Retrieve in Style: Unsupervised Facial Feature Transfer and Retrieval

Min Jin Chong[1]
mchong6@illinois.edu

Wen-Sheng Chu[2]
wschu@google.com

Abhishek Kumar[2]
abhishk@google.com

David Forsyth[1]
daf@illinois.edu

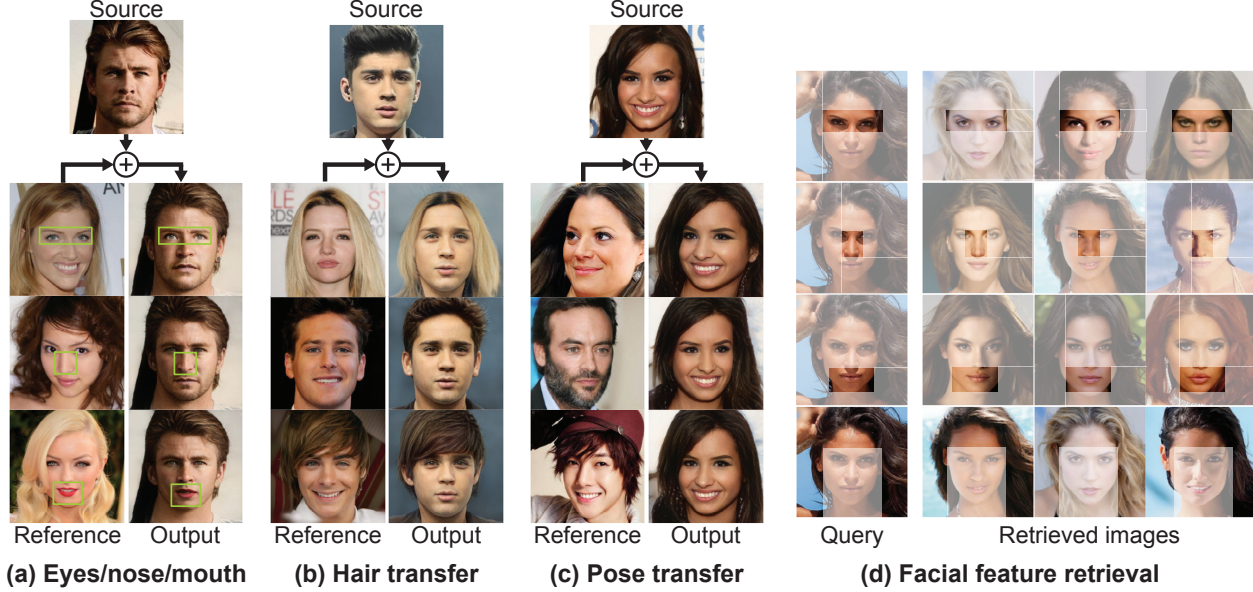[1]University of Illinois at Urbana-Champaign   [2]Google Research

Figure 1: We propose an *unsupervised* method to transfer local facial appearance from real reference images to a real source image, *e.g.*, **(a)** eyes, nose, and mouth. Compared to the state-of-the-art [10], our method enables photo-realistic transfers for **(b)** hair and **(c)** pose, and can be naturally extended for **(d)** semantic retrieval according to different facial features.

Abstract

We present Retrieve in Style (RIS), an unsupervised framework for facial feature transfer and retrieval on real images. Recent work shows capabilities of transferring local facial features by capitalizing on the disentanglement property of the StyleGAN latent space. RIS improves existing art on the following: 1) Introducing more effective feature disentanglement to allow for challenging transfers (i.e., hair, pose) that were not shown possible in SoTA methods. 2) Eliminating the need for per-image hyperparameter tuning, and for computing a catalog over a large batch of images. 3) Enabling fine-grained face retrieval using disentangled facial features (e.g., eyes). To our best knowledge, this is the first work to retrieve face images at this fine level. 4) Demonstrating robust, natural editing on real images. Our qualitative and quantitative analyses show RIS achieves both high-fidelity feature transfers and accurate fine-grained retrievals on real images. We also discuss the responsible applications of

RIS. Our code is available at https://github.com/mchong6/RetrieveInStyle.

1. Introduction

Recent advancements in Generative Adversarial Networks (GANs) [6, 18, 19] have shown capabilities to generate realistic high resolution images, particularly for faces. Under unconditional settings, it is often hard to interpret or control the outputs of GANs. Conditional GANs are more naturally amenable for semantic editing. However, the degree of meaningful control over the output images is largely dependent on how detailed the annotations are. This presents a challenge for fine-grained face editing as it is often difficult or impossible to annotate datasets with the degree of detail needed for fine-grained editing.

Existing works on face editing typically leverage additional information to guide conditional generation, such as manual labels [3, 8, 21, 42, 44, 45], segmentation masks [12, 22], attribute classifiers [14], rendering models

1

[20, 38], *etc*. However, the additional information requires extra computation and is not always available in practice. In addition, the fine-grained facial features (*e.g.*, a distinctive shape of eyes) are difficult to describe as labels or features. As an alternative, unsupervised discovery of latent directions in a pretrained GAN [13, 31, 39] allows for finding meaningful latent representations in a computationally efficient way. However, such approaches are less effective for fine-grained editing compared to supervised approaches.

Recently, Editing in Style (EIS) [10] proposed a mostly unsupervised method for facial feature transfer. While EIS allows semantic editing of spatially coherent facial features (*e.g.*, eyes, nose and mouth), it requires computing a semantic catalog over the whole dataset and separate hyperparameter tuning for each image. Such requirements make EIS non-scalable to large datasets as commonly encountered in retrieval domains. In addition, it remains challenging for EIS to control facial features that are difficult to describe as a spatial map, such as hair and head pose. More importantly, EIS works only on synthetic images and remains untested on how real images could be manipulated.

In this study, we propose Retrieve in Style (RIS), a simple and efficient unsupervised framework that tackles both fine-grained facial feature transfer and retrieval. Fig. 1 illustrates the capabilities offered by RIS. RIS improves EIS in several aspects. First, we discover the "submembership" property in the style space, showing that style channels corresponding to a particular feature (*e.g.*, eyes) are different for every image and thus must be computed individually instead of over the entire dataset. As the discovered channels are image-specific, RIS achieves more precise face editing for not only spatially coherent facial features (*e.g.*, eyes, nose, mouth) but also challenging ones (*i.e.*, hair, pose). Second, with the discovered "submembership", we show it is possible to eliminate EIS's requirements on per-image semantic catalog and per-image hyperparameter tuning, and offer better scalability to larger problems. Third, the image-specific representations naturally extend RIS for fine-grained facial feature retrieval that was not shown possible in EIS. Lastly, we demonstrate that RIS offers editing and retrieval of *real images* when combined with GAN inversion methods, while EIS worked with synthetic images. Although RIS is general and can be applied to a wide range of datasets, this study focuses on faces as there are established conventions on facial parts and its relevance in face retrieval applications (*e.g.*, [4, 11, 23, 25]).

Our contributions are:

1. RIS improves over EIS based on our finding of "submembership", obtaining better controllability over facial features that are spatially coherent (eyes, nose, mouth) and incoherent (hair pose), while requiring no hyperparameter tuning.
2. We obtain feature-specific representations (*e.g.*, eyes,

nose, mouth, hair), which enable face retrieval by fine-grained features that are difficult to describe or annotate even for humans. To our best knowledge, this is the first work to address the fine-grained retrieval problem without supervision.

3. We show that RIS generalizes to GAN-inverted images, allowing transfer and retrieval on real images that was not shown possible in earlier studies. Results on CelebA-HQ validates that RIS achieves high-quality retrieval on large, real-world datasets.

2. Related Work

StyleGAN: StyleGAN1 [19] and StyleGAN2 [19] achieve state-of-the-art unconditional image generation. StyleGAN's unique architecture is inspired by style transfer work by Huang *et al*. [15]. Contrary to previous GAN architectures that map a random noise vector $\mathbf{z}$ to an image, StyleGAN maps $\mathbf{z}$ to $\mathbf{w} \in \mathcal{W}$ via a non-linear mapping network. Feature maps in the generator are then controlled by $\mathbf{w}$ in the AdaIN module [15].

The $\mathcal{W}+$ latent space of StyleGAN has been shown to exhibit disentangled feature representations [1, 2, 19, 31]. Xu *et al*. [43] further showed that style coefficients $\boldsymbol{\sigma}$, where $\boldsymbol{\sigma} = FC(\mathbf{w})$ with FC being an affine layer, demonstrate more disentangled visual features compared to $\mathbf{w}$. The style coefficients $\boldsymbol{\sigma}$ are directly used to scale the layer-wise activations in the generator.

Latent space image editing: Radford *et al*. [28] show that the latent space of GANs is semantically meaningful – latent directions can be associated with semantics (*e.g.*, pose, smile), with directions obtained by either supervised (*e.g.* a pretrained attribute classifier, InterFaceGAN [31]) or unsupervised means (*e.g.* zooms and shifts, Jahanian [16]). Voynov [39] finds directions corresponding to changes that can be observed by a classifer. GANSpace [13] uses PCA to identify meaningful latent directions. Shen and Zhou [32] propose a closed-form factorization to obtain directions.

Feature activation image editing: Local edits can follow from manipulating GAN feature activations. GAN Dissection [5] uses a segmentation model to correspond internal GAN activations to semantic concepts, allowing them to add or remove objects. Feature Blending [34] recursively blends feature activations between source and images to allow local semantics transfer. These methods require a pretrained segmentation model or user-provided masks.

One might obtain edits as image-to-image translations. AttGAN [14] allows multi-attribute facial editing via a conditional GAN setup. StarGAN [8] proposes a single generator, multi-domain approach that uses conditional generation to achieve facial editing. GANimation [27] conditions the generator with Action Units annotations to allow smooth facial expression editing. MaskGAN [22] uses segmentation masks to enable interactive spatial image editing.

Face retrieval: Current facial retrieval systems generally match faces based on identities and lack the granularity to match on a facial feature level. Non deep-learning based retrieval systems such as Photobook [26] and CAFIIRIS [40] use features such as Eigenfaces [37], textual descriptions, and/or facial landmarks; but we expect learned features to have advantages. FaceNet [30] learns embeddings via a triplet loss where the Euclidean distances between embeddings correspond to facial similarity by training with identities. Other works [33, 35] formulate the problem as a classification task between identities. But these methods perform retrieval at the level of identity and by design, are invariant to details such as expressions and hairstyles. In contrast, RIS aims to improve the granularity of face retrieval. Instead of asking to "retrieve faces with similar features" we are asking to "retrieve faces with similar eyes, nose, mouth, *etc*. ".

GAN Inversion: GAN inversion encodes a real image to the latent space of a GAN. It is commonly done via gradient descent in the latent space [2, 19, 41] which leads to accurate reconstruction at the expense of scalability. An encoder-based approach [29, 43, 46] instead allows scalable GAN inversion.

3. Retrieve in Style

In this section, we describe the proposed Retrieve in Style (RIS) for both facial feature transfer and retrieval. We first review Editing in Style (EIS) [10] that our method is built upon. Then, we propose improvements to EIS for a more controllable and intuitive transfer, and show that our method can be naturally extended for fine-grained face retrieval, which was not possible in EIS.

3.1. Editing in Style

Unlike methods that manipulate the latent space via vector arithmetic [13, 16, 31, 32, 39], EIS formulates the semantic editing problem as copying style coefficients $\boldsymbol{\sigma}$ of StyleGAN [18] from a reference image to a source image, *i.e.*, the output image carries facial features from the reference images while preserving the remaining features from the source image. The authors show that semantic local transfer is possible on images generated by a pretrained StyleGAN with minimal supervision.

One key insight of EIS is that spatial feature activations of a StyleGAN generator can be grouped into clusters that correspond to semantically meaningful concepts such as eyes, nose, mouth, *etc*. Specifically, let $\mathbf{A} \in \mathbb{R}^{N \times C \times H \times W}$ be the activation tensor at a particular layer of StyleGAN, where N is the number of images, C the number of channels, H the height and W the width. Spherical K-way k-means [7] is applied spatially over $\mathbf{A}$, *i.e.*, clustering over $N \times H \times W$ vectors of size C. Each spatial location of $\mathbf{A}$ is associated with cluster memberships $\mathbf{U} \in \{0, 1\}^{N \times K \times H \times W}$, and then used to compute a contribution

Figure 2: **Submembership:** Contribution scores $\mathbf{M}_k$ from our method allow meaningful clustering. In this figure, each row is a cluster for $k = $ hair; images within a row are similar, showing that clustering is effective. Across rows, the images differ, showing that there is real variation in the hair.

score $\mathbf{M}_{k,c} \in [0, 1]^{K \times C}$:

$$\mathbf{M}_{k,c} = \frac{1}{NHW} \sum_{n,h,w} \mathbf{A}_{n,c,h,w}^2 \odot \mathbf{U}_{n,k,h,w}. \qquad (1)$$

Intuitively, $\mathbf{M}_{k,c}$ tells how much the c-th channel of style coefficients $\boldsymbol{\sigma} \in \mathbb{R}^C$ contributes to the generation for facial feature k. Note that $\boldsymbol{\sigma}$ directly scales the activations $\mathbf{A}$ in the modulation module — the larger the activations, the more k is affected by the channel c.

Transferring a facial feature k across two images is then performed via interpolation between style coefficients $\boldsymbol{\sigma}^S, \boldsymbol{\sigma}^R$ of the source and the reference images. The style coefficient of the edited image $\boldsymbol{\sigma}_k^G$ can be obtained by rewriting the style interpolation in Eq. (3) of [10]:

$$\boldsymbol{\sigma}_k^G = (1 - \mathbf{q}_k) \odot \boldsymbol{\sigma}^S + \mathbf{q}_k \odot \boldsymbol{\sigma}^R, \qquad (2)$$

where $\mathbf{q}_k \in [0, 1]^C$ is the interpolation vector for a given facial feature k. EIS finds $\mathbf{q}_k$ using a greedy optimization derived from $\mathbf{M}_{k,c}$ and manual hyperparameter tuning to determine which channels to ignore. Such hyperparameters can be sensitive to different reference images and lead to suboptimal transfers, as shown in Sec. 4. In addition, $\mathbf{M}_{k,c}$ is computed over N images and is fixed for all feature transfers. We argue in Sec. 3.2 that having a fixed $\mathbf{M}_{k,c}$ may not be ideal for transfer, as not all images share the same channels to describe the same facial feature.

3.2. Improving EIS for Facial Feature Transfer

Submemberships: EIS assumes that the channels that make a high contribution for a particular feature (say, eyes) are the same for each image. So to compute $\mathbf{M}_k$ in Eq. (1), EIS averages the scores over a large collection of images of size N. We hypothesize the high-contribution channels may
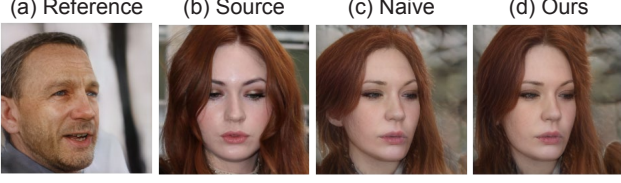
| (a) Reference | (b) Source | (c) Naive | (d) Ours |

Figure 3: **Pose transfer** from **(a)** reference to **(b)** source. **(c)** Naively copying style coefficients from the first 4 layers of StyleGAN2 [19] transfers primarily pose and partially hair (shorter hair on left, flatter hair top), showing their style co-efficients are entangled in the early layers. **(d)** Our method matches the pose of the reference image and preserves the hair faithfully from the source.

vary from image to image. This means averaging over N images can lose details specific to the source or reference.

We visualize the presence of this effect in Fig. 2. Performing Spherical k-means clustering over per image $\mathbf{M}_{\text{hair}}$ ($N = 1$) of images in a dataset yields semantically meaningful clusters. Images in each row belong to the same cluster. The hairstyles within the same row are similar, while hairstyles across rows are distinctively different. We further analyze the top active channels (each channel corresponds to a dimension of $\mathbf{M}_k$) for each cluster, and observe that each cluster has its own set of top active channels that are unique to it. Please refer to supplementary materials for more detailed analyses. This validates our hypothesis that high-contribution channels for a semantic feature are not the same across images. That is, the same feature k of different images are controlled by different groups of channels. We term these groups as "submembership", which is a crucial motivation for this work.

With "submembership" in mind, instead of computing $\mathbf{M}_{k,c}$ over a large batch of N images, we show that the responsible channels are more accurately computed over only the source and reference images, *i.e.*, $N = 2$. Specifically,

$$\mathbf{M}_{k,c} = \max\left(\sum_{h,w} \mathbf{A}[s]^2_{c,h,w} \odot \mathbf{U}[s]_{k,h,w}, \right.$$
$$\left. \sum_{h,w} \mathbf{A}[r]^2_{c,h,w} \odot \mathbf{U}[r]_{k,h,w}\right), \quad (3)$$

where s and r indicate the particular source and reference images of interest, respectively. Intuitively, to transfer from a reference to a source image, we are interested in channels that are important to source, reference, or both.

Obtaining interpolation vector: Instead of getting the interpolation vector $\mathbf{q}_k$ from the greedy optimization process (like in EIS) which is dependent on per-image hyper-parameters ρ and ϵ, we assume each channel of the style coefficient σ corresponds to one facial feature. This follows from the disentangled style space of StyleGAN and in practice, works well. Under this assumption, we obtain a

soft class assignment for each style coefficient channel with a softmax of all classes (rows of $\mathbf{M}$), obtaining:

$$\mathbf{q} = \operatorname*{Softmax}_{k}\left(\frac{\mathbf{M}}{\tau}\right), \quad (4)$$

where $\mathbf{M} \in [0, 1]^{K \times C}$ is the stacked contribution score of all facial features, τ is the temperature, $\mathbf{q} \in [0, 1]^{K \times C}$ is the interpolation vector. The interpolation vector for a particular feature k, $\mathbf{q}_k$ can be indexed from the row of $\mathbf{q}$. $\mathbf{q}_k$ can be thought of the mask for k that allows interpolation between σ^S and σ^R.

Pose transfer: Karras *et al*. [19] have shown that the first few layers of StyleGAN2 capture high level features such as pose. In Fig. 3, we show that copying the style coefficients of the first 4 layers of StyleGAN2 (which corresponds to the first 2048 style coefficient channels), transfers mostly pose and hair information from reference to source image, leaving other features like eyes and mouth untouched. By assuming that the first 4 layers *only* contain pose and hair information, we simply derive:

$$\mathbf{q}_{\text{pose}} = \mathbf{1} - \mathbf{q}_{\text{hair}}, \quad (5)$$

for only the first 4 layers with the rest zeroed out. Similarly, for all facial features other than hair, the first 4 layers are zeroed out to prevent pose changes. As shown in Fig. 3, $\mathbf{q}_{\text{pose}}$ captures pose information without affecting hair.

One significant advantage of our pose transfer is that it requires no labels or manual tuning. For example, GANSpace [13] requires manually choosing layer subsets; AttGAN [14] and InterFaceGAN [31] requires attribute labels, StyleRig [36] requires a 3D face model. Fig. 4 illustrates our full capability of facial feature transfer.

Latent Direction: Unlike EIS that limits facial feature transfer to style interpolation as in Eq. (3), we formulate the problem as traversing along the latent direction, based on work showing StyleGAN's latent space vector arithmetic property [28]. Then, we revise Eq. (3) to:

$$\sigma^G_k = \sigma^S + \alpha \mathbf{q}_k \odot (\sigma^R - \sigma^S), \quad (6)$$

where the latent direction is $\mathbf{n} = \mathbf{q}_k \odot (\sigma^R - \sigma^S)$ and the scalar step size is α. If we restrict $\alpha \in [0, 1]$, we will be performing a style interpolation. Under the property of vector arithmetic, we can instead use $\alpha \in \mathcal{R}$ which allows style extrapolation. We show in Fig. 5 that scaling α allows an increase or decrease in the particular facial property. For example, we are able to do smooth pose interpolation.

3.3. Facial Feature Retrieval

This section shows the style representation in Eq. (6) can be adapted to fine-grained facial feature retrieval, which is defined as follows. Given a query image I_Q and a retrieval dataset $\mathcal{X}$, we aim to retrieve the top-K closest images $\mathcal{T}_k \subset \mathcal{X}$ with respect to a facial feature (*e.g.*, eyes). As
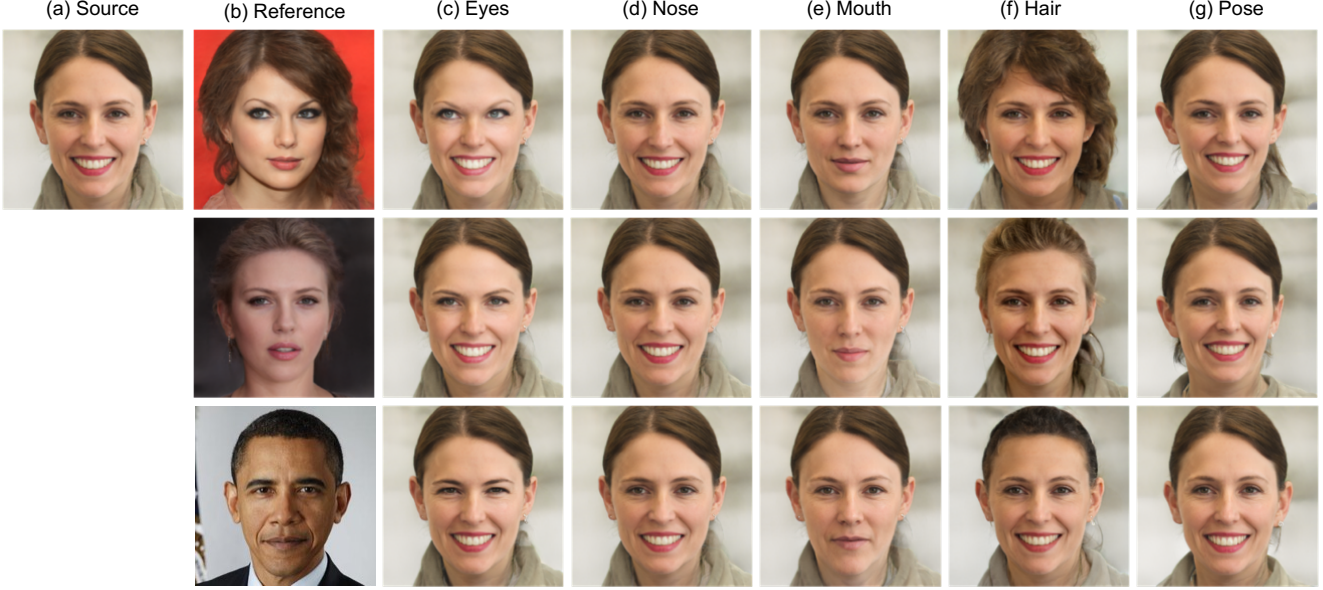
Figure 4: **Facial feature transfer:** Our method performs effective semantic editing on real images by transferring facial features from **(b)** a reference image to **(a)** a source image. Our method transfers spatially coherent features (*i.e.*, eyes, nose, mouth) as well as challenging features hair and pose. Note that real image editing is not possible with SoTA EIS [10].
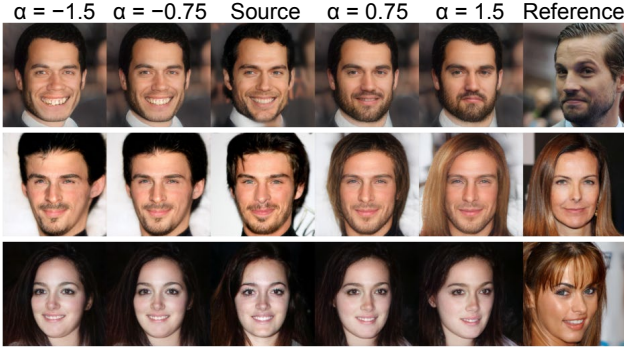

Figure 5: **Latent direction:** The α variable in RIS controls interpolation between the source and the reference images, showing a smooth transition of mouth (top row), hair (middle row) and pose (bottom row).

described in the previous section, RIS identifies the style channels that mediate the appearance of facial features for particular images. This suggests the style channels can be used to retrieve faces with appearance similar to the facial features in a query face. Face retrieval is usually done by matching on an identity embedding [30, 33, 35]. However, fine-grained facial feature retrieval is relatively unexplored as it is difficult to collect and annotate training data with fine granularity (*e.g.*, shape of the eyes or nose).

For each facial feature k, we have $\mathbf{q}_k \in [0,1]^{1 \times C}$ to encode, for a particular image, how much that channel contributes to that feature. Since $\mathbf{q}_k$ can be considered as a

mask, we construct a feature-specific representation:

$$\mathbf{v}_k^Q = \mathbf{q}_k^Q \odot \boldsymbol{\sigma}^Q. \tag{7}$$

Feature retrieval can be then performed by matching $\mathbf{v}_k$, as two images with similar $\mathbf{v}_k$ suggest a lookalike feature k.

We compute the representations $\mathbf{v}_k^R = \mathbf{q}_k^R \odot \boldsymbol{\sigma}^R$ where $\boldsymbol{\sigma}^R \in \boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}$ are the style coefficients for the images in $\mathcal{X}$. We then define the distance between the facial features of two style coefficients/face images as

$$\text{Distance}_k(I^Q, I^R) = d(\mathbf{v}_k^Q, \mathbf{v}_k^R), \tag{8}$$

where d is a distance metric (cosine distance in this study). We then rank the distances for nearest neighbor search for facial feature k. Intuitively, if there is a $\mathbf{M}_k$ and consequently, a $\mathbf{q}_k$ mismatch between two images, their distance will be large. Since Fig. 2 shows that similar features have similar $\mathbf{M}_k$, vice versa, it follows that smaller distances will reflect more similar features. We show this is true empirically and RIS works as in expectation from Fig. 7. Additionally, we observe better results if we normalize $\boldsymbol{\sigma}^Q$ and $\boldsymbol{\sigma}^R$ using layer-wise mean and standard deviation from $\boldsymbol{\Sigma}$.

Comparison between SoTA EIS [10] and RIS (ours): Both EIS and RIS share a unique way to perform unsupervised local face editing by attributing transfers to reference images. They differ in how they accomplish it. (1) EIS computes the contribution score $\mathbf{M}$ by averaging over a batch of N images. Based on the findings of $\mathbf{M}$'s submembership, RIS uses $N = 2$, which avoids manual per-image hyperparameter tuning and thus allows a more scalable and intuitive
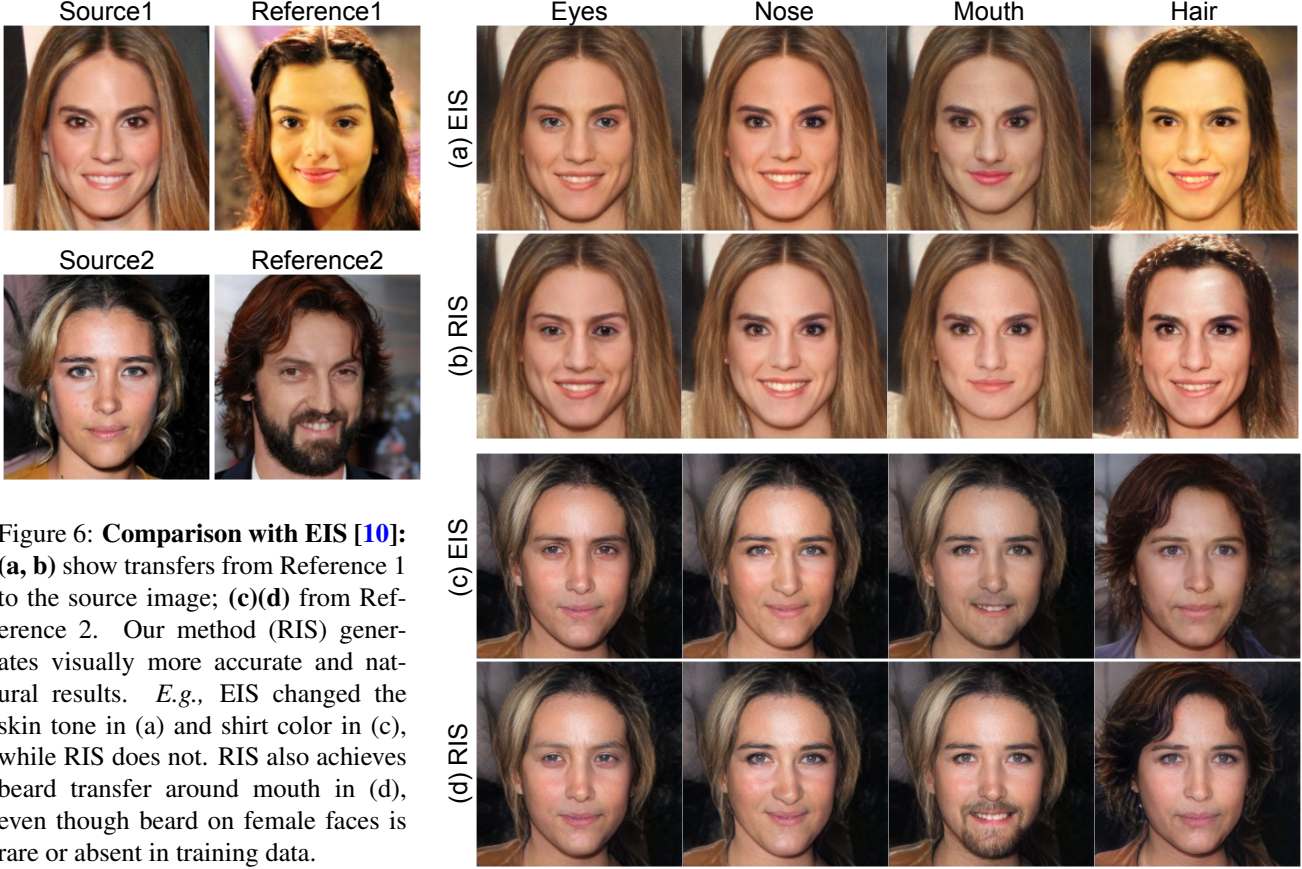
Figure 6: **Comparison with EIS [10]:** **(a, b)** show transfers from Reference 1 to the source image; **(c)(d)** from Reference 2. Our method (RIS) generates visually more accurate and natural results. *E.g.,* EIS changed the skin tone in (a) and shirt color in (c), while RIS does not. RIS also achieves beard transfer around mouth in (d), even though beard on female faces is rare or absent in training data.

transfer. As a result, RIS yields more precise transfer of eyes, nose, and mouth, and enables transferring novel features such as hair and pose that were not shown possible in EIS. (2) RIS redefines $\mathbf{M}$ as an image-specific representation, which allows for unsupervised fine-grained face feature retrieval. EIS assumes an averaged representation of $\mathbf{M}$, which will be shown in experiments to be less effective for feature retrieval.

4. Experiments

While other work based on StyleGAN, including EIS [10, 13], focus on manipulating generated images, we focus on the more relevant problem of manipulating *real images*. This is a more difficult problem as there are no guarantees that GANs performing well on generated images are stable enough to generalize to real images.

To show that RIS generalizes to real datasets, we use CelebA-HQ [17] with 30k images for all our experiments. Since feature-based retrieval requires the inversion of the entire dataset, we opt to use pSp [29], a SoTA encoder-based GAN inversion method, for all our experiments.

4.1. Facial Feature Transfer

In this section, we provide qualitative and quantitative analyses for facial feature transfer on real images. We fixed $\tau = 0.1$ and $\alpha = 1.3$ for all experiments, as we observed the temperature τ in Eq. (4) is insensitive to different source and reference images. We used $N = 200$ for EIS [10] following the authors' implementation.

Qualitative analysis: Fig. 6 shows a qualitative comparison between RIS (our method) and EIS on real images. It can be observed that RIS offers better localization ability. EIS (Fig. 6(a)) affects skin tone heavily across all transfers, notably changing lighting heavily for hair transfer. In contrast, RIS maintains relatively similar skin tones while transferring the targeted features. EIS also changes the eyes and nose of the source image while transferring mouth (Fig. 6(a)), indicating entanglement in their representations. While transferring mouth (which includes the chin region), EIS fails to reproduce the beard in the image Reference2 (Fig. 6(c)). On the other hand, RIS faithfully reproduces the beard (Fig. 6(d)). It is noteworthy that RIS is able to generate a female face with beard, representing an out-of-distribution generation that is absent in the training set. Please refer to supplementary materials for more comparisons.
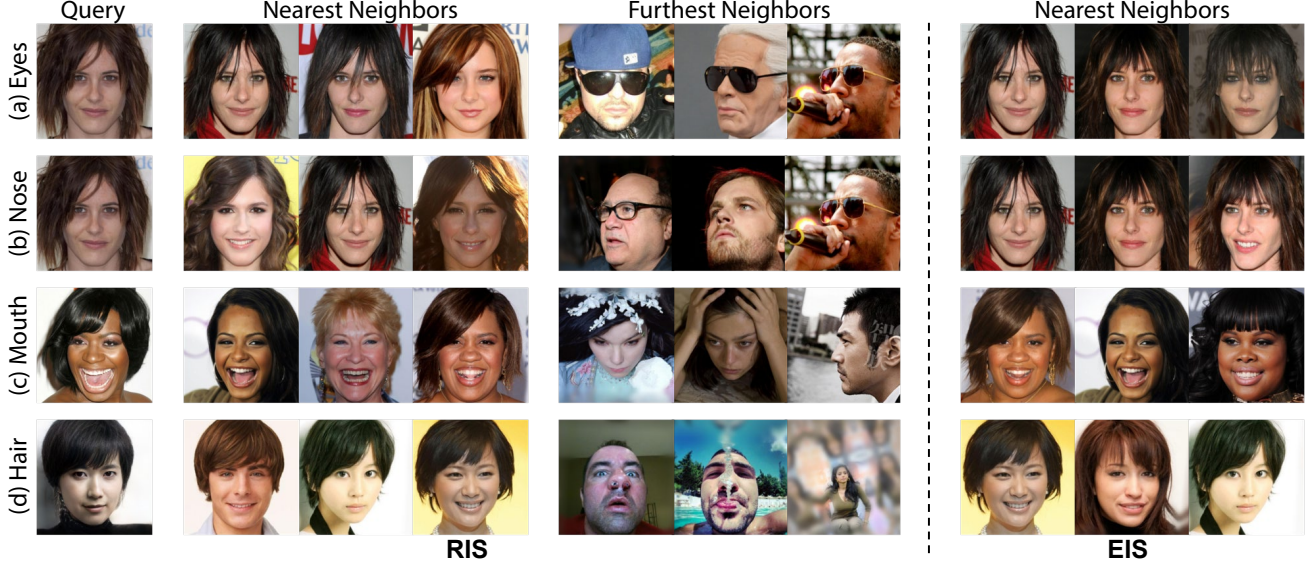
Figure 7: **Facial feature retrieval:** We compare fine-grained retrieval between our method RIS (submembership $\mathbf{M}_k$) and EIS [10] (universal $\mathbf{M}_k$) on real faces. We show 3 faces each from nearest and furthest retrieval (NR and FR). RIS retrieves semantically similar NRs on all facial features while showing variance on non-matching features. Note EIS retrieves very similar NR on eyes and nose with same query image indicating a lack of feature localization.

Method	FID_∞
StyleGAN2 [19]	2.44
EIS [10]	3.47
RIS (ours)	3.73

Table 1: **Image fidelity comparison:** RIS achieves a comparable FID_∞ compared to EIS and is only slightly worse compared to the base StyleGAN2. The larger FID_∞ can be attributed to our capability of OOD generation, *e.g.*, long-hair males or bald females as in the right image.

Quantitative analysis: To quantitatively validate our transfer results, we computed FID_∞ [9], an unbiased estimate of FID, for baseline StyleGAN2 [19], EIS [10] and RIS. Details on the setup are provided in the supplementary.

Table 1 shows the FID_∞ comparison. Both EIS and RIS achieved small FID_∞ differences compared to the base StyleGAN2. However, RIS yielded a slightly larger FID_∞, which can be explained by the ability of our method to even generate out-of-distribution samples, if needed for transferring features. Such samples are uncommon in the FFHQ dataset that trains the base StyleGAN2, and thus contribute to a larger FID_∞, *e.g.*, our method is capable of transferring long hair to a bearded male, or bald hair to a female, as shown in the right of Table 1.

4.2. Facial Feature Retrieval

We evaluate our retrieval performance qualitatively and quantitatively. We use GAN inverted CelebA-HQ images as the retrieval dataset, and cosine distance as the metric.

Qualitative analysis: As fine-grained facial retrieval is relatively unexplored, to the best of our knowledge, there are no proper metrics to evaluate this task. Instead, we re-purposed the averaged $\mathbf{M}_k$ in EIS for retrieval and use it as a baseline. Specifically, when computing retrieval representations in Eq. (7), we replaced the individual $\mathbf{q}_k$ with the $\mathbf{q}_k$ derived from EIS's averaged $\mathbf{M}_k$. Since large-scale hyperparameter tuning for every reference image is infeasible for EIS, we obtained $\mathbf{q}_k$ with a fixed hyperparameter choice that may not generalize to all images.

Fig. 7 shows qualitative comparisons between RIS and EIS. RIS has observably more disentangled representations. Specifically, for eyes retrieval, although the query has distinct eyes, RIS retrieves images with the same eyes but different identities, while EIS only retrieves the same identity. This suggests EIS representations are entangled between eyes and identity features. In addition, EIS retrieves almost the same images for different features (*i.e.*, eyes and nose), suggesting entanglement. For mouth retrieval, RIS recognizes the wide open mouth of the query, retrieving semantically similar (w.r.t. the mouth feature) yet diverse images. EIS, on the other hand, retrieves images with the same skin tone, suggesting a lack of feature localization. Lastly, for hair retrieval, RIS retrieves images with similar hair but with different genders, while EIS only retrieves only female images. Finally, furthest neighbors for RIS differ semantically from the query image.

Overall, RIS nearest neighbors exhibit significant variance on non-matching features while EIS nearest neighbors do not. Along with our superior results in Fig. 6, this further

	Attribute Matching Score (%)	
Class	Ours	EIS
Eyes	96.3	95.4
Nose	100.0	100.0
Mouth	81.1	75.8
Hair	97.5	97.1

Table 2: We compare AMS between RIS and EIS to measure retrieval accuracy *w.r.t.* a given facial feature using a pretrained attribute classifier. RIS outperforms EIS in all classes, with *mouth* retrieval being noticeably better.

reinforces that our individual $\mathbf{M}_k$ yields better disentanglement and feature focus compared to the averaged $\mathbf{M}_k$ in EIS. This also validates our hypothesis of submemberships.

TRSI-IoU. We use retrieval to evaluate how well RIS disentangles facial features. We focus on two retrieved set identity IoU (TRSI-IoU): retrieve two sets of images using two facial feature queries on the same face; TRSI-IoU is computed as intersection-over-union of the identities between these two sets. A full face retrieval method should have a TRSI-IoU close to 1 if the two queries are the same person, and 0 otherwise. Assume a method does not disentangle features, it is possible to approximately predict (say) mouth from eyes. In turn, retrieving using eyes (resp. mouth) will implicitly constrain mouth (resp. eyes), so the two retrieved sets will have many individuals in common; hence TRSI-IoU becomes relatively large. On the other hand, if a method properly disentangled (say) eyes and mouth, its identities should not overlap much; thus TRSI-IoU becomes relatively small. The minimum obtainable value of TRSI-IoU is difficult to know, but lower TRSI-IoU is good evidence a method disentangles better. Fig 8 shows boxplots of TRSI-IoU for RIS and EIS, evaluated for 100 queries and all pairs of facial features (chosen from eyes, nose, mouth, hair). RIS shows significantly lower TRSI-IoU, and the difference is statistically significant.

Attribute Matching Score. We used attribute classifiers pretrained on CelebA attributes [24] to further evaluate the quality of our retrieval. Note that these attributes are binary and not sufficiently detailed for fine-grained purposes. There is also a distinct lack of diversity in CelebA and its attributes, *e.g.*, lack of head coverings, curly hairs, *etc.*, which makes evaluation of RIS on generating faces of diverse and inclusive people not possible.

The intuition of our procedure is as follows: for retrieval of the k-th feature *hair*, hair-related attributes $\mathcal{A}_k$ (*e.g.*, "black_hair", "wavy_hair", *etc.*) should remain similar between query and retrieved images. Please see supplementary for the full list of attributes associated with k.

We retrieve top-5 images $\mathcal{T}_5^{(i)}$ according to a query image $I_Q^{(i)}$ for a feature k. We took an attribute classifier
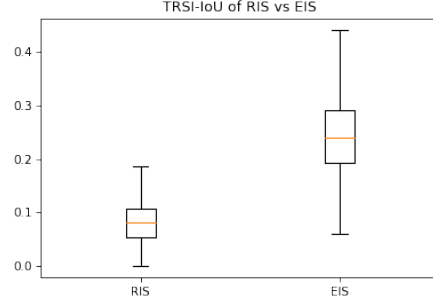

Figure 8: TRSI-IoU measures the extent of overlapping identities between two different feature queries on the same face. Methods that disentangle facial features better are expected with smaller TRSI-IoU (see text). We compare a boxplot of TRSI-IoU for RIS and EIS. RIS shows noticeable improvement in the median (red line) with much smaller interquartile range (boxes). This suggests our method better disentangles facial features.

$\mathcal{F}$, and got its prediction for the a-th attribute as $\widehat{\mathcal{F}}_a(\cdot) = [\mathcal{F}_a(\cdot) > T]$, *i.e.*, $\widehat{\mathcal{F}}_a(\cdot) = 1$ if the prediction is larger than threshold $T = 0.5$ and 0 otherwise. Then, Attribute Matching Score (AMS) is defined for the k-th facial feature:

$$\text{AMS}_k = \frac{\sum_{I_Q^{(i)} \in \mathcal{X}} \sum_{t^{(i)} \in \mathcal{T}_5^{(i)}} \sum_{a \in \mathcal{A}_k} \left[\widehat{\mathcal{F}}_a(I_Q^{(i)}) = \widehat{\mathcal{F}}_a(t^{(i)}) \right]}{|\mathcal{X}| \cdot |\mathcal{T}_5^{(i)}| \cdot |\mathcal{A}_k|}.$$

Table 2(b) compares AMS scores between EIS and RIS. As the classifier is trained on predefined attributes that do not contain fine granularity, it could be less descriptive to our particular task of fine-grained retrieval. Still, RIS outperforms EIS in all classes under this less-granular setting, with *mouth* retrieval being noticeably better.

5. Conclusion

We presented Retrieve in Style (RIS), a simple and efficient unsupervised method of facial feature transfer that works across both short-scale features (eyes, nose, mouth) and long-scale features (hair, pose) on real images without any hyperparameter tuning. RIS produces realistic, accurate feature transfers without modifying the rest of the image, and naturally extends to the fine-grained facial feature retrieval. Note that techniques for photorealistically manipulating images could be misused to produce fake or misleading information, and researchers should be aware of these risks. To the best of our knowledge, this is the first work that enables unsupervised, fine-grained facial retrieval, especially so on real images. Our qualitative and quantitative analyses verify the effectiveness of RIS.

References

[1] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan: How to embed images into the stylegan latent space? In *Proceedings of the IEEE international conference on computer vision*, pages 4432–4441, 2019. 2

[2] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan++: How to edit the embedded images? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8296–8305, 2020. 2, 3

[3] Jianmin Bao, Dong Chen, Fang Wen, Houqiang Li, and Gang Hua. Towards open-set identity preserving face synthesis. In *CVPR*, 2018. 1

[4] James C Bartlett, Susan Hurry, and Warren Thorley. Typicality and familiarity of faces. *Memory & Cognition*, 12(3):219–228, 1984. 2

[5] David Bau, Jun-Yan Zhu, Hendrik Strobelt, Bolei Zhou, Joshua B Tenenbaum, William T Freeman, and Antonio Torralba. GAN dissection: Visualizing and understanding generative adversarial networks. *arXiv preprint arXiv:1811.10597*, 2018. 2

[6] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018. 1

[7] Christian Buchta, Martin Kober, Ingo Feinerer, and Kurt Hornik. Spherical k-means clustering. *Journal of Statistical Software*, 50(10):1–22, 2012. 3

[8] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *CVPR*, 2018. 1, 2

[9] Min Jin Chong and David Forsyth. Effectively unbiased fid and inception score and where to find them. In *CVPR*, 2020. 7

[10] Edo Collins, Raja Bala, Bob Price, and Sabine Susstrunk. Editing in Style: Uncovering the local semantics of GANs. In *CVPR*, 2020. 1, 2, 3, 5, 6, 7, 10, 11, 12

[11] Michael R Courtois and John H Mueller. Target and distractor typicality in facial recognition? *Journal of Applied Psychology*, 66(5):639, 1981. 2

[12] Shuyang Gu, Jianmin Bao, Hao Yang, Dong Chen, Fang Wen, and Lu Yuan. Mask-guided portrait editing with conditional gans. In *CVPR*, 2019. 1

[13] Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. GANSpace: Discovering interpretable GAN controls. *arXiv preprint arXiv:2004.02546*, 2020. 2, 3, 4, 6

[14] Zhenliang He, Wangmeng Zuo, Meina Kan, Shiguang Shan, and Xilin Chen. AttGAN: Facial attribute editing by only changing what you want. *IEEE Transactions on Image Processing*, 28(11):5464–5478, 2019. 1, 2, 4

[15] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*, 2017. 2

[16] Ali Jahanian*, Lucy Chai*, and Phillip Isola. On the "steerability" of generative adversarial networks. In *ICLR*, 2020. 2, 3

[17] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017. 6

[18] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019. 1, 3

[19] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *CVPR*, 2020. 1, 2, 3, 4, 7

[20] Marek Kowalski, Stephan J. Garbin, Virginia Estellers, Tadas Baltrušaitis, Matthew Johnson, and Jamie Shotton. CONFIG: Controllable neural face image generation. In *European Conference on Computer Vision (ECCV)*, 2020. 2

[21] Guillaume Lample, Neil Zeghidour, Nicolas Usunier, Antoine Bordes, Ludovic Denoyer, and Marc'Aurelio Ranzato. Fader networks: Manipulating images by sliding attributes. In *NeurIPS*, 2017. 1

[22] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. MaskGAN: Towards diverse and interactive facial image manipulation. In *CVPR*, 2020. 1, 2

[23] Eric Lee, Thomas Whalen, John Sakalauskas, Glen Baigent, Chandra Bisesar, Andrew McCarthy, Glenda Reid, and Cynthia Wotton. Suspect identification by facial features. *Ergonomics*, 47(7):719–747, 2004. 2

[24] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015. 8, 11

[25] Ahmed M Megreya and A Mike Burton. Matching faces to photographs: Poor performance in eyewitness memory (without the memory). *Journal of Experimental Psychology: Applied*, 14(4):364, 2008. 2

[26] Alex Pentland, Rosalind W Picard, and Stan Sclaroff. Photobook: Content-based manipulation of image databases. *International journal of computer vision*, 18(3):233–254, 1996. 3

[27] Albert Pumarola, Antonio Agudo, Aleix M Martinez, Alberto Sanfeliu, and Francesc Moreno-Noguer. Ganimation: Anatomically-aware facial animation from a single image. In *Proceedings of the European conference on computer vision (ECCV)*, pages 818–833, 2018. 2

[28] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015. 2, 4

[29] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in style: a stylegan encoder for image-to-image translation. *arXiv preprint arXiv:2008.00951*, 2020. 3, 6

[30] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, 2015. 3, 5

[31] Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the latent space of GANs for semantic face editing. In *CVPR*, 2020. 2, 3, 4

[32] Yujun Shen and Bolei Zhou. Closed-form factorization of latent semantics in GANs. *arXiv preprint arXiv:2007.06600*, 2020. 2, 3

[33] Yi Sun, Xiaogang Wang, and Xiaoou Tang. Deeply learned face representations are sparse, selective, and robust. In *CVPR*, 2015. 3, 5

[34] Ryohei Suzuki, Masanori Koyama, Takeru Miyato, Taizan Yonetsuji, and Huachun Zhu. Spatially controllable im-

age synthesis with internal representation collaging. *arXiv preprint arXiv:1811.10153*, 2018. 2

[35] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. Deepface: Closing the gap to human-level performance in face verification. In *CVPR*, 2014. 3, 5

[36] Ayush Tewari, Mohamed Elgharib, Gaurav Bharaj, Florian Bernard, Hans-Peter Seidel, Patrick Pérez, Michael Zöllhofer, and Christian Theobalt. Stylerig: Rigging stylegan for 3d control over portrait images, cvpr 2020. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, june 2020. 4

[37] Matthew Turk and Alex Pentland. Eigenfaces for recognition. *Journal of cognitive neuroscience*, 3(1):71–86, 1991. 3

[38] Ben Usman, Nick Dufour, Kate Saenko, and Chris Bregler. PuppetGAN: Cross-domain image manipulation by demonstration. In *ICCV*, 2019. 2

[39] Andrey Voynov and Artem Babenko. Unsupervised discovery of interpretable directions in the gan latent space. *arXiv preprint arXiv:2002.03754*, 2020. 2, 3

[40] Jian-Kang Wu, Yew Hock Ang, PC Lam, SK Moorthy, and A Desai Narasimhalu. Facial image retrieval, identification, and inference system. In *Proceedings of the first ACM international conference on Multimedia*, pages 47–55, 1993. 3

[41] Jonas Wulff and Antonio Torralba. Improving inversion and generation diversity in stylegan using a gaussianized latent space. *arXiv preprint arXiv:2009.06529*, 2020. 3

[42] Taihong Xiao, Jiapeng Hong, and Jinwen Ma. Elegant: Exchanging latent encodings with gan for transferring multiple face attributes. In *ECCV*, 2018. 1

[43] Yinghao Xu, Yujun Shen, Jiapeng Zhu, Ceyuan Yang, and Bolei Zhou. Generative hierarchical features from synthesizing images. *arXiv e-prints*, pages arXiv–2007, 2020. 2, 3

[44] Xi Yin, Xiang Yu, Kihyuk Sohn, Xiaoming Liu, and Manmohan Chandraker. Towards large-pose face frontalization in the wild. In *CVPR*, 2017. 1

[45] Gang Zhang, Meina Kan, Shiguang Shan, and Xilin Chen. Generative adversarial network with spatial attention for face attribute editing. In *ECCV*, 2018. 1

[46] Jiapeng Zhu, Yujun Shen, Deli Zhao, and Bolei Zhou. Indomain gan inversion for real image editing. In *European Conference on Computer Vision*, pages 592–608. Springer, 2020. 3

6. Supplementary Material for *Retrieve in Style: Unsupervised Facial Feature Transfer and Retrieval*

Overview

Even though RIS framework is built upon a pretrained StyleGAN which generates fake images, we focus on applying RIS to real images in the main paper. For completeness, we show RIS on fake images in the supplementary. We further provide more results that could not fit in the main paper due to space constraints. In particular, we offer deeper discussion on these aspects:

1. We elaborate the **submembership analysis** on the contribution scores $\mathbf{M}_k$ [10] with respect to overlapping channels across different clusters.
2. We show **latent interpolation** between the source and reference images, verifying the smooth transition for the facial feature transfer.
3. We enumerate the **attribute classifier accuracy** available in the CelebA attribute dataset and their correspondence to describe facial features, confirming that the accuracy of retrieval performance is meaningful.

7. Submemberships

A central claim to the proposed method, Retrieve in Style (RIS), is the concept of submemberships, *i.e.*, highly contributing channels that vary from image to image. In order to validate the existence of submemberships as discussed in Sec. 3.1 of the main paper, we conducted the following experiment. We generated $N = 5000$ images and computed their $\mathbf{M}_k$ for a particular feature k. Then, we performed spherical $K = \{2, 5, 10, 20, 50, 100\}$-way clustering and averaged each cluster's $\mathbf{M}_k$. Denote $\mathbf{M}_k^i$ as the average contribution score of feature k for all images belonging to cluster i. With a slight abuse of notation, we obtain:

$$\mathbf{Z}_k^i = \operatorname{argsort}_n \mathbf{M}_k^i, \tag{9}$$

where $\operatorname{argsort}_n$ is a sorting operator that returns the indices of the top n leading values of $\mathbf{M}_k^i$ ($n = 100$ in our case). That is, $\mathbf{Z}_k^i$ represents the set of top-n most contributing channel for feature k cluster i. Suppose that there exists a universal $\mathbf{M}_k$ for all images, $\mathbf{Z}_k^i$ should have a high degree of intersection since the important channels for all clusters should be the same. We thus define an *intersection ratio* as the number of channels common in $\mathbf{Z}_k^i$ divided by the n. From Fig. 9, the intersection ratio for different features progressively decreases as the number of clusters increases. This means that as the clusters get more specific, the number of overlapping channels decreases, validating our hypothesis on submemberships.
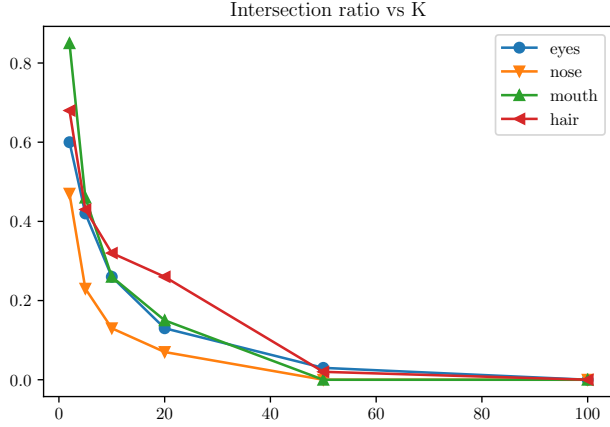
Figure 9: **Intersection Ratio**: This figure shows the intersection ratio (**y-axis**) computed against K, the number of clusters (**x-axis**). The common channels shared by all clusters decrease as the number of clusters increase. This means that for the same facial feature, images do not share the same contributing channels, validating the "submembership" effect discussed in Sec. 3.1 of the main paper.

8. Interpolation of Transfers

In this section, we show that the proposed RIS allows smooth interpolations for facial feature transfers for generated images, in addition to the results shown in Fig. 5 of the original paper. Fig. 10 shows natural and smooth transition for our interpolation on the target facial features, *i.e.*, eyes, nose, mouth, hair, and pose. Note that hair and pose transfers were not shown possible in the state-of-the-art EIS approach [10].

More results: Similar to the figures shown for facial feature transfer and retrieval as in the main paper, Figs. 11 and 12 provide more examples for facial feature transfer retrieval, respectively on generated images.

9. Attribute Classifier for AMS score

In this section, we provide details about attribute classifiers that were used to evaluate our Attribute Matching Score (AMS) in Sec. 4.2 of the original paper. In particular, we pretrained a attribute classifier based on 40 attributes on the CelebA dataset [24]. Subsets of features were manually selected to associate attributes with the facial features that the proposed method attempts to retrieve. Table 3 shows the full list of binary attributes for each facial feature. For completeness, Fig. 13 illustrates the accuracy of each of the 40 attributes of our pretrained model, with an average of 85.27% overall accuracy.

10. TRSI-IoU metric

The goal of TRSI-IoU is to measure how disentangled the facial feature representations are, and not the accuracy of retrieval (which is evaluated by Attribute Matching Score). For the task of fine-grained feature retrieval, it is pertinent to sufficiently disentangle the feature representations, *i.e.*, the retrieval results of eyes should not predict the retrieval results of nose. In an extreme case where features are fully entangled, the identities retrieved across different features become the same. This task is then trivially reduced to the conventional identity retrieval, a simpler and well-researched task compared to our goal of fine-grained feature retrieval. We observe that EIS retrieves the same images and identities for different features (as shown in Fig. 7(a) and (b) for EIS), which signify *significant entanglement* between facial features. TRSI-IoU is thus introduced to quantify this entanglement. The combination of AMS and TRSI-IoU gives a comprehensive evaluation of both accuracy and entanglement.

11. Inference speed

For both EIS and RIS, we perform 100 inference runs (includes both computing $\mathbf{M}$ and generating the edited image), and compute the mean and standard deviation of the runs on a single Titan Xp GPU. Measured in seconds, we observe for EIS: 0.0394 ± 0.00289, for RIS: 0.234 ± 0.00633. Although computing instance-level $\mathbf{M}$ adds ~ 0.2s latency, we believe RIS remains suitable for real world applications. Computing $\mathbf{M}$ for a dataset of 50K images for retrieval takes less than 10 minutes on a single Titan Xp GPU (avg 0.12s per image).

12. Effects of noise input

In all experiments, we fix the noise input to prevent variations caused by the random noise. We perform an experiment showcasing the effect of varied noise input on RIS, as shown in Fig. 14. From the absolute difference between different random runs, we observe that their delta is negligible.
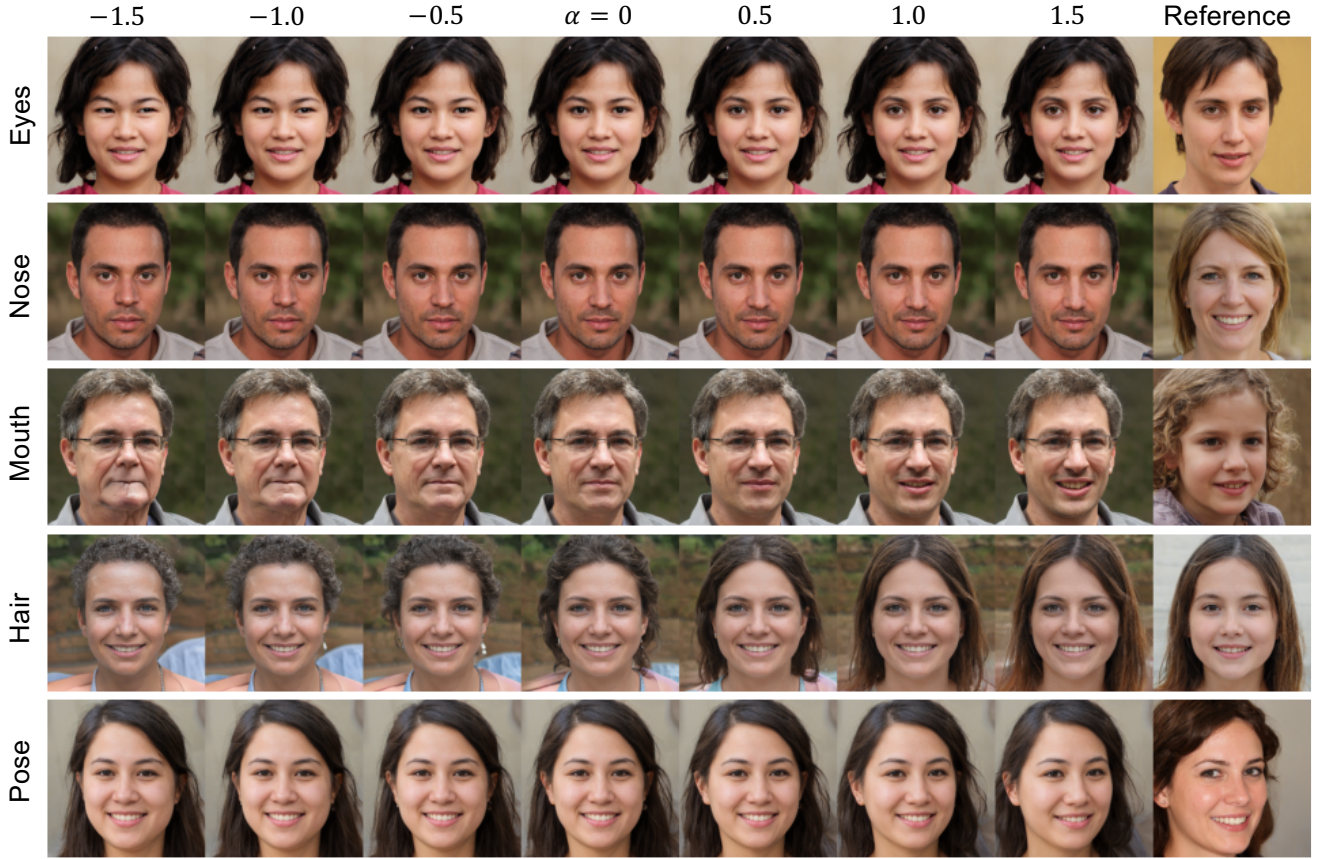
Figure 10: We scale $\mathbf{q}_k$ according to different α to allow interpolation between the source image (the left most column) and the reference image (the right most column) on a particular facial feature. With the side-by-side comparisons with different α, we observe that RIS is able to produce smooth and realistic transitions between the transfers. The larger value the α, the closer the facial features are similar to the reference images. Note that hair and pose transfers were not shown possible in the state-of-the-art EIS [10].

Facial Feature	CelebA Attributes
Eyes	Arched Eyebrows, Bags Under Eyes, Bushy Eyebrows, Narrow Eyes.
Nose	Big Nose, Pointy Nose.
Mouth	5 of Clock Shadow, Big Lips, Goatee, Mouth Slightly Open, Mustache, No Beard, Smiling, Wearing Lipstick.
Hair	Bald, Bangs, Black Hair, Blond Hair, Brown Hair, Gray Hair, Receding Hairline, Sideburns, Straight Hair, Wavy Hair.

Table 3: The relationship between facial features and CelebA attributes that we used to evaluate Attribute Matching Score (AMS) in Sec. 4.4 in the main paper.
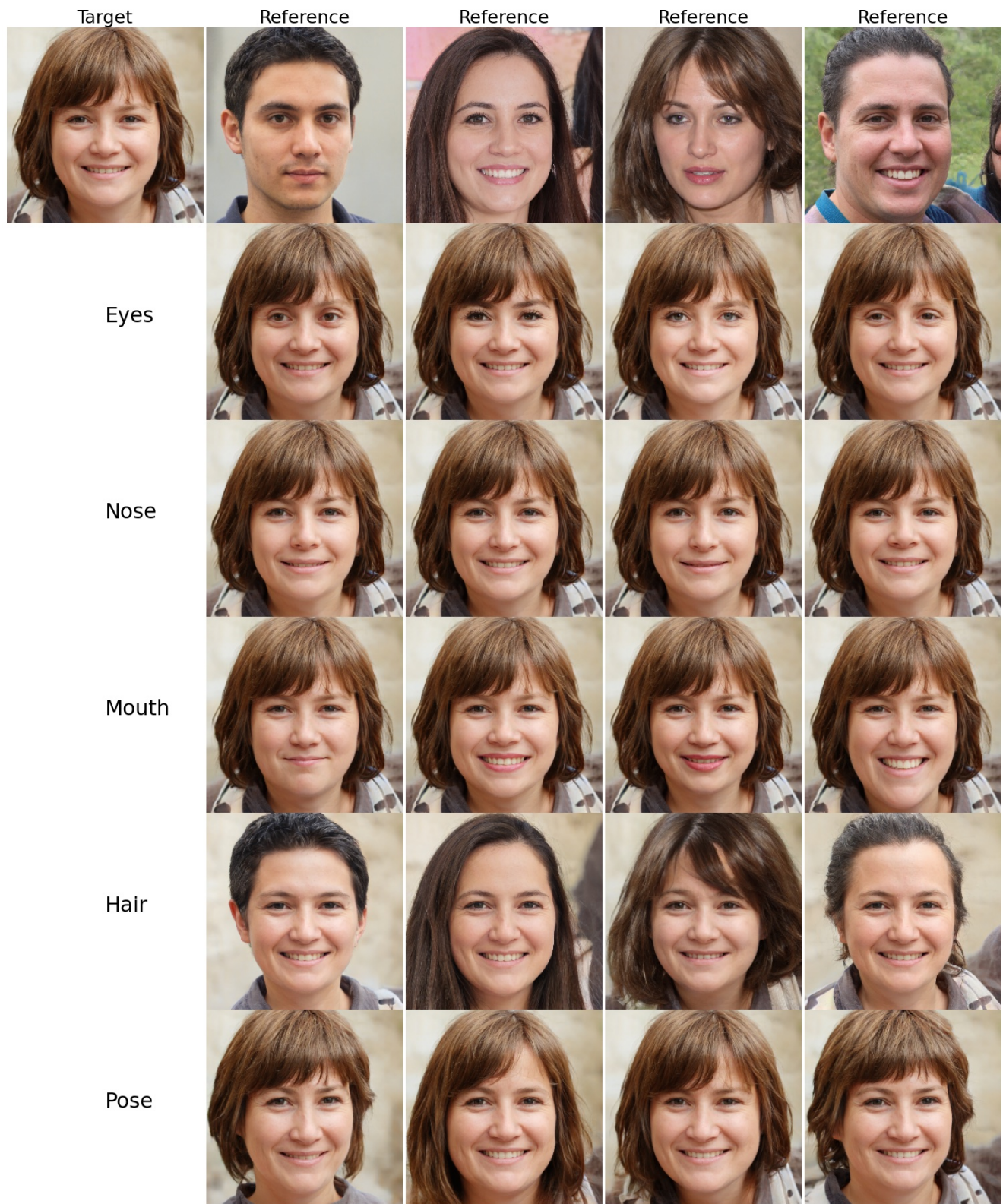
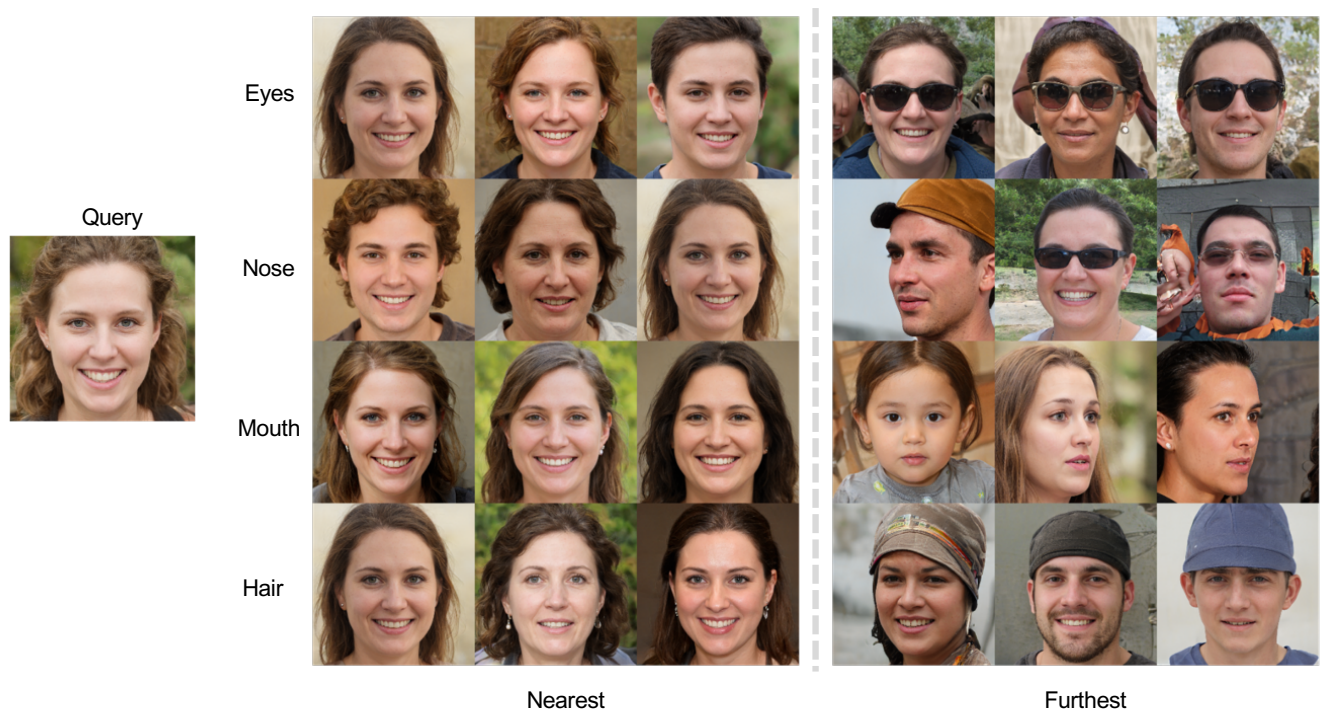Figure 11: Results of facial feature transfer on generated images.
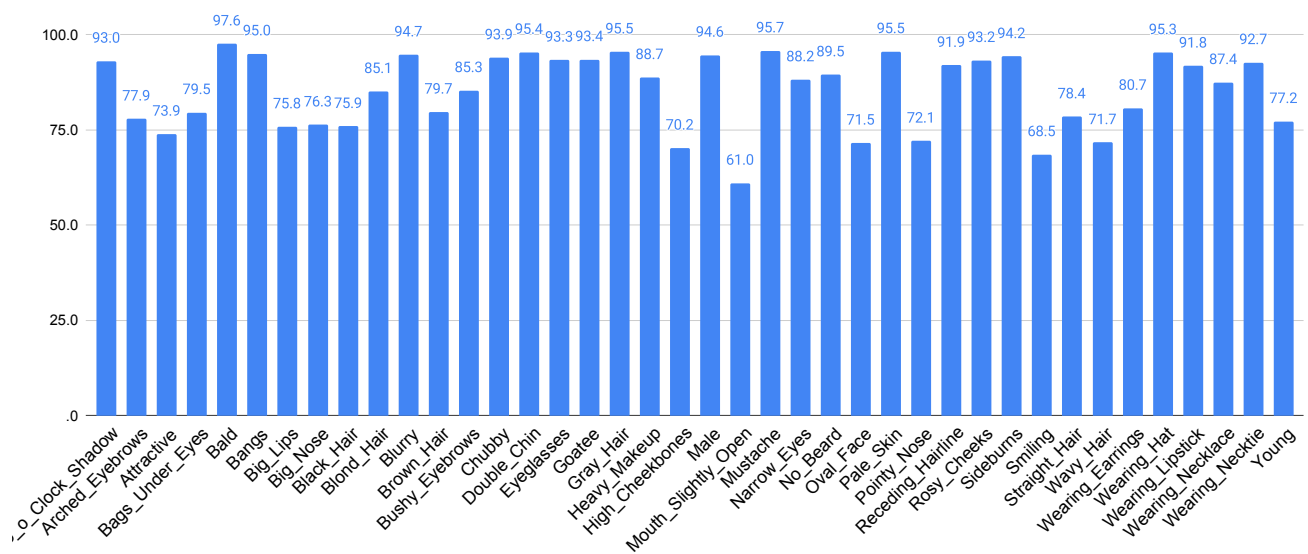
Figure 12: Results of retrieval on generated images.

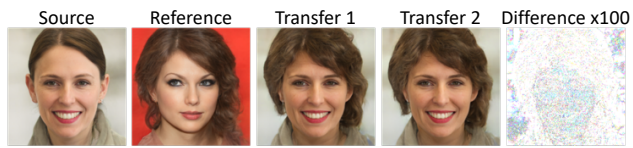Figure 13: Accuracy on 40 CelebA attributes (in %).

Figure 14: **Hair transfer with random noise input**: The effect of noise is negligible to our results even with 100x magnification.