Appendix of EleGANt: Exquisite and Locally Editable GAN for Makeup Transfer

A Landmark Embedding

We adopt the Landmark Embedding proposed by [11] to introduce spatial information into attention. It utilizes facial landmarks as anchor points to represent relative positions. Given N landmark points $\{L_n\}_{n=1}^{N}$ of the facial image, each pixel x_i on the image is assigned with a vector $\mathbf{p}_i \in \mathbb{R}^{2N}$ as positional embedding, which is computed by its relative positions to those landmark points:

$$\begin{aligned} \mathbf{p}_i^{(2n)} &= \mathrm{x}(x_i) - \mathrm{x}(L_n) \\ \mathbf{p}_i^{(2n+1)} &= \mathrm{y}(x_i) - \mathrm{y}(L_n) \end{aligned}\ ,\ n = 1, \dots, N \tag{1}$$

where $\mathrm{x}(\cdot)$ and $\mathrm{y}(\cdot)$ denote the x-coordinate and y-coordinate of a point respectively. The spatial feature $\mathbf{p}_i$ is then normalized by its 2-norm to ensure size-invariance, and is concatenated with the visual features at pixel x_i.

B TPS Transformation

We adopt STN [5] that uses a grid generator to compute a sampling grid $\mathcal{P} = \{p_i\}$ on an image to form a transformation. A 2D TPS transformation with N control points $\mathbf{C}, \mathbf{C}' \in \mathbb{R}^{2\times N}$ is parameterized by a $2 \times (N + 3)$ matrix:

$$\mathbf{T} = \begin{bmatrix} a_0\ a_1\ a_2\ \mathbf{u} \\ b_0\ b_1\ b_2\ \mathbf{v} \end{bmatrix} \tag{2}$$

where $\mathbf{u}, \mathbf{v} \in \mathbb{R}^{1\times N}$. We follow the formulation in [10] to describe the grid computation of 2D TPS. For a point $\mathbf{p} \in \mathbb{R}^{1\times 2}$, its sampling point $\mathbf{p}'$ is computed by a linear projection:

$$\mathbf{p}' = \mathbf{T} \begin{bmatrix} 1 \\ \mathbf{p} \\ \phi(||\mathbf{p} - \mathbf{c}_1||) \\ \cdots \\ \phi(||\mathbf{p} - \mathbf{c}_N||) \end{bmatrix} \tag{3}$$

where $\phi(r) = r^2 \log r$ is the radial basis kernel applied to the Euclidean distance between $\mathbf{p}$ and the control points $\mathbf{C}$. The coefficients of TPS are obtained by solving a linear system involving N correspondences between $\mathbf{C}$ and $\mathbf{C}'$:

$$\mathbf{c}_i' = \mathbf{T} \begin{bmatrix} 1 \\ \mathbf{c}_i \\ \phi(||\mathbf{c}_i - \mathbf{c}_1||) \\ \cdots \\ \phi(||\mathbf{c}_i - \mathbf{c}_N||) \end{bmatrix}\ ,\ i = 1, \dots N \tag{4}$$

subject to the following boundary conditions:

$$\begin{aligned} 0 &= \mathbf{u}\mathbf{1} \\ 0 &= \mathbf{v}\mathbf{1} \\ 0 &= \mathbf{u}\mathbf{C}_x^T \\ 0 &= \mathbf{v}\mathbf{C}_y^T \end{aligned} \tag{5}$$

where $\mathbf{C}_x$ and $\mathbf{C}_y$ are the x and y coordinates of $\mathbf{C}$, respectively. $\mathbf{T}$ has a closed-form solution in matrix form:

$$\begin{aligned} \mathbf{T} &= \begin{bmatrix} C' & \mathbf{0}^{2\times 3} \end{bmatrix} \mathbf{\Delta}_C^{-1} \\ \mathbf{\Delta}_C &= \begin{bmatrix} \mathbf{1}^{1\times N} & \mathbf{0} & \mathbf{0} \\ \mathbf{C} & \mathbf{0} & \mathbf{0} \\ \hat{\mathbf{C}} & \mathbf{1}^{N\times 1} & \mathbf{C}^T \end{bmatrix} \end{aligned} \tag{6}$$

where $\hat{\mathbf{C}} \in \mathbb{R}^{N\times N}$ is a square matrix comprising $\hat{\mathbf{C}}_{i,j} = \phi(||\mathbf{c}_i - \mathbf{c}_j||)$.

C Full Objective

Given the non-makeup domain X and the makeup domain Y, our proposed EleGANt learns the mapping bidirectionally between these two domains. Let $x \in X$ and $y \in Y$ denote the source image and the reference makeup image, respectively. $\hat{x} = \mathcal{G}(x, y)$ is the transferred result with the makeup style of y and the facial identity of x.

Adversarial Loss. Adversarial loss [2] is introduced to guide realistic generation. We apply two discriminators $\mathcal{D}_X$ and $\mathcal{D}_Y$ to discriminate between real and generated images in the domain X and Y respectively. The adversarial loss $L_{\mathcal{G}}^{adv}$ for generator and $L_{\mathcal{D}}^{adv}$ for discriminator are defined as

$$\begin{aligned} L_{\mathcal{G}}^{adv} = &- \mathbb{E}_{x\sim X, y\sim Y}\left[\log\left(\mathcal{D}_X(\mathcal{G}(y, x))\right)\right] \\ &- \mathbb{E}_{x\sim X, y\sim Y}\left[\log\left(\mathcal{D}_Y(\mathcal{G}(x, y))\right)\right] \end{aligned} \tag{7}$$

$$\begin{aligned} L_{\mathcal{D}}^{adv} = &- \mathbb{E}_{x\sim X}\left[\log \mathcal{D}_X(x)\right] - \mathbb{E}_{y\sim Y}\left[\log \mathcal{D}_Y(y)\right] \\ &- \mathbb{E}_{x\sim X, y\sim Y}\left[\log\left(1 - \mathcal{D}_X(\mathcal{G}(y, x))\right)\right] \\ &- \mathbb{E}_{x\sim X, y\sim Y}\left[\log\left(1 - \mathcal{D}_Y(\mathcal{G}(x, y))\right)\right] \end{aligned} \tag{8}$$

Cycle Consistency Loss. We use cycle consistency loss [12] for unsupervised learning with unpaired images. The cycle consistency loss $L_{\mathcal{G}}^{cyc}$ is defined as the L_1-distance between the original image and reconstructed image:

$$\begin{aligned} L_{\mathcal{G}}^{cyc} = &\mathbb{E}_{x\sim X, y\sim Y}\left[\|\mathcal{G}(\mathcal{G}(x, y), x) - x\|_1\right] \\ &+ \mathbb{E}_{x\sim X, y\sim Y}\left[\|\mathcal{G}(\mathcal{G}(y, x), y) - y\|_1\right] \end{aligned} \tag{9}$$

Perceptual Loss. To guarantee that the personal identity of the source image is preserved in the transferred image, we utilize perceptual loss [7] to maintain face identity. $L_{\mathcal{G}}^{per}$ is defined as

$$\begin{aligned} L_{\mathcal{G}}^{per} &= \mathbb{E}_{x\sim X,y\sim Y}\left[\|F_l(\mathcal{G}(x,y)) - F_l(x)\|_2\right] \\ &+ \mathbb{E}_{x\sim X,y\sim Y}\left[\|F_l(\mathcal{G}(y,x)) - F_l(y)\|_2\right] \end{aligned} \tag{10}$$

where $F_l(\cdot)$ is the l-th layer output of the pre-trained VGG-16 model.

Makeup Loss. To provide guidance for transferring specific makeup styles, we introduce makeup loss as extra supervision. $L_{\mathcal{G}}^{make}$ is defined as the L_2-distance between the transferred image and the pseudo ground truth (PGT) generated by our proposed strategy AC-PGT:

$$L_{\mathcal{G}}^{make} = \|\mathcal{G}(x,y) - PGT(x,y)\|_1 + \|\mathcal{G}(y,x) - PGT(y,x)\|_1 \tag{11}$$

Total Loss. The total loss can be expressed as:

$$\begin{aligned} &L_{\mathcal{G}} = \lambda_{adv}L_{\mathcal{G}}^{adv} + \lambda_{cyc}L_{\mathcal{G}}^{cyc} + \lambda_{per}L_{\mathcal{G}}^{per} + \lambda_{make}L_{\mathcal{G}}^{make} \\ &L_{\mathcal{D}} = \lambda_{adv}L_{\mathcal{D}}^{adv} \end{aligned} \tag{12}$$

where λ_{adv}, λ_{cyc}, λ_{per}, λ_{make} are trade-off parameters.

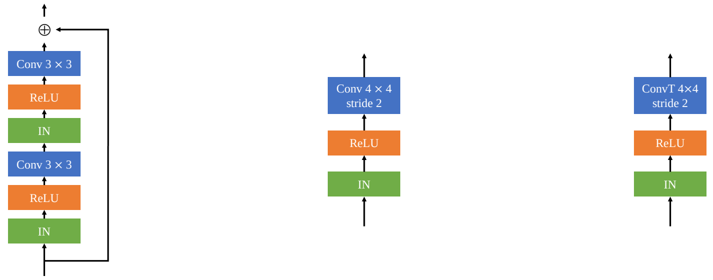
D Implementation Details

D.1 Training Settings

The optimizer of the generator the discriminator is Adam [8] with $\beta_1 = 0.5 = 0.5$ and $\beta_2 = 0.999$. The learning rate is initially 2e-4 and decreases to 1e-5 by cosine annealing decay. The model are trained for 50 epochs with batch size 1. We extract features from $relu\_4\_1$ layer of pretrained VGG-16 to calculate the perceptual loss. The trade-off parameters in the loss function are set as $\lambda_{adv} = 1$, $\lambda_{cyc} = 10$, $\lambda_{per} = 0.005$, $\lambda_{make} = 1$.

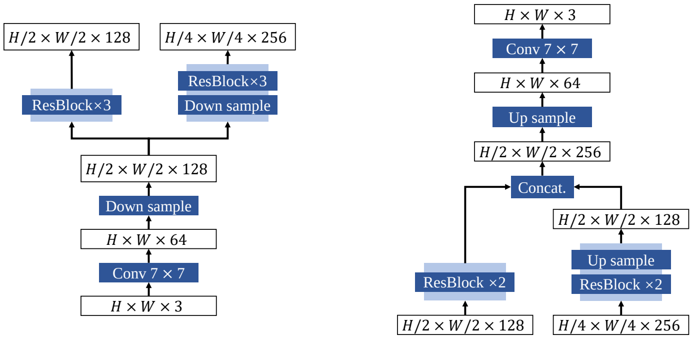
D.2 Network Architecture

We use three basic blocks to construct the generator of our EleGANt: Resblock, Down-sampling block, and Up-sampling block, whose architectures are shown in Fig. 1. The architectures of FAEnc and MADec in EleGANt with the shape of corresponding feature maps are illustrated in Fig. 2. In Makeup Transfer Module (MTM), there is one Attention Module and one Sow-Attention Module, each of which performs cross-attention once as described in Sec. 3.2. We adopt the discriminator in [4] that distinguishes overlapped patches of size 70 × 70 of the image between real and fake.



(a) Resblock (b) Down-sampling (c) Up-sampling

Fig. 1. Basic blocks used in our network. "IN" denotes Instance Normalization.



(a) FAEnc (b) MADec

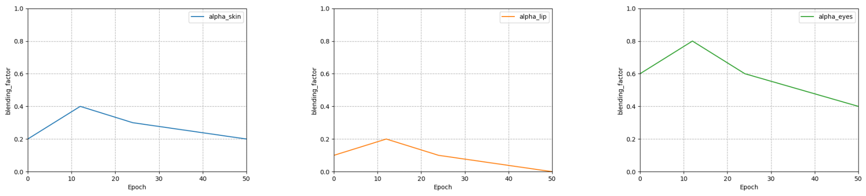
Fig. 2. Architecture of FAEnd and MADec in EleGANt. "Concat." denotes concatenation along the channel dimension. "$\times n$" indicates a stack of n blocks.

D.3 Pseudo Ground Truth

Our proposed AC-PGT can be expressed in formula:

$$PGT(x, y) = \sum_{i \in I} M_i^x \left(\alpha_i^D TPS\left(x, y, C_i^x, C_i^y\right) + \left(1 - \alpha_i^D\right) HM(x, y) \right) \tag{13}$$

where $HM(x, y)$ is the result of histogram matching which has the makeup of y and the identity of x, $TPS\left(x, y, C_i^x, C_i^y\right)$ denotes the result of warping y to fit x by TPS using control points C_i^x and C_i^y, M_i^x denotes the binary mask, and α_i^D is the blending factor for region i in $I = \{skin, lip, eyeshadow\}$. We set the control points as all facial landmarks for the skin region, the landmarks around the eyes for the eye shadow region, and the landmarks on the lip for the lip region. We design annealing functions for blending factors α_{skin}^D, α_{eyes}^D, and α_{lip}^D, which are shown in Fig. 3. In the early stages of training, α_i^D gradually increases to guide the generator to first learn the global transfer for overall colors and then learn the local one for makeup details. In the later stages, we decrease α_i^D to avoid the artifacts in PGT introduced by image warping and stitching being learned.



(a) α^D_{skin} (b) α^D_{lip} (c) α^D_{eyes}

Fig. 3. Annealing functions for the blending factors.

E Additional Results

E.1 User Study

We have recorded additional information about the participants of the user study. The distribution of their ages and genders are reported in Table. 1.

Table 1. Ages and genders of the participants in the user study

Age	[20,30)	[30,40)	[40,50)	[50, 60)
Ratio (%)	62.5	27.5	7.5	2.5

Gender	Female	Male
Ratio (%)	55	45

E.2 Qualitative Comparison

Fig. 4 presents additional qualitative results of BeautyGAN [9], LADN [3], PSGAN [6], SCGAN [1] and our EleGANt. BeautyGAN generates visually acceptable results when the images have the same poses, but it falls short when a large spatial difference exists between the two faces. There are severe artifacts and blurs in the results of LADN, and the transferred colors are also incorrect. PSGAN cannot transfer detailed makeup attributes and suffers from unnatural shadows and illuminations. SCGAN fails to synthesize makeup details, and there are rectangular color blocks around the eyes in the generated images due to an improper decomposition that manually splits the face with rectangles. Our EleGANt surpasses all existing methods: it is robust to misaligned head poses and different illuminations, and it can preserve and precisely transfer makeup details, representatively, the shapes and colors of the eye shadows.

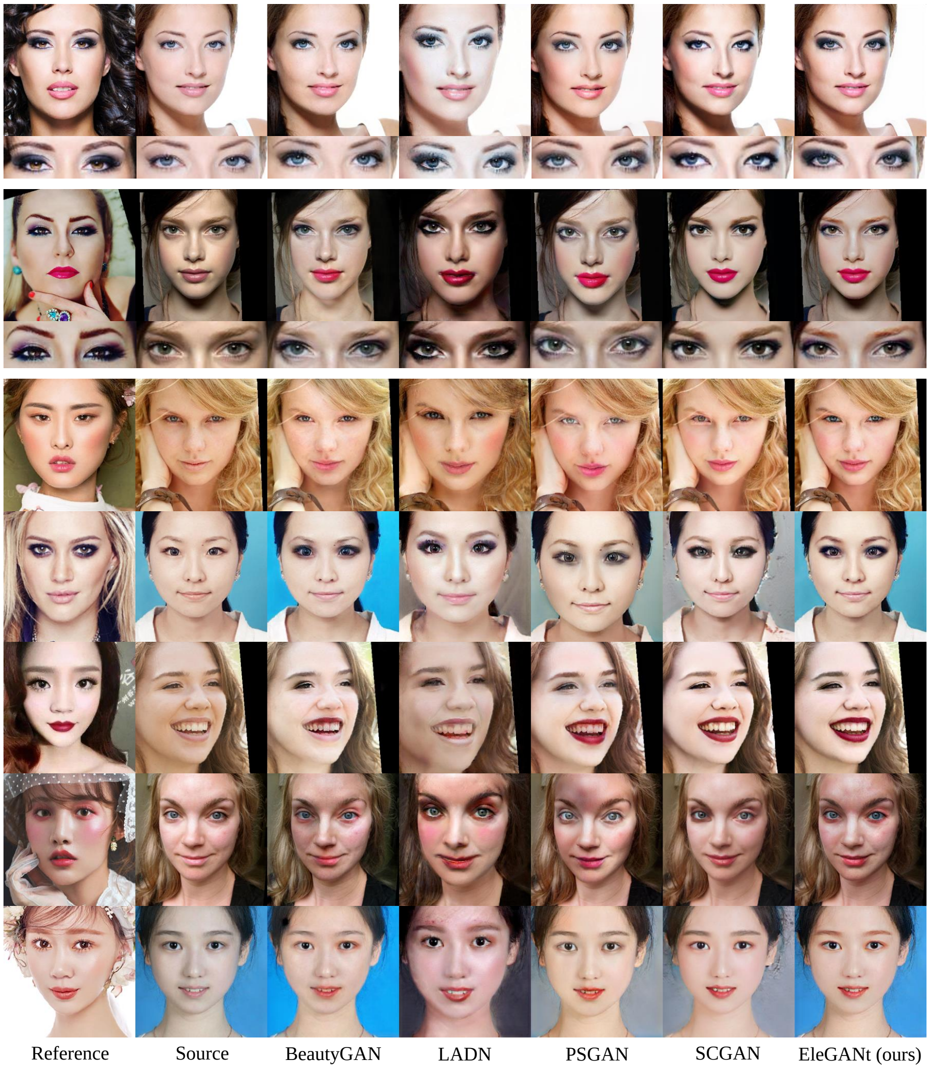


Fig. 4. Qualitative comparisons with existing methods.

References

1. Deng, H., Han, C., Cai, H., Han, G., He, S.: Spatially-invariant style-codes controlled makeup transfer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6549–6557 (2021)
2. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Proceedings of the International Conference on Neural Information Processing Systems (NIPS) (2014)
3. Gu, Q., Wang, G., Chiu, M.T., Tai, Y.W., Tang, C.K.: Ladn: Local adversarial disentangling network for facial makeup and de-makeup. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 10481–10490 (2019)
4. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1125–1134 (2017)
5. Jaderberg, M., Simonyan, K., Zisserman, A., Kavukcuoglu, K.: Spatial transformer networks. In: Proceedings of the International Conference on Neural Information Processing Systems (NIPS). pp. 2017–2025 (2015)
6. Jiang, W., Liu, S., Gao, C., Cao, J., He, R., Feng, J., Yan, S.: Psgan: Pose and expression robust spatial-aware gan for customizable makeup transfer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5194–5202 (2020)
7. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 694–711 (2016)
8. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: Bengio, Y., LeCun, Y. (eds.) Proceedings of the International Conference on Learning Representations (ICLR) (2015)
9. Li, T., Qian, R., Dong, C., Liu, S., Yan, Q., Zhu, W., Lin, L.: Beautygan: Instance-level facial makeup transfer with deep generative adversarial network. In: Proceedings of the 26th ACM International Conference on Multimedia. pp. 645–653 (2018)
10. Shi, B., Yang, M., Wang, X., Lyu, P., Yao, C., Bai, X.: Aster: An attentional scene text recognizer with flexible rectification. IEEE Transactions on Pattern Analysis and Machine Intelligence **41**(9), 2035–2048 (2018)
11. Wan, Z., Chen, H., An, J., Jiang, W., Yao, C., Luo, J.: Facial attribute transformers for precise and robust makeup transfer. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 1717–1726 (2022)
12. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 2223–2232 (2017)