusion models, offering an advanced technique for creating realistic and detailed synthesized images.

Fine-Grained Makeup Module The Fine-Grained Makeup Module comprises the Identity Preservation Block (IPB), which safeguards individual facial identity, the Style Preservation Block (SPB) for precise makeup style management, and a latent face mask that guides region-specific makeup application and facilitates the flawless integration of diverse makeup styles.

Style Preservation Block (SPB). The crux for DDA lies in the accurate replication of makeup styles. Traditional diffusion-based image synthesis methods [38, 41], which use text to describe artistic styles, are ineffective for the nuanced task of capturing makeup. Text descriptions might broadly categorize lipstick as "red" but cannot detail the specific shade, undertone, or finish and how it interacts with the skin tone. Therefore, we employ visual examples as style references instead of textual references. To focus solely on the makeup while minimizing the influence of other facial features, we apply makeup to a frontal face image and employ a facial mask to isolate the makeup-applied regions, creating clean and unobstructed makeup images as style references. Then we encode the style image using a pre-trained image encoder and use a trainable MLP to project it as style tokens.

Identity Preservation Block (IPB). A critical aspect of our DDA is the intrinsic maintenance of facial identity, especially when learning makeup styles, which can inadvertently morph facial features. To maintain facial identity, our DDA employs the Identity Preservation Module (IPB) to disentangle style from identity, ensuring that the addition of makeup styles does not alter the facial structure. We initially considered using a distinct facial encoder like ArcFace [7] for the IPB but faced challenges due to incompatible encoding spaces. Thus, we unify the encoding space for SPB and IPB, leveraging a shared MLP for feature blending. Then the global condition vector $\mathbf{c}$ is divided into independent content $\mathbf{c}_{con}$ and style $\mathbf{c}_{sty}$ components.

Mask Guidance. To accurately define the facial areas affected by the makeup style, we divide the feature space into three distinct regions under the guidance of resized facial masks: M_{face}, M_{lips}, and M_{eyes}, as shown in Fig. 4. For single makeup style learning, we define the influenced area $M_{changed}$ as $M_{changed} = M_{face} + M_{lips} + M_{eyes}$. In training process, we only focus the learning on $M_{changed}$ regions to effectively capture makeup style. This is achieved by computing the loss L_{simple} only over the features within the $M_{changed}$ area:

$$L^M_{simple} = \mathbb{E}_{\mathbf{z}_0,\epsilon\sim\mathcal{N}(\mathbf{0},\mathbf{I}),\mathbf{c},t}||(\epsilon - \epsilon_\theta(\mathbf{z}_t, \mathbf{c}, t)) * M_{changed}||^2. \tag{4}$$

In the inference phase, the application of the latent space is similarly restricted to the regions specified by the changed mask. The blended latent L'_y is a combination of the output latent of the diffusion process L_y in the makeup-altered regions and the original latent L_x in the unaffected regions. This blending is mathematically represented as:: $L'_y = L_y \bigodot M_{changed} + L_x \bigodot (1 - M_{changed})$. When learning to combine multiple makeup styles, the model needs to handle various makeup conditions simultaneously. By applying different makeup conditions to each respective facial mask, we can impose constraints that allow for a seamless integration of multiple makeup elements on a single portrait.

3.2 TinyBeauty Model

Previous makeup pipelines are trained on unpaired data and, therefore heavily rely on face prompts like landmarks and facial masks to achieve face alignment. This reliance on heavy pipelines results in bulky networks with unstable and subpar accuracy. Benefiting from DDA-generated paired data, the TinyBeauty Model sidesteps this laborious pre-processing by directly applying L1 loss aligning generated images closely with their targets. Furthermore, we observed that traditional GAN-based methods directly generate makeup images as outputs, introducing noise to regions that should remain unaffected, like the background and hair Fig. 2. To mitigate these unwanted artifacts, we design TinyBeauty to output only the makeup residuals, as depicted in Fig. 3. This innovative approach significantly diminishes artifacts and ensures the preservation of the original image's features and details.

Network Architecture. Benefiting from DDA-generated data, TinyBeauty can be designed as a hardware-friendly network optimized for resource-constrained devices. We construct our TinyBeauty entirely from the most basic building block - fully convolutional layers. As depicted in Fig. 3, we implement a U-Net-based [29] generator architecture containing only four convolutional layers and four residual blocks, comprising a total of fourteen convolutional operations. This architectural choice is highly efficient while retaining the ability to capture spatial dependencies. When hiring the lightweight network, another key advantage of TinyBeauty is that the outputted residuals can be applied to images of any resolution without losing the original texture information. As a result, our tiny makeup model maintains the capability to translate fine-grained makeup details with a compact network architecture comprising only 81KB parameters.

Eyeliner Loss. Previous methods [6, 21, 39, 40] fail to learn clear eyeliner which holds pivotal significance in makeup, due to using unpaired data that hampers the network's ability to match and learn such fine details. To enhance the delineation of eyeliners, which typically exhibit evident edges, we employ edge filters for their extraction. We specifically use the Sobel filter $\mathcal{S}$ to extract the edges effectively from the eyeliner area. These extracted edges are then leveraged

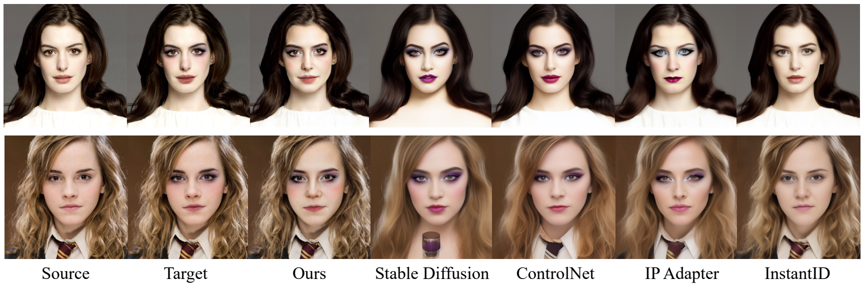


Fig. 5: Visual Comparison of Facial Makeup Using DDA. Comparative results highlight the superior performance of DDA in maintaining facial integrity and style consistency when compared with alternative methods.

by a specialized loss function $\mathcal{L}_s$ to accurately define the eyeliner's contours.

$$\mathcal{L}_s = ||\mathcal{S}(y) - \mathcal{S}(y')||_2^2 * M_{eyes}, \tag{5}$$

where y is makeup image, $y' = M(x) + x$ is the predicted makeup image, M_{eyes} is the mask around eyes.

Reconstruction Loss. Previous methods, lacking paired data, resorted to complex loss functions like histogram matching for makeup applications. In contrast, thanks to our DDA-generated paired data, we can directly use a global L1 loss to enforce constraints.

$$\mathcal{L}_{rec} = ||y - y'||_1. \tag{6}$$

In addition to the above losses, we also leverage the perception loss $\mathcal{L}_{per}$ and adversarial loss $\mathcal{L}_{adv}$, which have been widely utilized in previous researches [6, 21, 39, 40]. The supplementary material provides more details of these losses.

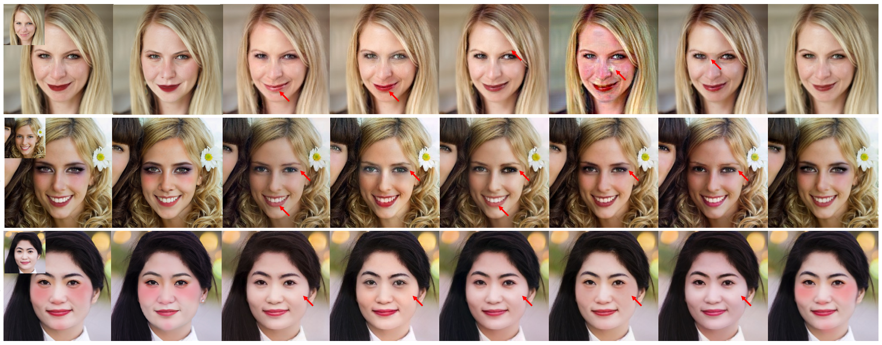
4 Experiments

4.1 Datasets and Evaluation Metrics

Data Seeds and Evaluation Dataset. We randomly select *5* images without makeup from the Flickr-Faces-HQ (FFHQ) [18] dataset as our DAL's data seeds, *i.e.* training data. We then manually annotate these data seeds with five predefined makeup styles using MEITU [1], along with the makeup-only images as the input of SPB. To enable a comprehensive and fair evaluation, 100 images from the FFHQ dataset [18] and 100 images from the MT dataset [21] are additionally annotated as test data when computing PSNR, FID, and LPIPS.

Evaluation Metrics. To evaluate TinyBeauty quantitatively, we employ widely recognized metrics to evaluate the similarity, diversity, and realism of the

[1] https://www.meitu.com/



Target	DDA	BeautyGAN	PSGAN	SCGAN	EleGANt	BeautyREC	Ours
Parameter(M) & Latency*(ms)		8.04 & 27.89	8.41 & N/A	35.33 & 195.61	10.27 & N/A	0.99 & 206.46	0.08 & 2.18

Fig. 6: Visual comparison of TineBeauty on the FFHQ [18] images, with the parameter size and inference latency of each model **on an iPhone13 device**. N/A means the model cannot be deployed on iPhone devices. (Find more results of MT Datasets and more makeup styles in Supplementary Material.)

images generated. We use the Peak Signal-to-Noise Ratio (PSNR) to measure the similarity between the generated data and ground truth (GT) data. A higher PSNR indicates a closer approximation to the GT. Additionally, we report the Fréchet Inception Distances (FID) and Learned Perceptual Image Patch Similarity (LPIPS) by comparing 100 generated images to 100 GT images. FID and LPIPS metrics evaluate the diversity and realism of the generated samples, with lower scores indicating a higher similarity to the GT data distribution.

4.2 Implementation Details

Diffusion-based Data Amplifier (DDA). Our diffusion fine-tuning experiment use SDv1.5 [25] as the base model. The image encoder utilized is OpenCLIP ViT-H/14 [10], which is pre-trained by IP-Adapter [41]. We set the length of both style token and identity token to 32 to balance the strength of style and identity. LoRA [15] in conjunction with U-net is employed to fine-tune the SD model with a learning rate of 1e-4 for 500 epochs. Notably, the five distinct makeup styles are concurrently trained within a unified model, demonstrating the model's capacity for multi-style learning in a single training session. The training process is executed on an NVIDIA V100 GPU, taking about 50 mins to fine-tune five makeup styles. We utilize the FaRL [44] to generate facial masks and resize them to a resolution of 64 × 64 to guide the network's training in the latent space.

TinyBeauty Model. The TinyBeauty model is trained using a dataset consisting of 4000 diffusion-generated images for 50 epochs. The training processes utilized a learning rate of $2e-4$ and the Adam [19] optimizer. The entire training procedure is executed on an NVIDIA V100 GPU, requiring approximately 12 hours. It is important to mention that **NO** specific face operations are conducted

Table 1: Results of Diffusion-based Data Amplifier (DDA), TinyBeauty, and competing methods on FFHQ [18] dataset and MT [21] dataset, in style1. The best and second best results of each column are indicated with bold font and underlined respectively. * means the models are trained with our proposed DAL scheme.

Method	FFHQ			MT		
	PSNR↑	FID↓	LPIPS↓	PSNR↑	FID↓	LPIPS↓
BeautyGAN [21]	26.50	45.25	0.0564	27.49	25.05	0.0434
PSGAN [17]	25.65	36.22	0.0594	28.05	18.72	0.0301
SCGAN [6]	27.55	36.98	0.0485	27.22	30.85	0.0467
EleGANt [40]	30.18	25.47	0.0396	32.77	12.55	0.0191
BeautyREC [39]	24.93	26.88	0.0538	27.39	21.37	0.0430
EleGANt* [40]	35.45	10.78	0.0148	34.65	11.57	**0.0164**
BeautyREC* [39]	31.06	15.59	0.0374	27.39	18.21	0.0232
DDA	**35.96**	10.28	0.0195	**34.79**	10.37	0.0231
TinyBeauty	35.39	**8.03**	**0.0146**	34.26	**9.33**	0.0181

on the input image during the inference phase, which makes our makeup process fast and hardware-friendly.

4.3 Comparison

To showcase the quality of the data generated using our DDA, we compare it with the text-conditioned SD model, ControlNet [43], IP-Adapter [41], and InstantID [38]. We train both previous methods in the same data as ours, with the result shown in Fig. 5. The results reveal two key insights. Firstly, the subtleties of makeup styles are not easily captured by textual descriptions alone. Consequently, it is challenging to precisely manipulate attributes, particularly color, in the synthesized images. Our DDA addresses this by directly encoding makeup images, thus producing results with consistent color and makeup content. Secondly, IP-Adapter falls short of meeting our stringent criteria for identity retention in the facial makeup task. Our DDA, however, effectively resolves this issue, generating high-quality makeup images by synergistically leveraging mask guidance, IPB, and RDM. Additionally, Fig. 1 illustrates the DDA's capability of fusing various makeup styles. Our DDA adeptly blends eye, lip, and blush makeup to create a myriad of diverse and intricate makeup applications.

To evaluate the effectiveness of our TinyBeauty on DDA-generated data, we conduct a comparison with several representative pre-trained facial makeup techniques, including BeautyGAN [21], PSGAN [17], SCGAN [6], EleGANt [40], and BeautyREC [39]. The experimental results, as depicted in Fig. 6, demonstrate the makeup data generated by our DDA and TinyBeauty are both visually pleasant and consistent with the ground-truth data. Beyond overall visual enhancement, TinyBeauty also provides auxiliary makeup functions such as automatic eyebrow completion and eyeliner application, as demonstrated in Fig. 9. The precision of

Table 2: Parameters of TinyBeauty and competing methods. (+) means the method uses facial pre-processing including face parsing*, and (-) means the method only has a single model.

Method	Param.(M)↓	FLOPs(G)↓	RunTime(ms)↓
Face Parsing* [42]	12.68	1.61	13.29
BeautyGAN [21]	8.04	24.70	27.89 (-)
PSGAN [17]	8.41	91.28	N/A (+)
SCGAN [6]	35.33	288.51	195.61 (+)
EleGANt [40]	10.27	127.94	N/A (+)
BeautyREC [39]	0.99	12.58	206.46 (+)
TinyBeauty	**0.08**	**0.69**	**2.18** (-)

Table 3: The voting result (%) of user study on test data. Participants are asked to rank the makeup quality based on accurate makeup color, clear details like eyebrows or eyeliner, pleasing effects, and minimal distortion.

Method	Rank-1	Rank-2	Rank-3	Rank-4	Rank-5	Rank-6
BeautyGAN [21]	0.18	0.55	0.98	4.59	14.14	**79.56**
PSGAN [17]	1.56	0.86	11.43	6.80	**64.75**	13.98
SCGAN [6]	1.00	1.76	27.71	**54.27**	13.76	1.46
EleGANt [40]	10.46	**84.07**	2.27	0.23	1.17	2.57
BeautyREC [39]	0.28	1.95	**55.06**	34.13	6.15	2.43
TinyBeauty	**86.56**	10.81	2.55	0.05	0.03	0.00

the makeup results stems from the combination of high-quality DDA-generated data and our targeted constraints on eyeliner design.

Quantitative Comparison. We conduct a comprehensive evaluation by comparing TinyBeauty not only with the pre-trained models provided by earlier studies [6,17,21,39,40], but also by retraining the two most advanced models [39,40] from those studies using our DAL. The comparative analysis, detailed in Tab. 1, reveals TinyBeauty's performance over the pre-established and newly trained models in terms of PSNR, FID, and LPIPS across both the FFHQ and MT datasets. Remarkably, despite its significantly smaller network size, TinyBeauty achieves comparable results to EleGANt in the same training set. The success lies in the high-quality paired data we generate, which enables even small networks to learn and apply stable makeup styles effectively. Moreover, our specialized design incorporating residual learning not only diminishes noise interference but also enhances TinyBeauty's performance, even surpassing larger predecessors. Notably, the MT dataset is not utilized during training, yet TinyBeauty yields a PSNR improvement of 1.49dB over prior methods on MT.

Model Efficiency Evaluation. As shown in Tab. 2, we measure FLOPs, parameter size, and inference time of each model. The models are packaged with CoreML [5] and tested on an iPhone13 without including face pre-processing time. We also test a face parsing model [42] on iPhone13 for reference, commonly utilized by other approaches but not integrated within our TinyBeauty method. Our model achieves an inference time of only 2.18ms for a 256x256 image, which is 13× faster than the fastest competing method on iPhone13, i.e., BeautyGAN [21]. Specifically, TinyBeauty is approximately 6 times faster than face parsing on iPhone hardware, which is used as a pre-processing step in previous makeup transfer pipelines [6,17,39,40]. In contrast, our TinyBeauty only utilizes a single model by simplifying the optimization problem facilitated by amplified paired data. These results highlight TinyBeauty's efficiency and mobile deployment readiness, achieved by implementing our DAL scheme.

User Study. We conduct a subjective evaluation on the FFHQ and MT test sets. Workers are shown makeup results from TinyBeauty and competing methods for raw images in random order. A total of 100 validated workers aged

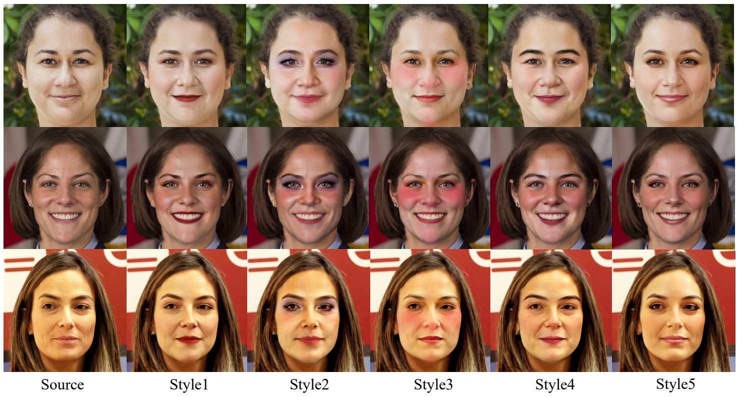


Fig. 7: Results of the diffusion-based data amplifier. Consistent makeup styles are generated while retaining the facial content and identity of the original image.

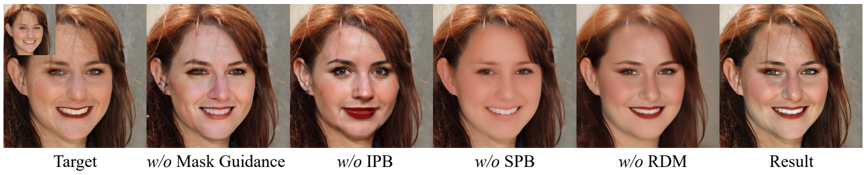


Fig. 8: Ablation study of RDM and modules in FMM of DDA.

19-48 from different cultural backgrounds participate in the fair and randomized evaluation. The average rankings are summarized in Tab. 3.

4.4 Ablation Study

Ablation Study of Diffusion-based Data Amplifier. DDA is the core component of DAL. Its performance directly affects the ability of DAL model training. The effectiveness of each module in the data amplifier is verified through combination experiments as shown in Fig. 8. The experimental findings indicate that the integration of mask guidance and the IPB enables the DDA to more precisely retain the subject's identity. Simultaneously, the SPB delivers more accurate guidance on the makeup style, while the RDM maintains the facial texture, thereby enhancing the overall realism of the image. The combination of these modules enables DDA to produce high-quality makeup images.

Component Analysis of TinyBeauty. To investigate the effectiveness of eyeliner loss, we conduct ablation studies. As shown in Fig. 11, we find that removing the eyeliner loss on the region around the eyes indicates that it indeed greatly promotes learning of clear eyeliner. This improvement is attributed to the fact that the network struggles to capture high-frequency signals such as eyeliner without specific constraints. By employing edge constraints to direct

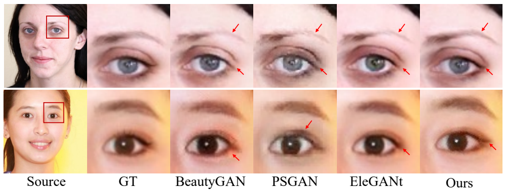


Fig. 9: Comparison of makeup details. TinyBeauty generates finer details on eyeliner and complete eyebrows compared to previous methods. (Zoom in for more details.)

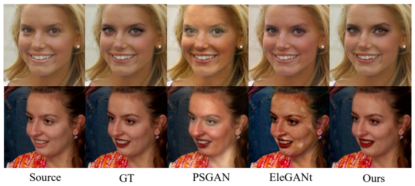


Fig. 10: Visual comparison of the in-the-wild images with large poses and expressions.

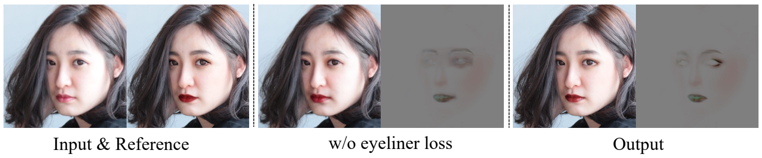


Fig. 11: Ablation study of eyeliner loss in TinyBeauty.

the network's learning process, our method enables it to discern and reproduce eyeliner details, which is not achievable by previous techniques.

4.5 Results on the in-the-wild images and videos

To further test the robustness of TinyBeauty, supplementary tests are conducted on both in-the-wild large-pose images and in-the-wild videos. The final makeup results are depicted in Fig. 10. Thanks to the large amount of paired data introduced by DAL, TinyBeauty has achieved significantly better performance than previous methods such as PSGAN [17] and EleGANt [40], which specializes in addressing facial makeup with large poses and expressions. Additionally, TinyBeauty demonstrates stronger makeup application abilities in videos, as shown in Supplementary Material.

5 Conclusion

In this paper, we introduce a revolutionary approach to facial makeup application with the development of Data Amplify Learning, and the implementation of a compact makeup model, TinyBeauty. By leveraging the innovative Residual Diffusion Model and Fine-Grained Makeup Module within our Data Amplifier, we effectively amplify limited paired data as the training data of TinyBeauty, a tiny 14-layer convolutional model replacing previous cumbersome pipelines. This results in a swift 2.18ms mobile makeup application and a 17.3% PSNR quality boost, establishing a new milestone for real-time, low-resource mobile makeup.

Acknowledgements

This work was supported by National Science Foundation of China (U20B2072, 61976137). This work was also partly supported by SJTU Medical Engineering Cross Research Grant YG2021ZD18.

References

1. Chen, R., Chen, X., Ni, B., Ge, Y.: Simswap: An efficient framework for high fidelity face swapping. In: MM '20: The 28th ACM International Conference on Multimedia, Virtual Event / Seattle, WA, USA, October 12-16, 2020. pp. 2003–2011. ACM (2020)
2. Chen, X., Ni, B., Liu, N., Liu, Z., Jiang, Y., Truong, L., Tian, Q.: Coogan: A memory-efficient framework for high-resolution facial attribute editing. In: Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XI. vol. 12356, pp. 670–686. Springer (2020)
3. Chen, X., Ni, B., Liu, Y., Liu, N., Zeng, Z., Wang, H.: Simswap++: Towards faster and high-quality identity swapping. IEEE Trans. Pattern Anal. Mach. Intell. **46**(1), 576–592 (2024)
4. Chen, X., Yan, X., Liu, N., Qiu, T., Ni, B.: Anisotropic stroke control for multiple artists style transfer. In: MM '20: The 28th ACM International Conference on Multimedia, Virtual Event / Seattle, WA, USA, October 12-16, 2020. pp. 3246–3255. ACM (2020)
5. Https://github.com/apple/coremltools
6. Deng, H., Han, C., Cai, H., Han, G., He, S.: Spatially-invariant style-codes controlled makeup transfer. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021. pp. 6549–6557. Computer Vision Foundation / IEEE (2021)
7. Deng, J., Guo, J., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. CoRR **abs/1801.07698** (2018)
8. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
9. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Advances in neural information processing systems **34**, 8780–8794 (2021)
10. (2021), https://github.com/mlfoundations/open_clip
11. Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A.H., Chechik, G., Cohen-Or, D.: An image is worth one word: Personalizing text-to-image generation using textual inversion. arXiv preprint arXiv:2208.01618 (2022)
12. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. Advances in neural information processing systems **27** (2014)
13. Guo, D., Sim, T.: Digital face makeup by example. In: 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA. pp. 73–79. IEEE Computer Society (2009)
14. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models (2020)
15. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. In: The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022. OpenReview.net (2022)

16. Isola, P., Zhu, J., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017. pp. 5967–5976. IEEE Computer Society (2017)
17. Jiang, W., Liu, S., Gao, C., Cao, J., He, R., Feng, J., Yan, S.: PSGAN: pose and expression robust spatial-aware GAN for customizable makeup transfer. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020. pp. 5193–5201. Computer Vision Foundation / IEEE (2020)
18. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019. pp. 4401–4410. Computer Vision Foundation / IEEE (2019)
19. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: Bengio, Y., LeCun, Y. (eds.) 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings (2015)
20. Li, C., Zhou, K., Lin, S.: Simulating makeup through physics-based manipulation of intrinsic image layers. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015. pp. 4621–4629. IEEE Computer Society (2015)
21. Li, T., Qian, R., Dong, C., Liu, S., Yan, Q., Zhu, W., Lin, L.: Beautygan: Instance-level facial makeup transfer with deep generative adversarial network. In: 2018 ACM Multimedia Conference on Multimedia Conference, MM 2018, Seoul, Republic of Korea, October 22-26, 2018. pp. 645–653. ACM (2018)
22. Liu, X., Wang, R., Chen, C., Yin, M., Peng, H., Ng, S., Li, X.: Face beautification: Beyond makeup transfer. CoRR **abs/1912.03630** (2019), `http://arxiv.org/abs/1912.03630`
23. Lyu, Y., Dong, J., Peng, B., Wang, W., Tan, T.: SOGAN: 3d-aware shadow and occlusion robust GAN for makeup transfer. CoRR **abs/2104.10567** (2021)
24. Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., Chen, M.: Glide: Towards photorealistic image generation and editing with text-guided diffusion models. arXiv preprint arXiv:2112.10741 (2021)
25. (2022), https://github.com/huggingface/
26. Qiu, T., Ni, B., Liu, Z., Chen, X.: Fast optimal transport artistic style transfer. In: MultiMedia Modeling - 27th International Conference, MMM 2021, Prague, Czech Republic, June 22-24, 2021, Proceedings, Part I. Lecture Notes in Computer Science, vol. 12572, pp. 37–49. Springer (2021)
27. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
28. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022. pp. 10674–10685. IEEE (2022)
29. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015 - 18th International Conference Munich, Germany, October 5 - 9, 2015, Proceedings, Part III. Lecture Notes in Computer Science, vol. 9351, pp. 234–241. Springer (2015)

30. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22500–22510 (2023)
31. Saharia, C., Chan, W., Chang, H., Lee, C., Ho, J., Salimans, T., Fleet, D., Norouzi, M.: Palette: Image-to-image diffusion models. In: ACM SIGGRAPH 2022 Conference Proceedings. pp. 1–10 (2022)
32. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. Advances in Neural Information Processing Systems **35**, 36479–36494 (2022)
33. Https://huggingface.co/lambdalabs/sd-image-variations-diffusers
34. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)
35. Https://huggingface.co/runwayml/stable-diffusion-v1-5
36. Https://huggingface.co/stabilityai/stable-diffusion-2-1-unclip
37. Wan, Z., Chen, H., Zhang, J., Jiang, W., Yao, C., Luo, J.: Facial attribute transformers for precise and robust makeup transfer. CoRR **abs/2104.02894** (2021)
38. Wang, Q., Bai, X., Wang, H., Qin, Z., Chen, A., Li, H., Tang, X., Hu, Y.: Instantid: Zero-shot identity-preserving generation in seconds (2024)
39. Yan, Q., Guo, C., Zhao, J., Dai, Y., Loy, C.C., Li, C.: Beautyrec: Robust, efficient, and component-specific makeup transfer. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023 - Workshops, Vancouver, BC, Canada, June 17-24, 2023. pp. 1102–1110. IEEE (2023)
40. Yang, C., He, W., Xu, Y., Gao, Y.: Elegant: Exquisite and locally editable GAN for makeup transfer. In: Avidan, S., Brostow, G.J., Cissé, M., Farinella, G.M., Hassner, T. (eds.) Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XVI. Lecture Notes in Computer Science, vol. 13676, pp. 737–754. Springer (2022)
41. Ye, H., Zhang, J., Liu, S., Han, X., Yang, W.: Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. CoRR **abs/2308.06721** (2023)
42. Yu, C., Wang, J., Peng, C., Gao, C., Yu, G., Sang, N.: Bisenet: Bilateral segmentation network for real-time semantic segmentation. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XIII. Lecture Notes in Computer Science, vol. 11217, pp. 334–349. Springer (2018)
43. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3836–3847 (2023)
44. Zheng, Y., Yang, H., Zhang, T., Bao, J., Chen, D., Huang, Y., Yuan, L., Chen, D., Zeng, M., Wen, F.: General facial representation learning in a visual-linguistic manner. CoRR **abs/2112.03109** (2021)