Image-based CLIP-Guided Essence Transfer

Hila Chefer[1], Sagie Benaim[2], Roni Paiss[1], and Lior Wolf[1]

[1] Tel Aviv University
[2] University of Copenhagen

Fig. 1: Results of our method on various targets and sources. The first row presents the target images to extract the essence from, the middle row shows the sources to transfer the essence to, and bottom row presents the results of our method.

Abstract. We make the distinction between (i) style transfer, in which a source image is manipulated to match the textures and colors of a target image, and (ii) essence transfer, in which one edits the source image to include high-level semantic attributes from the target. Crucially, the semantic attributes that constitute the essence of an image may differ from image to image. Our blending operator combines the powerful StyleGAN generator and the semantic encoder of CLIP in a novel way that is simultaneously additive in both latent spaces, resulting in a mechanism that guarantees both identity preservation and high-level feature transfer without relying on a facial recognition network. We present two variants of our method. The first is based on optimization, while the second fine-tunes an existing inversion encoder to perform essence extraction. Through extensive experiments, we demonstrate the superiority of our methods for essence transfer over existing methods for style transfer, domain adaptation, and text-based semantic editing. Our code is available at: https://github.com/hila-chefer/TargetCLIP.

1 Introduction

Style transfer, which typically refers to rendering the content of an image in the style of a different image, is a highly researched task in computer vision and computer graphics [12,5,35,19,10,16,28,32,33,44,6]. This work explores a related task, which we refer to as *essence transfer*. The essence of an image is defined to be the set of attributes that appear in the high-level textual description of the image. Our blending involves borrowing semantic features from a "target" image I_t and transferring them to a "source" image I_s, thus creating an output image $I_{s,t}$. We find that "essence" features capture properties such as skin complexion or texture, as do traditional style transfer methods, but also semantic elements such as gender, age, and unique facial attributes, when considering faces.

A rigorous definition of our goal is elusive, as the set of features defined as the essence may change from image to image, so we adopt a pragmatic approach. It has been shown [36,11] that latent spaces of high-level vision networks, (i.e., involving capabilities such as image understanding [48]) are additive. Ergo, subtracting the representation of two inputs yields a meaningful shift between the inputs encoding the difference between the two. Our method transforms source images I_s conditioned on a target image I_t. It forces the learned transformation to be doubly additive: once in the latent space of the generator, and once in the latent space of the understanding engine. Out of all possible ways of transforming a source image I_s according to a target image I_t, we obtain, for every I_t a transformation that is based on a constant shift over all I_s in the generator space and leads to a constant difference in the high-level description of the image.

For the generator network, we employ the powerful StyleGAN [23] generator. Additivity in the latent space of StyleGAN was demonstrated in [43,49,15], for linearly interpolating different images along semantic directions, as well as for the manipulation of semantic attributes ([36] for example). For the image recognition engine, we employ the CLIP network [39], which has shown impressive zero-shot capabilities across multiple domains such as image classification [39] and adaptation of generated images [11]. It was also shown to behave additively [36,11]. Since CLIP was trained in a contrastive manner, using textual descriptions, different images with the same high-level textual description are expected to receive a high similarity score, as their textual descriptions will be nearly identical. This allows our method to enforce consistency based on the semantic properties of the image, rather than pixel-level similarity.

We propose a method based on two loss terms. The first term ensures that the transformed image is semantically similar to the target image I_t in the latent space of CLIP. The second term links the constant shift in the latent space of the generator to a constant shift in the latent space of CLIP, leading to a semantically consistent edit that is independent of the identity of the source I_s. We propose two methods for essence transfer. The first is based on per target optimization, while the second fine-tunes an inversion encoder to perform essence transfer. While an optimization-based approach is more accurate in capturing the relevant semantic properties, our encoder version only requires a forward pass on the target image to produce the target-specific (source-agnostic) essence vector, which

defines the essence transfer operator for that image. We compare our method with state of the art style transfer and semantic editing methods and show that our novel double-additive formulation is necessary to successfully perform essence transfer. Finally, we decode to text the learned directions, demonstrating that the semantic edits we employ correspond to the attributes of the target image.

2 Related Work

Style Transfer Our work most closely relates to style transfer [10,16,28,32,44,33,25]. Unlike [12,20,18], we derive style from CLIP [39], a recently proposed method for the semantic association of text and images. In this space, two images are close, or similar, to each other if their textual association is close. Such similarity may consider unique style elements, such as texture or complexion. It may also consider semantic elements, such as gender and facial attributes. We argue that this notion of style, which we use here, is more general. CLIP has been used in several recent works to enable the fine-tuning of StyleGAN for domain adaptation [11,52] with impressive results, yet, as we show, the existing style transfer and domain adaptation methods fall short on the task of essence transfer, focusing usually on colors, textures and domain shifts, and can suffer from severe identity loss.

Image Manipulation Our work is also related to recent image manipulation works based on a pre-trained generator [7,45,15,36]. One set of works typically manipulates an image based on finding a set of possibly disentangled and semantic directions [43,49]. These works typically borrow the semantic meaning from the generator itself. A recent work called StyleCLIP [36] showed the remarkable ability to borrow the semantic meaning from the CLIP space, and inspired many additional works to use CLIP for semantic editing and domain adaptation [52,11,2]. Our method uses CLIP in a similar manner to that presented in [36]. However, unlike our method, StyleCLIP considers text-driven manipulations, thus it is limited by what can be described in words, and the knowledge obtained by CLIP.

GAN Inversion GAN inversion aims to extract a latent vector z that corresponds to a target image I, i.e. z holds that $G(z) = I$, where G is the generator. Most inversion methods can be split into two types; optimization-based methods [51,1,4,8,13,29,53], which employ an optimization process to find a latent z such that $G(z)$ is closest to a specific target image I, and encoders [47,3,14,21,24,34,37,38,50] which are trained to extract a latent z for any input image I. Most methods for StyleGAN inversion focus on the $\mathcal{W}, \mathcal{W}+$ latent spaces. The $\mathcal{W}$ space is more editable, yet suffers from degraded expressiveness [1], therefore $\mathcal{W}+$ has been adopted for inversion. We employ the e4e encoder [47] since it mitigates the distortion-editability trade-off by training an encoder in the $\mathcal{W}+$ latent space while encouraging the result to be close to the $\mathcal{W}$ space.

3 Assumptions and Problem Formulation

We now provide a formal definition of the essence transfer task, and an overview of the proposed method for any generator G and semantic encoder C.

We define the essence of an image I as the set of semantic features that constitute the high-level textual description of the image. The method employs four input components: (i) A generator G, which, given a vector z, generates an image $G(z)$, (ii) An image recognition engine C, which, given an image I, provides a latent representation of its high-level textual description, $C(I)$, in some latent space, (iii) A target image I_t, from which the essence is extracted, and (iv) A set of source images S, which are used to provide the statistics of images for which the method is applied. For clarity, we define S as a set of z vectors in a latent space of G, and each source image as $G(z)$.

Given these four inputs, our goal is to provide a generator H such that the image $H(z)$ transfers the high-level textual features of a target image I_t to the source image $I_s = G(z)$. If one wishes to transform an image I_s rather than a vector z using H, the image can be converted to a latent z using an inversion method [47,40,41]. We note that our formulation does not require, at any stage, the inversion of I_t. On the generator side, linearity is expressed by:

$$H(z) = G(z + b) \tag{1}$$

for some shift vector b, in the latent space of G. Linearity in the latent space of the image recognition engine is expressed as:

$$\forall z \in S \quad d = C(H(z)) - C(G(z)) \,, \tag{2}$$

for some fixed d. Put differently, modifying any vector z in G's latent space with b induces a constant semantic change in the latent space of C. This property is what is referred to as a "global semantic direction" by Patashnik et al. [36]. However, our method goes about obtaining said global direction differently. The source-agnostic behavior is obtained by minimizing the following over H and d:

$$\sum_{z \in S} dist(C(H(z)) - C(G(z)), d) \,, \tag{3}$$

where $dist(\cdot, \cdot)$ calculates the distance between two vectors in the latent space of the semantic encoder C. For example, CLIP uses cosine similarity to estimate vector similarities.

Equivalently, to obtain H, we can minimize the following over H:

$$\sum_{z1, z2 \in S} dist\left(C(H(z_1)) - C(G(z_1)), C(H(z_2)) - C(G(z_2))\right) \,. \tag{4}$$

So far, we defined the problem of learning a pair of semantic directions b, d, in two different latent spaces, such that (b, d) match. We wish to add a constraint that ties the shifts to I_t. To this end, we wish to maximize similarity in the semantic space provided by the recognition engine C between I_t and the generated images $H(z)$. That is, we wish to minimize the sum of distances $\sum_{z \in S} dist(C(H(z)), C(I_t))$.

4 Method

We now define our method, based on the formulation in Sec. 3. Note that in accordance with the training process of CLIP, we employ cosine similarity to

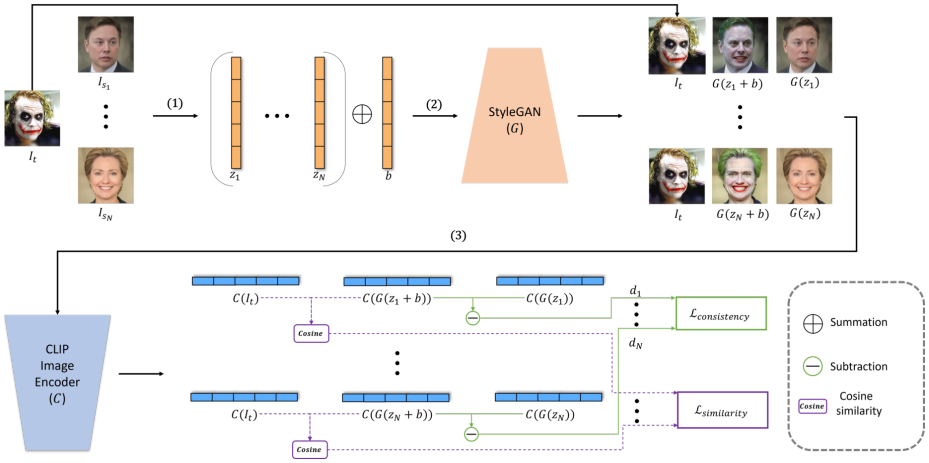

Fig. 2: An illustration of our loss calculation flow. Step (1) inverts the source images to obtain their latents $z_1, ..., z_N$, and adds the proposed essence vector b. In step (2), the StyleGAN generator decodes the source latents $z_1, ..., z_N$ and the manipulations $z_1 + b, ..., z_N + b$ to images. Step (3) encodes the sources $(G(z_1), ..., G(z_N))$, manipulations $(G(z_1 + b), .., G(z_N + b))$, and target (I_t) with CLIP. $\mathcal{L}_{consistency}$ demands that semantic changes be identical for $z_1, .., z_N$. $\mathcal{L}_{similarity}$ ensures that $G(z_1 + b), ..., G(z_N + b)$ are semantically similar to I_t.

estimate semantic similarity between two images. $\sum_{z \in S} dist(C(H(z)), C(I_t))$ becomes the following loss term, applied over a batch of source images, S:

$$\mathcal{L}_{similarity} = \frac{1}{N} \left(\sum_{z \in S} 1 - \frac{C(I_t) \cdot C(G(z + b))}{\|C(I_t)\|_2 \, \|C(G(z + b))\|_2} \right), \tag{5}$$

where $N = |S|$ is the batch size, and $z_1, .., z_N \in \mathcal{W}+$. The similarity loss estimates the semantic similarity between the image encodings of the target image and the manipulated images. By setting $N = 1$, this loss becomes identical to the semantic loss employed by other semantic editing methods based on CLIP [36].

The second concept we enforce is consistency. The goal of essence transfer is to modify the source image using a collection of semantic attributes that encapsulates the essence of the target image. These attributes are independent of the source image. We demand that the semantic edits induced by the direction b in the latent space of G be consistent across the source images, using CLIP's latent space. This is expressed in Eq. 4 above and translates to the following loss:

$$\mathcal{L}_{consistency} = \frac{1}{\binom{N}{2}} \left(\sum_{i_{src_1}, i_{src_2} \in I_s} 1 - \frac{\Delta i_{src_1} \cdot \Delta i_{src_2}}{\|\Delta i_{src_1}\|_2 \, \|\Delta i_{src_2}\|_2} \right) \tag{6}$$

where $\Delta i_{src} = C(G(i_{src} + b)) - C(G(i_{src}))$, as annotated in Eq. 2 as d, and, as before, N is the batch size. $\mathcal{L}_{consistency}$ guarantees that the direction encapsulated in b produces semantic edits that are identical across a batch of source images S. Fig. 2 illustrates the steps of obtaining $\mathcal{L}_{similarity}, \mathcal{L}_{consistency}$ from a batch of sources $I_{s_1}, ..., I_{s_N}$ and a target image I_t.

The optimization problem solved during training in order to recover H, as defined in Eq. 1, is given as:

$$b^* = \arg\min \mathcal{L}_{similarity} + \lambda_{consistency}\mathcal{L}_{consistency} + \lambda_{L_2} \|b\|_2 , \tag{7}$$

where $\lambda_{consistency}, \lambda_{L_2}$ are hyperparameters. We use the same hyperparameter values in all our experiments and all our methods: $\lambda_{consistency} = 0.5, \lambda_{L_2} = 0.003$.

In contrast to other methods [36], ours does not rely on any face recognition models for preventing identity loss. In order to maintain the identity of the source images $I_1, ..., I_N$ we employ a standard L_2 regularization to limit the magnitude of the effect that b has on source images.

Restating Eq. 1, after obtaining the essence vector b^* for a target image I_t, manipulating a source image I_s is done as follows:

$$I_{s,t} = G(z_s + b^*), \tag{8}$$

where z_s is the latent that corresponds to the source image I_s, which can be obtained by inverting the image I_s. Thus, we simply add the essence vector b^* to the latent representing the source image.

Essence Optimization The first method we propose is a simple optimization process of finding an essence vector b^* that minimizes the objective in Eq. 7. Unlike other optimization-based methods for semantic editing, our method is more stable in the sense that the same set of hyperparameters can be applied for each target, and no target-specific tuning is required. The implementation employs the Adam optimizer [26] for 1000 iterations with a learning rate of 0.2. Due to resource limitations, we use only $N = 4$ images for our double additivity losses (Eq. 5, 6). For difficult edits, i.e. edits containing unconventional or extreme semantic attributes such as blue skin, we found it beneficial to initialize the direction b in the optimization process to be the inversion of the target produced with the e4e encoder. This can be attributed intuitively to the fact that the inversion of the target contains, among other identity-specific attributes, the semantic attributes that constitute the high-level textual description of the image, i.e. its essence. Therefore, initializing the direction b to be the inversion of the target steers the optimization toward semantic properties that are related to the target image.

Essence Encoder For our second method, we fine-tune a pre-trained e4e encoder [47] over the pSp framework [40] to output the essence vector b^* of the input image instead of its inversion. Since the encoder is pre-trained for inversion, the initial output for each image I_t contains, among other features, the semantic features that comprise its essence. The goal of the fine-tuning process is to shift the weights of the encoder such that the output for each image I_t will be the semantic parts of the inversion that correspond to the essence vector.

Fig. 3: Examples of using our optimization-based method. Output images preserve the identity of the sources, while borrowing the semantic essence of the targets.

This fine-tuning is performed on a small dataset of 200 random images from the CelebA-HQ dataset [31], and evaluated on 50 random images from the CelebA-HQ test set. We use a learning rate of $1e-4$ for 3000 iterations, with a batch size of 1 target image and $N = 5$ source images for our double additivity losses (Eq. 5, 6). The objective and its hyperparameters are identical to the ones used for the optimization-based method (Eq. 7). Unlike other methods that train an encoder or a generator for each target text or image, such as [36,11], our encoder is fine-tuned once and can accommodate *any* target after the fine-tuning. Other methods require training for each target text or image from scratch, which takes at least a few minutes and in some cases hours, while our inference time per target is just a few seconds.

5 Experiments

We present qualitative and quantitative results that demonstrate the advantage of our method for the task of essence transfer over the most recent methods for style transfer and domain adaptation. For a complete evaluation, we make an effort to be inclusive and compare also with methods that have somewhat different goals, i.e. text-based image editing methods.

Qualitative results Fig. 3, Fig. 4 contain results of our optimization-based method and encoder-based method, using a wide variety of target and source images. All source images were inverted with e4e [47], and were not part of the training batch of sources used for optimization. The manipulation of the sources with the essence vector was done as detailed in Eq. 8. We present different choices of sources and targets in Fig. 3 and Fig. 4, in order to demonstrate the diversity of both our methods. For completeness, the complementary versions of the figures,

Fig. 4: Examples of using our essence encoder on various targets and sources. Output images preserve the identity of the sources while borrowing the semantic essence of the targets.

in which the sources and targets in Fig. 3 are edited with the encoder, and the sources and targets in Fig. 4 are edited with the optimizer, are also presented in the supplementary material.

As can be seen, our essence transfer results display the most notable semantic attributes of the target. For example, when using Doc Brown as target (first row in Fig. 3), the signature wild, white hair is transferred from the target to all sources, as are the wide open eyes. Our methods also preserve the identity of the sources well, despite training with only $N = 4$ ($N = 5$) images to enforce semantic consistency for our optimization (encoder) method. Additionally, the semantic edits are consistent across all sources, demonstrating that our method is indeed able to produce source-agnostic essence vectors.

Quantitative Results Our experiments use as sources a set of 68 images inverted with e4e. For each target image I_t and source image I_s we use our methods and the baseline methods to edit the source according to the target and produce $I_{s,t}$. We then evaluate the quality of the produced edits for each method. Since there are many works involving style transfer and domain adaptation, we focus on the most recent state of the art, including unpublished works. We focus on works that are applicable to our use-case, i.e., methods that are able to perform one-shot editing. Our baselines include BlendGAN [30] and JoJoGAN [6] for face stylization, StyleGAN-NADA [11] and Mind The Gap (MTG) [52] for domain adaptation, and as a CLIP-aided text-based image editing method, we include StyleCLIP's [36] global directions method. We note that since the StyleCLIP method is text-based, it can only be used in manipulations where the target is a well-known character. Despite its inherent limitation, we also present this comparison for completeness, since the global directions method resembles ours in

		Quality	Identity scores		Semantic scores	
		FID (↓)	Source (↑)	Target (↓)	BLIP (↑)	CLIP (↑)
Celebrities Test	StyleGAN-NADA [11]	215.7±26.1	23.0±4.7	33.0±7.1	**84.5±3.6**	**94.0±1.3**
	Mind The Gap [52]	180.4±19.3	27.2±5.6	39.4±8.1	75.8±5.6	75.4±7.0
	JoJoGAN [6]	186.1±12.7	36.0±6.1	50.7±6.9	72.6±7.3	71.8±6.2
	BlendGAN [30]	177.8±12.6	37.6±6.5	**5.2±7.7**	60.8±6.2	58.4±5.2
	StyleCLIP [36]	166.9±9.0	**70.7±26.0**	6.2±6.8	54.8±6.6	55.7±5.0
	Our encoder	188.7±23.2	39.0±6.5	31.9±5.7	69.0±6.0	72.6±5.5
	Our optimization	**163.6±16.7**	43.5±6.8	17.0±6.6	66.9±6.0	74.4±3.2
FFHQ Test	StyleGAN-NADA [11]	220.2±41.8	24.1±5.5	28.3±9.2	**81.1±4.2**	**91.0±3.2**
	JoJoGAN [6]	175.2±15.2	42.3±4.0	41.7±11.4	76.0±6.0	67.1±7.4
	BlendGAN [30]	175.1±14.5	37.6±5.3	**2.4±6.0**	64.4±6.7	54.7±7.8
	Our encoder	175.6±23.5	42.5±5.5	30.8±6.9	72.8±4.9	66.7±6.1
	Our optimization	**161.1±17.2**	**45.2±8.6**	17.0±7.2	74.1±4.9	74.8±5.8

Table 1: Quantitative comparison with baselines. The StyleCLIP baseline can only be applied to well-known characters (celebrities test), and Mind the Gap provides no public code at this time, thus can only be applied to the celebrities test (see main text). 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.

that it outputs a target-dependent and source-agnostic direction in the StyleGAN $\mathcal{S}$ space to perform a manipulation according to an input textual description. Since our methods strive to transfer the features of the high-level textual description of the target, we find this comparison to be relevant as well.

The goal of essence transfer is twofold. First, we wish to transfer the semantic properties that constitute the high-level textual description from a target image I_t to a source image I_s. Second, we wish to maintain the identity of I_s as much as possible. We therefore suggest two types of metrics to evaluate the quality of a proposed essence transfer result, $I_{s,t}$. The first type employs the ArcFace [9] network for face recognition to ensure that the manipulation maintains the identity of I_s as much as possible, while avoiding an identity shift towards I_t, i.e. we calculate:

$$\text{ID-score}_{source}(I_{s,t}) = \langle R(I_s), R(I_{s,t})\rangle,\ \text{ID-score}_{target}(I_{s,t}) = \langle R(I_t), R(I_{s,t})\rangle,$$

where R denotes a pre-trained ArcFace face recognition representation, and $\langle\cdot,\cdot\rangle$ computes cosine similarity. Since neither of our methods uses face recognition in the training process, this metric faithfully measures how well our manipulations preserve the source identity. Intuitively, since we add semantic features from the target, we shift the identity of the source to some extent. For example, modifying the gender of the source induces an inherent change in one of the identity attributes of the source. The combination of scores $\text{ID-score}_{source}, \text{ID-score}_{target}$ reveals whether the manipulation was able to remain close to the identity of the

source or shifted toward the identity of the target. A successful essence transfer is expected to maintain a *high* ID-score$_{source}$ score, and a *low* ID-score$_{target}$ score, indicating that the manipulation's identity fits the source better than the target.

Next, to estimate the semantic quality of the manipulation, we use the latent spaces of BLIP [27] and CLIP [39], as follows:

$$\text{Semantic-score}(I_{s,t}) = \langle C(I_t), C(I_{s,t}) \rangle,$$

where C notates a pre-trained BLIP or CLIP image encoder, and $\langle \cdot, \cdot \rangle$ computes cosine similarity. Since our method, as well as most baselines [11,52,36], use the latent space of CLIP in the training process, BLIP provides an important alternative for estimating the semantic similarity between the target image I_t and the manipulation $I_{s,t}$. For each target I_t, the overall identity scores and semantic scores are calculated as an average of the scores for all source images. The aggregated scores for the models are calculated as an average of the score for each target, i.e. we average the results across the sources for each target, and then average across the targets to obtain the model's final score. We also present the standard deviations as an indication of the method's consistency.

Additionally, in accordance with previous works on style transfer, we present the Fréchet inception distance (FID) [17] as implemented in [42] to estimate the quality of the manipulations, which is calculated as follows:

$$\text{FID-score}(I_{s_1,t}, ..., I_{s_{68},t}) = \text{FID}(\{I_{s_1,t}, ..., I_{s_{68},t}\}, \{I_1, ..., I_{7{,}000}\}),$$

where $\{I_{s_1,t}, ..., I_{s_{68},t}\}$ are the manipulations of the sources induced by the target t, and $\{I_1, ..., I_{7,000}\}$ is a set of $7,000$ randomly chosen images from the FFHQ [22] dataset, which provides the background distribution of natural faces. Since some of our baselines are trained to adapt the domain of the target, we calculate the FID score only for the targets describing a human face, in order to avoid biasing the results against these baselines. Note that this calculation produces a relatively high FID score for all methods, since the produced dataset $\{I_{s_1,t}, ..., I_{s_{68},t}\}$ is inherently limited in its diversity, due to the fact that all images share semantic properties transferred from t, leading to a shift from the distribution of unedited faces, which are more diverse. However, methods that suffer from mode collapse or overfitting are expected to achieve a much higher (lower is better) score than those that preserve the original identity of the source images, since identity preservation will lead to greater diversity among the results.

We present two experiments. For the first, we construct a comparison in a setting that is more similar to the setting the baselines were trained for, i.e. we construct a dataset of 31 images of celebrity faces with 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 out-of-domain animated characters (see the supplementary material for all examples used in this experiment). For the text-based baseline, we employ the same course of action as in the StyleCLIP paper, where the textual prompt for the manipulation is of the form "an image of {*name of target*}". Our second experiment involves targets with less extreme semantic features. We use the first 50 images of the FFHQ [22]

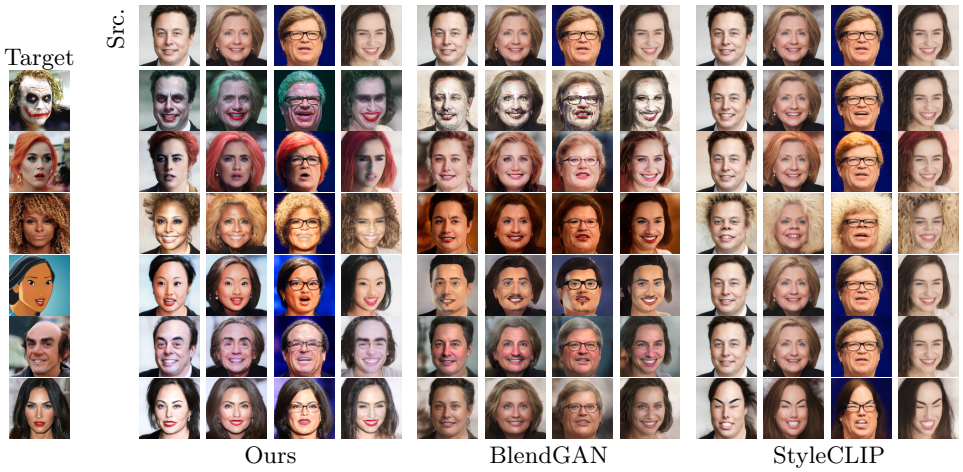

Fig. 5: Comparison to methods that only partially transfer the semantic properties. First three rows are manipulations with our optimizer, and the last three are with our encoder.

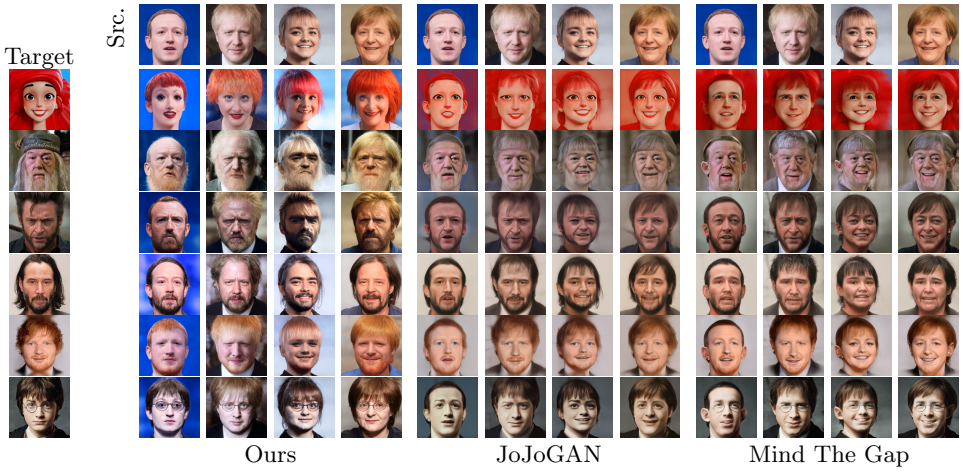

Fig. 6: Comparison to methods that suffer from high loss of source identity. First three rows are manipulations with our optimizer, and last three with our encoder.

dataset as targets, and the same 68 source images as before. Since our targets are no longer well-known characters, the baseline for text-based image editing is no longer applicable. Additionally, for the Mind The Gap baseline [52], no official code was released- although the authors kindly provided results for the

first experiment, but not the second one - therefore this baseline is not presented in our second experiment.

Tab. 1 presents the results of both our experiments. We divide the methods into three types. (i) Methods that demonstrate underfitting, i.e., fail to transfer the essence of the target. These methods perform very well on the identity metrics and very poorly on the semantic metrics. As can be seen in Tab. 1, both BlendGAN and the StyleCLIP demonstrate this phenomenon. Marked in red in the tables are the similarity scores for the methods by BLIP and CLIP. Both are significantly lower than the semantic scores of the other methods. See Fig. 5 for examples of this case from our first experiment. BlendGAN focuses on modifying almost only the colors, and StyleCLIP either hardly changes the semantic properties or distorts the sources. (ii) methods that demonstrate overfitting, i.e. methods that suffer from identity preservation issues. These methods transfer most or all of the semantic features of the target, and eliminate the source identity in the process or create a blended identity of source and target. This results in very high semantic compatibility scores, but on the other hand, a failure in identity preservation. As can be expected, the methods designed for domain adaptation, i.e., StyleGAN-NADA and Mind The Gap fall in this category, as does JoJoGAN. The values marked in orange in Tab. 1 demonstrate that both StyleGAN-NADA and Mind The Gap obtain very low source identity scores (significantly lower than the other methods), while JoJoGAN receives the highest (lower is better) target identity score in both experiments (50.7% on the celebrities test and 41.7% on the FFHQ test, surpassing all other baselines by more than 10%). This indicates that StyleGAN-NADA and Mind The Gap fall short on identity preservation, while JoJoGAN results in an image derived from the identity of the target instead of the source. For example, the Ariel target (first row in Fig. 6) demonstrates that the baselines result in a unified identity with the semantic features of the target. Similarly, the Keanu Reeves and Ed Sheeran targets (fourth and fifth rows in Fig. 6) result in blended identities with the baselines. We omit StyleGAN-NADA from Fig. 6 for brevity, as the other two methods scored higher in terms of identity preservation. The full comparison, as well as comparisons from our second experiment can be found in the supplementary material.

Lastly, (iii) methods that successfully transfer the semantic properties of the target (have high BLIP and CLIP similarity scores) while also preserving the identity of the source more than the target (i.e., ID-score$_{source}$ > ID-score$_{target}$), which both our methods fall under. When analyzing the quality of the manipulated images, our optimization-based method scores the best overall FID score in both experiments by a significant margin, indicating that it is able to produce high-quality manipulations. In addition, our optimization demonstrates a very low (lower is better) target identity score, suggesting that our essence transfer does not borrow from the identity of the target in order to obtain the semantic changes. While our encoder preserves the identity quality (ID-score$_{source}$ > ID-score$_{target}$), notice that it achieves a higher target identity score, indicating that our encoder is not as successful as our optimizer in identity preservation. This can be attributed to two facts. First, our encoder is based on a pre-trained inversion encoder that

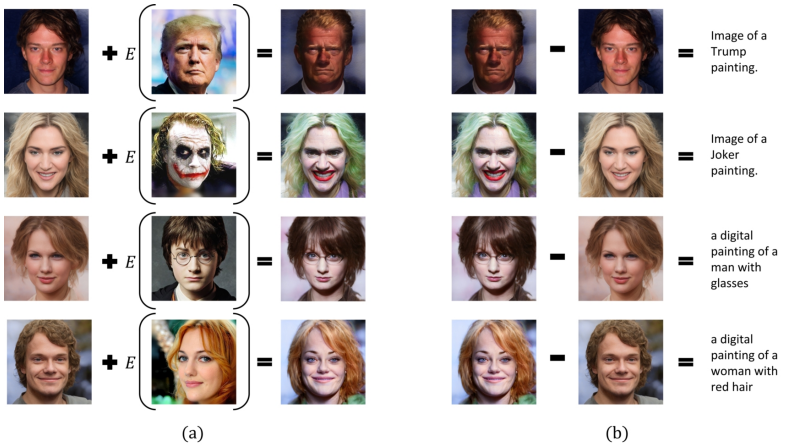

Fig. 7: Examples of essence decoding. (a) presents the targets, sources, and manipulation results, with E representing the essence extraction, i.e. we add the essence of the right image to the left image, and (b) demonstrates the decoding of the essence vectors for each example.

encapsulates the target identity by design. Second, while our optimizer learns an essence vector for each target, the encoder is only fine-tuned on a small set of images and is not optimized for each target at inference time.

Essence Interpretability To demonstrate that our methods indeed produce essence vectors that correspond to the semantic attributes of the target image I_t, we present results of decoding the essence vectors to text. We observe semantic differences by applying a decoder on the vector $d = C(I_{s,t}) - C(I_s)$ where C represents the semantic image encoder of CLIP or BLIP, i.e. we decode the difference between the source image after and before the manipulation. Fig. 7 shows examples for four different edits and their interpretations. The Donald Trump and the Joker edits (first and second row of Fig. 7) were performed with our optimization-based method, encoded with CLIP, and decoded with [46], and the rest were performed using our encoder approach and encoded/decoded with BLIP. As can be seen, the textual interpretations of each direction correspond well to the semantic properties of the targets, and for the Donald Trump and Joker edits, the directions are decoded as Trump and the Joker, demonstrating the ability of our method to capture essential semantic features of the targets used. For targets with less distinct semantic properties, the decoding shows that the apparent gender is transferred along with other significant semantic properties such as hair color and eye glasses. See the supplementary material for more examples of essence interpretability using decoding.

Limitations of the encoder-based approach Unlike the optimizer, the encoder is not re-trained for each target. This results in an accuracy-runtime trade-off, i.e., while the encoder produces an essence vector in a few seconds, in some cases

Fig. 8: A comparison between our optimizer (a), and encoder (b).

- where the target contains unconventional semantic properties - it produces a result that does not encapsulate all the semantic attributes one would expect to be included in the essence. In contrast, since the optimization is performed from scratch for each target, it takes longer (a few minutes) to produce the result, but it is more accurate. Fig. 8 presents two examples of such challenging targets, where the optimization-based method is superior to the encoder. For Donald Trump (first row in Fig. 8), optimization results in an essence that includes all notable semantic properties- the wrinkles, lips, and unique hair color - while the result of the encoder fails to capture the unique attributes with the same accuracy. Similarly, for Katy Perry (second row in Fig. 8), optimization captures the unconventional hair color, while the encoder fails to do so. As evident from Tab. 1, while both the encoder and optimizer receive high semantic scores, the optimizer allows for results with higher quality (lower FID, lower target ID score).

Ablation Study We refer the reader to the supplementary material for an ablation study that examines the impact of each loss term of our method.

6 Conclusions

We define a novel task referred to as essence transfer. Unlike style transfer or domain adaptation, essence transfer draws semantic features that correspond to the high-level textual description of an image. Essence transfer is particularly challenging since the set of attributes that constitute the high-level description may differ from image to image. We propose an optimizer and an encoder, both based on double-additivity in the latent spaces of StyleGAN and CLIP, and measure our method against state of the art methods adapted from style transfer and domain adaptation. Our extensive experiments demonstrate that our novel formulation is significantly preferable to the baselines in terms of identity preservation, the quality of the produced images, and the identification of the essential attributes of an image.

Acknowledgment This project has received funding from the European Research Council (ERC) under the European Unions Horizon 2020 research and innovation programme (grant ERC CoG 725974).

References

1. Abdal, R., Qin, Y., Wonka, P.: Image2stylegan: How to embed images into the stylegan latent space? 2019 IEEE/CVF International Conference on Computer Vision (ICCV) pp. 4431–4440 (2019)
2. Abdal, R., Zhu, P., Femiani, J.C., Mitra, N.J., Wonka, P.: Clip2stylegan: Unsupervised extraction of stylegan edit directions. ArXiv **abs/2112.05219** (2021)
3. Alaluf, Y., Patashnik, O., Cohen-Or, D.: Restyle: A residual-based stylegan encoder via iterative refinement. 2021 IEEE/CVF International Conference on Computer Vision (ICCV) pp. 6691–6700 (2021)
4. Bau, D., Strobelt, H., Peebles, W.S., Wulff, J., Zhou, B., Zhu, J.Y., Torralba, A.: Semantic photo manipulation with a generative image prior. ACM Transactions on Graphics (TOG) **38**, 1 – 11 (2019)
5. Bousmalis, K., Silberman, N., Dohan, D., Erhan, D., Krishnan, D.: Unsupervised pixel-level domain adaptation with generative adversarial networks. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 95–104 (2017)
6. Chong, M.J., Forsyth, D.: Jojogan: One shot face stylization. ArXiv **2112.11641** (2021)
7. Collins, E., Bala, R., Price, B., Susstrunk, S.: Editing in style: Uncovering the local semantics of gans. In: CVPR. pp. 5771–5780 (2020)
8. Creswell, A., Bharath, A.A.: Inverting the generator of a generative adversarial network. IEEE Transactions on Neural Networks and Learning Systems **30**, 1967–1974 (2019)
9. Deng, J., Guo, J., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 4685–4694 (2019)
10. Efros, A.A., Leung, T.K.: Texture synthesis by non-parametric sampling. In: Proceedings of the Seventh IEEE International Conference on Computer Vision, Kerkyra, Greece. vol. 2, pp. 1033–1038 vol.2 (1999). https://doi.org/10.1109/ICCV.1999.790383
11. Gal, R., Patashnik, O., Maron, H., Chechik, G., Cohen-Or, D.: Stylegan-nada: Clip-guided domain adaptation of image generators (2021)
12. Gatys, L.A., Ecker, A.S., Bethge, M.: Image style transfer using convolutional neural networks. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (6 2016)
13. Gu, J., Shen, Y., Zhou, B.: Image processing using multi-code gan prior. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 3009–3018 (2020)
14. Guan, S., Tai, Y., Ni, B., Zhu, F., Huang, F., Yang, X.: Collaborative learning for faster stylegan embedding. ArXiv **abs/2007.01758** (2020)
15. Härkönen, E., Hertzmann, A., Lehtinen, J., Paris, S.: Ganspace: Discovering interpretable gan controls. arXiv preprint arXiv:2004.02546 (2020)
16. Hertzmann, A., Jacobs, C.E., Oliver, N., Curless, B., Salesin, D.H.: Image analogies. In: Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques. Association for Computing Machinery, New York, NY, USA (2001). https://doi.org/10.1145/383259.383295, `https://doi.org/10.1145/383259.383295`
17. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: Proceedings

of the 31st International Conference on Neural Information Processing Systems. p. 6629–6640. NIPS'17, Curran Associates Inc., Red Hook, NY, USA (2017)
18. Huang, X., Belongie, S.: Arbitrary style transfer in real-time with adaptive instance normalization. In: 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy. pp. 1510–1519 (2017). https://doi.org/10.1109/ICCV.2017.167
19. Huang, X., Belongie, S.J.: Arbitrary style transfer in real-time with adaptive instance normalization. 2017 IEEE International Conference on Computer Vision (ICCV) pp. 1510–1519 (2017)
20. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: European Conference on Computer Vision. pp. 694–711. Springer (2016)
21. Kang, K., Kim, S., Cho, S.: Gan inversion for out-of-range images with geometric transformations. 2021 IEEE/CVF International Conference on Computer Vision (ICCV) pp. 13921–13929 (2021)
22. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 4396–4405 (2019)
23. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of stylegan. In: CVPR. pp. 8110–8119 (2020)
24. Kim, H., Choi, Y., Kim, J., Yoo, S., Uh, Y.: Exploiting spatial dimensions of latent in gan for real-time image editing. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 852–861 (2021)
25. Kim, S.S.Y., Kolkin, N., Salavon, J., Shakhnarovich, G.: Deformable style transfer. In: European Conference on Computer Vision (ECCV) (2020)
26. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. CoRR **abs/1412.6980** (2015)
27. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation (2022)
28. Li, Y., Fang, C., Yang, J., Wang, Z., Lu, X., Yang, M.H.: Universal style transfer via feature transforms. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. p. 385–395. NIPS'17, Curran Associates Inc., Red Hook, NY, USA (2017)
29. Lipton, Z.C., Tripathi, S.: Precise recovery of latent vectors from generative adversarial networks. ArXiv **abs/1702.04782** (2017)
30. Liu, M., Li, Q., Qin, Z., Zhang, G., Wan, P., Zheng, W.: Blendgan: Implicitly gan blending for arbitrary stylized face generation. In: Advances in Neural Information Processing Systems (2021)
31. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: Proceedings of International Conference on Computer Vision (ICCV) (December 2015)
32. Luan, F., Paris, S., Shechtman, E., Bala, K.: Deep photo style transfer. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, USA. pp. 6997–7005 (2017). https://doi.org/10.1109/CVPR.2017.740
33. Luan, F., Paris, S., Shechtman, E., Bala, K.: Deep painterly harmonization. Computer Graphics Forum **37**(4), 95–106 (2018). https://doi.org/https://doi.org/10.1111/cgf.13478, `https://onlinelibrary.wiley.com/doi/abs/10.1111/cgf.13478`
34. Luo, J., Xu, Y., Tang, C., Lv, J.: Learning inverse mapping by autoencoder based generative adversarial nets. In: International Conference on Neural Information Processing. pp. 207–216. Springer (2017)

35. Ojha, U., Li, Y., Lu, J., Efros, A.A., Lee, Y.J., Shechtman, E., Zhang, R.: Few-shot image generation via cross-domain correspondence. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 10738–10747 (2021)
36. Patashnik, O., Wu, Z., Shechtman, E., Cohen-Or, D., Lischinski, D.: Styleclip: Text-driven manipulation of stylegan imagery. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2085–2094 (2021)
37. Perarnau, G., van de Weijer, J., Raducanu, B., Álvarez, J.M.: Invertible conditional gans for image editing. ArXiv **abs/1611.06355** (2016)
38. Pidhorskyi, S., Adjeroh, D.A., Doretto, G.: Adversarial latent autoencoders. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 14092–14101 (2020)
39. Radford, A., Kim, J.W., Hallacy, C., et al.: Learning transferable visual models from natural language supervision. arXiv preprint arXiv:2103.00020 (2021)
40. Richardson, E., Alaluf, Y., Patashnik, O., Nitzan, Y., Azar, Y., Shapiro, S., Cohen-Or, D.: Encoding in style: a stylegan encoder for image-to-image translation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2021)
41. Roich, D., Mokady, R., Bermano, A.H., Cohen-Or, D.: Pivotal tuning for latent-based editing of real images. arXiv preprint arXiv:2106.05744 (2021)
42. Seitzer, M.: pytorch-fid: FID Score for PyTorch. `https://github.com/mseitzer/pytorch-fid` (August 2020), version 0.2.1
43. Shen, Y., Zhou, B.: Closed-form factorization of latent semantics in gans. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1532–1540 (2021)
44. Sunkavalli, K., Johnson, M.K., Matusik, W., Pfister, H.: Multi-scale image harmonization. ACM Trans. Graph. **29**(4) (Jul 2010). https://doi.org/10.1145/1778765.1778862, `https://doi.org/10.1145/1778765.1778862`
45. Tewari, A., Elgharib, M., Bharaj, G., Bernard, F., Seidel, H., Pérez, P., Zollhofer, M., Theobalt, C.: Stylerig: Rigging stylegan for 3d control over portrait images. In: CVPR (2020)
46. Tewel, Y., Shalev, Y., Schwartz, I., Wolf, L.: Zero-shot image-to-text generation for visual-semantic arithmetic. In: CVPR (2021)
47. Tov, O., Alaluf, Y., Nitzan, Y., Patashnik, O., Cohen-Or, D.: Designing an encoder for stylegan image manipulation. arXiv preprint arXiv:2102.02766 (2021)
48. Ullman, S.: High-level vision: Object recognition and visual cognition. MIT press (2000)
49. Voynov, A., Babenko, A.: Unsupervised discovery of interpretable directions in the gan latent space. In: International Conference on Machine Learning. pp. 9786–9796. PMLR (2020)
50. Wang, T., Zhang, Y., Fan, Y., Wang, J., Chen, Q.: High-fidelity gan inversion for image attribute editing. ArXiv **abs/2109.06590** (2021)
51. Zhu, J.Y., Krähenbühl, P., Shechtman, E., Efros, A.A.: Generative visual manipulation on the natural image manifold. In: ECCV (2016)
52. Zhu, P., Abdal, R., Femiani, J.C., Wonka, P.: Mind the gap: Domain gap control for single shot domain adaptation for generative adversarial networks. ArXiv **2110.08398** (2021)
53. Zhu, P., Abdal, R., Qin, Y., Wonka, P.: Improved stylegan embedding: Where are the good latents? ArXiv **abs/2012.09036** (2020)