PAPER

FSAMT: Face Shape Adaptive Makeup Transfer

Haoran LUO[†a)], **Tengfei SHAO**[†], **Shenglei LI**[†], ***Nonmembers***, ***and*** **Reiko HISHIYAMA**[†], ***Member***

SUMMARY Makeup transfer is the process of applying the makeup style from one picture (reference) to another (source), allowing for the modification of characters' makeup styles. To meet the diverse makeup needs of individuals or samples, the makeup transfer framework should accurately handle various makeup degrees, ranging from subtle to bold, and exhibit intelligence in adapting to the source makeup. This paper introduces a "3-level" adaptive makeup transfer framework, addressing facial makeup through two sub-tasks: 1. Makeup adaptation, utilizing feature descriptors and eyelid curve algorithms to classify 135 organ-level face shapes; 2. Makeup transfer, achieved by learning the reference picture from three branches (color, highlight, pattern) and applying it to the source picture. The proposed framework, termed "Face Shape Adaptive Makeup Transfer" ($FSAMT$), demonstrates superior results in makeup transfer output quality, as confirmed by experimental results.

key words: *makeup transfer, GAN, face classification, style transfer*

1. Introduction

Despite the pandemic, consumer behavior has expanded the beauty market. Projections [1] estimate the global makeup market growing from $41.85 billion in 2022 to $61.34 billion by 2029. In the retail-beauty nexus, customers often try makeup at stores. However, this poses issues: 1. Time: Visits and trials consume hours; 2. Expertise: Most consumers lack professional knowledge on suitable product combinations.

Makeup transfer is a solution which can swiftly display makeup effects. SCGAN [2] addresses makeup transfer's spatial alignment using two extraction and one assignment modules; BeautyGAN [3] uses pixel-level histogram loss for quality generation; BeautyGlow [4] decomposes face picture latent vectors into makeup-specific vectors; PSGAN [5] offers shadow-controllable transfer; CA-GAN [6] presents a quantitative makeup style analysis; RamGAN's [7] attention module permits makeup transfer for varying poses. Yang et al. [8] introduced an Illumination-Aware image decomposition method, which can utilize 3D morphable models through regression-based inverse rendering to extract coarse materials; With EleGANt [9], a new Sow-Attention module is introduced, which utilizes attention within shifted and overlapping windows to decrease computational expenses.

However, methods above lack extreme-style makeup transfer ability. LADN [10] employs local adversarial discriminators for detail transfer; Nguyen [11] uses separate branches for patterns and color transfer. While effective, challenges like side profiles, strong lighting, or substantial source-reference differences can lead to uneven results.

This paper introduces an adaptive makeup transfer framework adept at steadily handling extreme makeup. We tackle two main tasks: 1. "Makeup Adaptation" simulates 135 facial classifications based on five face shapes and 27 eye shapes, using feature descriptors and eyelid curves. It then adapts professional makeup from matching facial types; 2. For "Makeup Transfer," three branches (color, highlights, and patterns) are employed. The color branch features an adaptive binning method and a refined discriminator, reducing distortions. The highlight branch is supported by a new dataset, Highlight Face Dataset (HFD). The pattern branch utilizes a cutting-edge feature extractor, Resnet50MultiScale(R50) [12] - Vision Transformer(ViT) [13] - Wasserstein Domain Adaptation (WDA) [14] (termed RMW). Through this framework, whether it is stickers, tattoos, metal decorations, or ordinary eyeshadows, highlights, blushes, and mascaras, they can be accurately transferred to the target picture. Our framework is named as "Face Shape Adaptive Makeup Transfer" ($FSAMT$), consistently outperforms peers in tests and user survey. Figure 1 depicts the framework, in this pipeline, the adaptive module selects the most suitable makeup for the user based on our face shape matching algorithm, while the makeup transfer module ensures precise, efficient, and detailed transfer results. These steps not only guarantee optimal makeup effects but also significantly reduce makeup trial costs, making it easily applicable in various commercial scenarios. At the same time, although this model is outstanding in its ability to transfer heavy style makeup, due to the improvement of the color migration module, it also has reliable adaptability to common makeup.

2. Related Work

Prior to makeup transfer, selecting appropriate makeup for a target is crucial. Few studies address makeup transfer adaptation. Jiang [5] introduced "makeup customization," emphasizing transfer flexibility over customization to facial attributes. Although research has tackled facial shape classification by extracting facial points and categorizing shapes, basing makeup adaptability solely on face shape is incomplete. Makeup choices also hinge on facial organ charac-

Manuscript received October 13, 2023.
Manuscript revised January 10, 2024.
Manuscript publicized April 1, 2024.

[†]Graduate School of Creative Science and Engineering, Waseda University, Tokyo, 169–8555 Japan.
a) E-mail: tinywheel@fuji.waseda.jp

DOI: 10.1587/transinf.2023EDP7212

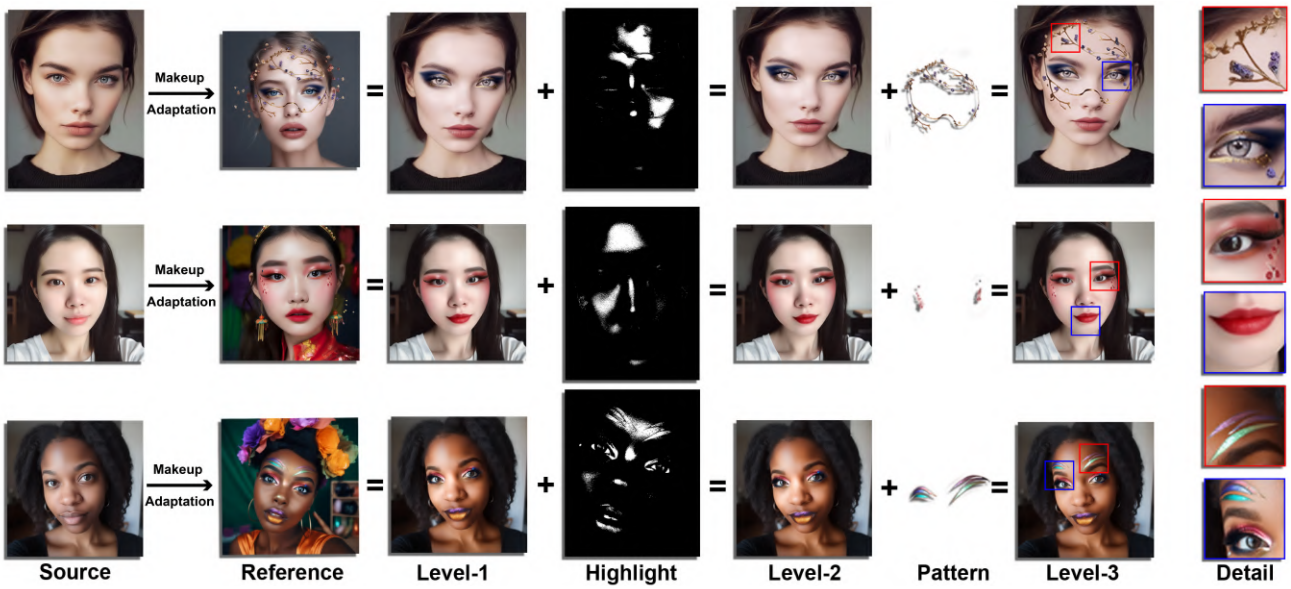


Fig. 1 Face shape adaptive makeup transfer (FSAMT) pipeline

teristics, with eye makeup being notably intricate. Zhang Wei [15] identified eyes into four categories considering size, eyelid form, and corners. Despite varied methodologies in other studies, eye shape categorization remains narrow.

In makeup adaptation section, we developed SEDNet: a blend of Squeeze-and-Excitation Network (SENet) and DenseNet [16] for facial shape classification. The SENet functions as an attention mechanism. This network processes a normalized vector from PRNet to discern five face shapes. We also categorized 27 eye shapes via three parameters: length, roundness, and tilt. Eyelid contours are then determined mathematically. Combining five face and 27 eye shapes, we offer 135 organ-level classifications, enriching makeup adaptation.

In makeup transfer section, we presented an integrated makeup transfer module. Models such as SCGAN [2], BeautyGAN [3], BeautyGlow [4], PSGAN [5], CA-GAN [6], ElegANt [9] have been verified to have some degree of effect on the color or pattern of makeup transfer, but unlike all of them, our input is not manually provided but recommended by the adaptation module. Here, we emphasized capturing facial patterns and refined features, like highlights. We produced pictures with synthetic highlight effects and their respective binary annotations as training for the highlight branch. The three transfer branches operated concurrently, converging into a single output. In addition, by using methods like modifying the histogram matching function and introducing adversarial learning to train feature extractors, $FSAMT$ also has more accurate transfer capabilities compared to other models.

3. Makeup Adaptation Model

In the Makeup Adaptation model, the categorization of the source picture P_S and the reference picture P_R is performed at the organ-level face shapes. Consequently, the matching of the source and reference pictures is executed based on their respective classifications. This procedure encompasses two parallel modules, namely the classification of face shapes and eye shapes. Given the existence of 27 eye shapes and 5 face shapes, the number of categories amounts to a total of 135, calculated as the product of 27 and 5.

3.1 Face Shape Classification

In the task of classifying facial and eye shapes, the primary goal is to extract facial feature points. In this chapter, we used PRNet [17] to extract the 3D coordinates of the face. The PRNet input is a 256*256 RGB picture, and the output is a 3D position map of the same resolution. Each position has a 3D coordinate (x, y, z), which in this paper is a three-dimensional feature vector of (256,256,3).

Next, a neural network called SEDNet was constructed for the purpose of predicting face shape. Following a convolution operation, the feature vector obtained from PRNet will serve as the input for the SED block. Within each SED block, there are multiple SED layers, which encompass a dense layer and an SE layer, as depicted in Fig. 2. In the context of each SED layer, the fundamental unit consisting of BN-ReLU-Conv constitutes the non-linear function of the dense layer. The computation process is illustrated by Eq. (1):

$$x_l = H_l(x_0, x_1, \ldots, x_{l-1}) \tag{1}$$

where H_l is a nonlinear function, x_l represents the output of the l-th Dense Block, and x_0 represents the input. In this way, the SED layer uses the feature maps of all previous layers to help the feature map extraction of the current layer, which can enhance the feature reuse ability of the model and make the model more compact and effective.

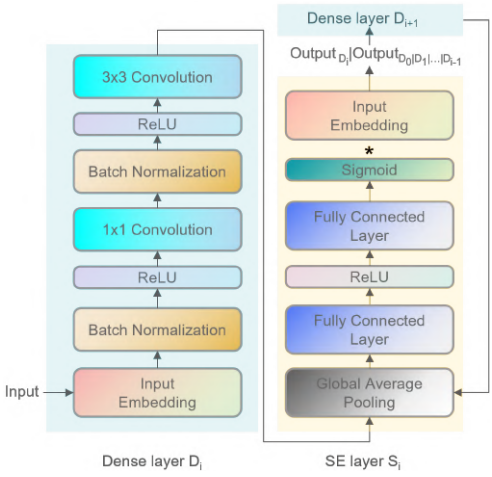


Fig. 2 SED layer of SEDNet

In the training phase, due to the lack of a specific dataset for face shape classification, we utilized Baidu's cloud-based face recognition service to assemble 5,000 aligned facial images from the CelebA dataset [18], evenly distributed across five face shapes. After manual verification, these images were divided in a 7:3 ratio for training and testing. Our SEDNet model demonstrated superior performance, achieving 95.7% accuracy, outperforming conventional machine learning and basic neural network approaches in face shape classification.

3.2 Eye Shape Classification

The basic method of eye shape classification [19], [20] is as follows: Based on the feature points of human eyelids, the contour curves of the upper and lower eyelids are fitted, and different eye shapes are drawn by the eyelid contour curve equation.

3.2.1 Obtain Eyelid Sampling Points

After retaining 12 key points for a single eye as EF_i, we derive the Eyelid Sampling Points (P_i). We connect these points sequentially to form an eye contour polygon. Using the pupil point EF_{12} as a center, and drawing a line parallel to the X-axis through it as the 0 ° reference, we divide the circle into 48 equal segments, yielding sampling points P_i ($i \in [0,47]$). The line connecting inner corner EF_0 and outer corner EF_6 segments the sampling points into upper eyelid points ($i \in [0,m]$) and lower eyelid points ($i \in [m+1,47]$). Figure 3 illustrates the sampling points on the upper and lower eyelids.

3.2.2 Divide Eye Shapes

For the eye feature points EF_i ($i \in [0, 11]$) defined in the previous step, each contains a set of coordinate vectors (X_{EF_i}, Y_{EF_i}), the shape parameters of the following four eyes are obtained:

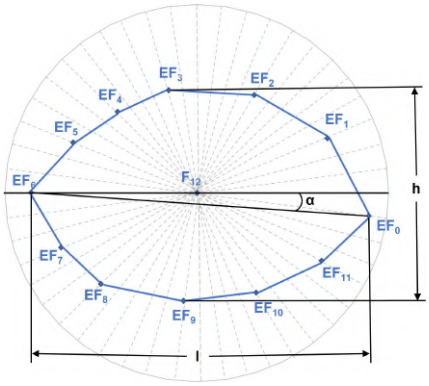


Fig. 3 Sampling points of upper and lower eyelid curves

Table 1 Classification and range of eye shape parameters

Shape Parameter	Semantic Description		
l	l<0.038 short eye	0.038≤l<0.086 average eye	l≥0.086 long eye
d	d<0.455 leaf shape eye	0.455≤d<0.492 oval eye	d≥0.492 round eye
α	α<0.305 drooping eye	0.305≤α<0.364 horizontal eye	α≥0.364 upper squint eye

$$
\begin{aligned}
l &= X_{EF_0} - X_{EF_6} \\
h &= \max\left(Y_{EF_1}, \ldots, Y_{EF_5}\right) - \min\left(Y_{EF_7}, \ldots, Y_{EF_{11}}\right) \\
d &= \frac{h}{l} \\
\alpha &= \arctan\left[\left(Y_{EF_6} - Y_{EF_0}\right) / \left(X_{EF_0} - X_{EF_6}\right)\right]
\end{aligned}
\tag{2}
$$

In the Eq. (2): X_{EF_i} represents the X coordinate of the i-th feature point EF_i representing the eye; Y_{EF_i} represents the Y-coordinate of the i-th feature point EF_i representing the eye ; l is the horizontal distance between the outer corner point and the inner corner point, indicating the length of the eye; h is the maximum longitudinal distance of the eye, indicating the height of the eye; d is the ratio of the height and length of the eye, indicating the degree of ellipse of the eye; α is the tilt angle of the eyes.

We capture 500 pictures of eyes from the celebA and calculate four distinct shape parameters for each eye. Simultaneously, we divide the semantic description of the three shape parameters, namely l, d, and α. The parameter l_i encompasses the following: short eye (l_0), average eye (l_1), and long eye (l_2). The parameter d encompasses the following: leaf-shaped eye (d_0), oval eye (d_1), and round eye (d_2). Lastly, α_i encompasses the following: drooping eye (α_0), horizontal eye (α_1), and upper squint eye (α_2). In total, there are nine shape parameters. Subsequently, the category C_i [l_i, d_i, α_i] ($i \in [0,2]$) to which each eye belongs is determined through manual judgment. The numerical range for each shape parameter is obtained by averaging the eye parameters within each category of the nine shape parameters, as depicted in Table 1. Considering that each eye shape parameter is associated with three distinct semantic descriptions, a

total of 27 different eye shapes can be derived (3*3*3).

3.2.3 Fitting the Eyelid Contour Curve

In this paper, it is stipulated that the set S of human eye pictures to be predicted is shown in Eq. (3).

$$S = \begin{pmatrix} S_{11} & S_{21} & \dots & S_{i1} \\ S_{12} & S_{22} & \dots & S_{i2} \\ \vdots & \vdots & \vdots & \vdots \\ S_{1j} & S_{2j} & S_{3j} & S_{ij} \end{pmatrix} \tag{3}$$

Among them, $i \in [0,26]$ corresponds to 27 eye shapes; $j \in [0,N]$, N represents the number of pictures of each eye shape. And for each eye picture S_{ij} to be classified, it can be expressed in the form of Eq. (4).

$$S_{ij} = (P_0, P_1, P_2 \dots P_m, P_m + 1, \dots P_{47}) \tag{4}$$

Among them, P_0-P_m represents the feature point of the upper eyelid, P_{m+1}-P_{47} represents the feature point of the lower eyelid, and for each eyelid feature point P_i, it can be expressed as $P_i = (x_i, y_i)$. Among them, x_i and y_i represent the abscissa and ordinate of the feature point respectively.

To model the eyelid contour, the least square method (LSM) is used to fit a horizontal quadratic curve. The basic functions for fitting both eyelids are represented by: $\Phi = \text{span}\left\{\varphi_0(x) = 1, \varphi_1(x) = x, \varphi_2(x) = x^2, \cdots, \varphi_n(x) = x^n\right\}$.

Given n<m and n<48-m, the eyelid contour equation is derived. Using upper eyelid sampling points P_i (i = 0 to m), inner and outer corner points EF_0 and EF_6, along with lower eyelid sampling points (i = m+1 to 47) and the same corner points, we fit the curve to closely match the sampling points (as in Eq. (5)).

$$S(x) = \sum_{t=0}^{n} \beta_t \varphi_t(x) \tag{5}$$

After experiments, we set the fitting degree as n=4 in this paper. The curve obtained by fitting satisfies the requirement of passing through the inner and outer corner points. $\beta_t (t = 0, 1, \dots, 4)$ represents the coefficients of the curve equation, and $\varphi_t(x)$ is the basic function with an exponential term of t. The β_t of each eye shape can be obtained by bringing the inner and outer corner points (EF_0 and EF_6) and eyelid sampling points of the 27 eye shapes into the normal equation. Figure 4 shows an example when $0.305 \le \alpha < 0.364$ (horizontal eye) is fixed, the fitting result of adjusting roundness in long eye.

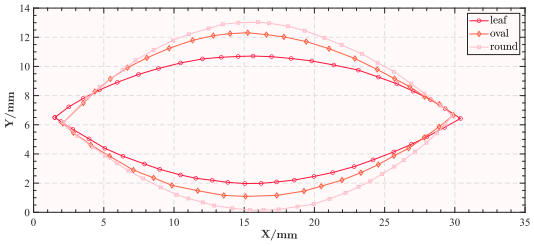


Fig. 4 Adjust roundness in long eye

In all source picture sets P_S and reference picture sets P_R, based on 135 different organ-level face shape classification algorithms, the face shape of each picture in P_S and P_R is calculated respectively. Use P_{S_i}, P_{R_i} ($i \in [0,134]$) to denote the picture collections belonging to the i-th face shape in P_S and P_R, respectively. Then put P_{S_i}, P_{R_i} ($i \in [0,134]$) under the directory F_i. Therefore, for each face category F_i, there is $F_i = (P_{S_i}, P_{R_i})$.

4. Makeup Transfer Model

In the "Makeup Transfer" section, we divide the learning of all makeup attributes into three branches (color, highlight, and pattern). We continue to utilize the PRNet approach for UV mapping while incorporating bilinear interpolation to optimize texture sampling. This approach delivers smoother results during sampling, reducing distortions and blurriness in edge regions.

The overall architecture of makeup transfer is shown in Fig. 5. For the input source picture P_{S_i} and reference picture P_{R_i}, use "-" and "+" to mark "non-makeup" and "makeup-applied" respectively. Then $P_{S_i}^-$ and $P_{R_i}^+$ can be obtained. The objective function $\mathcal{M}$ of makeup transfer can be expressed as Eq. (6):

$$\mathcal{M}\left(P_{S_i}^-, P_{R_i}^+\right) = P_{S_i}^+ \tag{6}$$

$P_{S_i}^+$ is the source picture of the consistent makeup effect obtained through migration from the adapted reference.

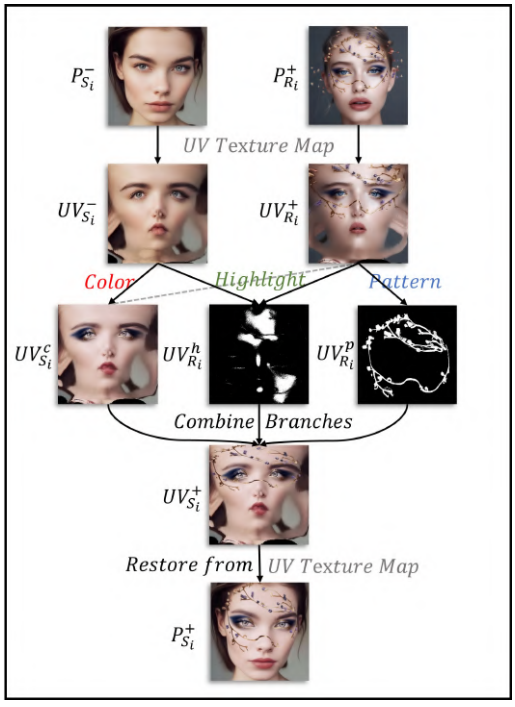


Fig. 5 Makeup transfer architecture

$P^-_{S_i}$ and $P^+_{R_i}$ are first converted into UV texture maps $UV^-_{S_i}$, $UV^+_{R_i}$. Then, pass $UV^-_{S_i}$, $UV^+_{R_i}$ to three parallel branches for color, highlight and pattern transmission respectively. The processing results of the three branches $UV^c_{S_i}$, $UV^h_{R_i}$, $UV^p_{R_i}$ are further combined into an output texture map $UV^+_{S_i}$, and finally through UV^{-1} operation restores $UV^+_{S_i}$ to $P^+_{S_i}$.

4.1 Color Transfer

In this paper, we propose a novel color transfer approach for makeup transfer tasks, utilizing an improved GAN architecture and incorporating additional features to enhance the performance of the model.

Our approach is derived from the concept behind BeautyGAN, however, we have introduced numerous novel elements to enhance the outcomes. The primary element of our methodology is a makeup swapping network C that operates based on color, facilitating the exchange of makeup colors on cosmetic regions between the source and reference images, denoted as $UV^C_{S_i}, UV^C_{R_i} := C\left(UV^-_{S_i}, UV^+_{R_i}\right)$. In order to train network C, we utilize a loss function that is similar to the one employed in BeautyGAN, but with several innovative modifications: 1. We replace the VGG-16 [21] model with a VGG-19 model in order to more effectively capture facial features during perceptual loss calculation; 2. We develop a more resilient discriminator to enhance the effectiveness of the adversarial loss; 3. We improve the histogram matching function by incorporating adaptive binning, thereby enabling a more accurate matching of color distributions.

The original paper used Spectral Normalization to improve the discriminator of GAN. This paper further optimizes the approach by proposing an improved operational flow combining Gradient Penalty and Adaptive Spectral Normalization for the discriminator involved in the Adversarial Loss L_a. The steps are as follows:

1. Apply spectral normalization after each convolutional layer of the discriminator. To achieve Adaptive Spectral Normalization, the normalization factor needs to be dynamically adjusted based on the loss value during the training process. Specifically, the normalization factor is multiplied by the reciprocal of the loss value for adaptive adjustment.

2. Calculate the gradient penalty during the training process. We first generate a random weight θ within the range of 0 and 1. Then, generate an interpolated sample using the following Eq. (7):

$$x_{interpolated} = \theta \cdot x_{real} + (1 - \theta) x_{fake} \tag{7}$$

Where x_{real} is a sample from the real data distribution, and x_{fake} is a sample generated by the generator.

3. Calculate the gradient of the interpolated sample $x_{interpolated}$: $g = \nabla D(x_{interpolated})$, where D denotes the discriminator.

4. Calculate the gradient penalty: $L_{GP} = (\|g\|_2 - 1)^2$, where $\|g\|_2$ is the L2 norm of the gradient.

5. Add the gradient penalty term to the loss function of the discriminator(Eq. (8)):

$$L_{a_{total}} = L_a + \lambda L_{GP} \tag{8}$$

Where L_a is the original discriminator loss, and λ is the hyperparameter controlling the weight of the gradient penalty. Finally, train the discriminator with the new total loss function $L_{a_{total}}$.

Compared with original method, by applying gradient penalty and adaptive spectral normalization, the risk of gradient explosion or vanishing in the discriminator output can be reduced by constraining the gradient. Meanwhile, a higher convergence rate can be achieved while maintaining training stability.

In our research, we advanced the conventional histogram matching approach by adapting the number of bins to better capture the color distribution of images, especially where distributions are non-uniform. Utilizing Contrast Limited Adaptive Histogram Equalization (CLAHE) [22], we initially analyze the color histograms of both the source and reference images for local color adaptation. We then apply K-means clustering for adaptive binning, grouping colors based on their distribution. The optimal number of clusters (K) is determined using the elbow method, confined within a range of 20 to 40 to balance detail and computational efficiency, as established through extensive testing.

For color mapping, we align the cumulative density functions (CDF) of both images, modifying the source image ' s pixels to match the reference image's color profile. This adaptive binning approach leads to more accurate color matching, enhancing the realism in makeup transfer. Our model demonstrates superior performance in both detail and accuracy, as evidenced in our experimental results.

4.2 Highlight Transfer

A picture's tonality is shaped by exposure, highlights, shadows, and extremes of white and black. Despite sharing a color preset, pictures vary in perceived brightness. Thus, for realistic makeup replication, facial highlights are crucial. This paper introduces a highlight capture branch. Post UV mapping, the texture map $UV^+_{R_i}$ is channeled to this branch. Here, a U-Net discerns and isolates the highlight effect from both source and reference pictures: $UV^h_{R_i} := P(UV^+_{R_i})$. Ultimately, $UV^h_{R_i}$ is merged pixel-wise with outputs from the color and pattern branches.

To train our highlight capture branch, we crafted the Highlight Face Dataset (HFD) using assorted makeup face pictures from Makeup Datasets and Pixels. After standardizing their sizes, we utilized facial landmarks to deduce highlight regions like the nose bridge and cheekbones. We then superimposed synthetic specular highlights onto these areas, yielding 2000 enhanced pictures.

For each synthetically highlighted picture, a binary segmentation annotation was created: highlighted regions were white (value 1), while the rest was black (value 0). These annotations acted as the "ground truth" for the highlight capture branch's training. To streamline labeling, we implemented auto-labeling techniques such as highlight recogni-

tion, binarization, morphological operations, and Connected Component Analysis (CCA). The procedure is illustrated in Fig. 6.

First, we apply a Laplacian filter to compute pixel second-order derivatives, identifying picture highlight edges. The filtered pictures are then binarized using Otsu's method, retaining key highlight edges. Morphological operations, specifically dilation, are then employed for data augmentation to refine the highlight regions by reducing noise and filling gaps. Lastly, using CCA, we label the highlight regions in the processed binary picture, grouping like-valued adjacent pixels. We utilized the Two-Pass algorithm for this, resulting in labeled pictures where each highlight area has a distinct label. This procedure efficiently produces a binary segmentation annotation for pictures with highlight effects.

4.3 Pattern Transfer

In the pattern transfer branch, we aim to transfer detailed patterns, like facial stickers and metal decorations, from reference to source pictures. To extract the binary segmentation mask $UV^p_{R_i}$ from the texture map $UV^+_{R_i}$, we employ the RMW model. Central to this is a feature extractor adept at multi-scale feature fusion and unsupervised domain adaptation. Through W-DA, the extractor discerns the feature mapping relationship across various domains, enhancing its analytical strength. We also used tools like Midjourney to generate portraits with extreme makeup styles, complementing the CPM-Synt-1 and CPM-Synt-2 datasets.

Table 2 shows the basic architecture of the RMW model, which is mainly divided into four functional modules. In the branch of pattern migration, their respective uses and connections with each other are as follows:

1. Resnet50MultiScale: Encoder. Responsible for extracting multi-scale features from input pictures. This means extracting multi-scale feature representations from face pictures with certain makeup to capture different levels of details. In the pattern transfer task, this will help to capture the "local features" of makeup.

2. MultiScaleViT: This part is based on Vision Transformer's multi-scale feature fusion. We first input the four-layer feature maps extracted by Resnet50MultiScale into the pretrained ViT model. The ViT model performs global feature extraction on the feature maps of each layer. Then, we concatenate these global features with the previously extracted feature maps along the channel dimension. This helps to establish connections between features at different scales, so that the makeup transfer task can more accurately acquire and understand the "global features" C_{F_i} of makeup.

3. FPN [23]: Decoder. After receiving the feature representation C_{F_i} from the previous layer, for use within the FPN, we temporarily decompose it into four feature maps of different scales. After processing through the FPN, a final 1*1 convolutional layer maps M to the output channel count (number of categories) and a per-pixel sigmoid activation function is applied to achieve the purpose of generating high-resolution segmentation masks.

4. Wasserstein Domain Adaptation (W-DA): Within makeup transfer tasks, disparities can exist between the source (makeup-applied faces) and target (non-makeup faces) domains. We introduce W-DA to bridge this gap, serving as an auxiliary module to enhance feature extraction using adversarial learning. This process employs a domain discriminator, a neural network with several fully connected layers. During training, feature representations C_{F_i} from both domains are input to the discriminator, which minimizes the Wasserstein distance between them. A Gradient Reversal Layer (GRL) is used, multiplying gradients negatively before backpropagation, urging the extractor to mislead the discriminator and blur domain distinctions.

$$UV^+_{S_i} = UV^+_{R_i} \odot UV^p_{R_i} + UV^c_{S_i} \odot \left(1 - UV^p_{R_i}\right) + UV^+_{R_i} \odot \left(UV^h_{R_i}\right) \quad (9)$$

$$I^m_s = \mathcal{UV}^{-1}\left(PM, UV^+_{S_i}\right) \quad (10)$$

To obtain the final picture, we use Eqs. (9)–(10). Equation (9)'s left side represents the desired combined UV texture map. The right side has three components: 1. $UV^+_{R_i} \odot UV^p_{R_i}$: This denotes element-wise multiplication of the reference makeup pattern with the pattern mask, overlaying the reference pattern onto specific source picture areas; 2. $UV^c_{S_i} \odot \left(1 - UV^p_{R_i}\right)$: Here, the color-transferred texture map is applied to the source picture's remaining areas. The term $(1 - UV^p_{R_i})$ computes the complement of the pattern mask, indicating areas without the reference makeup pattern; 3. $UV^+_{R_i} \odot UV^h_{R_i}$: This adds the reference makeup's

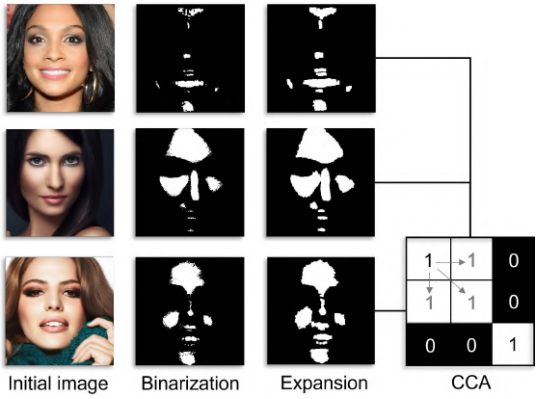


Fig. 6 Overview of highlight transfer

Table 2 Overview of RMW

Module	Input	Output
Resnet50MultiScale	Picture($UV^+_{R_i}$)	FeatureMaps($F_4 - F_7$)
MultiScaleViT	FeatureMaps($F_4 - F_7$)	ConcatenateFeatures(C_F)
FPN	ConcatenateFeatures(C_F)	SegmentMask($UV^p_{R_i}$)
W-DA	Picture($UV^+_{R_i}$), ConcatenateFeatures(C_F)	DomainLoss(L_D, L_F)

highlight areas to the source picture by multiplying the UV texture map of the reference makeup with the highlight areas' binary mask.

Merging these components results in a UV texture map $UV_{S_i}^{+}$ with the transferred makeup and highlights for the source pictures. Inputting the UV position map PM gives the final output I_s^m after the inverse operation (Eq. (10)).

5. Experiment

5.1 Experiment Parameters

We choose to train each branch independently based on the following two considerations: First, each branch is unique in its functions and tasks, such as color transfer, highlight transfer, and pattern transfer. To ensure that each branch achieves optimal performance on its specific task, we decided to train them independently. Second, independent training can avoid the possible adverse effects of the training results of one branch on another branch. After all three branches are trained independently, we use an ensemble strategy to integrate the outputs of all branches to achieve the final makeup transfer.

Due to having three independent makeup transfer branches, we have configured different training sets and training parameters for different branches. Ultimately, we used the no-makeup and makeup-applied pictures produced in the paper containing the LADN model as test data, and compared the performance of six models, including BeautyGAN, LADN, PSGAN, WMT(the model raised in Nguyen's research [11]), EleGANt and the proposed FSAMT in this paper.

For the color transfer branch, we trained the color transfer network on the MT dataset. The training parameters are kept basically consistent with those in BeautyGAN. First, all picture pairs are aligned and resized to a resolution of 256x256. The adversarial loss weight $\lambda_a = 1$, cycle consistency loss weight $\lambda_c = 10$. Perceptual loss weight $\lambda_p = 0.003$. Histogram matching loss weight $\lambda_h = 1$. The learning rates for both the generator and discriminator are 0.0002, with a batch size set to 1. We used the Adam optimizer, with momentum parameters set to (0.5, 0.999).

For the highlight transfer branch, we trained and tested on our own dataset HFD. We divided the dataset of 2000 pictures into a 7:3 ratio, with 1400 pictures for training and 600 pictures for testing. U-Net was used as the main architecture, with binary cross entropy loss and Adam optimizer, with a learning rate of 0.0002. The momentum parameters (beta1 and beta2) were set to 0.9 and 0.999, respectively. The batch size was set to 8, and the number of training epochs was set to 50.

For the pattern transfer branch, we used the augmented CPM-Synt-1 and CPM-Synt-2 dataset to train our RMW model. 80% of the data is for training with a batch size of 16. We used the Adam optimizer with a learning rate of 1e-4 and weight decay of 1e-5. The number of training epochs was set to 350. During training, we monitored the loss functions for both the feature extractor and the domain discriminator to ensure model convergence and avoided overfitting. To ensure the Lipschitz constraint of the domain discriminator in W-DA, weight clipping was applied.

5.2 Quantitative Experiment

For qualitative analysis, effective ground truth is necessary. The CPM-Synt-2 dataset consists of synthetic pictures stored in triplets: makeup pictures, non-makeup pictures, and ground truth. Constructed based on the assumption of "makeup transfer stability," where a good makeup transfer method produces consistent makeup style outputs using the same reference picture. The dataset creators, Nguyen et al., randomly selected two non-makeup pictures from the MT dataset, transferred them to the same makeup style using BeautyGAN, and manually added stickers to obtain ground truth.

To assess the overall effectiveness of makeup transfer, we have devised three quantitative evaluation metrics in order to comprehensively evaluate the capabilities of the model. These metrics include the Structural Similarity Index (SSIM), the Learned Perceptual picture Patch Similarity (LPIPS), and the Frechet Inception Distance (FID). By employing the triplets from the CPM-Synt-2 dataset, we have applied six models and conducted a quantitative analysis of the transferred results in comparison to the ground truth using these four criteria. The detailed experimental findings can be found in Table 3.

The structural similarity index measure (SSIM) takes into account the structural information, brightness, and contrast of images. A value closer to "1" denotes a higher degree of similarity between two images. According to the table, the FSAMT model achieves the highest SSIM value (0.2527). On the other hand, the Learned Perceptual Image Patch Similarity (LPIPS) serves as an evaluation method based on disparities in the human visual system. A lower value implies a closer evaluation to that of the human visual system. The FSAMT model attains the lowest LPIPS value, which is approximately 4.6% lower than the average value. Moving on to the Fréchet Inception Distance (FID) metric, a lower value signifies better picture quality and finer details. Remarkably, the FSAMT model outperforms other models with a significantly lower FID value, approximately 20.1% lower than the average value. Drawing from the outcomes of these four evaluation metrics, we can assert that the FSAMT model exhibits superior overall performance in the makeup transfer task compared to other models.

Table 3 Results of 4 evaluation methods

	Range	BeautyGAN	PSGAN	EleGANt	LADN	WMT	FSAMT
PSNR	$(0, +\infty)$	27.9318	28.2161	**28.3562**	28.3164	28.302	28.2976
SSIM	(0, 1)	0.1889	0.2049	0.2216	0.2087	0.2157	**0.2384**
LPIPS	(0, 1)	0.5414	0.5427	0.5338	0.5432	0.5327	**0.5164**
FID	$(0, +\infty)$	234.4531	210.0957	202.7207	200.3964	186.4172	**162.580**

Fig. 7 Comparison of transfer results

Although models such as LADN and WMT, which have the ability to transfer patterns, may not perform as well as other models in some evaluation metrics, this could be due to a trade-off between performance and interpretability. While models may sacrifice performance in certain evaluation metrics to maintain their ability to transfer heavy makeup styles, models that cannot transfer patterns may focus on optimizing specific evaluation metrics and perform better in these metrics. Since makeup transfer is a visually appealing task, qualitative evaluation and user surveys can effectively complement the limitations of quantitative analysis.

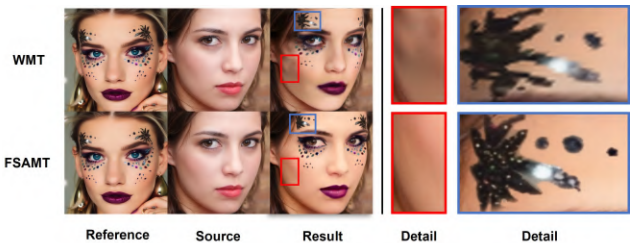


Fig. 8 Comparison of different models

5.3 Qualitative Experiment

Using the LADN model's makeup-free and made-up pictures as source and reference, we compared the makeup transfer results of six previously discussed models. While BeautyGAN, PSGAN, and EleGANt possess limited pattern transfer capabilities, making a complete comparison challenging, their contributions to the color transfer branch task warranted their inclusion in this assessment (Fig. 7).

LADN, WMT, and FSAMT demonstrate varying pattern transfer abilities. Our proposed FSAMT model stands out by preserving pattern and texture details while maintaining picture clarity. Moreover, its color consistency is commendable. Despite BeautyGAN and PSGAN avoiding color artifacts, with PSGAN exhibiting accurate color restoration, they falter in transferring intense makeup styles. EleGANt, though not fully transferring patterns, discerns and modifies pattern colors within limits. To delve deeper into FSAMT's ability, we juxtaposed its detailed transfer results against WMT, a model with commendable performance in similar tasks. Figure 8 reveals issues like color inconsistencies, fuzzy pattern edges, and detail loss in magnified face pictures with WMT. Contrarily, FSAMT manifests crisper pattern edges, distinct textures, and reduced facial color anomalies.

5.4 User Survey

We designed the questionnaire from three aspects: makeup color, makeup pattern and overall makeup. For the first two, we ask the respondents to focus on the transfer accuracy, while for the last item, we suggest them to give an evaluation intuitively from the perspective of human vision. Each aspect consists of ten pictures, and the score range is a discrete value between 0-10. Therefore, thirty responses are included in one questionnaire. In order to ensure fairness, we shuffled the pictures generated by the six models, and invited 50 respondents(an equal number of males and females) to fill out the questionnaire. The results of the survey are shown in Table 4. The table reveals FSAMT as the top performer in all three categories. From the survey, we observed: 1. Some respondents, despite rating high for "color" and "pattern," scored lower for "overall"; 2. Some were hesitant to rate without a reference picture. 3. Overall, men are more likely to score high. We deduced three reasons: First, accurate makeup transfer doesn't always equate to enhanced aesthetics. If reference makeup doesn't fit the source picture, more accurate transfers might yield less appealing results. Second, respondents lacked confidence in their aesthetic judgment, only feeling sure when comparing two pictures. These insights led to the necessary of the Makeup Adaptation model.

Table 4 User survey(M-Male, F-Female)

	Color		Pattern		Overall	
	M	F	M	F	M	F
BeautyGAN	7.61	7.45	-	-	7.70	7.58
PSGAN	7.57	7.41	-	-	7.62	7.51
EleGANt	7.59	7.63	-	-	7.78	7.82
LADN	7.03	6.89	6.28	6.35	6.65	6.42
WMT	7.56	7.43	7.87	7.76	7.41	7.29
FSAMT	**7.65**	**7.71**	**8.11**	**7.97**	**8.03**	**7.92**

It addresses these issues by selecting optimal makeup for the source picture and operates end-to-end without needing comparison. Its value is further discussed in the Ablation Study chapter. Additionally, female, due to their more accurate recognition and keen insight into makeup compared to men (also benefiting from familiarity with cosmetics), tend to be more discerning.

6. Ablation Study

We conducted three ablation studies to evaluate the contribution of different components in the makeup transfer pipeline: 1. Validation of the makeup adaptation model; 2. Validation of the color transfer branch in makeup transfer model; 3. Validation of the pattern transfer branch in makeup transfer model. We used a combination of quantitative and qualitative methods to help us determine which models, structures, or parameters have a greater impact on picture quality.

To assess the Makeup Adaptation Model's influence, we conducted an A/B experiment. In group A, without the adaptation model, we selected 100 source and reference pictures for random, unordered matching, then processed them through the Makeup Transfer Model. Group B kept the same source pictures as A but used the adaptation model to match face shapes before transferring makeup. The procedure is illustrated in Fig. 9. We used a user survey to gauge satisfaction, showing participants the original and makeup-transferred pictures and asking them to rate the makeover's aesthetics on a 1-10 scale.

Based on the survey questionnaires returned by 30 participants, we obtained the following key data: 1. The average score of Group A is 6.62, and that of Group B is 7.53, indicating that Group B is approximately 13.7% higher than Group A; 2. The score variance of Group A is 16.27, and that of Group B is 9.16, showing that Group B is approximately 43% lower than Group A; 3. The probability of obtaining an extremely low score (less than 3) in Group A is 8%, while it is only 2% in Group B. Overall, the performance of Group B is significantly better than that of Group A, with smaller score differences between each group of pictures, more aesthetically pleasing makeup after transfer, and less occurrence of unsuitable makeup. Overall, the makeup adaptation model effectively provides guidance for makeup selection.

In the color transfer module, we first removed the discriminator with Gradient Penalty and Adaptive Spectral Normalization, and instead used the original discriminator (*A*). Then, we replaced the adaptive binning method with fixed number of bins for histogram matching (*B*).

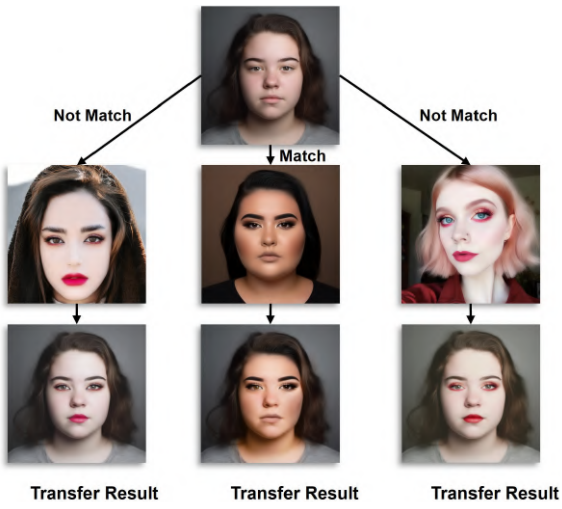


Fig. 9 Control group for makeup adaptation model

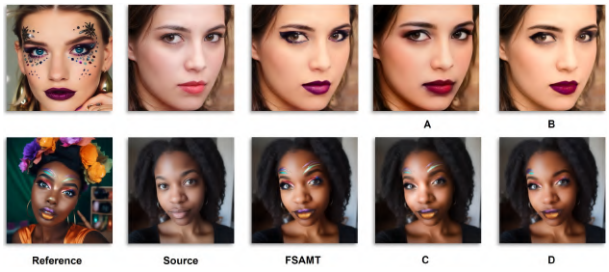


Fig. 10 Ablation study result

Table 5 Quantitative analysis of four operations

	Baseline	Color		Pattern	
Group	FSAMT	A	B	C	D
SSIM	**0.2216**	0.2075	0.2143	0.2191	0.2003
LPIPS	**0.5217**	0.5412	0.5460	0.5338	0.5477
FID	**159.6872**	188.0048	177.2079	171.6127	190.5107
nFID	**0**	1	0.4661	0.3175	0.5527

In the pattern transfer module, we removed the Resnet50MultiScale module and replaced it with a simple feature extractor (*C*). Then, we removed the MultiScaleViT module and only used the features output by Resnet50MultiScale (*D*). We retrained the model on the same dataset after each modification and recorded performance metrics. Figure 10 shows the changes brought by each operation during the experimental process as an example.

Table 5 shows performance variations across three metrics after distinct operations. For analysis, we normalized FID to a 0-1 range, termed nFID. Notably, *D* most affects SSIM and LPIPS, highlighting the significance of the MultiScaleViT module. *A* primarily influences nFID at 0.8123, showing that omitting the discriminator with specific features

increases differences between generated and actual pictures. While *B* moderately impacts SSIM and LPIPS, its lowest nFID contribution (0.5041) suggests the adaptive binning method's efficiency in histogram matching. *C* has the least influence on all three metrics. Next, we combine the results of ablation experiments and the structural features of the model to explain the reasons why the improvements on the model work from two aspects of color and pattern transfer.

In color transfer, our method successfully addresses the challenge of authentic makeup transfer while preserving facial features, especially in cases of heavy makeup where eyeshadow blends with skin tones. This is achieved through the combined use of CLAHE, which enhances local picture details through histogram equalization in small regions, and adaptive binning, which adjusts histogram intervals according to data distribution. Together, these techniques ensure accurate color transfer and realistic makeup application.

For pattern transfer, the integration of Resnet50MultiScale and ViT is crucial. Resnet50MultiScale effectively captures details from large facial areas to fine points like eyeliner, as evidenced by the loss of texture clarity in Group *C* when omitted. ViT, meanwhile, excels in merging multi-scale features, crucial for coherent facial makeup integration and avoiding disjointed effects, as demonstrated in Group *D*.

7. Conclusion

This study introduces a end-to-end makeup transfer framework *FSAMT*. It adaptively adjusts and precisely transfers selected makeup. Experiments demonstrate its efficacy across quantitative and qualitative analysis. However, it may obscure parts of eyebrows or eyelashes during pattern transfers. Future work could integrate additional facial parameters to enhance face shape classification granularity.

Acknowledgments

This work was supported by JST W-SPRING, Grant Number JPMJSP2128.

References

[1] Fortune Business Insights. Makeup market size, share & covid-19 impact analysis, 2023.

[2] H. Deng, C. Han, H. Cai, G. Han, and S. He, "Spatially-invariant style-codes controlled makeup transfer," *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, pp.6549–6557, 2021.

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

[4] H.-J. Chen, K.-M. Hui, S.-Y. Wang, L.-W. Tsao, H.-H. Shuai, and W.-H. Cheng, "Beautyglow: On-demand makeup transfer framework with reversible generative network," *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, pp.10042–10050, 2019.

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

[6] R. Kips, P. Gori, M. Perrot, and I. Bloch, "Ca-gan: Weakly supervised color aware gan for controllable makeup transfer," *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, pp.280–296, Springer, 2020.

[7] J. Xiang, J. Chen, W. Liu, X. Hou, and L. Shen, "Ramgan: Region attentive morphing gan for region-level makeup transfer," *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, Oct. 23–27, 2022, Proceedings, Part XXII*, pp.719–735, Springer, 2022.

[8] X. Yang, T. Taketomi, and Y. Kanamori, "Makeup extraction of 3d representation via illumination-aware image decomposition," *Computer Graphics Forum*, vol.42, no.2, pp.293–307, Wiley Online Library, 2023.

[9] C. Yang, W. He, Y. Xu, and Y. Gao, "Elegant: Exquisite and locally editable gan for makeup transfer," *European Conf. Computer Vision*, pp.737–754, Springer, 2022.

[10] Q. Gu, G. Wang, M.T. Chiu, Y.-W. Tai, and C.-K. Tang, "Ladn: Local adversarial disentangling network for facial makeup and demakeup," *Proc. IEEE/CVF Int. Conf. Computer Vision*, pp.10481–10490, 2019.

[11] T. Nguyen, A.T. Tran, and M. Hoai, "Lipstick ain't enough: beyond color matching for in-the-wild makeup transfer," *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, pp.13305–13314, 2021.

[12] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *Proc. IEEE Conf. computer vision and pattern recognition*, pp.770–778, 2016.

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

[14] J. Shen, Y. Qu, W. Zhang, and Y. Yu, "Wasserstein distance guided representation learning for domain adaptation," *Proc. AAAI Conf. Artificial Intelligence*, vol.32, no.1, 2018.

[15] Z. Wei, "Oriental eye shape classification and cosmetology," *Medical Aesthetics and Cosmetology*, (5):38–39, 1 1995.

[16] G. Huang, Z. Liu, L. Van Der Maaten, and K.Q. Weinberger, "Densely connected convolutional networks," *Proc. IEEE Conf. computer vision and pattern recognition*, pp.4700–4708, 2017.

[17] Y. Feng, F. Wu, X. Shao, Y. Wang, and X. Zhou, "Joint 3d face reconstruction and dense alignment with position map regression network," *Proc. European Conf. computer vision (ECCV)*, pp.534–551, 2018.

[18] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," *Proc. IEEE Int. Conf. computer vision*, pp.3730–3738, 2015.

[19] S. Yaorui and B. Fanliang, "Eye type classification based on convolutional neural network and semantic features," *Electronic Measurement Technology*, 42(3):16–20, 1 2019.

[20] S. Jinguang and R. Wenzhao, "Curve similarity eye type classification" *Computer Science and Exploration*, 11(8):1305–1313, 1 2017.

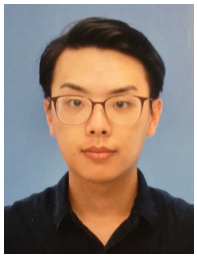
[21] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[22] K. Zuiderveld, "Contrast limited adaptive histogram equalization," *Graphics gems*, pp.474–485, 1994.

[23] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," *Proc. IEEE Conf. computer vision and pattern recognition*, pp.2117–2125, 2017.

Haoran Luo is currently a Ph.D. student at the Graduate School of Creative Science and Engineering, Waseda University. He is now sponsored by JST for research. His research interests include: sentiment analysis, semantic extraction, image style transfer and smart city construction, etc.

Tengfei Shao completed his master's degree in engineering at Waseda University. He is currently a Ph.D. student at the Department of Industrial and Management Systems Engineering, Waseda University, and sponsored by JST SPRING for research. His research interests include artificial intelligence, knowledge graphs, and complex networks.



Shenglei Li is currently a Ph.D. student at the Graduate School of Creative Science and Engineering, Waseda University and sponsored by JST SPRING for research. He obtained his B.E. degree from the Department of Civil Engineering, Southwest Jiaotong University, and his M.E. degree from the Graduate School of System and Information Engineering, Tsukuba University. His research interest lies in Artificial Intelligence, Human-computer Interactions, and Smart city implementations

Reiko Hishiyama is a professor at the Graduate School of Creative Science and Engineering of Waseda University, where she directs the Intelligent Information System Laboratory. She received her Doctor of Informatics degree in 2005 from Kyoto University, Japan. Her current research interests include artificial intelligence, autonomous multi-agent systems, knowledge representation, autonomy-oriented computing, and related areas.