SKIN TONE DISENTANGLEMENT IN 2D MAKEUP TRANSFER WITH GRAPH NEURAL NETWORKS

Masoud Mokhtari[⋆] *Fatemeh Taheri Dezaki*[⋆] *Timo Bolkart*[†]
Betty Mohler Tesch *Rahul Suresh* *Amin Banitalebi-Dehkordi*

BeautyTech, Amazon

ABSTRACT

Makeup transfer involves transferring makeup from a reference image to a target image while maintaining the target's identity. Existing methods, which use Generative Adversarial Networks, often transfer not just makeup but also the reference image's skin tone. This limits their use to similar skin tones and introduces bias. Our solution introduces a skin tone-robust makeup embedding achieved by augmenting the reference image with varied skin tones. Using Graph Neural Networks, we establish connections between target, reference, and augmented images to create this robust representation that preserves the target's skin tone. In a user study, our approach outperformed other methods 66% of the time, showcasing its resilience to skin tone variations.

Index Terms— Graph Neural Networks, Generative Models, Style Transfer, Feature Disentanglement

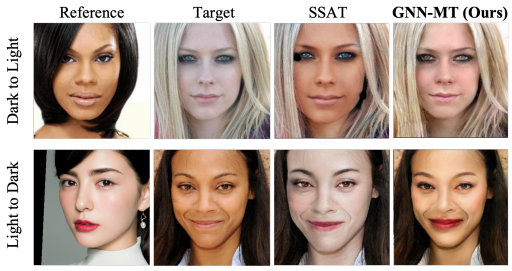

Fig. 1. Sample results using our approach on MT Dataset [1]: Our model excels in preserving skin tone for both dark-to-light (top) and light-to-dark (bottom) cases.

1. INTRODUCTION

Virtual try-on of makeup allows for quick and easy exploration of a wide range of products, creating a seamless digital shopping experience. Of these virtual try-on techniques, makeup transfer is among the most popular, which involves extracting facial makeup components from a reference image and applying them to a target image. These components are often composed of eyes, lips and skin makeup such as blushes. The quality of the transferred image is evaluated based on how well the texture and color of the makeup are transferred, how well the non-makeup components of the target image are preserved and how accurately the makeup is applied to the correct regions of the target image.

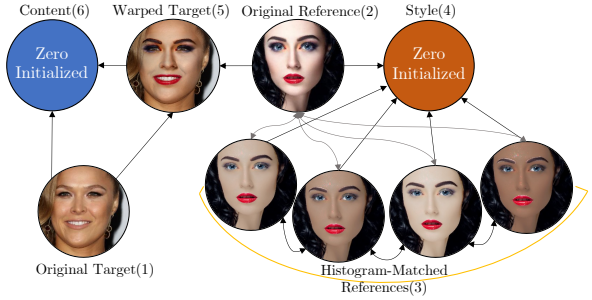

Fig. 2. Makeup Graph: Using a graph and GNN processing, we create a skin tone-robust style embedding for the reference image and a content embedding for the target image. The Style node (4) aggregates messages from nodes with matching makeup but diverse skin tones to capture the robust style. Meanwhile, the Content node (6) retains the target's identity by observing its original version (1) and a version with the reference image's makeup applied through face warping (5), despite some artifacts.

While prior Generative Adversarial Networks (GAN) based works [2] show varying levels of performance in terms of transferred makeup accuracy, they struggle to disentangle facial makeup color from skin tone. More specifically, when the target and reference skin tones do not match, the target image takes on the reference's skin tone. In this paper This study aims to disentangle skin tone in makeup transfer between unpaired images using a specialized Graph Neural Network (GNN-MT) illustrated in Fig. 2. Graph-learned style and content embeddings drive end-to-end training for makeup image transfer. This streamlined approach enhances the effectiveness of our makeup transfer framework. Since our reference style embedding is robust to skin tone, GNN-MT's transferred images affect the target image's skin tone considerably less than prior works while accurately transferring the

*Equal contribution.
†Contributions while at Amazon Research Development Center Germany.

makeup components.

In summary, our main contribution is a novel framework for 2D makeup transfer that improves skin tone preservation by incorporating GNNs and synthetic histogram matching into GAN-based makeup transfer.

2. PRIOR WORKS

Li *et al*. [1] introduce BeautyGAN, a makeup transfer framework that relies on a component-wise makeup objective function where the lips, eyes and skin area of the reference and target images are extracted before histogram matching these components to produce pseudo labels. LADN is another work that focuses on transferring facial patterns and makeup from a reference image onto a target image by using local patch-wise discriminators [3]. Jian *et al*. [4] enhance makeup spatial accuracy in PSGAN by incorporating pixel-wise attention based on the relational position of pixels with facial landmarks. Similarly, SSAT [5] uses a semantic attention mechanism along with multi-scale feature fusion to combine content and semantic information, while Yang *et al*. [6] introduce an attention mechanism to capture the correspondence between the target and the reference at both high and low resolutions, capturing high-frequency pattern information as well as low-frequency textural information. In SCGAN, Deng *et al*. [7] use a dual branch network where one branch extracts a pose-invariant embedding from the components of the reference image, while the other branch extracts a content embedding from the target image. CPM [8] extracts color information in the UV space to be pose-invariant. Recently, Li *et al*. [9] employs transformers for this task, while Xiang *et al*.[10] computes attention masks to transfer makeup from the corresponding regions of the reference image to the target image.

While these prior works show varying levels of success for makeup transfer accuracy, they struggle to keep the target image's identity when the skin tone of the target and reference image differ. In this work, we aim to address this shortcoming by extracting a skin tone-robust embedding from the reference image via constructing a graph structure and learning its node embeddings using GNNs.

3. METHOD

3.1. Problem Formulation

We assume that our makeup dataset is unpaired, meaning that the images in the makeup category are not associated with the images in the non-makeup category on a before/after makeup basis. Consequently, we denote the set of non-makeup images with $X \in \mathbb{R}^{H\times W\times 3}$ and the set of makeup images with $Y \in \mathbb{R}^{H\times W\times 3}$ where H and W are the height and width of the RGB images. At each training iteration, we create a pseudo pair by sampling $x_i \in X$ and $y_j \in Y$ where i and j are uniformly sampled from $[1, |X|]$ and $[1, |Y|]$, respectively. Our goal is to learn a function $g : \{\mathbb{R}^{H\times W\times 3}, \mathbb{R}^{H\times W\times 3}\} \mapsto \mathbb{R}^{H\times W\times 3}$ that transfers the makeup components of the makeup reference image onto the non-makeup target image while preserving the target's identity including its skin tone.

3.2. Graph Construction

One of the main components of our framework is how we construct a graph that enables obtaining a skin tone-robust embedding for the reference and target images. As shown in Fig. 2, the graph is constructed such that a style node is connected to skin tone-augmented versions of the reference, while a content node is connected to the original and a face-warped version of the target image. This way, the style node receives information from images that have the same makeup but different skin tones, while the content node receives information from images that have the identity of the target image but with different styles.

More formally, let us denote this graph with $G_{\text{makeup}}(V, E)$ where V is the set of nodes in the graph and E is the set of edges such that if there is an edge from $v_i \in V$ to $v_j \in V$, then $(v_i, v_j) \in E$. $|V|$ denotes the number of nodes in this graph which is equal to $5 + |V_{\text{hm}}|$, where the five nodes consisting of the target, warped target, content, style and the original reference are fixed, but the number of histogram matched reference nodes denoted by $|V_{\text{hm}}|$ is variable and set by a hyperparameter.

3.2.1. Reference Image Histogram Matching

To generate the skin tone augmented versions of the reference image from Fig. 2, we first generate a number of synthetic portrait images where the depicted subject's skin tone varies from light to dark. As illustrated in Fig. 3, for each augmented reference image, one of the synthetic images are randomly chosen. A BiSeNet-based face parser [11] is then used to segment the synthetic and the original reference images. Using the segmentation map, the skin area of both images is extracted. Lastly, the skin area of the original reference image is histogram matched to that of the synthetic image before mixing the resulting skin image with the original reference image. Note that the adoption of synthetic faces for skin tone diversification allows us to generate a large variety of shades (38 in our experiments), which is difficult to collect through natural image datasets.

3.3. Graph Representation Learning

GNNs efficiently capture graph-structured data and can be used for a variety of tasks including node, edge or graph classification and regression [12]. We use a GNN to enable message passing [13, 14] among the nodes in the Makeup Graph and learn the embeddings of the Style and Content nodes, which are used as inputs to separate style and content

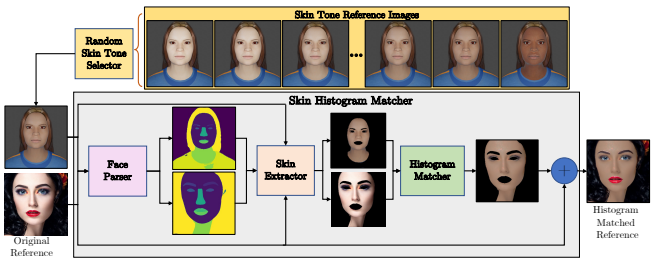

Fig. 3. Reference Image Skin Tone Augmentation: Synthetic images are used to augment the reference images and generate a wide variety of skin tones via histogram matching.

branches shown in Fig. 4. More formally, for each target/reference pair, multiple GNN layers are used to process the node features in the Makeup Graph G_{makeup} as:

$$h_{l+1} = \text{ReLU}(\text{GNN}_l(G_{\text{makeup}}), h_l), \quad l \in [1, L], \quad (1)$$

where L denotes the number of layers in the network, and $h_0 \in \mathbb{R}^{|V| \times H \times W \times 3}$ is the initial node embeddings. Additionally, for the GNN layers, while we use the general framework of Graph Convolutional Networks (GCN) [15], we modify their message function to use a two-layer Convolutional Neural Network (CNN) that preserves the spatial dimension of the input features.

3.4. Overall Model Architecture

As shown in Fig. 4, we build GNN-MT's generator as an extension of SCGAN's generator [7]. After creating the Makeup Graph for a target/reference pair, we use GNNs to obtain embeddings for the nodes in the graph. The embedding for the Content node is then fed into the content branch, while the learned embedding for the Style node is fed into the style branch. These embeddings are then combined via feature fusion before being decoded into the output image.

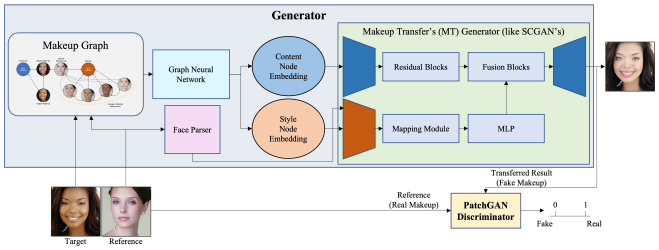

Fig. 4. GNN-MT Model Architecture: Building on SCGAN's generator, we enhance it with a Makeup Graph processed by GNNs. This yields target content and reference skin tone-robust embeddings. These embeddings then drive SCGAN's content and style branches for makeup transfer. During training, the PatchGAN discriminator [16] ensures realistic image generation.

3.5. Objective Functions

For conciseness, we skip the details of most commonly used loss functions in the GAN-based makeup transfer literature and summarize them in the list below. The details for the more specialized loss functions are provided in the subsections that follow.

- To train the GAN, we use the adversarial loss from [2], which we denote by L^D_{adv} and L^G_{adv}, which are the discriminator and generator losses, respectively.
- To encourage accurate transfer of makeup, we use the component-wise makeup loss first introduced by Li *et al*. [1], in which histogram matching of facial components is performed to produce pseudo makeup labels. We denote this by L_{makeup}.
- To enable learning without paired data, we use the cycle consistency objective function introduced by [17], which we denote by L_{cycle}.
- To further improve the preservation of the target's identity, we minimize a perceptual loss (Mean Squared Error (MSE) between intermediate VGG [18] features) and a pixel-wise reconstruction loss (MSE) between the original target image and the output transferred image. We denote the sum of these losses by $L_{\text{reconstruction}}$.

3.5.1. *Style Node Consistency Loss*

In the Makeup Graph illustrated in Fig. 2, we expect the Style node to capture a skin tone-robust style embedding using the messages it receives from the augmented versions of the reference image. Therefore, we expect this learned embedding to be equally close to the original reference image and its augmented versions in the Euclidean space. To enforce this, we introduce the Style Node Consistency Loss as shown in Equation 2.

$$L_{\text{style\_node}} = \|F_l(h_{\text{style}}) - F_l(y^k_{\text{hm}})\|_2, \quad (2)$$

with $k \in [1, |V_{\text{hm}}| + 1]$, and where $F_l(.)$ is the l-th layer feature of a pre-trained VGG [18] model, h_{style} is the learned embedding for the Style node, and $\{y_{\text{hm}}\}$ is the set of augmented versions of the reference image.

3.5.2. *Content Node Loss*

To ensure that the Content node embedding matches the identity features of the target image, we use the term:

$$L_{\text{content\_node}} = \|F_l(h_{\text{content}}) - F_l(x)\|_2, \quad (3)$$

where h_{content} is the learned embedding of the Content node, and x is the original target image.

3.5.3. Total Loss

Putting it all together, the final objective is defined as:

$$
\begin{aligned}
L^{D}_{\text{total}} &= \lambda_{\text{adv}} L^{D}_{\text{adv}} \\
L^{G}_{\text{total}} &= \lambda_{\text{adv}} L^{G}_{\text{adv}} + \lambda_{\text{cycle}} L_{\text{cycle}} + \lambda_{\text{style\_node}} L_{\text{style\_node}} \\
&\quad + \lambda_{\text{content\_node}} L_{\text{content\_node}} + \lambda_{\text{makeup}} L_{\text{makeup}} \\
&\quad + \lambda_{\text{reconstruction}} L_{\text{reconstruction}},
\end{aligned} \tag{4}
$$

where λ terms are the weights given to each term.

4. EXPERIMENTS

4.1. Dataset

To train and evaluate our framework, we use the Makeup Transfer (MT) dataset [1], which is an unpaired dataset consisting of 1,115 non-makeup and 2,719 makeup images. We use the same training and test splits as [1] and [7]. Additionally, to further showcase the generalize-ability of GNN-MT, we use the Facial Cosmetic Content (FCC) dataset [19] as an independent test set. This dataset consists of 13,112 light makeup, 4,422 strong makeup, and 4,148 non-makeup images extracted from online YouTube videos. For both datasets, we down-sample the images to a fixed size of 256×256 for fair comparison with prior works that follow the same down-sampling strategy.

4.2. Qualitative and Quantitative Results

In Figs. 5 and 6, we qualitatively show that for both dark to light and light to dark cases, GNN-MT preserves the skin tone of the target image more than prior works. Additionally, we see that while the skin tone is preserved, facial makeup features such as blushes are preserved by our model (e.g. bottom row of Fig. 5). Moreover, we see that our model performs better in terms of hair color preservation compared to prior works as most evident by the first two rows of Fig. 1. It must also be noted that all models perform relatively worse on the FCC dataset compared to the MT-dataset, and we attribute this to the domain shift between the two datasets where the FCC dataset includes frames of lower quality videos with more in-the-wild image variations not seen during training. Nonetheless, our model still performs better in terms of skin tone preservation on the FCC dataset.

Additionally, as shown in Table 1, we quantitatively compare GNN-MT's performance to prior works. We include 4 prior works in a user study where we randomly choose 50 target and reference images from the MT dataset. 20 users are asked to select the best picture based on the image's realism, resemblance to reference makeup looks, and preservation of target's skin tone. We also employ other evaluation metrics including FID and the identity preservation metric used by Li et al. [9] on a randomly selected subset of MT Dataset.

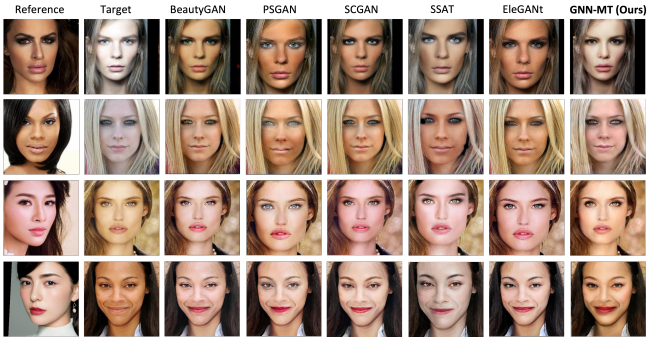

Fig. 5. Qualitative results on the MT dataset. We see that the GNN-MT model is more robust to cases where the skin tone of the target and reference do not match and preserves the skin tone of the target better than prior works.

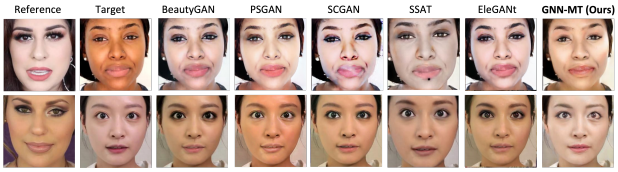

Fig. 6. Qualitative results on the FCC dataset for models trained on the MT dataset. While all models perform relatively worse when compared to Fig. 5, GNN-MT does a better job of preserving the skin tone of the target.

Table 1. Quantitative results show our model is preferred over prior works in the user study (User Pref.) while also showing better FID and identity preservation performance (ID Pres.).

Model	User Pref [%] ↑	FID ↓	ID Pres. ↑
SCGAN	2	64.23	0.8153
PSGAN	10	59.25	0.6348
BeautyGAN	22	56.59	0.8980
SSAT	*excluded*	60.65	0.5889
EleGANt	*excluded*	62.32	0.6626
GNN-MT (Ours)	**66**	**50.12**	**0.9425**

5. CONCLUSION

We presented a GNN-based method to transfer facial makeup from a reference image to a target image, which improves skin tone preservation of the subject by disentangling skin tone from makeup. While our model outperforms the state of the art in terms of skin tone preservation for the task of makeup transfer, it has some shortcomings that will be addressed in future works. For one, while our reference image augmentations are suitable for skin tone preservation, we still need a mechanism to preserve the illumination of the target.

6. REFERENCES

[1] Tingting Li, Ruihe Qian, Chao Dong, Si Liu, Qiong Yan, Wenwu Zhu, and Liang Lin, "BeautyGAN: Instance-level facial makeup transfer with deep generative adversarial network," in *Multimedia Conference on Multimedia Conference (MM)*, 2018, pp. 645–653.

[2] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger, Eds. 2014, vol. 27, Curran Associates, Inc.

[3] Qiao Gu, Guanzhi Wang, Mang Tik Chiu, Yu-Wing Tai, and Chi-Keung Tang, "LADN: Local adversarial disentangling network for facial makeup and de-makeup," in *International Conference on Computer Vision (ICCV)*, 2019, pp. 10480–10489.

[4] Wentao Jiang, Si Liu, Chen Gao, Jie Cao, Ran He, Jiashi Feng, and Shuicheng Yan, "PSGAN: Pose and expression robust spatial-aware gan for customizable makeup transfer," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 5193–5201.

[5] Zhaoyang Sun, Yaxiong Chen, and Shengwu Xiong, "Ssat: A symmetric semantic-aware transformer network for makeup transfer and removal," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 2, pp. 2325–2334, Jun. 2022.

[6] Chenyu Yang, Wanrong He, Yingqing Xu, and Yang Gao, "Elegant: Exquisite and locally editable gan for makeup transfer," in *Computer Vision – ECCV 2022*, Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, Eds., Cham, 2022, pp. 737–754, Springer Nature Switzerland.

[7] Han Deng, Chu Han, Hongmin Cai, Guoqiang Han, and Shengfeng He, "Spatially-invariant style-codes controlled makeup transfer," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 6549–6557.

[8] Thao Nguyen, Anh Tran, and Minh Hoai, "Lipstick ain't enough: Beyond color matching for in-the-wild makeup transfer," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[9] Mingxiu Li, Wei Yu, Qinglin Liu, Zonglin Li, Ru Li, Bineng Zhong, and Shengping Zhang, "Hybrid transformers with attention-guided spatial embeddings for makeup transfer and removal," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2023.

[10] Jianfeng Xiang, Junliang Chen, Wenshuang Liu, Xianxu Hou, and Linlin Shen, "Ramgan: Region attentive morphing gan for region-level makeup transfer," in *ECCV 2022*, Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, Eds. 2022, pp. 719–735, Springer Nature Switzerland.

[11] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang, "BiSeNet: Bilateral segmentation network for real-time semantic segmentation," in *European Conference on Computer Vision (ECCV)*, 2018, vol. 11217, pp. 334–349.

[12] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini, "The graph neural network model," *Transactions on Neural Networks*, vol. 20, no. 1, pp. 61–80, 2009.

[13] Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl, "Neural message passing for quantum chemistry," in *International Conference on Machine Learning (ICML)*, 2017, vol. 70, pp. 1263–1272.

[14] Renjie Liao, *Deep Learning on Graphs: Theory, Models, Algorithms and Applications*, Ph.D. thesis, University of Toronto (Canada), 2021.

[15] Thomas N. Kipf and Max Welling, "Semi-supervised classification with graph convolutional networks," in *International Conference on Learning Representations (ICLR)*, 2017.

[16] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros, "Image-to-image translation with conditional adversarial networks," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 5967–5976.

[17] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *International Conference on Computer Vision (ICCV)*, 2017, pp. 2242–2251.

[18] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations (ICLR)*, 2015.

[19] M. Saquib Sarfraz, Constantin Seibold, Haroon Khalid, and Rainer Stiefelhagen, "Content and colour distillation for learning image translations with the spatial profile loss," in *British Machine Vision Conference (BMVC)*, 2019, p. 287.