Image-based CLIP-Guided Essence Transfer

Supplementary Material

1 Qualitative Results

In this section, we provide the complementary versions of Fig. 3, Fig. 4 from the paper, i.e. we present the sources and targets from Fig. 3 manipulated with the encoder, and the sources and targets from Fig. 4 manipulated with the optimizer. Fig. 1 complements Fig. 3 from the paper, and Fig. 2 complements Fig. 4 from the paper. The Dumbledore target in Fig. 1 is slightly different than the one used for the optimizer since the target we used for the optimizer cannot be aligned, which is a precondition for our encoder to process an image. As can be seen in both Fig. 1 and Fig. 2, both our methods produce meaningful manipulations, and successfully transfer notable semantic attributes from the targets to the various sources, while preserving the identity of the source images.

Fig. 1: Examples of using our encoder with the same sources and targets as in Fig. 3 in the main paper. The manipulations are consistent and induce the same semantic manipulation for a wide range of diverse sources.

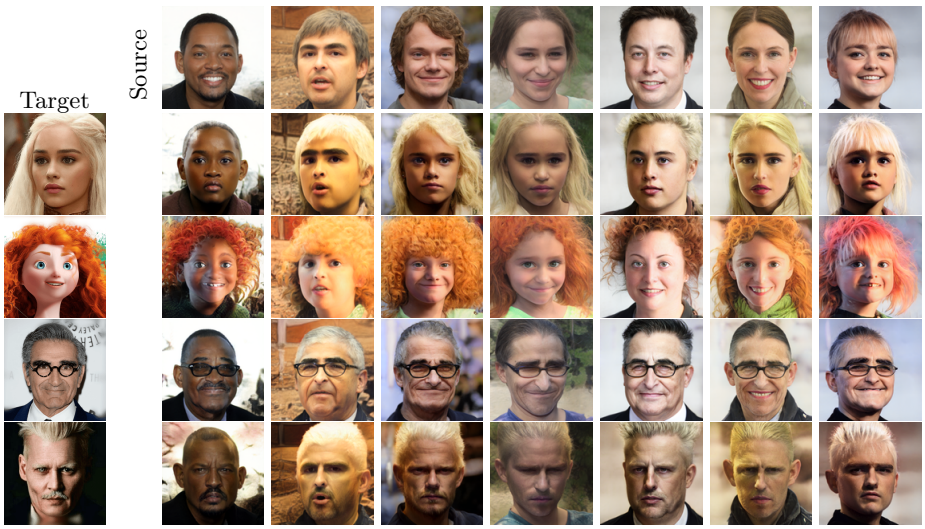

Fig. 2: Examples of using our optimizer with the same sources and targets as in Fig. 4 in the main paper. The manipulations are consistent and induce the same semantic manipulation for a wide range of diverse sources.

2 Celebrity Targets

Fig. 3 presents all the targets used in our celebrities test. As can be seen, our targets demonstrate notable or extreme semantic properties, such as unusual hair colors and styles, beards, glasses, as well as a variety of ages, genders, and ethnicities, and also contain out-of-domain animated characters.

Fig. 3: Well-known characters used as targets for our celebrities test. The chosen targets are diverse in terms of gender, ethnicity, and the attributes that they demonstrate.

3 Qualitative Comparisons

In this section, we enclose the qualitative comparison to StyleGAN-NADA that was omitted from Fig. 6 of the main text (see Fig. 4), as well as qualitative comparisons from our FFHQ experiments, which do not appear in the main text.

As can be seen from Fig. 4, StyleGAN-NADA falls short on preserving the source identity, and results in either a uniform identity or a mixed identity of the source and the target, as mentioned in the main text.

Fig. 5 presents a comparison between our method and methods that suffer from identity loss on the FFHQ test. The demonstrated results are very similar to those obtained on our celebrities test, where both JoJoGAN and StyleGAN-NADA suffer from a high loss of the source identity. Fig 6 shows a comparison between our method and BlendGAN on the FFHQ test. As in the celebrities test, BlendGAN does not transfer the semantic attributes of the target to the sources.

Fig. 4: Comparison to StyleGAN-NADA from our celebrities test. First three rows are manipulations with our optimizer, and last three with our encoder.

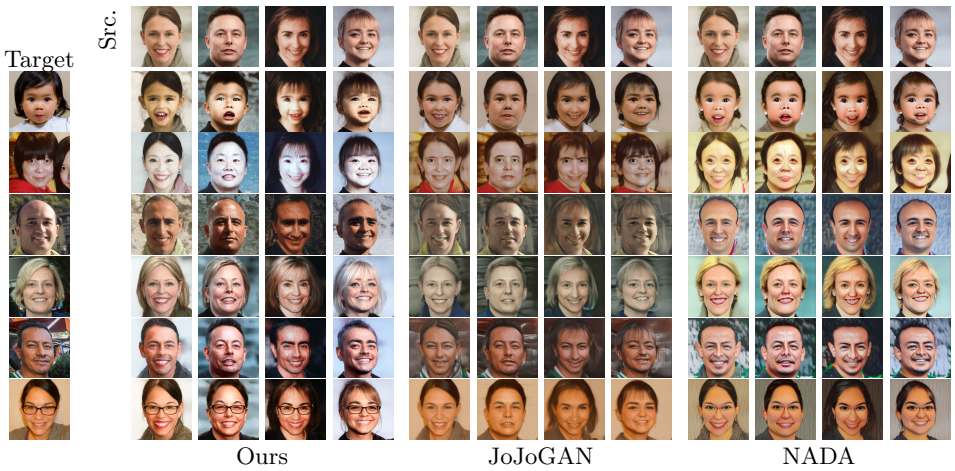

Fig. 5: Comparison to methods that suffer from a high loss of source identity on our FFHQ test. First three rows are manipulations with our optimizer, and last three with our encoder.

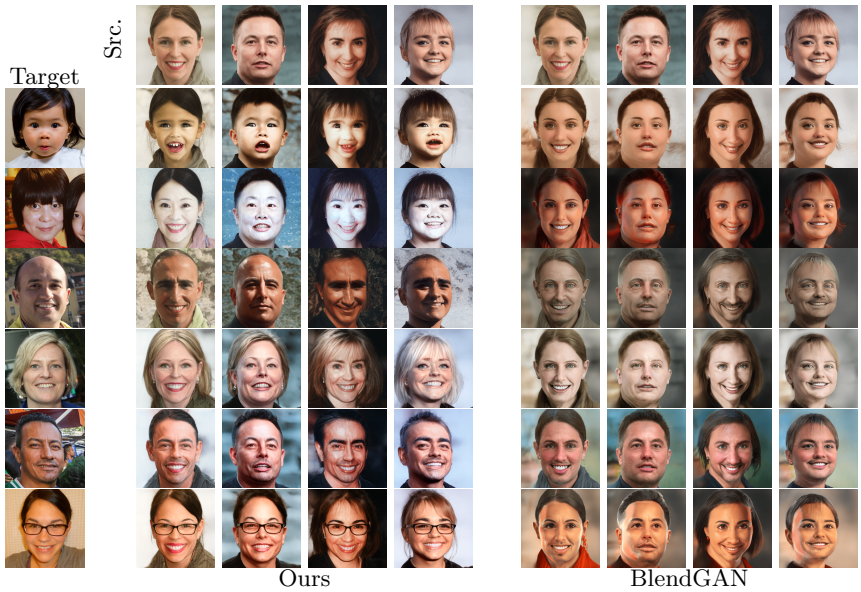

Fig. 6: Comparison to methods that only partially transfer the semantic properties from our FFHQ test. First three rows are manipulations with our optimizer, and the last three are with our encoder.

4 Essence Decoding

Fig. 7 presents additional examples of essence decoding with BLIP for various sources and targets using our optimizer. The decoding demonstrates that the transferred attributes correspond to the most notable semantic attributes of the target images.

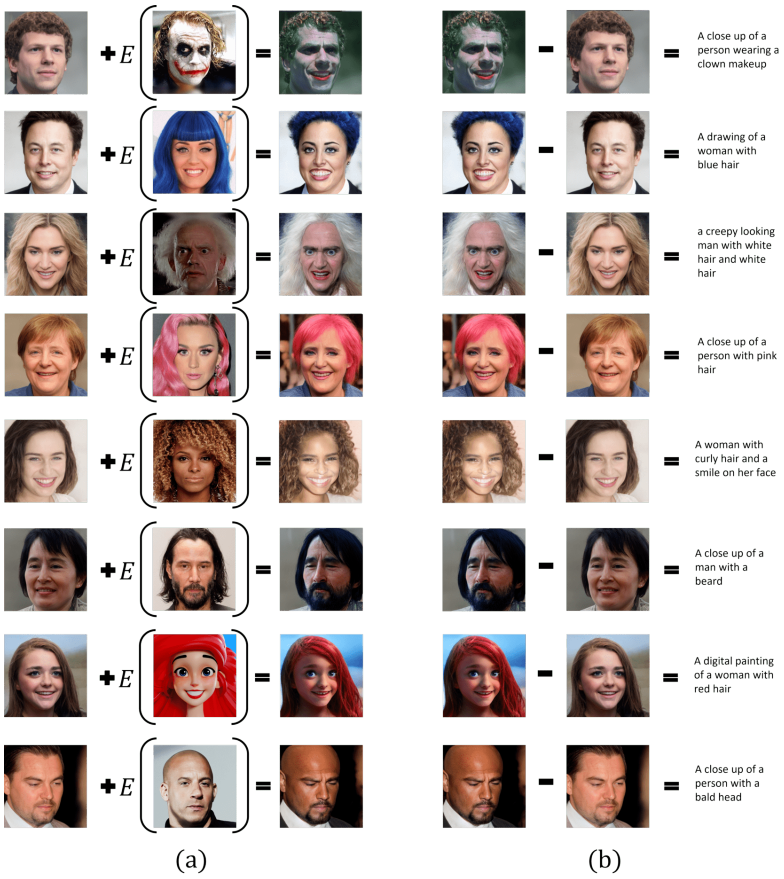

Fig. 7: Examples of essence decoding. (a) presents the targets, sources, and manipulation results, with E representing the essence extraction, i.e. we add the essence of the right image to the left image, and (b) demonstrates the decoding of the essence vectors for each example.

5 Ablation Study

	Quality	Identity scores		Semantic scores	
	FID (↓)	Source (↑)	Target (↓)	BLIP (↑)	CLIP (↑)
Ours	173.4±24.6	42.6±11.1	17.4±8.1	67.5±6.7	73.5±5.1
Our w/o Eq. 6	165.6±19.5	45.6±9.8	15.5±6.6	63.9±7.8	69.5±4.1
Our w/o Eq. 5	**133.0±0**	**77.4±0**	**2.2±5.0**	51.2±6.6	53.2±3.4
Our w/o L_2	233.3±29.9	6.8±3.8	26.6±8.9	**78.2±6.2**	**88.3±4.0**

Table 1: Quantitative comparison of different variations of our optimization-based method on the celebrities test. Results that indicate identity loss of the source are marked in orange; results that indicate that no semantic attributes were transferred are marked in red.

Our ablation tests are conducted with the optimizer, which is also the most successful variant of our method, using the celebrity targets set. We choose to use the celebrity targets dataset due to its diverse nature and challenging semantic attributes. We use the first 4 images of StyleCLIP's test set as our training sources.

As can be seen from Fig. 8 of the main text, removing the similarity loss (Eq. 5) results in nearly no semantic change to the sources. As explained, this can be attributed to the fact that Eq. 5 is the only one among our loss terms that demands semantic similarity to the target. When removing the L_2 regularization, the result is a blurry unified identity presenting some of the semantic attributes of the target. Fig. 8 of the main text also demonstrates that our consistency loss (Eq. 6) is necessary to produce semantically accurate directions, otherwise, the optimization deviates towards a partial subset of the semantic properties that are easy to control for the source images.

Tab. 1 encloses the results of our full ablation tests. As can be seen, removing our L_2 loss results in severe identity loss, with the overall lowest source identity score by a very large margin. This result is consistent with Fig. 8 in the main text. Removing Eq. 5 results in very low semantic scores and great identity preservation, as can be expected since there is almost no change in the source image. Note that in the case where we remove Eq. 5, the algorithm is target-agnostic, therefore the FID and source identity scores have a standard deviation of 0. When removing Eq. 6, notice how both semantic scores are degraded. This can be attributed to the fact that as mentioned, the consistency loss is necessary to ensure proper transfer of the semantic attributes, especially when the target presents challenging attributes such as unusual hair colors or styles, glasses, and makeup (see Fig. 8). While the identity scores are slightly improved without Eq. 6, recall that as mentioned in the main text, there is an inherent trade-off between the identity scores and the semantic scores- as the semantic scores

increase, more attributes are transferred, thus the identity preservation score is lower as a direct result. Importantly, the identity preservation property from the main text ID-score$_{source}$ > ID-score$_{target}$ is well preserved with our method, and the target identity score is very similar and low in both cases (with and without Eq. 6), indicating that the added semantic changes by the consistency do not cause the identity to shift more towards the target than the source, i.e. our consistency loss in Eq. 6 is necessary to transfer the essence features that *do not* break the balance between the source and target identities, and maintains high quality, diverse results, as reflected by the higher semantic scores.

Fig. 8 contains additional examples from the described ablation experiment. As can be seen, in accordance with Tab. 1, removing Eq. 6 results in inaccurate or partial semantic transfer. For example, with Harry Potter as target (first row in Fig. 8), the signature glasses are not transferred without our consistency loss (without Eq. 6), while our full method captures the glasses in their original shape, as well as the hair color, and the pale skin. In other cases, removing Eq. 6 leads to inaccurate transfer, for example, without Eq. 6, the pink and blue hair (lines 2,5 of Fig. 8) transfer blonde-green and orange-pink hair, respectively, while our full method faithfully captures all hair colors. Removing Eq. 5 leads to nearly no semantic change as expected, and removing the L_2 regularization results in severe identity loss and distorted results.

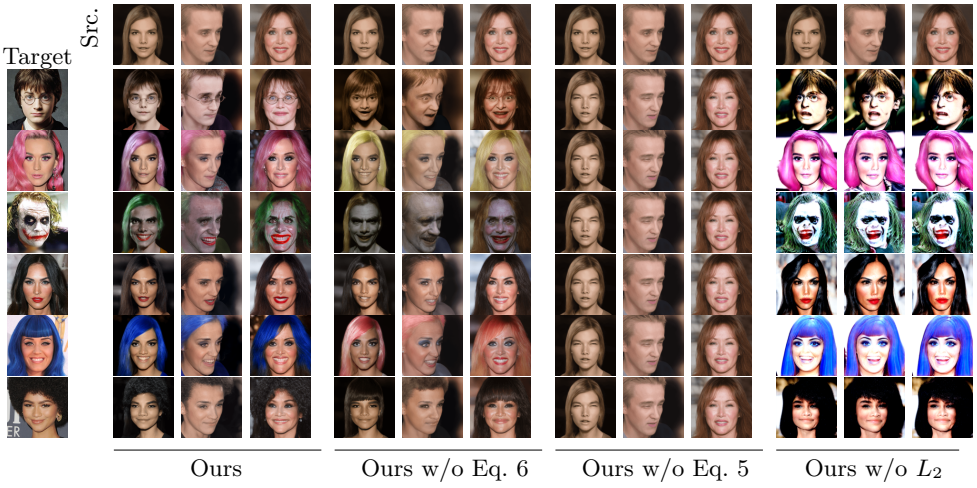

Fig. 8: A comparison between different variations of our method. Removing the consistency loss (Eq. 6) results in inaccurate or partial transfer of the attributes, removing the similarity loss (Eq. 5) results in nearly no semantic changes to the source, and removing the L_2 regularization results in severe identity loss.

6 Sensitivity Test

To evaluate our method's sensitivity to the selection of the number of training sources N, we conduct an experiment where we gradually increase N on our optimizer and encoder, using random targets from the celebrities test. As can be seen from Fig. 9, for small values of N ($N < 4$) the semantic BLIP and CLIP scores are low indicating that the semantic features of the target were not transferred to the sources. Starting from $N = 4$, the semantic scores are high, at the expanse of some identity loss. There is some evident advantage to using more than $N = 4$ sources for training, however, a larger training set would require additional computational resources for training. Additionally, in accordance with the results in the main paper, the optimizer is superior to the encoder in terms of identity preservation (low target ID score, even for large values of N).

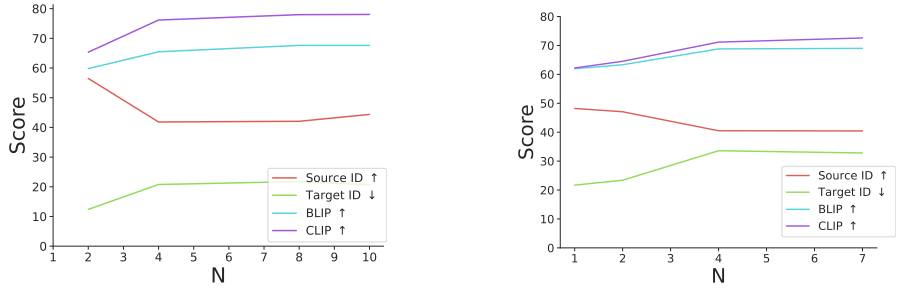

Fig. 9: Sensitivity tests to the number of training sources N with (a) the optimizer, (b) the encoder.