IFSCC 2025 full paper (223)

"AI Redefining the Aesthetics of Makeup: Unveiling Patterns of Beauty and Achieving Fair and Diverse Expressions"

Shun Obikane[1,2,*], Haruna Tagawa[2], Rie Nakamura[1], Ryo Sasaki[1] and Yoshimitsu Aoki [2]

[1]Research Laboratories, KOSÉ Corporation, Tokyo, Japan;
[2]KEIO University, Kanagawa, Japan

1. Introduction

The diversity of makeup culture has evolved across the world, embracing creativity, individuality, and cultural richness. Today, makeup is not simply a means of beautification; it is a powerful medium of self-expression that reflects personal identity and societal trends. As the makeup landscape becomes increasingly diverse and dynamic, there is a growing need for technologies that can help us understand, analyze, and expand this complex world in an objective and inclusive manner. Artificial intelligence (AI) offers promising opportunities to support this understanding. However, conventional AI approaches often rely heavily on human-labeled data. Such reliance not only incurs high costs in terms of time and labor but also introduces the risk of bias, as human annotations inevitably reflect subjective preferences and cultural assumptions. These biases can limit the fairness, diversity, and creativity that are essential to makeup culture. To address these challenges, we propose a computational aesthetics approach that extracts meaningful aesthetic features from makeup images without requiring human labeled data. Our model focuses on two fundamental aspects of makeup "shape and color" and learns the relationships between images through comparisons, inspired by the way humans naturally perceive and interpret visual patterns. By avoiding the need for pre-defined labels, the model enables the discovery of objective similarity relationships, thereby preserving the authentic diversity of makeup styles. Our approach has several advantages. First, Our method supports clustering of makeup styles based on shared visual characteristics, allowing for the organization of vast image datasets into meaningful groups. Second, it facilitates trend analysis by capturing the continuous transitions and discontinuities inherent in makeup evolution, without artificially imposing rigid categories. Third, it enables personalized recommendations, helping users find preferred styles and explore new possibilities in an intuitive and unbiased manner. Furthermore, by promoting equal access to a wide range of makeup expressions, our model contributes to creating a more inclusive environment where creativity is not constrained by pre-existing stereotypes or biases. For consumers, it enhances the makeup experience by enabling more flexible and authentic self-expression. For companies, it offers a cost-effective and objective tool for real-time trend discovery, based on the analysis of large-scale makeup images shared on social media platforms

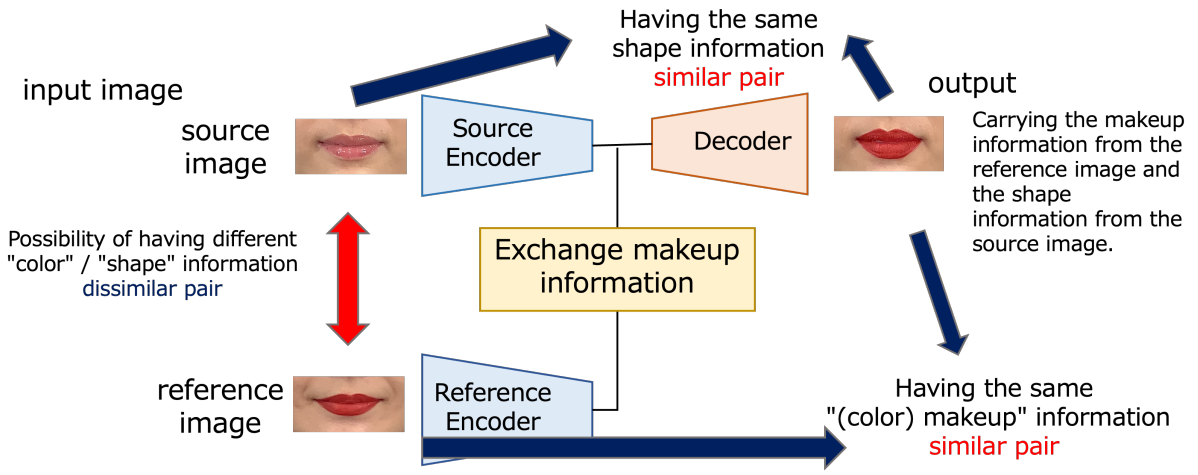

Figure 1. Overview of our feature extraction method. Our method extracts features relevant to makeup by performing makeup transfer, where the makeup style from a reference image is applied to a source image. We divide makeup similarity into shape and color, assigning the source encoder to capture shape-related features and the reference encoder to capture color-related ones. Our model learns by focusing on the shared and differing attributes among the source, reference, and generated images.

worldwide. Ultimately, we believe that the fair and unbiased application of AI to makeup culture is essential for fostering both innovation and inclusivity. By bridging sensory information and computational analysis, our model helps to preserve the dynamic, evolving nature of makeup culture while expanding its possibilities for individuals and industries alike.

2. Materials and Methods

2-1. Overview of our method

Our approach is based on representation learning to capture aesthetic relationships between makeup images, with a particular focus on shape and color [1]. Rather than relying on human annotations, we leverage image similarity through makeup transfer[2-4] to extract meaningful features in an unsupervised manner [5-6]. This enables three core capabilities: (1) unsupervised clustering of makeup styles, (2) analysis of style transitions, and (3) personalized, bias-free style recommendation.

2-2 . Training Strategy

Our training process of consists of three sequential steps. In the first step, the model is pre-trained to adapt to the characteristics of the target dataset. Since our task focuses on makeup, this step involves training the model specifically on facial images. The second step involves training a model for makeup transfer, which is a technique that transfers cosmetic attributes - such as lip color or eyeshadow - from a reference image to another face image. In the final step, we utilize the output images generated through makeup transfer to create pseudo pairs of similar and dissimilar images. These pairs are then used to train the model to learn similarity relationships in terms of color and shape, without relying on manually labeled data.

Model Architecture: Our model consists of three main components, as illustrated in Figure 1. This architecture follows a structure commonly used in makeup transfer models. To extract information from images, it uses two encoders. The source encoder takes as input a face image to which makeup is to be applied, and extracts its relevant features. The reference encoder

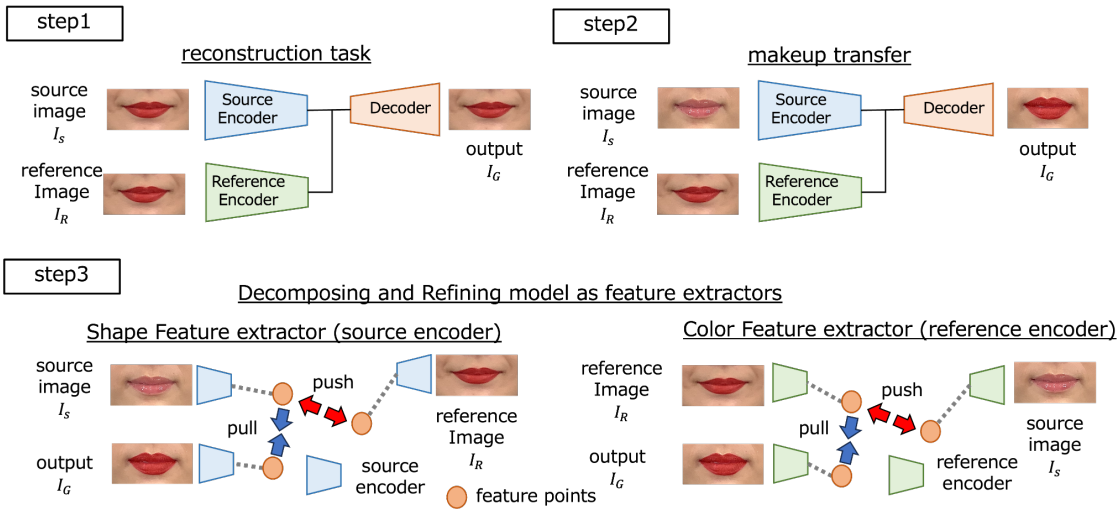

Figure 2. Strategies for each of the three learning steps. This figure illustrates the overall training procedure of our method, which consists of three sequential stages. In Step 1, the model is trained to reconstruct the input image. In Step 2, it learns to perform makeup transfer from a reference image to a source image. In Step 3, the source and reference encoders are separated and refined into final feature extractors specialized for shape and color, respectively.

processes a reference image that contains the desired makeup style, and extracts the corresponding cosmetic features. The information obtained from both encoders is then passed to a decoder, which integrates the extracted features and generates the final output image. As a result, the output is a version of the source image enhanced with the makeup features taken from the reference image. This architecture enables the model to naturally transfer makeup characteristics from one face to another. These two encoders are trained to serve distinct roles: the source encoder is refined to extract shape-related features, while the reference encoder is developed to focus on color-related features.

Step 1: Pre-training

First, we perform pretraining to ensure the model can effectively process facial images. As shown in Figure 2, the same image is input into both encoders, and the decoder is trained to reconstruct the original image. This helps the model accurately capture facial features.

Step 2: Learning the Makeup Transfer Model

In Step 2, the model learns makeup transfer, where makeup information from a reference image is applied to a source image as shown in Figure 2. The goal is to generate a source image that reflects the makeup style of the reference image. To train this process, we define four loss functions: adversarial loss, perceptual loss, makeup loss, and feature matching loss. Among these, feature matching loss plays a particularly important role in the final step (Step 3), and will therefore be explained in more detail later.

Adversarial loss is a loss function commonly used in generative models such as GANs (Generative Adversarial Networks) [2-4]. In this framework, two models are involved: one is the generator, which produces images, and the other is the discriminator, which tries to determine whether a given image is real or generated. The generator is trained to fool the discriminator by producing images that are so realistic that the discriminator cannot tell they are fake. At the same time, the discriminator is trained to improve its ability to distinguish real from fake. Through this adversarial learning process, the generator gradually improves, ultimately learning to produce highly natural and realistic images.

Perceptual loss evaluates the visual similarity between images based on high level features such as appearance and texture. In our model, using source encoder, its intermediate features are used to compare the generated image with the original. This helps the model produce images that are more consistent with human perception.

Makeup loss is a loss function that uses the result of applying makeup from a reference image to a source image as if it were a ground truth, even though it is not an actual label image. Such results are created using algorithms like histogram matching [7] and are treated as pseudo labels during training. This allows the model to receive a general guideline for the desired output, even without real ground truth images, helping it generate results that better align with the intended makeup style.

Feature matching loss plays the most key role. First, we organize the relationship between the source image, the reference image, and the result generated by the makeup transfer model. For clarity, we denote the source image as I_s, the reference image as I_R, and the generated image as I_G. The generated image I_G is an image of the same person as I_s, but with the makeup attributes of I_R.

From this relationship:

・**In terms of shape**, and represent the same individual, meaning they share the same structural features. Therefore, and can be treated as a pseudo similar pair for shape learning.

・**In terms of color**, and share the same makeup attributes, meaning they have similar color features. Thus, and, can be treated as a pseudo similar pair for color learning.

To quantify the similarity, we employ cosine similarity $\cos(x, y) = \frac{x \cdot y}{|x||y|}$ as the loss function. Specifically, let F_s denote the encoder that processes the source image and F_R denote the encoder that processes the reference image within the generator G. The objective is to minimize L_{shape} and L_{color} based on the cosine similarity between the corresponding feature representations.

$$L_{shape} = 1 - cos\big(F_s(I_s), F_s(I_G)\big)$$
$$L_{color} = 1 - cos\big(F_R(I_R), F_s(I_G)\big)$$

Through this design, even during Step 2, the model can effectively generate and learn feature representations where similar shapes and colors are properly aligned.

Step 3: Similarity Learning Using Pseudo Pairs

In the final training step, we refine the model into feature extractors by utilizing the makeup transfer model developed in the previous stages. First, we decompose the generator G into two encoders: the encoder F_s for the source image and the encoder F_R for the reference image. We then train each feature extractor individually using pseudo pairs.

Training of the Feature Extractor F_S for Shape Information

From the perspective of shape, the source image I_s and the generated imageI_G represent the same individual and thus share the same shape information. In contrast, the reference image I_R and I_G are likely to correspond to different individuals, implying differences in shape (although highly similar cases may occasionally occur).

Thus,

I_s and I_G are treated as a pseudo similar pair for shape learning,

$$L_{shape}^{similar} = 1 - cos\big(F_s(I_s), F_s(I_G)\big)$$

I_R and I_G are treated as a pseudo dissimilar pair for shape learning.

$$L_{shape}^{dissimilar} = 1 + cos\big(F_s(I_R), F_s(I_G)\big)$$

To train these encoders, we solve two separate optimization problems: one to minimize the loss $L_{shape}^{similar}$ that brings similar pairs closer, and another to minimize the loss $L_{shape}^{dissimilar}$ that pushes dissimilar pairs apart. These two losses are alternately optimized during training, allowing the model to balance the forces that bring similar features together and keep dissimilar features apart. Because dissimilar pairs may occasionally be similar, the loss function for dissimilarity is not applied at that learning step to prevent destabilizing the feature space. Incorporating dissimilar pairs into the learning process prevents the feature representations from collapsing too tightly, maintaining feature dispersion across the space. This strategy enables the model to handle tasks like makeup representation, where the differences in shape are often extremely subtle.

Training of the Feature Extractor F_R for Color Information

The training process for color information follows the same structure. From the perspective of color, the reference image I_R and the generated image I_G share the same makeup attributes, and thus are considered similar. In contrast, the source image I_S and I_G generally have different color characteristics, although very similar examples may occasionally occur.

Thus,

I_R and I_G are treated as a pseudo similar pair for color learning,

$$L_{color}^{similar} = 1 - cos\big(F_R(I_R), F_s(I_G)\big)$$

I_s and I_G are treated as a pseudo dissimilar pair for color learning.

$$L_{color}^{dissimilar} = 1 + cos\big(F_R(I_S), F_s(I_G)\big)$$

Again, the two types of loss are alternately optimized during training. When a dissimilar pair is found to be overly similar, the update is skipped at that learning step. This training strategy ensures that the feature space remains adequately dispersed, allowing the model to distinguish fine-grained variations in makeup color and better adapt to the subtleties of real-world makeup styles.

3. Results

We evaluated the performance of the proposed method using two different datasets, conducting both quantitative and qualitative evaluations. Since it is difficult to define clear and appropriate labels for measuring similarity in makeup, we first perform quantitative evaluation using a dataset with more general and well-defined labels. For the quantitative evaluation, we use a dataset of Ukiyo-e, a traditional form of Japanese painting, and assess the model's performance by predicting the artist of each artwork. Given that artworks also reflect abstract and aesthetic concepts, this task aligns well with the nature of our method, which is designed to learn similarity without relying on explicit labels. For the qualitative evaluation, we use a facial image dataset, focusing on lip makeup. Through visual inspection based on human perception, we evaluate how effectively the model captures makeup similarity.

Implementation detail

All learning steps were optimized using the Adam optimizer with parameters $\beta_1 = 0.500$ and $\beta_2 = 0.999$. The learning rate for the generator in Step 1 was set to 5.0×10^{-2}, for both the generator and discriminator in Step 2 to 1.0×10^{-3}, and for Step 3 to 5.0×10^{-4}. The batch size for all steps was fixed at 32. The number of training epochs was set to 200 for Step 1, 100 for Step 2, and 50 for Step 3. The encoder architecture is based on a structure similar to VGG [], while the decoder consists of a five-layer deconvolutional network. The dimensionality of the feature vectors generated by our model is fixed at 512 dimensions. The network was implemented using PyTorch and trained on a single NVIDIA RTX A5000 GPU.

3-1 Quantitative evaluation

For the quantitative evaluation, we utilized the ARC Ukiyo-e Faces Dataset [8], which consists of facial images derived from Ukiyo-e, a traditional genre of Japanese painting. In this evaluation setting, the artist associated with each artwork is used as the classification label. To construct the dataset, we selected 17 artists, each with more than 80 available images. Among these, 3,195 images were used for training, and 850 images were reserved for testing. The model independently generates feature representations for shape and color. For the purpose of evaluation, we integrated these features by computing their centroid, thereby obtaining a single unified feature representation for each image. The quality of the features generated by our method is evaluated through two tasks: recommendation and clustering analysis. Since our method is an unsupervised approach that does not rely on labels, the evaluation is conducted under the assumption that labels are available. In unsupervised learning, the target performance is generally considered to be the level achieved under supervised conditions. Therefore, we adopt ArcFace [9], a representative supervised method, as a benchmark for the supervised setting. For comparisons with other unsupervised methods, we evaluate against several representative approaches: DeepCluster [10], UEL [11], and UDMLSS [12].

For the recommendation task, we evaluate the model by using a test image as a query key to retrieve similar training images, and measure the performance of the retrieved results. We assesses the local structure of the learned feature space. We adopt Precision@K as the evaluation metric. Precision@K measures the proportion of correctly retrieved items among the top K retrieved results. A higher Precision@K indicates that more relevant images (i.e., images by the same artist) are ranked closer to the query image. Table 1 presents a comparison between the proposed method and existing approaches. We vary K=1,5,30 to assess how closely images by the same artist are located around each test image in the feature space. As shown in the results, the proposed method achieves better Precision@K scores than existing unsupervised methods, and even surpasses the performance under the supervised setting using ArcFace. This suggests that, for abstract tasks such as author identification, rigid reliance on label information may overly constrain the learning process. Table 2 summarizes the evaluation of each learning stage. Precision@K is again used for consistency. As the results indicate, performance improves progressively through each training step, confirming the effectiveness of the staged learning strategy. In particular, the model after Step 1, which corresponds to a standard autoencoder setting, achieves a baseline performance. This demonstrates the necessity of further refining feature representations based on semantic meaning rather than simple reconstruction. Furthermore, when evaluating at a broader neighborhood range (e.g., K=30), we observe a significant improvement after Step 3. Since Step 2 only focuses on pulling similar features closer together, Step 3 plays a critical role in appropriately dispersing feature representations, ensuring that the feature space maintains sufficient separation and structure.

Next, we evaluate the proposed method on a clustering task, which assesses the global structure of the learned feature space. For this purpose, we employ spectral clustering[13] as the clustering algorithm. Table 3 presents the results of our method compared with the same baseline methods used in the recommendation task. As shown, our approach achieves a performance comparable to ArcFace, a supervised method trained with explicit labels. Moreover, it outperforms other representative unsupervised methods, demonstrating the robustness and discriminative quality of the learned features. These results indicate that our

method produces high-quality representations suitable for both recommendation and clustering tasks. Despite being an unsupervised approach, the feature design achieved performance levels on par with supervised models and superior to other unsupervised techniques. Furthermore, in some tasks, the absence of label constraints appears to be beneficial, allowing the model to learn more flexible and unbiased representations of visual similarity.

Table 1. Comparison of our method with previous method on the recommendation task. The evaluation metric used is Precision@K, with results reported for K=1,5,10. The "Model" column indicates the name of each method, and the "Type" column specifies whether the method uses labeled data. As shown in the table, our method achieves better performance than existing methods and even surpasses the target performance typically achieved under supervised settings. These results highlight not only the strength of our approach itself but also the unique advantages of label-free learning.

Model	Type	K=1	K=5	K=30
ArcFace [9]	Supervised	0.34	0.24	0.04
DeepCluster [10]	Unsupervised	0.06	0.08	0.01
UEL [11]	Unsupervised	0.30	0.27	0.08
UDMLSS [12]	Unsupervised	0.42	0.29	0.17
Ours	Unsupervised	**0.50**	**0.51**	**0.41**

Table 2. Results of validating the effectiveness of each learning step on the recommendation task. This result showing the performance improvements at each training step. The evaluation metric used is Precision@K, with results reported for K=1,5,10. As shown in the figure, each successive training step leads to improved score, demonstrating the effectiveness of the proposed learning strategy

Learning step	K=1	K=5	K=30
Step1	0.45	0.40	0.15
Step 2	0.49	0.41	0.21
Step 3	**0.50**	**0.51**	**0.41**

Table 3. Results for the clustering task. The evaluation metric used is Normalized Mutual Information (NMI), which measures how well the data points are grouped into coherent clusters. As shown in the results, our method achieves higher clustering performance compared to existing methods and reaches a level of performance comparable to that of supervised approaches.

model	ArcFace[9]	Deep Cluster[10]	UEL[11]	UDMLSS[12]	Ours
type	Supervised	Unsupervised	Unsupervised	Unsupervised	Unsupervised
score	0.419	0.155	0.363	0.388	**0.397**

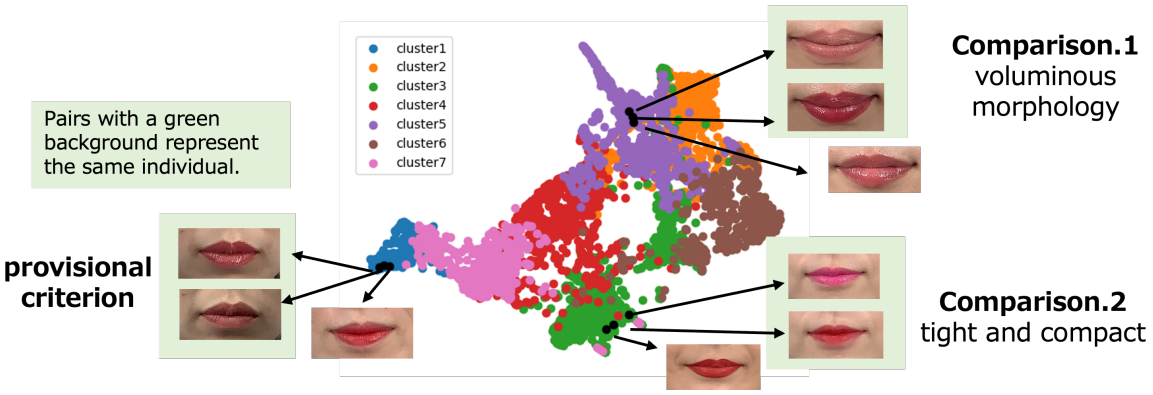

Figure 3. Clustering results based on shape features extracted for qualitative evaluation. The left plot shows a 2D embedding of shape features clustered using spectral clustering. Sample lip images from selected clusters are shown on the right. The clusters represent similarities in lip shape and capture continuous variations in structural characteristics without relying on explicit labels.

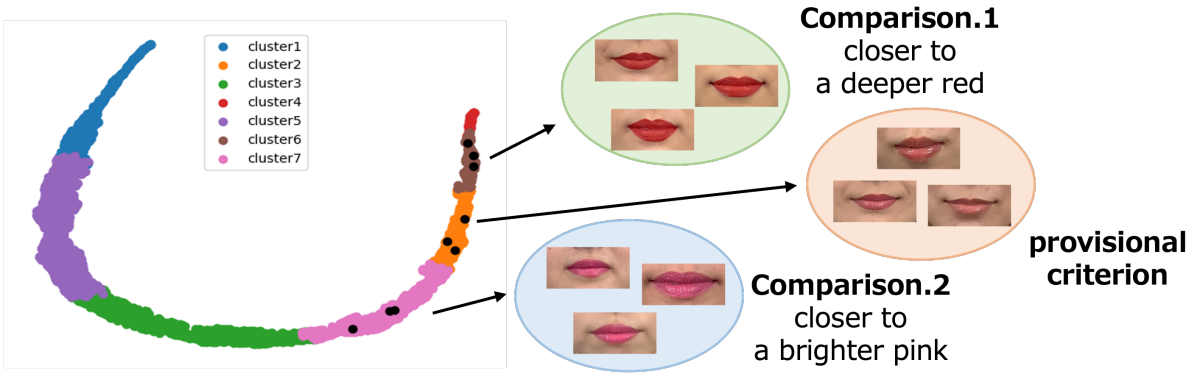

Figure 4. Clustering results based on color features extracted for qualitative evaluation. The left plot shows a 2D embedding of color features clustered by spectral clustering. Sample lip images from selected clusters are shown on the right. The clusters reflect similarities in lip color and capture continuous transitions across different color tones.

3-2 Qualitative evaluation

To evaluate the effectiveness of the proposed method, we conducted experiments using a dataset comprising 4,237 facial images. The analysis focused on the lip region, examining two main aspects: shape, referring to the contour and structure of the lips, and color, referring to the lip makeup tone. Two types of features were extracted using the proposed model: shape features from the source encoder and color features from the reference encoder. We applied spectral clustering to each type of feature separately, setting the number of clusters to 7. The resulting cluster structures were then qualitatively analyzed to investigate the semantic properties of the learned representations.

Figure 3 shows the clustering results based on shape features extracted from the source encoder. The features were projected into a two-dimensional space using UMAP (Uniform Manifold Approximation and Projection) for visualization. We highlight three representative clusters (clusters 1, 3, and 5), along with sample lip images from each cluster. The color of each point indicates the cluster label, and samples with a green background denote images from the same individual. The observed differences between clusters reflect meaningful

variations in lip contour and volume. For instance, compared to cluster 1, cluster 3 contains lips that are tighter and more compact, while cluster 5 includes fuller and more voluminous lips. These findings confirm that the source encoder effectively captures structural shape information relevant to lip appearance.

Figure 4 presents the clustering results using color features extracted from the reference encoder, visualized using UMAP [14] in the same manner as in Figure 3. We focus on clusters 2, 6, and 7, and show representative lip images for each cluster. This analysis emphasizes the importance of semantic continuity in color space—meaning that adjacent clusters should show gradual transitions in color tone and saturation. For example, cluster 2 represents orange-toned lips, while its neighboring clusters, cluster 6 and cluster 7, correspond to deeper red tones and lighter pink tones, respectively. Notably, cluster 2 lies between the other two in terms of hue and intensity, functioning as a semantic intermediary in the color space. These results demonstrate that the reference encoder captures smooth, meaningful gradients in lip color appearance.

4. Discussion

Furthermore, our model presents practical value for both cosmetic companies and digital beauty platforms. For example, the system can be deployed to analyze large volumes of user-generated makeup images from social media in real time, enabling early detection of emerging trends. Additionally, the ability to automatically cluster and recommend makeup styles based on shape and color features offers opportunities for personalized product recommendations, virtual try-on systems, and customized beauty experiences without relying on culturally biased labeling. For product developers, this approach enables data-driven design insights that reflect actual consumer behavior and visual preferences, while for consumers, it supports exploration of their identity through makeup in a bias-free and empowering manner. Importantly, our approach aligns with emerging principles of ethical AI and algorithmic fairness in the beauty domain. By explicitly avoiding human-annotated labels—which often embed cultural, gender, or socioeconomic biases—we allow aesthetic structures to emerge organically from the data, without privileging any particular ideal or stereotype. This opens a pathway to more inclusive and culturally adaptive beauty technologies, where individual users are not forced into narrow categories, but can instead explore diverse identities and preferences on their own terms. Such considerations are especially crucial in makeup and fashion, where cultural norms and self-expression vary significantly across regions, communities, and age groups.

5. Conclusion

We proposed a label-free, unsupervised learning framework to extract aesthetic features from makeup images, focusing on shape and color. Our approach enables clustering, trend analysis, and personalized recommendations without relying on human annotations. The model demonstrated strong performance in both recommendation and clustering tasks, surpassing existing unsupervised methods and matching supervised baselines. Importantly, by avoiding culturally biased labels, the method promotes fairness and inclusivity in beauty analysis. This framework offers practical value for both consumers and the cosmetics industry—supporting self-expression, enhancing product recommendations, and enabling real-time trend discovery from large-scale image data. Cultural shifts often manifest as changes in the structure of data itself, particularly in aesthetic domains. By observing these structural

transitions over time, we can not only trace past and present trends, but also anticipate future directions in visual culture. Looking ahead, the model could be extended to dynamic makeup analysis, virtual beauty tools, or cross-cultural aesthetic studies, contributing to a more ethical and diverse future for beauty technologies.

Reference

[1] *Jain, A.K., Vailaya, A.*, Image retrieval using color and shape. In: Pattern recognition, 29(8). pp. 1233–1244 (1996)
[2] *Li, T., Qian, R., Dong, C., Liu, S., Yan, Q., Zhu, W., Lin, L.*, Beautygan: Instancelevel facial makeup transfer with deep generative adversarial network. In: Proceedings of the 26th ACM international conference on Multimedia. pp. 645–653 (October 2018)
[3] *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: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020)
[4] *Xiang, J., Chen, J., Liu, W., Hou, X., Shen, L.*, Ramgan: Region attentive morphing gan for region-level makeup transfer. In: In Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXII. pp. 719–735 (2022)
[5] *Benitez-Garcia, G., Shimoda, W., Yanai, K.*, Style image retrieval for improving material translation using neural style transfer. In: Proceedings of the 2020 Joint Workshop on Multimedia Artworks Analysis and Attractiveness Computing in Multimedia. pp. 1–6 (2020).
[6] *Matsuo, S., Yanai, K.*, Cnn-based style vector for style image retrieval. In Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval pp.309–312 (2016)
[7] *G., R.C., W., R.E.*, Digital Image Processing. Pearson/Prentice Hall, 4th edn. (2017)
[8] *Tian, Y., Clanuwat, T., Suzuki, C., Kitamoto, A.*, Ukiyo-e analysis and creativity with attribute and geometry annotation. In: Proceedings of the International Conference on Computational Creativity (2021)
[9] *Deng, J., Guo, J., Xue, N., Zafeiriou, S.*, Arcface: Additive angular margin loss for deep face recognition. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4685–4694 (2019)
[10] *Caron, M., Bojanowski, P., Joulin, A., Douze, M.*, Deep clustering for unsupervised learning of visual features. In: European Conference on Computer Vision (2018)
[11] *Ye, M., Zhang, X., Yuen, P.C., Chang, S.F.*, Unsupervised embedding learning via invariant and spreading instance feature. In: IEEE International Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
[12] *Cao, X., Chen, B.C., Lim, S.N.*, Unsupervised deep metric learning via auxiliary rotation loss. arXiv preprint arXiv:1911.07072 (2019)
[13] *Von Luxburg, U.*, A tutorial on spectral clustering. Statistics and computing 17, 395–416 (2007)
[14] *McInnes, L., Healy, J., Melville, J.* , UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction,”arXiv:1802.03426, 2018.