Compressing Facial Makeup Transfer Networks by Collaborative Distillation and Kernel Decomposition

Bianjiang Yang1, Zi Hui2, Haoji Hu2*, Xinyi Hu2, Lu Yu2
1 Chu Kochen Honors College, Zhejiang University, Hangzhou, China
2 College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou, China
Emails: {yangbj, zihui, haoji_hu, xinyih, yulu}@zju.edu.cn

Abstract—Although the facial makeup transfer network has achieved high-quality performance in generating perceptually pleasing makeup images, its capability is still restricted by the massive computation and storage of the network architecture. We address this issue by compressing facial makeup transfer networks with collaborative distillation and kernel decomposition. The main idea of collaborative distillation is underpinned by a finding that the encoder-decoder pairs construct an exclusive collaborative relationship, which is regarded as a new kind of knowledge for low-level vision tasks. For kernel decomposition, we apply the depth-wise separation of convolutional kernels to build a light-weighted Convolutional Neural Network (CNN) from the original network. Extensive experiments show the effectiveness of the compression method when applied to the state-of-the-art facial makeup transfer network – BeautyGAN [1].

Index Terms—Facial Makeup Transfer, Network Compression, Knowledge Distillation, Convolutional Kernel Decomposition.

I. INTRODUCTION

Facial makeup transfer aims to translate the makeup style from a given reference face image to another non-makeup face image without the change of face identity. It is an interesting but challenging area because it needs to handle different local styles/cosmetics including eye shadow, lipstick and foundation in a coherent and natural way. Existing research works on automatic makeup transfer can be classified into two categories: traditional image processing approaches such as physics-based manipulation [2], [3], and deep learning based methods which typically build upon deep neural networks [1], [4]. A recent deep learning based network, BeautyGAN [1], further borrows the idea of style transfer networks [5], [6] and implements the encoder-decoder architecture for makeup transfer.

Although facial makeup transfer network has achieved good performance, its capability is still restricted by the massive computation and high storage of the network architecture, which constrains its application on mobile devices. Model compression and acceleration methods provide a possible solution to save computation and storage of deep neural networks, which include parameter pruning [7], quantization [8], and knowledge distillation [9], etc. However, most compression methods only focus on high-level tasks, e.g., classification and detection. Compressing models for low-level vision tasks, such as facial makeup transfer, is still less explored.

In this paper, taking BeautyGAN [1] as an example, we address the problem of compressing facial makeup transfer networks to save computation and storage. Firstly, we apply Collaborative Knowledge Distillation (CKD) [10] to compress the encoder of BeautyGAN. The main idea of CKD is underpinned by a finding that the encoder-decoder pairs construct an exclusive collaborative relationship, which is regarded as a new kind of knowledge for low-level vision tasks. Thus, to overcome the feature size mismatch when applying collaborative distillation, a feature loss is introduced to drive the student network to learn a linear embedding of the teachers features. Secondly, because BeautyGAN is not a standard encoder-decoder based network, it contains several residual blocks which are difficult to distill by CKD. To further reduce computation, we use the idea of kernel decomposition, which is proposed by MobileNets [11]. Kernel decomposition are based on a streamlined architecture that uses depth-wise separable convolutions to build light-weighted deep neural networks. Although extensive researches have shown the effectiveness of kernel decomposition in high-level vision tasks [11], [12], it is rarely implemented in low-level tasks such as facial makeup transfer. We decompose the residual blocks of BeautyGAN into a depth-wise convolution followed by a 1 × 1 point-wise convolution, thus reducing the computation by a large margin.

The major contributions of this paper lie in three aspects:

- We introduce an efficient two-step compression method for BeautyGAN by compressing the encoder with CKD and the residual blocks with kernel decomposition.
- We successfully demonstrate the effectiveness of kernel decomposition in facial makeup transfer network, indicating its usability in low-level vision tasks.
- Experiments on the MT dataset [1] demonstrate that the compressed network performs favorably against the original BeautyGAN.

II. THE PROPOSED METHOD

The overall framework of the proposed method is illustrated in Figure 1. The original BeautyGAN at the top is the teacher network which sequentially consists of an encoder, 6 residual blocks and a decoder. The student network is illustrated at the bottom which consists of a smaller encoder, 9 decomposed residual blocks and a decoder. The four loss functions in BeautyGAN, i.e., adversarial loss, cycle loss, perceptual loss...
and makeup loss, are also adopted by the student network. Collaborative distillation is implemented to distill the knowledge of the teacher’s encoder to the encoder of the student network. The distillation process is guided by the feature loss between the two encoders.

Detailed architectures of the original BeautyGAN and our proposed student network are summarized in Table I. The protocol of layer representation is ‘layer type-kernel size-number of input channels-number of output channels (stride)’. In addition, conv, res, deconv and separated res represent the convolutional layer, residual block layer, deconvolutional layer and separated residual blocks, respectively.

**TABLE I:** The architectures of the original BeautyGAN and our proposed student network.

| BeautyGAN | The Student Network |
|-----------|---------------------|
| Non-makeup Reference | Non-makeup Reference |
| conv7-3-64(1) | conv7-3-64(1) |
| conv4-64-128(2) | conv4-64-128(2) |
| Merge | Merge |
| conv4-256-256(2) | conv4-256-256(2) |
| 6 × res3-256-256(1) | 9 × separated res3-256-256(1) |
| deconv4-128-64(2) | deconv4-128-64(2) |
| Separate | Separate |
| makeup | remove-makeup |
| conv3-64-64(1) | conv3-64-64(1) |
| conv3-64-64(1) | conv3-64-64(1) |
| conv3-64-3(1) | conv3-64-3(1) |

A. The Loss Functions

As shown in Figure 1, the same with BeautyGAN, we implement the four losses, i.e., the adversarial loss $\mathcal{L}_{adv}$, cycle loss $\mathcal{L}_{cyc}$, perceptual loss $\mathcal{L}_{per}$ and makeup loss $\mathcal{L}_{makeup}$, to train the proposed student network. Please refer to [1] for details of these losses.

B. Collaborative Knowledge Distillation

In this paper, we distill the encoder of BeautyGAN by collaborative knowledge distillation (CKD) which was originally proposed in style transfer network compression [10]. BeautyGAN is an encoder-resnet-decoder based network, since the knowledge of the encoder is leaked into the decoder, we can compress the original encoder $E$ to the small encoder $E'$ by reducing the number of filters for each layer but maintaining the same architecture. Since the output dimensions of $E'$ and $E$ are different, CKD solves this mismatch problem by adding a $1 \times 1$ convolutional layer with linear activation function to each feature map of $E'$. Suppose that the $i$th feature of $E$ is represented as $F_i$ and its counterpart of $E'$ is represented as $F'_i$. Basically the dimension of $F'_i$ is smaller than that of $F_i$. Then CKD proposes a *feature loss* as follows:

$$L_{feat} = \sum_{i=1}^{n} \| F_i \times Q_i F_i \|_2. \quad (1)$$

where each $Q_i$ is a learnable matrix to make the dimension of $F'_i$ match the dimension of $F_i$. For our implementation, $n = 4$ because we distill 2 convolutional layers for both the non-makeup and reference branches. As shown in Table I.
first convolutional layers of the teacher and student’s encoders are ‘conv7-3-64(1)’ (BeautyGAN) and ‘conv7-3-16(1)’ (student network), so the dimension of $Q_i$ (i = 1, 2) is $16 \times 64$ to match the outputs of the convolutional layers of BeautyGAN and student network. Similarly, the second convolutional layers of the teacher and student’s encoders are ‘conv4-64-128(2)’ (BeautyGAN) and ‘conv4-16-32(2)’ (student network), so the dimension of $Q_i$ (i = 3, 4) is $32 \times 128$.

The final loss function is the weighted combination of the feature loss and the four losses of BeautyGAN:

$$L = L_{feat} + \alpha L_{adv} + \beta L_{cyc} + \gamma L_{per} + \sigma L_{makeup}. \quad (2)$$

### C. Residual Blocks Decomposition

Following the idea of MobileNets [11], we decompose the standard convolution of residual blocks in BeautyGAN into the combination of depth-wise and point-wise convolutions. Figure 2(a) and (b) show the architectures of the standard residual block of BeautyGAN and the decomposed block of the proposed student network, respectively. We decompose the standard convolution of residual blocks into a depth-wise and point-wise convolution.

The computation cost of a standard convolutional layer is $K \times C \times H_i \times W_i$ to make the number of the output channels consistent with the original. Suppose that a standard convolution takes an $H_i \times W_i \times C_i$ feature map as input and outputs an $H_o \times W_o \times C_o$ feature map, where $H, W$ and $C$ refer to the height, width and channel of the feature maps, respectively. If the kernel size is $K \times K$, the computation cost of a standard convolutional layer is $C_{stand} = K^2C_oC_iH_iW_i$. After decomposition, the computation cost is changed to $C_{decom} = K^2C_oH_iW_i + C_iC_oH_iW_i$, which is the sum of the depth-wise and point-wise convolutions. Thus, the computation reduction $R$ is calculated as:

$$R = \frac{N_{decom}}{N_{stand}}\frac{C_{decom}}{C_{stand}} = \frac{N_{decom}}{N_{stand}} \left( \frac{1}{K^2} + \frac{1}{C_o} \right) \quad (3)$$

Here, $N_{stand}$ and $N_{decom}$ refer to the number of standard and decomposed residual blocks in the teacher and student networks respectively. We set $N_{stand} = 6$ for BeautyGAN and $N_{decom} = 9$ for the student network (Refer to Section III for ablation study of $N_{decom}$). If we put $K = 3$ and $C_o = 256$ into Equation (3), we can obtain $R = 0.173$, which means that the computation of the decomposed residual blocks has been reduced to 17.3% of the uncompressed ones.

### III. Experiments

Data for experiments is from the MT dataset provided in BeautyGAN [1], which consists of 3,834 high-resolution face images with and without makeup. We use 3,600 images for training and 234 images for testing. The parameters in Equation (2) are set to the same as BeautyGAN, i.e., $\alpha = 1$, $\beta = 10$, $\gamma = 0.005$ and $\sigma = 1$. We set $N_{stand} = 6$ which is the same as the original BeautyGAN, and $N_{decom} = 9$ for the student network. We first train the student network from scratch for 80 epochs till convergence, then the feature loss of CKD is added to the student network and it is trained for another 60 epochs. The initial learning rate is $2 \times 10^{-4}$ and it decays linearly after 30 epochs.

#### A. The Compression Effects

Table II compares the computation and storage of the proposed student network and BeautyGAN. It can be seen that compared with the original BeautyGAN, The proposed student model consumes around 1/3 parameters, 1/2 MACs and 81.5% inference time.

| Model      | #Params | Model Size | MACs  | Inference Time |
|------------|---------|------------|-------|----------------|
| BeautyGAN  | 9.23M   | 30098KB    | 66.89G| 0.4952s        |
| Proposed   | 3.13M   | 12312KB    | 38.26G| 0.4037s        |

Figure 3 shows the makeup images generated by BeautyGAN and the proposed student network, respectively. It can be seen that the perceptual quality of both models is comparable. For some images such as the first and the last rows, makeup images generated by the student network are more similar with other is generated by the proposed student network. Together showing the corresponding non-makeup and reference images.

To further compare the quality of the generated images, we conduct a user study by randomly selecting 10 image pairs, of which one is generated by BeautyGAN, and the other is generated by the proposed student network. Together showing the corresponding non-makeup and reference images, we ask the 111 users to choose makeup images which are better in perceptual quality. The result is that of all 1,110 votes, there are 41.08% for BeautyGAN and 58.92% for the student network, which also indicates that the proposed student network generates comparable even better makeup images than the original BeautyGAN.

The perceptual quality can be quantitatively evaluated by two metrics – the makeup distance $D_{makeup}$, which is calculated by the histogram loss between the makeup and reference images [1], and the face distance $D_{face}$, which is...
the perceptual loss calculated by extracting feature maps of makeup faces and non-makeup faces from the 18th layer of the pretrained VGG16 model [6]. For the proposed student network, the makeup distance averaged over all testing images is 15.09, compared with 15.13 for BeautyGAN. In addition, the average face distance of the proposed student network is 0.163, compared with 0.188 for BeautyGAN. Thus, the proposed student model performs better for both metrics.

B. Ablation Study

The first ablation study is to compare the performance of the student models without and with the distillation process. Figure 4 shows three generated images without/with distillation. It can be seen that both models generate makeup images with similar perceptual quality. Table III shows $D_{\text{makeup}}$, $D_{\text{face}}$ and the inference time averaged over all testing images. It can be seen that the distilled model obtains smaller $D_{\text{makeup}}$ but bigger $D_{\text{face}}$ than the non-distilled model, which also indicates that the performance of these two models is comparable. Because the distilled model uses a smaller encoder to replace the encoder of the original BeautyGAN, it has less inference time than the non-distilled one. This ablation study shows that the main functionality of distillation is to save computation rather than generate perceptual pleasing images.

| $N_{\text{decom}}$ | 6 | 8 | 9 | 10 |
|-------------------|---|---|---|----|
| $D_{\text{makeup}}$ | 16.5514 | 16.5628 | 16.5538 | 16.4596 |
| $D_{\text{face}}$ | 0.1148 | 0.1401 | 0.1072 | 1.7742 |
| Inference Time(s) | 0.3926 | 0.4233 | 0.4424 | 0.4535 |

IV. Conclusion

To save computation and storage of facial makeup transfer networks, we introduce a two-step compression method including collaborative knowledge distillation and kernel decomposition for BeautyGAN. Extensive experiments show the effectiveness of the proposed facial makeup transfer network.

V. ACKNOWLEDGMENT

This work is supported by the Natural Key R&D Program of China (Grant No. 2017YFB1002400) and Natural Science Foundation of Zhejiang Province (Grant No. LY16F010004).

REFERENCES

[1] 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 ACM MM, 2018, pp. 645–653.
[2] C. Li, K. Zhou, and S. Lin, “Simulating makeup through physics-based manipulation of intrinsic image layers,” in CVPR, 2015, pp. 4621–4629.
[3] W. S. Tong, C. K. Tang, M. S. Brown, and Y. Q. Xu, “Example-based cosmetic transfer,” in Proceedings of the Pacific Conference on Computer Graphics and Applications, 2007, pp. 211–218.
[4] S. Liu, X. Ou, R. Qian, W. Wang, and X. Cao, “Makeup like a superstar: Deep localized makeup transfer network,” ArXiv preprint: 1604.07102, 2016.
[5] L. A. Gatys, A. S. Ecker, M. Bethge, A. Hertzmann, and E. Shechtman, “Controlling perceptual factors in neural style transfer,” in CVPR, 2017, pp. 3985–3993.
[6] J. Johnson, A. Alahi, and L. Fei-Fei, “Perceptual losses for real-time style transfer and super-resolution,” in ECCV, 2016, pp. 694–711.
[7] H. Peng, J. Wu, S. Chen, and J. Huang, “Collaborative channel pruning for deep networks,” in ICML, 2019, pp. 5113–5122.

[8] M. Courbariaux, I. Hubara, D. Soudry, R. E. Yaniv, and Y. Bengio, “Binarized neural networks: Training deep neural networks with weights and activations constrained to +1 or −1,” ArXiv preprint: 1602.02830, 2016.

[9] G. Hinton, O. Vinyals, and J. Dean, “Distilling the knowledge in a neural network,” ArXiv preprint: 1503.02531, 2015.

[10] H. Wang, Y. Li, Y. Wang, H. Hu, and M. H. Yang, “Collaborative distillation for ultra-resolution universal style transfer,” in CVPR, 2020, pp. 1860–1869.

[11] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, and H. Adam, “MobileNets: Efficient convolutional neural networks for mobile vision applications,” ArXiv preprint: 1704.04861, 2017.

[12] D. Osokin, “Real-time 2D multi-person pose estimation on CPU: Lightweight OpenPose,” ArXiv preprint: 1811.12004, 2018.