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 $\mathbf{M}_k$ [1] 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.

1. 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:

$$\mathbf{Z}_k^i = \text{argsort}_n \, \mathbf{M}_k^i, \tag{1}$$

where 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. 1, 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.

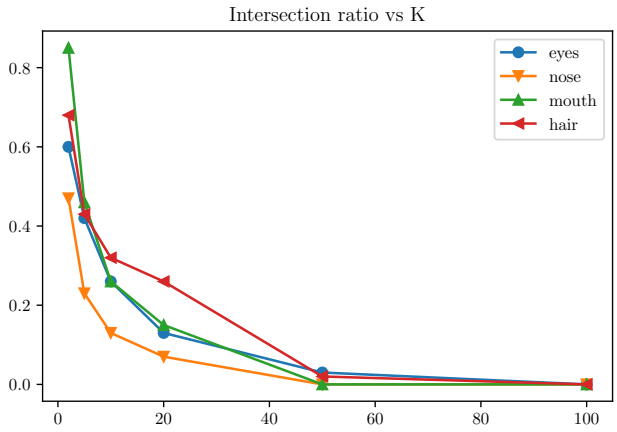

Figure 1: **Intersection Ratio**: This figure shows the intersection ratio (**y-axis**) computed against K, the number of clusters (**x-axis**). The common channels shared by all clusters decrease as the number of clusters increase. This means that for the same facial feature, images do not share the same contributing channels, validating the "submembership" effect discussed in Sec. 3.1 of the main paper.

2. 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. 2 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 [1].

More results: Similar to the figures shown for facial feature transfer and retrieval as in the main paper, Figs. 3 and 4 provide more examples for facial feature transfer retrieval, respectively on generated images.

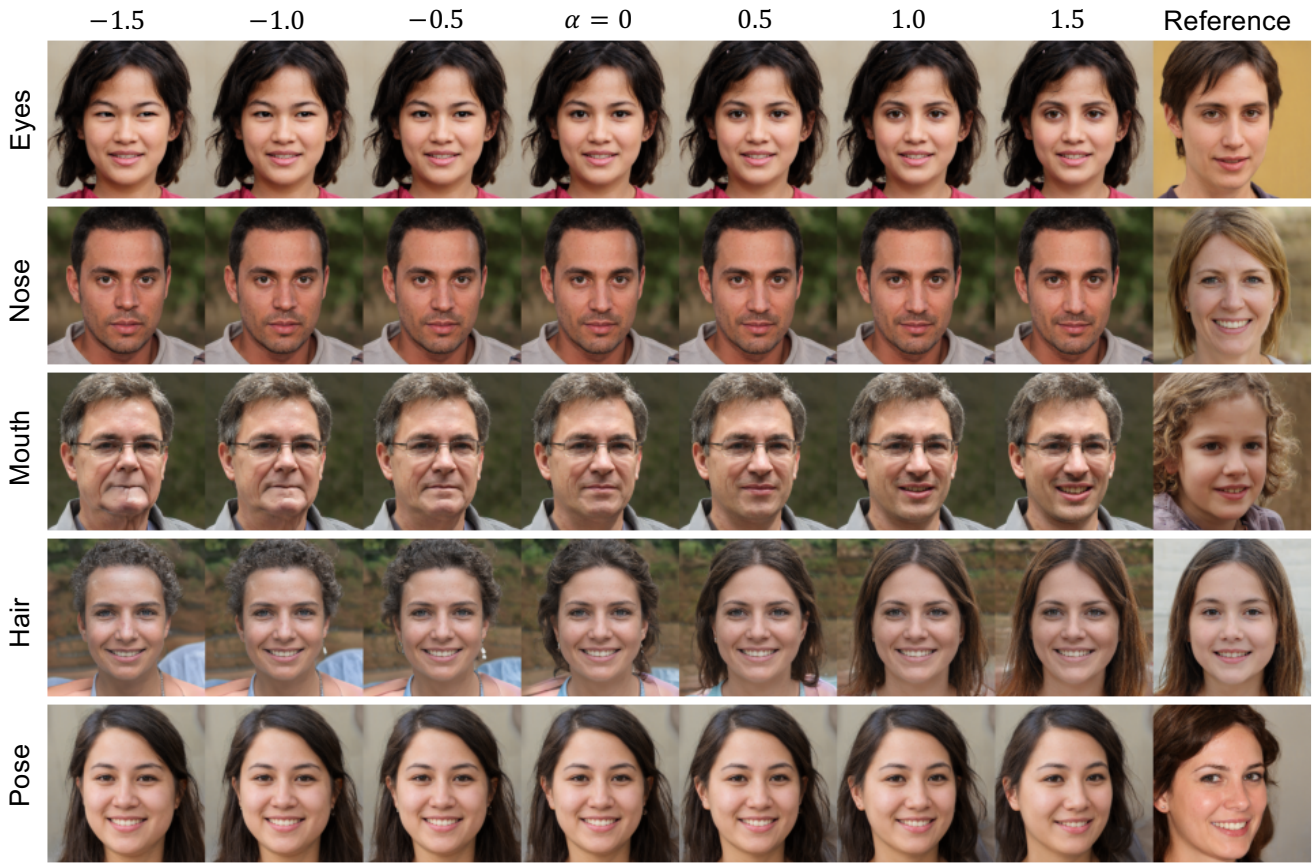

Figure 2: We scale $\mathbf{q}_k$ according to different α to allow interpolation between the source image (the left most column) and the reference image (the right most column) on a particular facial feature. With the side-by-side comparisons with different α, we observe that RIS is able to produce 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 [1].

3. 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 [2]. Subsets of features were manually selected to associate attributes with the facial features that the proposed method attempts to retrieve. Table 1 shows the full list of binary attributes for each facial feature. For completeness, Fig. 5 illustrates the accuracy of each of the 40 attributes of our pretrained model, with an average of 85.27% overall accuracy.

4. 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.

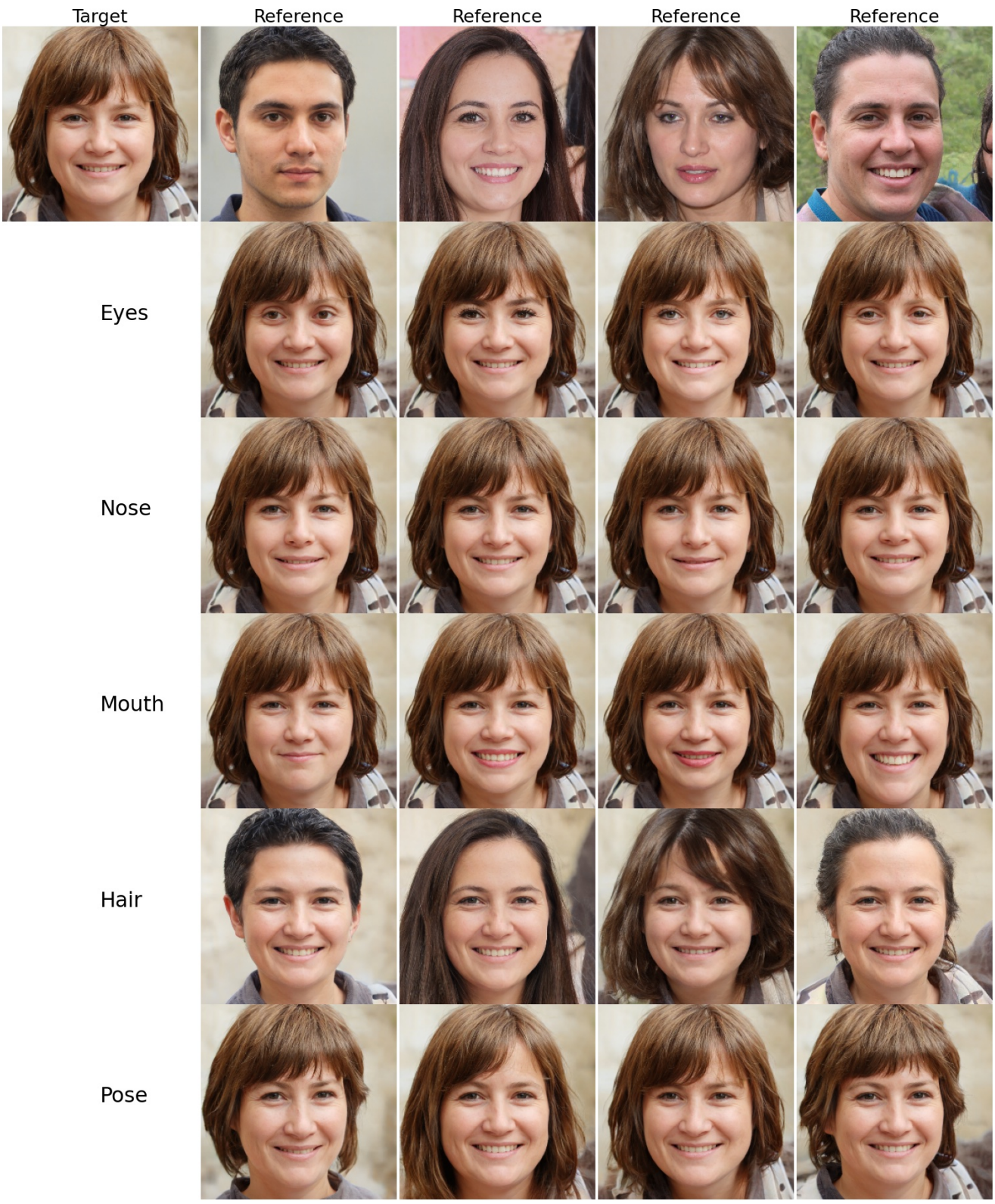

Figure 3: Results of facial feature transfer on generated images.

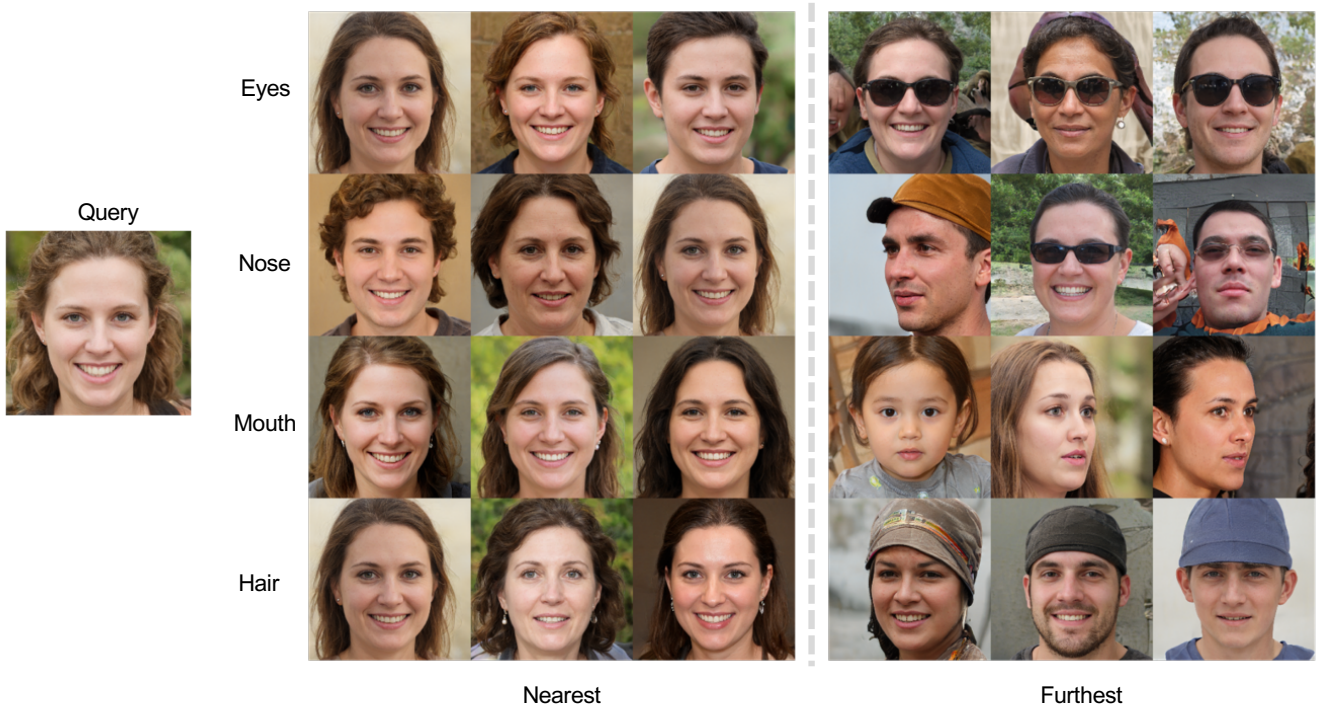

Figure 4: Results of retrieval on generated images.

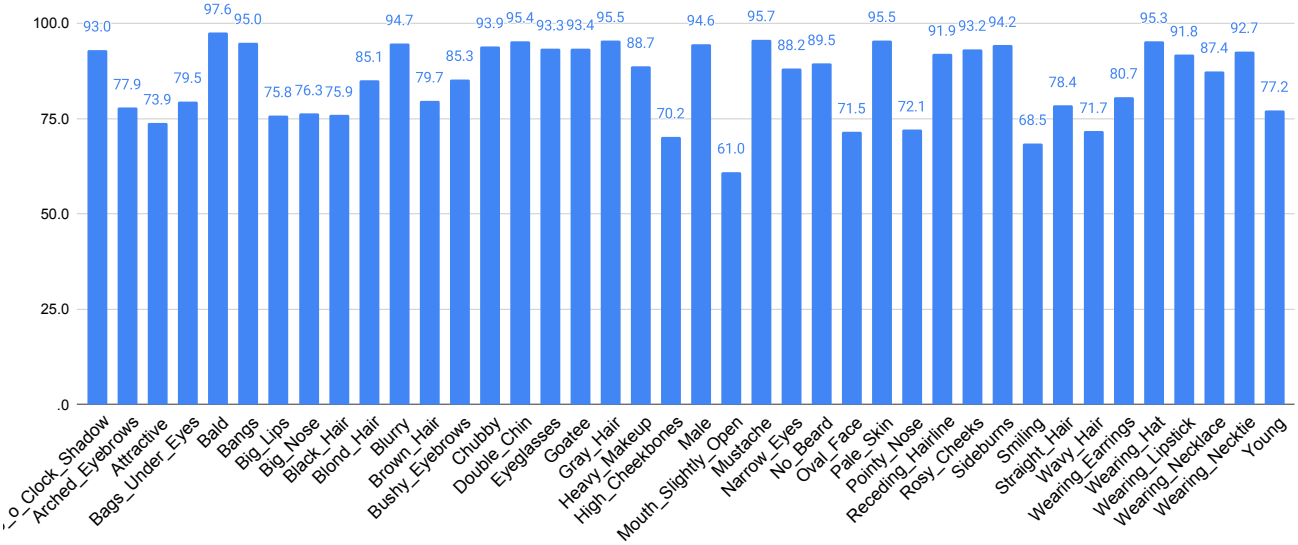

Figure 5: Accuracy on 40 CelebA attributes (in %).

5. 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 $\sim$0.2s latency, we believe RIS remains suitable for real world applications. Computing $\mathbf{M}$ for a dataset of 50K im-

ages for retrieval takes less than 10 minutes on a single Titan Xp GPU (avg 0.12s per image).

6. 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. 6. From the absolute difference between different random runs, we observe that their delta is negligible.

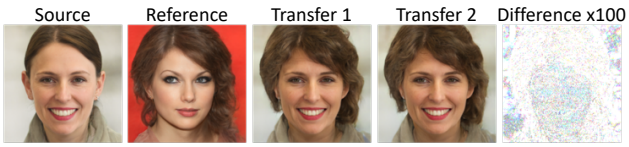

Figure 6: **Hair transfer with random noise input**: The effect of noise is negligible to our results even with 100x magnification.

References

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

[2] 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. 2

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 1: 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.