Makeup Style Transfer on Low-quality Images with Weighted Multi-scale Attention

Daniel Organisciak
Department of Computer and Information Sciences
Northumbria University
Email: daniel.organisciak@northumbria.ac.uk

Edmond S. L. Ho
Department of Computer and Information Sciences
Northumbria University
Email: e.ho@northumbria.ac.uk

Hubert P. H. Shum
Department of Computer Science
Durham University
Email: hubert.shum@durham.ac.uk
Corresponding Author

***Abstract*—Facial makeup style transfer is an extremely challenging sub-field of image-to-image-translation. Due to this difficulty, state-of-the-art results are mostly reliant on the Face Parsing Algorithm, which segments a face into parts in order to easily extract makeup features. However, this algorithm can only work well on high-definition images where facial features can be accurately extracted. Faces in many real-world photos, such as those including a large background or multiple people, are typically of low-resolution, which considerably hinders state-of-the-art algorithms. In this paper, we propose an end-to-end holistic approach to effectively transfer makeup styles between two low-resolution images. The idea is built upon a novel weighted multi-scale spatial attention module, which identifies salient pixel regions on low-resolution images in multiple scales, and uses channel attention to determine the most effective attention map. This design provides two benefits: low-resolution images are usually blurry to different extents, so a multi-scale architecture can select the most effective convolution kernel size to implement spatial attention; makeup is applied on both a macro-level (foundation, fake tan) and a micro-level (eyeliner, lipstick) so different scales can excel in extracting different makeup features. We develop an Augmented CycleGAN network that embeds our attention modules at selected layers to most effectively transfer makeup. Our system is tested with the FBD data set, which consists of many low-resolution facial images, and demonstrate that it outperforms state-of-the-art methods, particularly in transferring makeup for blurry images and partially occluded images.**

I. Introduction

Current state-of-the-art makeup style transfer methods [1], [2], [3], [4] display a common trend when applied to low quality images: only lipstick colour is consistently transferred. Eyeliner and mascara occasionally make the transition. Foundation, eye shadow, blusher, fake tan, concealer, powder, and contours are largely disregarded. Bags under the eyes, that want to be concealed, are ignored. For real-world applications, such as makeup-invariant face verification [5] or beautification [6], this does not suffice. We postulate that the fault lies with the hard attention that current makeup transfer methods use to handle the difficulty of the task.

Convolutional neural networks struggle to generalise to different data sets [7]. This also extends to GANs. For example, PULSE [8] applied to external data, converted a downsampled image of Barack Obama into a white man. To minimise generalisation error and to help to defend against model bias, it is beneficial for the algorithm to be able to train on a large variety of data, whereas current state-of-the-art makeup style transfer methods can only train on high-quality lab data sets. In practical applications of makeup transfer, low-resolution faces are prominent, because faces often take up a small proportion of an image. After cropping, they appear at a low resolution. As this scenario will frequently occur, it is important that models can handle it. Because we cannot reliably depend on models to generalise to this low-resolution setting, we design our framework to be able to train directly on low-resolution data.

A naïve approach to transfer makeup style is to use off-the-shelf image-to-image translation techniques, such as CycleGAN [9]. However, this performs poorly because the two domains are highly overlapping; both domains comprise of face images, with the greatest difference usually appearing on the lips and around the eyes. It is challenging to describe a makeup style since it consists of multiple non-requisite components. A face only wearing lipstick and the same face only wearing eyeliner should both belong in the makeup domain. This is difficult for standard models to gauge without being directly pointed to.

CycleGAN's ineffectiveness for makeup style transfer effectuates a part-by-part solution to apply a style from one face image to another. Current works develop a hard attention module where face parts likely to consist of makeup (eyes, lips and cheeks) are segmented and optimised upon separately [1], [2], [3]. However, to segment the image, these methods are dependent on the Face Parsing Algorithm (FPA) [10]. We demonstrate that FPA is limited in handling low-resolution face images due to the lack of detailed features to identify face parts. As a consequence, state-of-the-art makeup style transfer algorithms cannot be applied successfully to lower-resolution faces.

We discard the hard attention and develop an analogous soft attention module that to transfer makeup style in a holistic manner, rather than piece by piece. To tackle the problem of identifying salient parts of low resolution images without FPA, a novel weighted multi-scale spatial attention module is proposed. The module consists of spatial attention with multiple convolutional kernels. These convolutional layers determine salient areas of the image at different scales, which are then

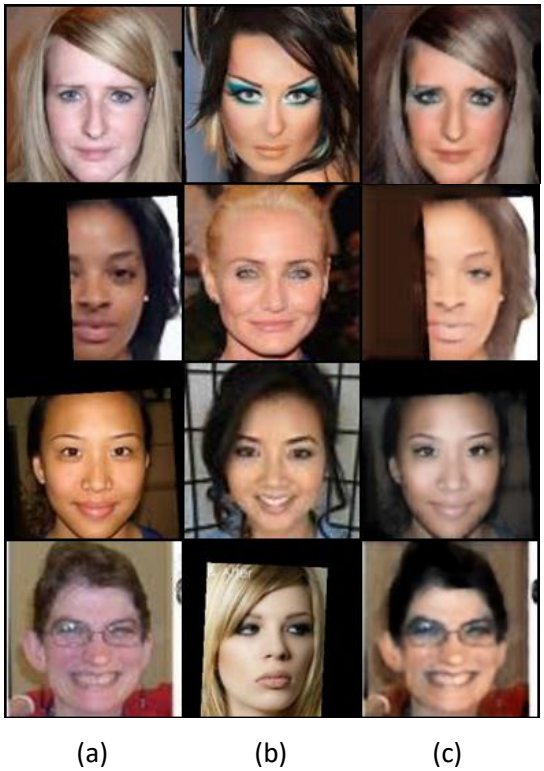

Fig. 1. (a) low quality source images; (b) makeup images from which to extract the makeup style; (c) the inferred result. Our method is capable of handling noisy, partially cropped, real-world data.

processed by an intermediate channel attention module, which determines the importance of different scales and assigns respective weights. This attention module serves two main purposes. Firstly, attention at different scales can focus on transferring different aspects of a makeup style: smaller scales capture fine-grained information such as fake eyelashes and lipstick, whereas larger scales focus on transferring foundation and fake tan that appear across the entire face. Secondly, faces can appear at any size in an image, so it is important to be able to effectively handle different resolutions. Our attention module can dynamically adjust which convolutional kernels are assigned high weights, and therefore extract more information from lower quality images. This results in a better face representation, and a better encoded makeup style. As shown in Figure 1, our framework is able to transfer makeup style under a wide range of difficult conditions, including low-resolution images, cropped faces and radical makeup styles.

Quantitative results show that the weighted spatial attention module outperforms the state of the art at transferring makeup style on low quality data. We also provide qualitative examples to show that our model compares favourably to state of the art on difficult tasks.

In this paper, the following contributions are provided:

1) We propose a new weighted multi-level spatial attention module to capture high-level and fine-grained style information. Such a mechanism is employed to encode the makeup style from the reference image and apply it to the source image with generative adversarial networks.
2) We demonstrate that state-of-the-art makeup style transfer techniques such as [2], [4] struggle to handle lower resolution data encountered in the real world. To handle this issue, an end-to-end framework based on Augmented CycleGAN is designed, with attention modules included within the generator, discriminator, and encoder.
3) We design a metric to quantitatively evaluate makeup style transfer.

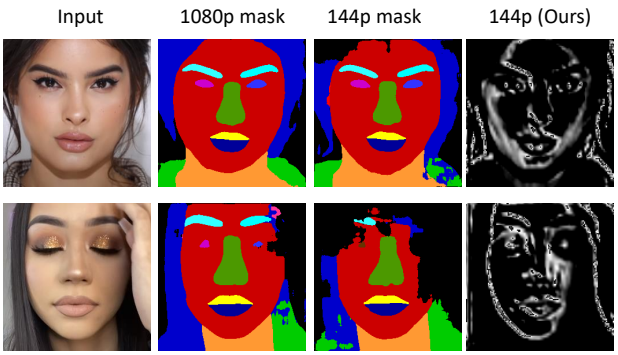

Fig. 2. Hard attention, used in current state-of-the-art frameworks, on different resolutions. In the top row, due to facial pose and good lighting, the low resolution image can be segmented well. However, on row 2, closed eyes and occlusion from the hand causes segmentation failure. Multi-scale attention (lighter means higher weight) is more capable of handling these challenges and gives a more detailed attention map. Note that current state of the art is dependent on the attention maps, whereas ours still attains reasonable performance without attention.

The rest of the paper is organised as follows. Related studies are outlined in Section II. Section III describes the spatial attention mechanism and explores the full network for many-to-many image translation. Section IV demonstrates our results compared to state-of-the-art methods. The paper is concluded in Section V.

II. Related Work

A. Attention

Attention is incorporated into Convolutional Neural Networks (CNNs) to highlight salient regions of an image that the network should focus on. There are several ways in which this can be done.

Self-Attention Networks [11] split an input into three streams: a *key*, *query*, and *value*. The higher the similarity between the key and query, the higher the weight attributed to the value.

Spatial Attention [12] is calculated via a residual convolutional block and aims to identify pixel regions the most contribute to minimising the loss, whereas *Channel Attention* [13] focuses on identifying the most important channels at each convolutional layer, which results in a better final feature representation. Harmonious Attention Networks [14] combine Spatial Attention with Channel Attention, and add a *hard attention* module, for the task of person re-identification. This

combination of attention modules improves performance with negligible change in computational complexity.

Attention has also been applied within Generative Adversarial Networks (GANs) for image generation. Zhang *et al.* [15] propose Self-Attention GAN, which makes use of a key, query, value set to perform attention similarly to traditional transformers [11]. Tang *et al.* propose AttentionGAN [16] for image-to-image translation, where the attention masks focus on key areas to translate.

B. Makeup Style Transfer

For digital makeup, an early work from Guo and Sim [17] can be used to transfer makeup style from one portrait to another. The source and target images are decomposed into three different layers: face structure layer, skin detail layer, and the colour layer. By altering the skin detail and colour layers, makeup style can be transferred. Xu et al. [18] proposed using face landmark detection to locate more important regions on the face and edit the skin colour and local details for each landmark. Li et al. [19] presented a physically-based model to alter the optical properties in the reflectance layers extracted from an image to simulate the digital makeup effects. More recent work by Liu et al. [20] proposed an end-to-end deep learning framework to i) recommend the suitable reference makeup style for the input image, ii) transfer the commonly used cosmetics (such as foundation, eye shadow and lip gloss) for different facial parts locally using the proposed *Deep Transfer Network*. The aforementioned approaches facilitate makeup style transfer based on underlying models or facial landmarks. Sub-optimal results will be produced if the model extraction and landmark detection are inaccurate.

CycleGAN [9] and other image-image translation works [21], [22], [23] variants demonstrated encouraging results on image-to-image translation tasks. PairedCycleGAN [1] improves the preservation of face identity by incorporating both a makeup transfer and a makeup removal networks. The face is separated into three parts, the eyes, lips and skin, and a generator-discriminator pair is trained for each part to capture unique characteristics. In addition to the typical cycle consistency loss and perceptual loss for ensuring the quality of the style transfer and realism of the resultant images, BeautyGAN [2] further includes the *makeup loss* to improve the appearance of the lips, eye shadow and face regions. Zhang *et al.* [4] are able to not only transfer the makeup style from one image to another, but control the strength with which it is applied, or apply a hybrid of two different makeup styles. BeautyGlow [3] decompose makeup and non-makeup images into latent vectors, then combine them in the latent space. They then generate the new makeup image from the combination vector. Unfortunately, none of these methods have demonstrated the ability to work effectively on low-resolution images that are frequently encountered in real world applications of makeup style transfer.

III. Weighted Multi-Scale Spatial Attention

Many current state-of-the-art makeup style transfer frameworks are reliant on hard attention to segment the face into parts. Figure 2 shows that the same image taken at a low resolution can result in drastically worse performance. As real-world applications would benefit from efficacy on low-quality data, the state of the art should capably handle this data.

We propose a new *weighted multi-scale spatial attention* module that is composed of *spatial attention* and *channel attention* as an alternative to the face parsing algorithm. The spatial attention extracts saliency information from the image at different scales. The channel attention learns the relative importance of these scales to give these attention maps an associated weight. This design is motivated by the observation that one makeup style is composed of different aspects of makeup: foundation covers a large area over the face while eyeliner is only visible across a few pixels. Using multiple scales of spatial attention allows us to capture all of the information necessary for makeup style transfer. Computing spatial attention with a large kernel size may result in eyeliner being overlooked. Our module combines three different kernel sizes to help avoid this issue.

In the rest of this section, we will formally define the problem, then describe the proposed attention module that can capture makeup information of each image. The full procedure is outlined in Figure 3.

A. Problem Formulation

Given a set of images without makeup, X, and a set of images with makeup, Y, we aim to convert any image pair $(x \in X, y \in Y)$ into a new image $\tilde{x}_y$, that has transferred the makeup style from image y onto image x. Most image-to-image translation tasks apply an arbitrary style to image $x \in X$ to obtain $\tilde{x} \in Y$. For each $x \in X$, we only get one $\tilde{x} \in Y$. However, we want to apply the specific makeup style from $y \in Y$.

To accomplish this, we will use the Augmented CycleGAN [23] model as a baseline. Here, we provide an outline of how it obtains a many-to-many mapping.

As visualised in Figure 4, we simultaneously train two generators and two encoders with associated discriminators:

$$\left.\begin{aligned} G_{XY} &: X \times Z_Y \longrightarrow Y, \quad D_Y : \quad Y \longrightarrow \{0,1\}, \\ G_{YX} &: Y \times Z_X \longrightarrow X, \quad D_X : \quad X \longrightarrow \{0,1\}, \\ E_X &: X \times Y \longrightarrow Z_X, \quad D_{Z_X} : \quad Z_X \longrightarrow \{0,1\}, \\ E_Y &: Y \times X \longrightarrow Z_Y, \quad D_{Z_Y} : \quad Z_Y \longrightarrow \{0,1\}. \end{aligned}\right\} \quad (1)$$

We optimise this with a standard GAN Loss, $\mathcal{L}^{Y}_{\text{GAN}}$; an encoder generator equivalent, $\mathcal{L}^{Z_X}_{\text{GAN}}$; an image cycle-consistency loss, $\mathcal{L}^{X}_{\text{CYC}}$; and an encoder cycle-consistency loss, $\mathcal{L}^{Z_Y}_{\text{CYC}}$. These are combined with hyperparameters γ_1 and γ_2 to obtain the loss function in the non-makeup to makeup direction:

$$\begin{aligned} &\mathcal{L}^{Y}_{\text{GAN}}(G_{XY}, D_Y) \quad + \quad \mathcal{L}^{Z_X}_{\text{GAN}}(E_X, G_{XY}, D_{Z_X}) \\ &+\gamma_1 \mathcal{L}^{X}_{\text{CYC}}(G_{XY}, G_{YX}, E_X) + \gamma_2 \mathcal{L}^{Z_Y}_{\text{CYC}}(G_{XY}, E_Y). \end{aligned} \quad (2)$$

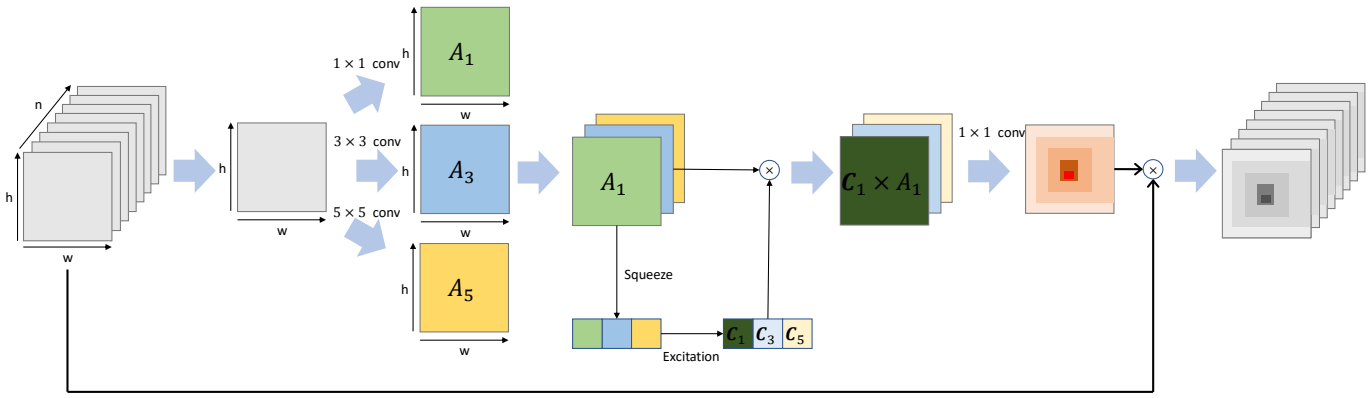

Fig. 3. Our proposed weighted, multi-scale attention module. a) the input is squeezed along the channel dimension to obtain the representation matrix; b) the representation matrix is convolved through different sized kernels to extract the intermediate attention maps consisting of different scale information; c) the intermediate attention maps are concatenated and passed through a squeeze and excitation mechanism to assign each map a weight; d) this weighted multi-scale representation is passed through a final convolutional layer to obtain an $h \times w \times 1$ representation; this representation is multiplied by the input

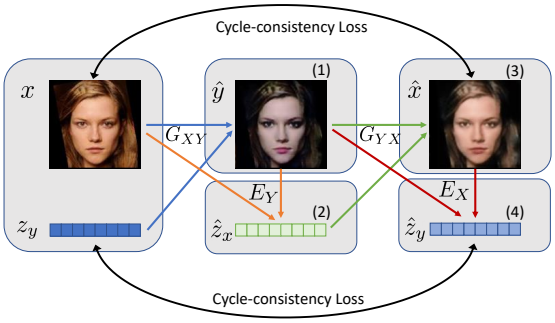

Fig. 4. An overview of the Augmented CycleGAN baseline: a four step algorithm (denoted by blue, orange, green and red arrows, respectively) to maintain cycle-consistency for both the input image x and the input latent code z_y.

A symmetric equation is simultaneously optimised in the opposite direction. In our experiments, we assign $\gamma_1 = 1$, $\gamma_2 = 0.5$. γ_1 is given a higher weight because the task to reconstruct a 32-dimensional latent code through a cycle is easier than to reconstruct an image. However, we find that assigning γ_2 a reasonably large weight encourages the network to pursue more dramatic changes, even if they are less realistic. We prefer this because unaltered images are not at all useful for practical applications. The full details of the individual losses can be found in the Appendix.

B. Multi-scale Spatial Attention

Spatial Attention aims to identify the most salient pixels. We develop a multi-scale spatial attention map to determine saliency at different granularities.

First, given a facial image, $p = (i, j), 0 \leq i \leq h, 0 \leq j \leq w$, are averaged across all channels, $s_p = \frac{1}{N}\sum_{k=1}^{N} p_k$, where N is the total number of channels to obtain a representative $h \times w$ feature map $\mathbf{Z}$.

We define the intermediate attention map, A_n, via $A_n = l_n(\mathbf{Z})$ where l_n is the convolutional layer with kernel $n \times n$.

We process $\mathbf{Z}$ with three different convolutional layers, with $1 \times 1, 3 \times 3$, and 5×5 kernels simultaneously to obtain A_1, A_3, and A_5. Smaller kernels extract more detailed fine-grain information, such as eyeliner and lipstick that may be present. Larger kernels will be more adept at learning the importance of information that may cover larger areas, such as blusher applied to the cheeks.

The intermediate multi-scale attention is obtained by concatenating the attention maps in the channel dimension,

$$A = [A_1, A_3, A_5], \tag{3}$$

and bilinearly upsampled, so A has dimensions $h \times w \times 3$.

Finally, a 1×1 convolutional layer processes A to produce the final $h \times w \times 1$ multi-scale attention map, A_{MS}, which combines information from the weighted intermediate attention maps. A_{MS} is then multiplied by the initial input; this forces the spatial attention network to learn to assign greater values to more salient pixels.

Our network is entirely self-contained and end-to-end. Unlike current state of the art, the framework is not reliant on any external models, algorithms, or software that can inhibit performance.

C. Channel Attention

Channel attention identifies which channels are most salient and weights them appropriately. As seen in Equation 3, A_1, A_3, and A_5 are concatenated along the channel dimension, so they act as three channels, which allows us to perform channel attention with reduction rate $r = 3$. Thus, the intermediate attention maps are weighted associated to the kernel size importance.

Each intermediate attention map A_k of size $h \times w$ is squeezed into a channel descriptor, $c \in \mathsf{R}$, via Global Average

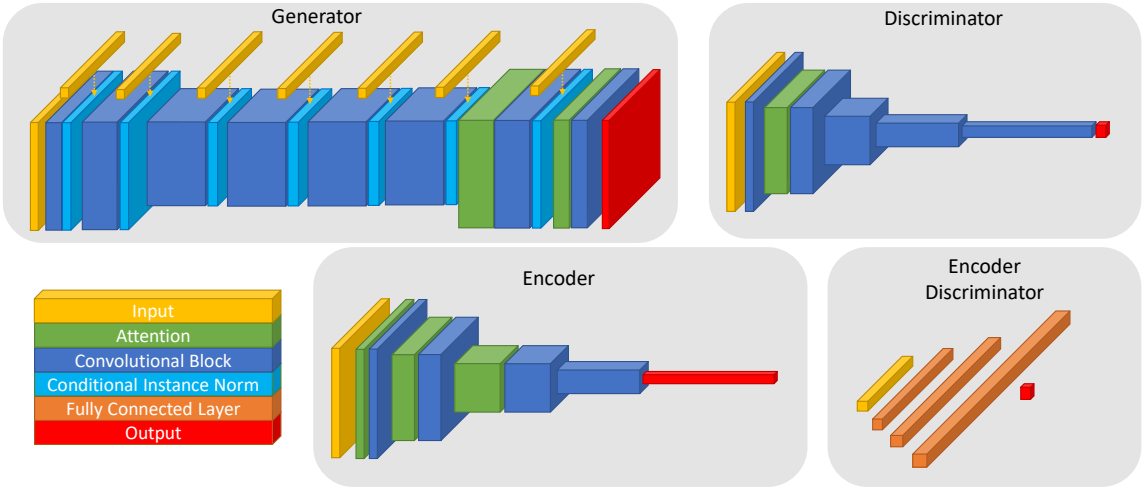

Fig. 5. The full architecture of all of our networks

Pooling. For the k-th attention map, the channel descriptor c_k is found via $c_k = \text{squeeze}(A_k) := \frac{1}{h\ w} \sum_{i=1}^{h} \sum_{j=1}^{w} A_k^{i,j}$.

As we are working with three channels, these channel descriptors form a vector $\mathbf{z} = [c_1, c_2, c_3]$. $\mathbf{z}$ is then processed by an excitation neural network that learns a non-linear interaction between channels to obtain channel importance from $\mathbf{z}$. The excitation network consists of a dimensionality reduction layer (so relatively few parameters are added to the overall network) followed by a dimensionality increasing layer,

$$\mathbf{C} = \text{excite}(\mathbf{z}) := \sigma\left(\mathbf{W_2}\delta(\mathbf{W_1}\mathbf{z})\right), \tag{4}$$

where $\mathbf{W_1} \in \mathsf{R}^{1\ \ 3}, \mathbf{W_2} \in \mathsf{R}^{3\ \ 1}$ are the parameters of the dimensionality-reduction and -increasing layers respectively, δ is the ReLU function and σ is a sigmoid activation. Note that this formulation allows multiple high-importance channels rather than a one-hot output. $\mathbf{C}$ represents the importance scores of each channel and is multiplied by the input.

D. Network Architecture

The architecture of our full system as described in Equation 1 can be found in Figure 5. Rather than applying the attention model indiscriminately, we select where to apply attention to get maximum benefit without sacrificing efficiency.

1) Encoder: We first explore the encoder as it is the network that benefits most from the attention module. The attention module is applied in early layers where the feature maps most resemble the input image. In fact, we apply attention directly to the image before feeding it into the network. We find that this significantly improves the encoder's ability to identify the makeup style to be encoded and incorporate it into the final feature representation.

2) Discriminator: The discriminator attempts to discover whether an image that has had the encoded makeup applied to it is real or fake, in the context of the makeup domain. By incorporating attention, the discriminator is more capable of identifying makeup in real images. In order to fool the discriminator, it becomes more important for the generator to apply makeup. However, if we add too much attention, or incorporate it too early, the discriminator becomes too good at identifying fake makeup images, so the generator can't fool it. The ramification is that the system attempts only to maintain cycle consistency and very little image editing is performed. We settle on one weighted multi-scale attention module after the first convolutional block.

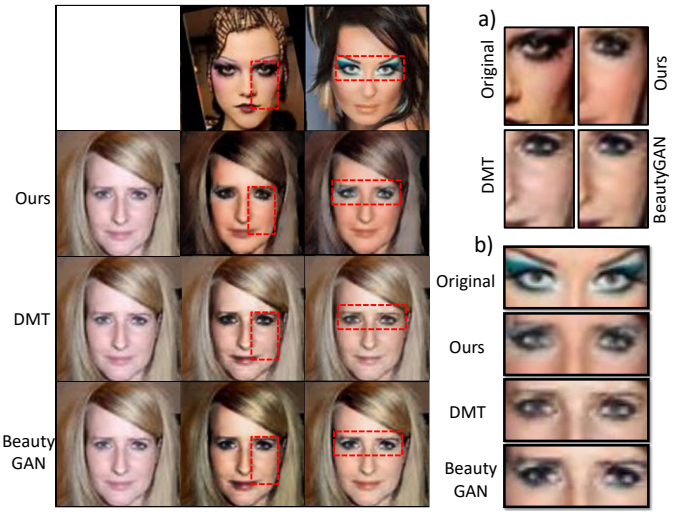

Fig. 6. Comparison with DMT and BeautyGAN on challenging makeup styles: a) our method is the only one that captures skin tone, and best approximates the colour contours in the original image; b) our method best transfers fake eyelashes and comes closest to transferring the butterfly wings.

3) Generator: The inputs are a source image and a 32-dimensional latent code obtained from the encoder. This latent code is injected into convolutional feature maps at every layer via conditional instance normalisation [24]. Because the final layers contribute most to the generated image, we incorporate attention towards the end of the image generation process - before and after the final injection of the latent code of the makeup style. The attention module before the final style

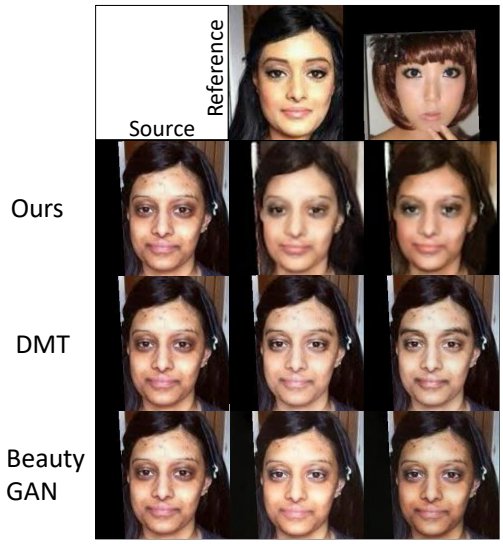

Fig. 7. Comparison with DMT and BeautyGAN demonstrating our ability to cover blemishes compared with state of the art.

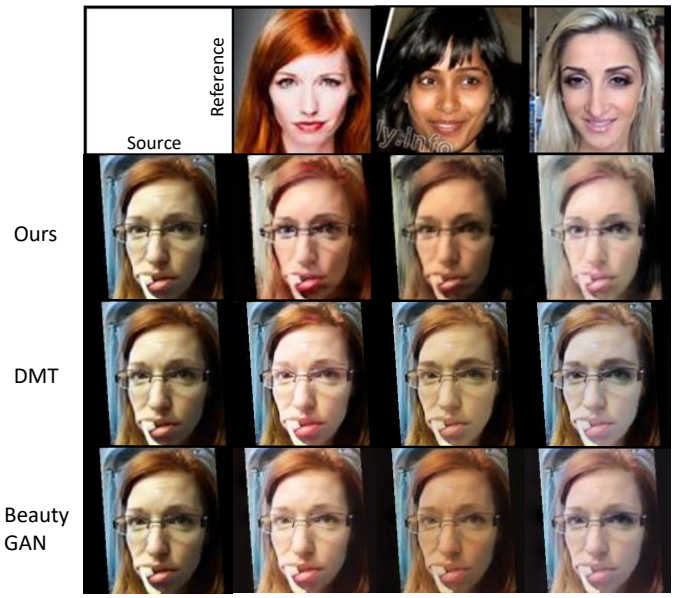

Fig. 8. Comparison on source image with occluded lips and eyes.

injection attends to areas where makeup should be applied, guiding the injected code towards those areas. The attention module at the final layer attends to important areas of the face where makeup has been applied.

IV. Results

The experiments were performed on a workstation with four NVIDIA GeForce RTX 1070 Ti GPUs with 8GB of VRAM each. The training process took around 8 hours, while the testing process was within 5 seconds.

A. Set Up

We train and evaluate our model on the FBD data set [5], a data set for makeup invariant face verification. It contains 2527 paired makeup and non-makeup images. We follow their pre-processing: the Viola Jones face detector [25] is applied to localise faces, then faces are cropped and aligned as per Shan *et al.* [26]. Pre-processing facial images to align facial landmarks is common; however, it has a propensity to introduce noise to images due to automatically scaling, skewing, cropping and zooming.

We also collect our own data set of 10 subjects from YouTube makeup tutorials, at 1080p and 144p, to allow us to perform quantitative evaluation, as explained in Section IV-C.

We compare against two state-of-the-art models: DMT [4] and BeautyGAN [2], because they are the best performing makeup style transfer frameworks that provide pre-trained models to test against.

Note that we are handicapped as DMT and BeautyGAN are trained on the MT data set [2], a lab created data set with mostly forward facing, high quality images with good lighting. In contrast, our models are trained on the low quality data set, which sometimes results in noise being added into the generated image.

B. Comparative Studies

In Figure 6, we create a challenging task to transfer radical makeup styles onto a source image. Our framework outperforms the state-of-the-art methods, best transferring the skin-tone from fake tan and powder and in the makeup images. Ours also best approximates the colour distribution of the makeup face, applying blusher on the cheeks, whereas the other methods only apply a generic pale skin tone across the entire face. Other methods only apply a bold line around the perimeter of the eyes, whereas ours accurately applies fake eyelashes. We also best succeed at transferring the unconventional butterfly wings found in the source image. This shows that our soft attention is more accurately able to adapt to unconventional and extreme styles, compared to hard attention that is only capable of handling styles similar to what it has seen during training.

In Figure 7, we demonstrate our framework's ability to cover blemishes compared with state of the art. The first reference image shows the same subject as in the source image. By applying makeup, the subject has clearly chosen to conceal the blemishes on her forehead. Our framework is the only method capable of accurately transferring the makeup style in order to cover blemishes as desired. This highlights a major flaw with current state of the art methods. Because they use hard attention to identify regions such as lips and eyes, they cannot adjust to different challenges. Our method is far more flexible in being able to handle outlier cases due to the holistic, soft attention approach.

Figure 8 provides a comparison on an image where both the lips and eyes are partially occluded. Ours best transfers the lip colour across all three images and is the only method able to apply fake lashes behind glasses in the first column. In the second column, DMT applies an unnatural pale green

TABLE I
COMPARISON WITH STATE OF THE ART METHODS

Method	Eyes	Skin	Lips	Total
BeautyGAN [2]	0.230	0.086	**0.215**	0.532
DMT [4]	0.238	**0.084**	0.218	0.541
Ours	**0.197**	0.089	0.229	**0.515**

Lower numbers are better

TABLE II
ABLATION STUDY ON ATTENTION

Method	Eyes	Skin	Lips	Total
Ours	0.197	**0.089**	**0.229**	**0.515**
w/o Multi-scale Attention	**0.188**	0.105	0.236	0.529
w/o Any Attention	0.274	0.126	0.247	0.647

Lower numbers are better

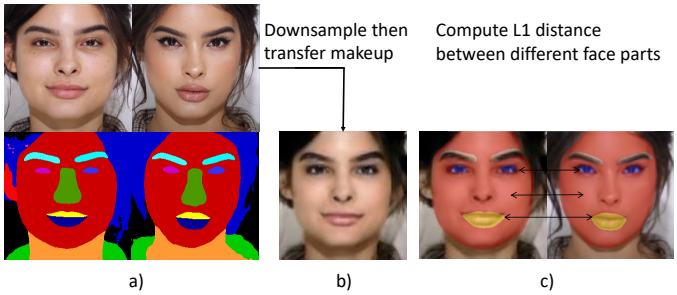

Fig. 9. We design a quantitative evaluation metric for low resolution makeup style transfer: a) extract segmentation masks from 1080p images; b) downsample images to 144p and transfer makeup style; c) apply segmentation masks to the real and fake makeup image, compute colour histograms for each face part then calculate the L1 distance between similar face parts.

tinge that, far from beautifying the image, makes the subject appear unwell.

Readers are referred to the supplementary material to see further results. Our framework can consistently apply makeup style on low-quality images without suffering from the problems that current state-of-the-art methods experience.

C. Quantitative Studies

1) Proportionate Face Distance Metric: The collected YouTube data set contains makeup and non-makeup images of subjects at 144p and 1080p. CelebAMask-HQ [27] was applied to the 1080p images (second column in Figure 2) to extract segmentation masks, and use them as ground truths of the 144p images for quantitative comparison. Each non-makeup image was then augmented with the makeup style from its own video. We did not add makeup styles from other videos to ensure that external factors, such as natural skin tone and different lighting, did not affect the results.

To perform quantitative analysis, we use the extracted masks to segment the eyes, face, and lips of the real 144p makeup images and the generated images. We then obtain the colour histograms of each segmented face part, and calculate the L1 distance, D, between colour histograms of equivalent face parts. This process is visualised in Figure 9. Because the skin takes up most of the pixels of the face, without any additional consideration, whichever method performed best at transferring skin tone would be adjudged to have performed best overall. We ensure a fair comparison by assigning a weight to each face part based on the inverse of the number of pixels that each face part has. This proportionate face distance is therefore found via

$$D_{\text{total}} = P\left(\frac{1}{p_{\text{eyes}}}D_{\text{eyes}} + \frac{1}{p_{\text{skin}}}D_{\text{skin}} + \frac{1}{p_{\text{lips}}}D_{\text{lips}}\right), \quad (5)$$

where P is the total number of pixels in the face mask of the generated image, p_i is the number of pixels in the masks of face part i, and D_i is the distance of the colour histograms of face part i, with $i \in \{\text{eyes}, \text{skin}, \text{lips}\}$.

2) Comparison with State of the Art: We compare against state-of-the-art methods on our YouTube data set in Table I. Our model outperforms the other methods at transferring makeup style on low quality images with a proportionate face distance of 0.515, 0.017 lower than the next best model. We obtain similar performance at transferring the makeup on the skin and lips, but significantly outperform both methods at transferring makeup around the eyes. Eyeliner and fake eye-lashes are usually represented on an image by a small number of pixels. The lowest scale of our attention module incorporates this information into the learned makeup style.

3) Ablation Studies: Table II shows the impact of dropping components of our attention module. Our model performs best at transferring the total face makeup, attaining the strongest performance on skin and lips. Removing multi-scale attention gives better eye makeup transfer, to the detriment of the rest of the face. Due to the presence of larger convolutional kernels, our multi-scale attention better identifies the importance of the skin and outscores regular spatial attention by 0.016.

The model without any attention at all is considerably weaker at transferring all three face parts because, without attention, the background has a large contribution on the encoded makeup style. As a result, the makeup style injected into the generator contains superfluous information.

V. CONCLUSION AND DISCUSSION

In this paper, we have developed an end-to-end framework for transferring makeup style that attains state-of-the-art performance on low-quality images. The framework does not suffer from the issues commonly seen among state-of-the-art methods, such as focusing only on lips, due to the developed novel weighted multi-scale spatial attention module.

One limitation of our method is that occasionally it can go too far with translating skin colour, and overly affect the background. Our framework favours riskier, more dramatic changes over safer ones. From an application perspective, it is more useful to have an extreme change than no change. If we wish to augment data for makeup invariant face recognition, extreme changes propose tough new challenges during training whereas little change does not assist training.

ACKNOWLEDGEMENT

The project is funded in part by the Royal Society (Ref: IES\R2\181024 and IES\R1\191147).

REFERENCES

[1] H. Chang, J. Lu, F. Yu, and A. Finkelstein, "PairedCycleGAN: Asymmetric style transfer for applying and removing makeup," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2018, pp. 40–48.

[2] T. Li, R. Qian, C. Dong, S. Liu, Q. Yan, W. Zhu, and L. Lin, "BeautyGAN: Instance-level facial makeup transfer with deep generative adversarial network," in *Proceedings of the 26th ACM International Conference on Multimedia*, ser. MM '18. New York, NY, USA: Association for Computing Machinery, 2018, p. 645–653.

[3] H. Chen, K. Hui, S. Wang, L. Tsao, H. Shuai, and W. Cheng, "Beautyglow: On-demand makeup transfer framework with reversible generative network," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019, pp. 10 034–10 042.

[4] H. Zhang, W. Chen, H. He, and Y. Jin, "Disentangled makeup transfer with generative adversarial network," *arXiv preprint arXiv:1907.01144*, 2019.

[5] L. Zhang, H. P. H. Shum, L. Liu, G. Guo, and L. Shao, "Multiview discriminative marginal metric learning for makeup face verification," *Neurocomputing*, vol. 333, pp. 339–350, 2019.

[6] J. Li, C. Xiong, L. Liu, X. Shu, and S. Yan, "Deep face beautification," in *Proceedings of the 23rd ACM international conference on Multimedia*, 2015, pp. 793–794.

[7] R. Geirhos, C. R. M. Temme, J. Rauber, H. H. Schütt, M. Bethge, and F. A. Wichmann, "Generalisation in humans and deep neural networks," in *Advances in Neural Information Processing Systems 31*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds., 2018, pp. 7538–7550.

[8] S. Menon, A. Damian, S. Hu, N. Ravi, and C. Rudin, "PULSE: Self-supervised photo upsampling via latent space exploration of generative models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 2437–2445.

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

[10] B. M. Smith, L. Zhang, J. Brandt, Z. Lin, and J. Yang, "Exemplar-based face parsing," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3484–3491.

[11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[12] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, "Residual attention network for image classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3156–3164.

[13] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[14] W. Li, X. Zhu, and S. Gong, "Harmonious attention network for person re-identification," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[15] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," in *International Conference on Machine Learning*, 2019, pp. 7354–7363.

[16] H. Tang, H. Liu, D. Xu, P. H. S. Torr, and N. Sebe, "AttentionGAN: Unpaired image-to-image translation using attention-guided generative adversarial networks," 2019.

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

[18] L. Xu, Y. Du, and Y. Zhang, "An automatic framework for example-based virtual makeup," in *2013 IEEE International Conference on Image Processing*, Sep. 2013, pp. 3206–3210.

[19] C. Li, K. Zhou, and S. Lin, "Simulating makeup through physics-based manipulation of intrinsic image layers," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015, pp. 4621–4629.

[20] S. Liu, X. Ou, R. Qian, W. Wang, and X. Cao, "Makeup like a superstar: Deep localized makeup transfer network," *2016 International Joint Conference on Artificial Intelligence (IJCAI)*, 2016.

[21] J.-Y. Zhu, R. Zhang, D. Pathak, T. Darrell, A. A. Efros, O. Wang, and E. Shechtman, "Toward multimodal image-to-image translation," in *Advances in neural information processing systems*, 2017, pp. 465–476.

[22] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1125–1134.

[23] A. Almahairi, S. Rajeswar, A. Sordoni, P. Bachman, and A. Courville, "Augmented cycleGAN: Learning many-to-many mappings from unpaired data," *arXiv preprint arXiv:1802.10151*, 2018.

[24] V. Dumoulin, J. Shlens, and M. Kudlur, "A learned representation for artistic style," *ICLR*, 2017.

[25] P. Viola and M. J. Jones, "Robust real-time face detection," *International journal of computer vision*, vol. 57, no. 2, pp. 137–154, 2004.

[26] S. Shan, Y. Chang, W. Gao, B. Cao, and P. Yang, "Curse of mis-alignment in face recognition: problem and a novel mis-alignment learning solution," in *Sixth IEEE International Conference on Automatic Face and Gesture Recognition, 2004. Proceedings.* IEEE, 2004, pp. 314–320.

[27] C.-H. Lee, Z. Liu, L. Wu, and P. Luo, "MaskGAN: Towards diverse and interactive facial image manipulation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

APPENDIX

A. Loss Functions

Generator Losses: The loss for the image generator is similar to a typical conditional GAN.

$$\mathcal{L}^{Y}_{\text{GAN}}(G_{XY}, D_Y) = \mathop{\mathsf{E}}_{\substack{y \\ p_d(y)}} \Big[\log D_Y(y)\Big] + \mathop{\mathsf{E}}_{\substack{x \\ z_y}} \mathop{\mathsf{E}}_{\substack{p_d(x) \\ p(z_y)}} \Big[\log(1 - D_Y(G_{XY}(x, z_y)))\Big], \tag{6}$$

while the loss for the encoder network, which generates a 32-dimensional latent code from input faces is

$$\mathcal{L}^{Z_X}_{\text{GAN}}(E_X, G_{XY}, D_{Z_X}) = \mathop{\mathsf{E}}_{\substack{z_x \\ p(z_x)}} \Big[\log D_{Z_X}(z_x)\Big] + \mathop{\mathsf{E}}_{\substack{x \\ z_y}} \mathop{\mathsf{E}}_{\substack{p_d(x) \\ p(z_y)}} \Big[\log(1 - D_{Z_X}(\tilde{z}_x))\Big], \tag{7}$$

where $\tilde{z}_x = E_X(x, G_{XY}(x, z_y))$. Other symbols are defined in preliminaries of Section III-A.

Cycle-consistency Loss: Both losses have an associated cycle-consistency restraint. For image generation, the loss is similar to cycle-consistency loss of CycleGAN.

$$\mathcal{L}^{X}_{\text{CYC}}(G_{XY}, G_{YX}, E_X) = \mathop{\mathsf{E}}_{\substack{x \\ z_y}} \mathop{\mathsf{E}}_{\substack{p_d(x) \\ p(z_y)}} \big|\big|\tilde{x} - x\big|\big|_1, \tag{8}$$

where $\tilde{x} = G_{YX}\big(\tilde{y}, E_X(x, \tilde{y})\big)$ and $\tilde{y} = G_{XY}(x, z_y)$.

For the encoder, we reconstruct makeup style z_y via

$$\mathcal{L}^{Z_Y}_{\text{CYC}}(G_{XY}, E_Y) = \mathop{\mathsf{E}}_{\substack{x \\ z_y}} \mathop{\mathsf{E}}_{\substack{p_d(x) \\ p(z_y)}} \big|\big|\tilde{z}_y - z_y\big|\big|_1, \tag{9}$$

where $\tilde{z}_y = E_Y(x, G_{XY}(x, z_y))$. Once more, symbols are defined in Section III-A.