FlexIT: Towards Flexible Semantic Image Translation

Guillaume Couairon, Asya Grechka, Jakob Verbeek, Holger Schwenk, Matthieu Cord

FlexIT: Towards Flexible Semantic Image Translation

Guillaume Couairon
Facebook AI Research
gcouairon@fb.com

Asya Grechka
Meero
asya.grechka@meero.com

Jakob Verbeek
Facebook AI Research
jjverbeek@fb.com

Holger Schwenk
Facebook AI Research
schwenk@fb.com

Matthieu Cord
LIP6
matthieu.cord@lip6.fr

Figure 1. FlexIT transformation examples. From top to bottom: input image, transformed image, and text query.

Abstract

Deep generative models, like GANs, have considerably improved the state of the art in image synthesis, and are able to generate near photo-realistic images in structured domains such as human faces. Based on this success, recent work on image editing proceeds by projecting images to the GAN latent space and manipulating the latent vector. However, these approaches are limited in that only images from a narrow domain can be transformed, and with only a limited number of editing operations. We propose FlexIT, a novel method which can take any input image and a user-defined text instruction for editing. Our method achieves flexible and natural editing, pushing the limits of semantic image translation. First, FlexIT combines the input image and text into a single target point in the CLIP multimodal embedding space. Via the latent space of an autoencoder, we iteratively transform the input image toward the target point, ensuring coherence and quality with a variety of novel regularization terms. We propose an evaluation protocol for semantic image translation, and thoroughly evaluate our method on ImageNet. Code will be made publicly available.

1. Introduction

The old saying goes: *"You can't make a silk purse from a sow's ear."* Or can you? Truly flexible and powerful semantic image editing is elusive, and current work is limited in terms of possible input images and edit operations. Research in deep generative image models has seen significant progress in recent years, with GANs in particular generating near photo-realistic samples in domains such as human and animal faces [27] or object-centric images [4]. Moreover, recent "style-based" GANs, like StyleGAN [28–30], have an impressively disentangled latent space, where performing copy-pastes between two latent vectors transfers the corresponding styles in the image space.

Consequently, significant research efforts have been put into using pre-trained GANs for semantic image edition. Through specific latent-space manipulation, high-level attributes such as age or gender can be identified and edited in a realistic manner [1, 23, 42, 59]. These approaches, however, present several caveats. First, contrary to generated latents, inferred latent codes representing real images have been shown to react poorly to latent editing operations [19]. Although recent methods [19, 47, 57] improve editability,

input images are still highly limited to the distribution of the generative network. Moreover, edit operations are also limited to the semantics identified in the latent space via a pre-trained classifier [1,42,59] or through a semi-automatic manner [23, 50]. These semantics are specific to the single domain the GAN was trained on, such as age or apparent gender in the case of faces. Some flexibility w.r.t. the input images can be obtained by training a GAN to directly modify the images, known as image-to-image translation. These methods learn a transformation between two domains, using paired data [24, 39, 51] or unpaired data [7, 58]. However, these models only learn a single transformation, or combinations thereof [52], specific to the training data, limiting the scope of their applicability.

We tackle these challenges with a unified framework which modifies an input image based on a user-defined text query of the form $(S \rightarrow T)$, like *cat* $\rightarrow$ *dog*. For this *semantic image translation* task, the goal is to make minimal image modifications while transforming the image as requested. We leverage CLIP [41], which combines text and image representations in one powerful multimodal embedding space. This space is used to define our target point, based on the embeddings from the user input. We perform a per-image optimization procedure, using specific strategies to ensure image quality and relevance to the transformation query. Our method requires only fixed pre-trained components, and can thus be used off-the-shelf without requiring any training. The image is optimized in the latent space of an auto-encoder, rather than a GAN, which greatly enlarges the scope of possible input images. This allows for truly flexible image edits; as Fig. 1 shows, even a sow's ear can be changed into a silk purse.

We propose an evaluation protocol for the task of semantic image translation. Evaluation is based on three criteria: (i) the transformed image should correctly correspond to the text query, (ii) the output image should look natural, and (iii) visual elements irrelevant to the text query should remain unchanged. We thoroughly evaluate our model on ImageNet, and demonstrate quantitatively and qualitatively the superiority of our method against baselines, broadening the horizon of text-driven image editing.

2. Related Work

Image editing. Deep generative networks, like GANs, have given rise to numerous image editing applications, ranging from photography retouching [43], image inpainting [54], object insertion [17], domain translation [55, 58], colorization [24], super-resolution [26, 36], among many others. Automatic user-driven image editing aims at providing the user control to modify an image, by tweaking segmentation masks [38], scene graphs [10], or class labels [6]. Allowing the user to provide unstructured free-form text queries is more challenging. Close to our objective, ManiGAN [37] aims at performing text-based edits by training a model to refine the details of an image based on its textual description. Their quantitative evaluation protocol uses transformation queries on the COCO dataset by considering unaligned (image, caption) pairs, resulting in possibly incoherent transformation queries. We carefully design our evaluation protocol to avoid such cases.

Image latent space. While GANs are highly effective as generative models, inference of the latent variable given an image is intractable. Even though joint learning of an inference network has been proposed, see *e.g.* [11, 14], the mode-seeking training dynamics of GANs are not suited for good reconstruction performance beyond the training distribution (or even within it, if modes are dropped). Variational autoencoders [33], on the other hand, offer an inference network by construction, and their likelihood-based training objective ensures accurate reconstructions.

Vector-quantized variational autoencoders (VQ-VAE) [2, 49], which discretize the latent space, have been found to offer both good reconstructions as well as compelling samples. In particular, VQ-GAN [15, 53] further improves reconstructions by including an adversarial loss term to train the autoencoder. In our work, we adopt the VQ-GAN autoencoder, and edit images in its latent space.

Latent space manipulation. The introduction of "style-based" GANs, such as StyleGAN [28–30] significantly improved the disentanglement of the latent space, leading to a surge of research into its interpretation and manipulation. By using an auxiliary classifier, a simple approach consists in finding linear boundaries in the latent space separating binary attributes [18, 42, 59], which allows to edit attributes by "walking" in the orthogonal latent direction. StyleFlow [1] proposes a non-linear approach by learning the latent transformations using normalizing flows. Other methods [23,50] operate without a pre-trained classifier and find the transformations in an unsupervised manner, requiring a manual labelling process to interpret and annotate the "discovered" transformations. Rather such restricted sets of possible edit dimensions, we target more general transformations described by free-text.

Semantic alignment with CLIP. To align images and text, CLIP [41] learns encoders that map both modalities to a shared latent space in which they can easily be compared and combined. Vision encoders are based on ResNets [20] and Vision Transformers [13].

CLIP, trained on 400M web-crawled image/text pairs with a simple contrastive InfoNCE loss [48], can provide a robust differentiable signal for image synthesis and editing, used in conjunction with diffusion models [32], and vector strokes generators [16]. Similarly to us, StyleCLIP [40] transforms images based on text queries via alignment in

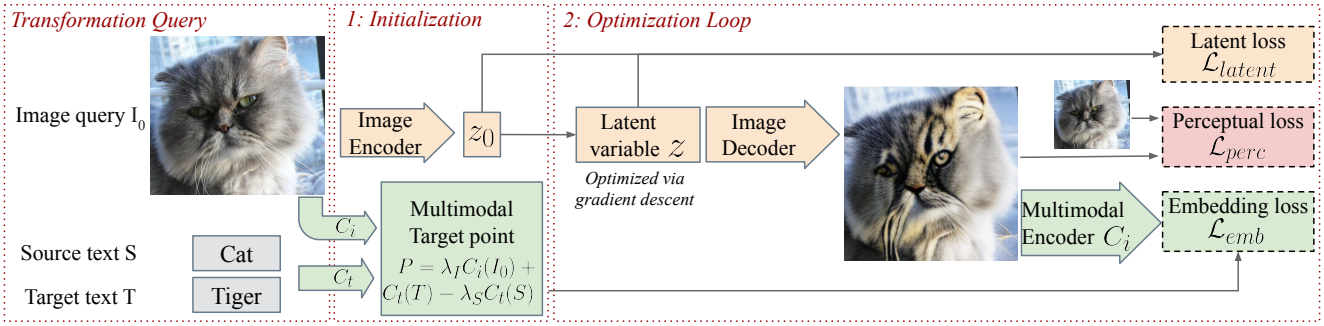

Figure 2. FlexIT optimization framework: components involving the multimodal latent space colored in green; those involving the image latent space in yellow; those involving the LPIPS distance in pink. Given a transformation query (I_0, S, T), we first compute a target point P in the multimodal embedding space, and we encode I_0 in the image latent space to get z_0. Then, for a fixed number of steps, we update the latent variable z (initialized with z_0) to get closer to the target point P. We add two regularization terms: the LPIPS perceptual distance between the input image and the output image, and a latent distance between z and z_0. All networks are frozen, only z is updated.

CLIP's latent space. However it relies on the latent space of StyleGAN2 to optimize the image, which requires training a separate generative and latent space inference model per application domain.

3. FlexIT framework for semantic editing

An overview of our image transformation approach is depicted in Figure 2. It relies on three pre-trained components. First, we edit the input image in a latent space, with the requirement that a wide range of images can be encoded and decoded back to an RGB image with minimal distortion. We chose the VQGAN autoencoder [15] for that purpose. Second, we embed the text query and input image in a multimodal embedding space, to define the optimization target for the modified image. We use the CLIP [41] multimodal embedding spaces. Finally, to ensure that the modified image remains similar to the input, we control its distance to the input image with the LPIPS perceptual distance [56] computed with a VGG [44] backbone.

Optimization scheme. The core idea of the FlexIT method is to edit the input image in a latent space, guided by a high-level semantic objective defined in the multimodal embedding space. Let E be the image encoder, D the image decoder and (C_t, C_i) the multimodal encoders for text and image respectively. Given an input image I_0 and a textual transformation $S \rightarrow T$, we first initialize FlexIT by computing the initial latent image representation as $z_0 = E(I_0)$ and the target multimodal point P as

$$P = C_t(T) + \lambda_I C_i(I_0) - \lambda_S C_t(S). \quad (1)$$

We choose to use a multimodal embedding space since it allows text and image modalities to be combined together in a meaningful way: semantic transformations defined by textual embeddings can be applied to images with linear operations [25]. In this context, our target point P can be seen as an image embedding that has been semantically modified with textual embeddings, by removing the source class information $(-\lambda_S E_t(S))$ and adding the target class information $(+E_t(T))$. Since we don't know what is the optimal linear combination of image and text embeddings, we consider λ_I and λ_S as parameters which will be validated on our development set.

To find an output image which, when encoded in the multimodal embedding space, gets as close as possible to the target point, we optimize the embedding loss:

$$\mathcal{L}_{emb}(z) = \|C_i(D(z)) - P\|_2^2. \quad (2)$$

We add two regularization terms to the embedding loss, to encourage that only the content related to the transformation query is changed. Without regularization, the optimization scheme can alter any part of the image if this helps in getting closer to the multimodal target point, which we have found to yield unnatural artifacts. The distance to the input image I_0 is controlled with a LPIPS distance:

$$\mathcal{L}_{perc}(z) = d_{LPIPS}(D(z), I_0). \quad (3)$$

To enforce staying in parts of the latent space that are well decoded by our image decoder, we use a regularization term with respect to the initial latent code z_0. We use a ℓ_2 norm at each spatial position i of the latent code, and sum these norms across spatial positions to obtain the loss:

$$\mathcal{L}_{latent}(z) = \sum_i \|z^i - z_0^i\|_2. \quad (4)$$

This $\ell_{2,1}$ loss encourages sparse z^i changes, *i.e.* limiting changes in spatial locations, which is aligned with our objective to transform a localized part of the input image.

Finally, note that λ_I in Eq. (1) also acts as a regularization parameter, by encouraging the input and output image to be close in the multi-modal embedding space.

The total loss we optimize can be written as:

$$\mathcal{L}_{total}(z) = \mathcal{L}_{emb}(z) + \lambda_p \mathcal{L}_{perc}(z) + \lambda_z \mathcal{L}_{latent}(z). \quad (5)$$

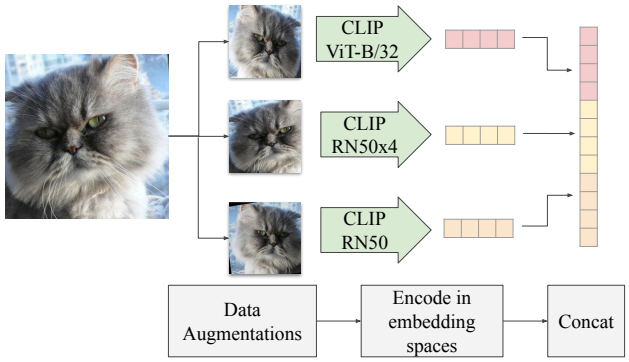

Figure 3. Architecture of our robust CLIP-based image encoder, which combines three different encoders by concatenation.

After initialization, the latent image variable z is updated via gradient descent with a fixed learning rate μ for a fixed number of steps N, while keeping all network weights frozen. Following the implementation of the Fast Gradient Method [12], we normalize the gradient before the update.

Image optimization space. The distance to the multi-modal target point is a differentiable loss that can be optimized via gradient descent. A straightforward approach consists in performing gradient descent directly in the pixel-space. However, this type of image representation lacks a prior on low-level image statistics. By optimizing over a latent variable instead, the image is obtained as the output of a neural-network based decoder. Choosing an autoencoder, like that of VQGAN, lets us (i) make use of the decoder's low-level priors, which guides the optimization problem towards images that exhibit at least low-level consistency; and (ii) encode and decode images in its latent space with little distortion. The spatial dimensions in the VQGAN latent space allows to edit specific parts of the image independently, contrary to GANs which typically rely on more global latent variables. Although GANs generate realistic images with stronger priors, it is problematic to optimize their latent space for two reasons: first, GANs work well on narrow distributions (such as human faces), but do not work as well when trained on a much wider distribution; second, even with a GAN trained on a wide distribution such as that of ImageNet, it is hard to faithfully reconstruct an image using its latent space.

We report on experiments with optimization over raw pixels and GAN latent spaces in Section 4.3.

Implementation details. In FlexIT, we run the optimization loop for $N = 160$ steps, which we found enough to transform most images. We use a resolution of 288 for encoding images with VQGAN, which compresses the images in a latent space with dimensions (256, 18, 18).

We take advantage of various pre-trained CLIP models, and combine their embeddings with concatenation, as shown in Figure 3. By default, we use three image embedding networks with different ResNet and ViT architectures, which implement complementary inductive biases. To encode an image with a single CLIP network, we average the embeddings of multiple augmentations of the input image (8 by default). We have empirically observed that using multiple augmentations per network stabilizes optimization in the early stages.

For the regularization coefficients, we use $\lambda_z = 0.05$, $\lambda_p = 0.15$, $\lambda_S = 0.4$, $\lambda_I = 0.2$ as our default values. These coefficients are set using our ImageNet-based development set, and are fixed for all experiments.

These implementation choices are analysed in Sec. 4.4.

4. Experiments

Below, we first describe our evaluation protocol in detail. We then present qualitative and quantitative results, and an in-depth analysis of various components of our approach.

4.1. Evaluation Protocol

Evaluation dataset.

We did not find a satisfying evaluation framework to study the problem of semantic image translation: existing dataset and metrics focus on narrow image domains, or random text transformation queries [37, 40]. To overcome this, we have decided to build upon the ImageNet dataset [9] for its diversity and its high number of classes: by defining which class labels can be changed into one another (like *cat* $\rightarrow$ *tiger*), we can build a set of sensible object-centric transformation queries. We have selected a subset of the 273 ImageNet labels that we manually split into 47 clusters according to their semantic similarity. For instance, there is a cluster containing all kinds of vegetables. Details on the subset selection and grouping are presented in the appendix. We only consider transformations $S \rightarrow T$ where S and T are in the same cluster, in order to avoid nonsensical transformations between unrelated objects, *e.g.* laptop $\rightarrow$ butterfly.

For each target label T we construct eight transformation queries by randomly sampling eight other classes $\{S_i\}$ within the same cluster, and sample a random image from each S_i from the ImageNet validation set. This gives a total of 2,184 transformation queries that we split into a development set and a test set of equal size. We use the development set to tune various hyper-parameters of our approach, and report evaluation metrics on the test set.

Metrics.

We evaluate the success of the transformation by means of the **Accuracy** of an image classifier, which is possible since we use ImageNet class labels as the transformation targets.

We use a DeiT [46] classifier, which has an ImageNet

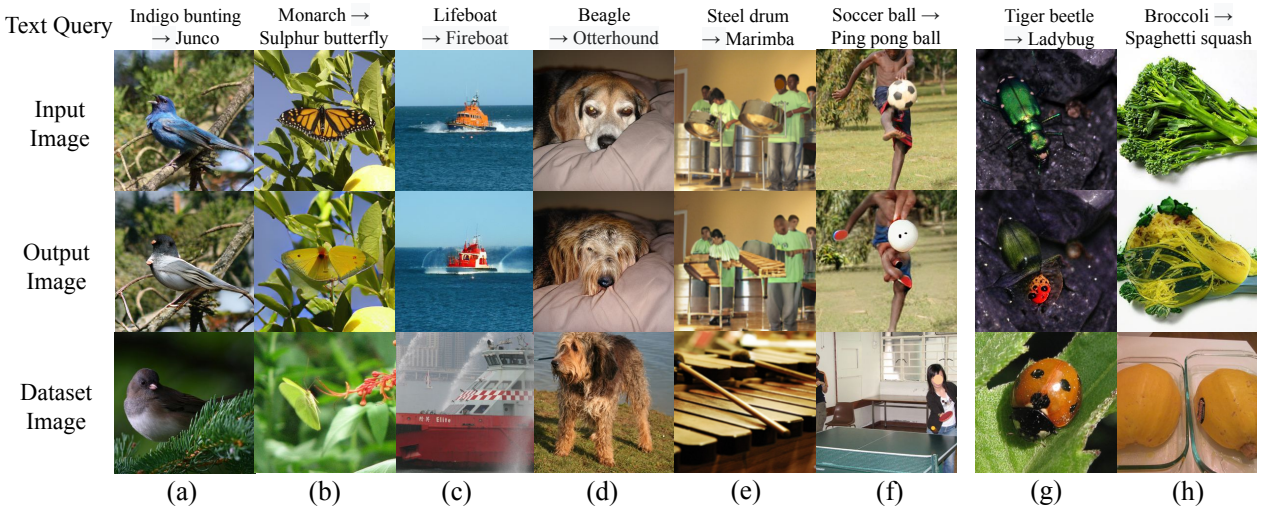

Figure 4. Transformation examples with FlexIT on ImageNet images. Columns (a)-(e) show examples of successful transformations. Column (f) shows an interesting behavior where another object has been added in the image to add more context (a table tennis racket in the hand of the person). The last two columns show the most frequent modes of failure: only part of the input object is transformed (g), or parts of the input object that should be changed are not changed: in column h, the transformed images still has a broccoli shape with green parts instead of an orange and round spaghetti squash).

validation accuracy of 85.2%. We judge a transformation successful if, for the transformed image, class T has the highest probability among the 273 selected classes.

To assess naturalness of transformed images, we use the Fréchet Inception distance (FID) [21]. To avoid numerical instability related to estimating the feature distribution with a small number of samples, we use the "Simplified FID" (**SFID**) [31] which does not take into account the off-diagonal terms in the feature covariance matrix. In addition to the SFID, we use a class-conditional SFID score (**CSFID**) which is an average of the SFID scores computed for each target class separately.[1] Because we compute these scores with a low number of examples for many classes, the CSFID score has a high bias, low variance profile on our dataset [8], and we have found it to be reliable and stable. The CSFID metric is a measure of both image quality and transformation accuracy, as it measures the feature distribution distance between the transformed images and the reference images from the target class in the training set. Editing should not change parts of the image that are irrelevant to the transformation defined in the text, *e.g.* the background. We use the **LPIPS** perceptual distance [56] to measure deviation from the input image. It is a weighted ℓ_2 distance of deep image features, and has been demonstrated to correlate well with human perceptual similarity. During training, we used the LPIPS distance using VGG features rather than AlexNet, so as to reduce bias in the evaluation results. The LPIPS distance cannot differentiate between edits that are relevant to the text query, and those which are not; and we don't know the minimal LPIPS distance between an image and its closest successful transformation. Still, we argue that it should be as low as possible.

More details on the metrics are presented in appendix.

4.2. Results

Qualitative results of FlexIT transformations on ImageNet images are presented in Figure 4, including successful transformations as well as several failure cases. To demonstrate the generality of our approach, we also show examples of color transformations for images from the Stanford Cars dataset [34] in Figure 5.

Semantic image translation is inherently a trade-off between having the most relevant and natural output image (as measured by Accuracy, CSFID and SFID), while staying as close as possible to the input image (as measured by LPIPS). We consider two extreme configurations as baselines, which only optimize one of these two criteria: (i) The COPY baseline, which simply copies the input image without any modification, and (ii) the RETRIEVE baseline that outputs a random validation image labelled with the target class T. We add the ENCODE baseline that simply passes the input image through the VQGAN autoencoder.

We also evaluate StyleCLIP [40], the most relevant text-driven image transformation algorithm from the literature. We consider the version most similar to our method that embeds images with an ImageNet-trained StyleGAN2,[2] and iteratively updates the StyleGAN2 latent representation to maximize the similarity with a given text in the CLIP latent space. We have also trained ManiGAN [37] on ImageNet with the official implementation.

[1] Referred to as within-class FID in [3].

[2] We used the publicly available model from https://github.com/justinpinkney/awesome-pretrained-stylegan2, and train our own e4e encoder [47] to embed images into this latent space.

Figure 5. Example transformations on the *Cars* dataset: input images (first row), FlexIT results (second row), StyleCLIP results based on a StyleGAN2 backbone pre-trained on LSUN Cars. dataset (last row). Although GAN-based images have better details like the wheels, they are farther away from the input images.

	LPIPS ↓	Acc.%↑	CSFID ↓	SFID ↓
COPY	0.0	0.45	106.0	0.20
ENCODE	17.5	1.6	107.5	2.99
RETRIEVE	72.4	90.6	27.2	0.23
ManiGAN [37]	21.7	2.0	123.8	17.0
StyleCLIP [40]	33.4	8.0	146.6	35.8
FlexIT (Ours)	24.7	51.3	57.9	6.8

Table 1. Evaluation of FlexIT and baselines on ImageNet images.

Results are reported in Table 1. As expected, the copy baseline is ideal on LPIPS and SFID, but fails to adapt to the transformation target T, and thus fails on Accuracy and CSFID. The auto-encoding baseline fails on Accuracy and CSFID for the same reason, but demonstrates the non-trivial impact of using the VQGAN latent space on LPIPS and SFID. The RETRIEVE baseline provides ideal metrics for Accuracy, CSFID and SFID, as it returns natural images of the target class. It fails on LPIPS, however, since the output image is unrelated to the input.

Our FlexIT approach combines a low LPIPS (24.7 *vs.* 17.5 for ENCODE) with an accuracy of 51.3% and a CSFID of 57.9, which is closer to the CSFID of RETRIEVE (27.2) than that of ENCODE (107.5). The StyleCLIP scores are poor, with high SFID and CSFID scores which was expected as StyleCLIP has been designed to work well where GANs shine. The StyleGAN2 model we use, trained on ImageNet, is agnostic to class information and cannot synthesize realistic images for all ImageNet classes. ManiGAN works well when trained on narrow domains with color change transformation requests, but we find that it does not produce convincing edits when trained on ImageNet.

To provide insight into which transformations work well, and which less so, we group the ImageNet clusters into 13 bigger groups (see appendix for details) and report the average CSFID and failure rate (1 - accuracy) scores for each group in Figure 6. Generally, transformations among natural objects are more successful than transformations among man-made objects. We believe that this is mostly because the latter appear in a wider variety of shapes and contexts which leads to more difficult transformations.

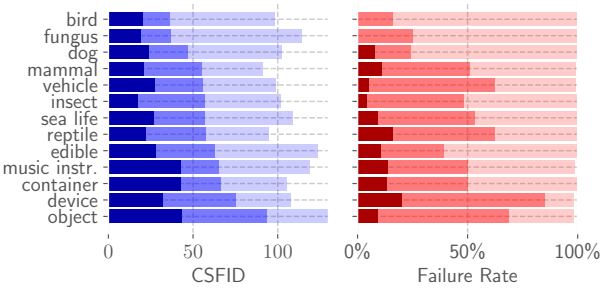

Figure 6. Groupwise CSFID and Failure Rate (1-Accuracy), lower is better for both metrics. Dark colors: best possible values obtained with RETRIEVE baseline; medium colors: scores obtained with FlexIT; light colors: values obtained with COPY baseline.

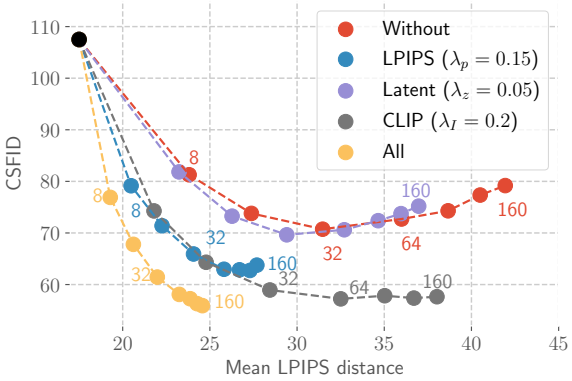

Figure 7. CSFID obtained without regularization, with individual LPIPS, Latent and CLIP regularizers, and using all. Each curve corresponds to 160 steps of optimization on the dev. set.

4.3. Ablation studies

Regularizers. In Figure 7, we show the evolution of CSFID along the optimization steps, where we consider our method without regularization, with each regularization scheme separately, and with all regularizers (default configuration). Compared to not using regularization, the LPIPS regularization substantially improves the CSFID score along the optimization path, while also reducing LPIPS as expected. The CLIP regularizer has a similar effect, but is able to reduce CSFID further while the LPIPS distance is only slightly reduced compared to our method without any regularization. Using all regularizers allows us to obtain the lowest CSFID scores at low LPIPS. We believe that these two regularizers are complementary: while

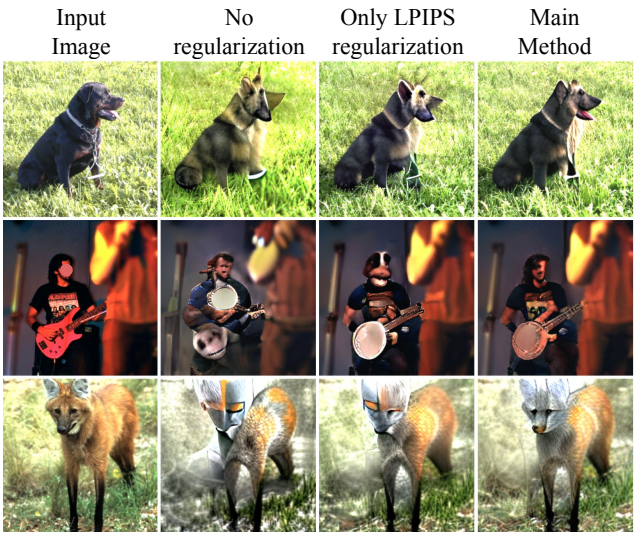

Figure 8. Example transformations with different regularizers. Textual queries from top to bottom: Rottweiler → German shepherd, Electric guitar → Banjo, Red wolf → Grey fox.

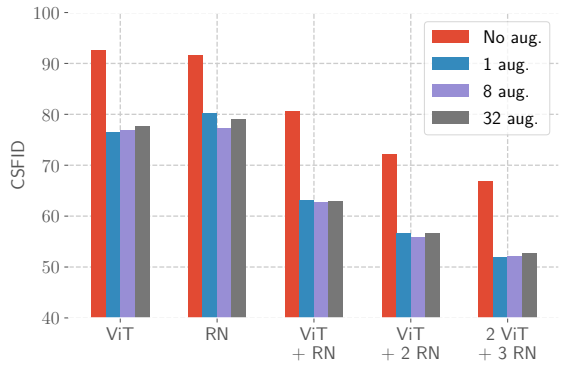

Figure 9. CSFID for different CLIP networks combinations and number of data augmentations options. Default setting: ViT+2RN.

the LPIPS loss mitigates image deviation for local features, the CLIP loss provides semantic guidance which helps to reconstruct recognizable objects. Corresponding qualitative examples are shown in Figure 8.

CLIP embedding module. We study how different choices of CLIP image encoders impact the CSFID score. Our default configuration involves two ResNet-based networks and one ViT-based network to embed the image in the CLIP space. We experiment with a single ViT or ResNet, a combination of ViT with a single ResNet, and also using all available pre-trained CLIP networks, which comprises a ViT-B/16, a ViT-B/32, a ResNet50, ResNet50x4 and ResNet50x16, see [41] for details on the modules. For each CLIP network configuration, we experiment with either not using data augmentation, or using $d \in \{1, 8, 32\}$ augmentations. We apply basic geometric augmentations that are commonly used to train image classification networks (more details in appendix). Each of the N_{nets} CLIP

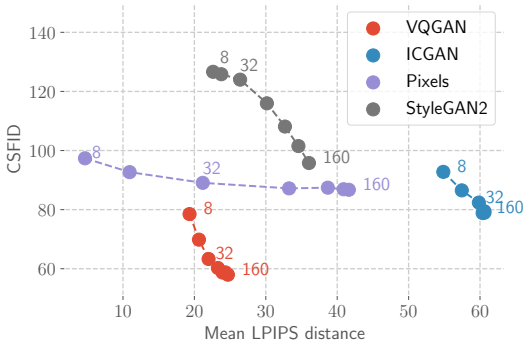

Figure 10. CSFID and LPIPS scores across iterations, using different latent spaces, or raw pixels, for optimization.

networks sees a different augmentation in each of the N_{steps} steps of the optimization process, resulting in a total of $d \times N_{\text{nets}} \times N_{\text{steps}}$ augmentations of the input image.

From the results in Figure 9, we see that while the ViT and ResNet embedding networks lead to similar results, they are complementary and combining them leads to a substantial improvement. Adding additional networks leads to further improvements. Second, using data augmentation is very beneficial, and leads to a reduction in CSFID of 10 or more points for all network configurations. Using more than one augmentation does not improve results substantially: it suffices to a different augmentation for each network at each optimization step. In our other experiments we use the three smallest (and fastest) CLIP networks as our default setting.

Image optimization space. We compare our choice of optimizing in the VQGAN latent space with using the latent spaces of StyleGAN2 [30] and IC-GAN [6], as well as optimizing directly in the pixel space. IC-GAN [6] generates images similar to an input image, and uses a latent variable to allow for variability in its output. As IC-GAN does not offer direct inference of the latents for a given image, we take 1,000 samples from the latent prior, and keep the one yielding minimal LPIPS distance to the input image. We found that optimization to further reduce the LPIPS w.r.t. the input image from this point on was not effective. For StyleGAN2 [30], we use the same network pre-trained on ImageNet as we used for StyleCLIP. To embed the evaluation images into this latent space, we first obtain an initial prediction of the vector with the e4e encoder [47], as in StyleCLIP, and then perform an additional 1,000 optimization steps to better fit the input image, following the GAN inversion procedure described in [29].

The results in Figure 12 show that using the VQGAN latent space allows to substantially decrease the CSFID score along the iterations, while only slightly increasing LPIPS. Using the raw pixel space is not effective to decrease the CSFID. IC-GAN has relatively good image synthesis abilities but it is hard to faithfully encode images in its latent space, yielding high LPIPS scores above 50. The StyleGAN2 latent space ($\mathcal{W}+$) is bigger, allowing generated im-

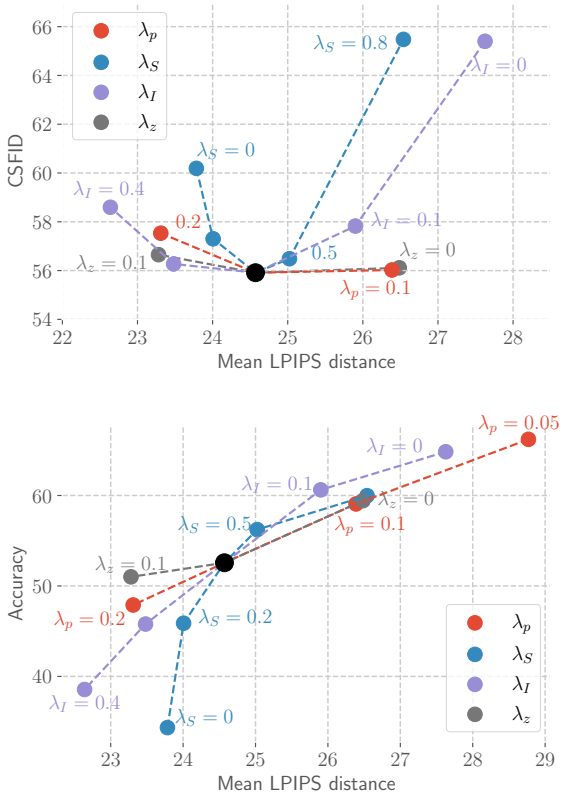

Figure 11. Effect on CSFID and Accuracy of hyper-parameters; default settings represented by the black dot, where all lines cross.

ages to be closer to the input images; however its CSFID scores are not competitive with the other approaches.

4.4. Hyperparameter study

In Figure 11, we illustrate the effect of our hyper-parameters on the LPIPS, CSFID, and Accuracy metrics. For the three regularization parameters $\lambda_p, \lambda_z, \lambda_I$, we observe that (i) the LPIPS distance with respect to the input image is smaller as the regularization gets stronger, as expected; (ii) less regularization allows more image modifications, yielding better accuracy scores, as illustrated in the bottom panel; (iii) there is a global minimum in CSFID scores when we make each hyperparameter vary independently (top panel). Regularization constraints are indeed useful to prevent inserting unnatural visual artifacts; however, too much regularization penalizes our algorithm as the distribution of output images gets closer to the input distribution, and thereby farther from the target distribution.

The parameter λ_S, similarly to the regularization parameters, has a an optimal value which minimizes the CSFID. It is beneficial to give a hint to the optimization algorithm which semantic content should be changed, however focusing too much this objective reduces image realism.

For our main experiments, we set our hyper-parameters to minimize the CSFID score on the development set. This is a natural choice given the convex shape of the CSFID scores, whereas optimizing for accuracy would remove the regularizers which is detrimental for image quality.

5. Conclusion

Contributions. We propose FlexIT, a novel method for semantic image translation. By relying on an autoencoder latent space, rather than specialized GAN latent spaces, it can operate on a wide range of images. Using a general pre-trained multi-modal embedding space provides flexibility, giving FlexIT the ability to process free-text transformation queries without training. We also propose an evaluation protocol for semantic image translation, based on ImageNet, which we use to thoroughly evaluate our approach and its components.

Limitations. Our method works best for semantic translation when the input image provides guidance, but has difficulties synthesizing realistic novel objects from scratch. Also, while we studied transformations that change the class or color of the main object in a scene, other transformations of interest could consider changing the action of a subject (person walking *vs*. running), changing object attributes, adding or deleting objects, or consider more elaborate textual descriptions which require non-trivial grounding in the image ("change the color of car parked next to the bicycle."). However, progress in this direction will require to identify the right data and evaluation metrics.

Broader impacts. As our algorithm relies on CLIP for editing, it could potentially inherit biases embedded in the CLIP model. The authors of CLIP have demonstrated that similarly to other neural network models, CLIP is subject to fairness issues such misclassifying human faces into non-human or crime-related categories, and producing gender biased associations where some labels that describe high status occupations are disproportionately more attached to images of men than that of women. Our editing method could reflect such biases if prompted transformations such as doctor $\rightarrow$ newscaster, although we have not observed experimental evidence of this. A potential bias mitigation strategy would be to add constraints with CLIP prompts, for instance by enforcing that the probability of the labels *man* and *woman* remain the same before and after editing.

Our model provides new capabilities to an expanding set of image editing and synthesis tools based on deep generative models. As any generative image model, synthetic images generated by our method can potentially be used in unintended ways with undesirable effects. We believe however that open publication of research in this area contributes to a good understanding of such techniques, and can aid the community in efforts to develop method that detect unauthentic content.

References

[1] Rameen Abdal, Peihao Zhu, Niloy Mitra, and Peter Wonka. Styleflow: Attribute-conditioned exploration of stylegan-generated images using conditional continuous normalizing flows. *ACM Trans. Graph.*, 2021. 1, 2

[2] Oriol Vinyals Ali Razavi, Aaron van den Oord. Generating diverse high-fidelity images with VQ-VAE-2. In *NeurIPS*, 2019. 2

[3] Yaniv Benny, Tomer Galanti, Sagie Benaim, and Lior Wolf. Evaluation metrics for conditional image generation. *IJCV*, 129:1712–1731, 2021. 5

[4] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *ICLR*, 2019. 1, 11

[5] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *NeurIPS*, 2020. 11

[6] Arantxa Casanova, Marlène Careil, Jakob Verbeek, Michal Drozdzal, and Adriana Romero-Soriano. Instance-conditioned GAN. In *NeurIPS*, 2021. 2, 7, 11

[7] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. StarGAN v2: Diverse image synthesis for multiple domains. In *CVPR*, 2020. 2

[8] Min Jin Chong and David Forsyth. Effectively unbiased fid and inception score and where to find them. In *CVPR*, 2020. 5

[9] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 4

[10] Helisa Dhamo, Azade Farshad, Iro Laina, Nassir Navab, Gregory D Hager, Federico Tombari, and Christian Rupprecht. Semantic image manipulation using scene graphs. In *CVPR*, 2020. 2

[11] J. Donahue, P. Krähenbühl, and T. Darrell. Adversarial feature learning. In *ICLR*, 2017. 2

[12] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *CVPR*, 2018. 4

[13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2020. 2

[14] V. Dumoulin, I. Belghazi, B. Poole, A. Lamb, M. Arjovsky, O. Mastropietro, and A. Courville. Adversarially learned inference. In *ICLR*, 2017. 2

[15] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *CVPR*, 2021. 2, 3

[16] Kevin Frans, L. B. Soros, and Olaf Witkowski. CLIPDraw: exploring text-to-drawing synthesis through language-image encoders. *arXiv preprint*, arXiv:2106.14843, 2021. 2

[17] O. Gafni and L. Wolf. Wish you were here: Context-aware human generation. In *CVPR*, 2020. 2

[18] Lore Goetschalckx, Alex Andonian, Aude Oliva, and Phillip Isola. GANalyze: Toward visual definitions of cognitive image properties. *arXiv preprint*, arXiv:1906.10112, 2019. 2

[19] Asya Grechka, Matthieu Cord, and Jean-Francois Goudou. MAGECally invert images for realistic editing. In *BMVC*, 2021. 1

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 2

[21] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *NeurIPS*, 2017. 5

[22] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *NeurIPS*, 2017. 11

[23] Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. GANSpace: Discovering interpretable GAN controls. In *NeurIPS*, 2020. 1, 2

[24] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017. 2

[25] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. *arXiv preprint*, arXiv:2102.05918, 2021. 3

[26] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, 2016. 2

[27] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. In *NeurIPS*, 2020. 1

[28] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. In *NeurIPS*, 2021. 1, 2

[29] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019. 1, 2, 7

[30] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila. Analyzing and improving the image quality of StyleGAN. In *CVPR*, 2020. 1, 2, 7, 11

[31] Chung-Il Kim, Meejoung Kim, Seungwon Jung, and Eenjun Hwang. Simplified Fréchet distance for generative adversarial nets. *Sensors*, 20(6):1548, 2020. 5, 11

[32] Gwanghyun Kim and Jong Chul Ye. DiffusionCLIP: Text-guided image manipulation using diffusion models. *arXiv preprint*, arXiv:2110.02711, 2021. 2

[33] D. Kingma and M. Welling. Auto-encoding variational Bayes. In *ICLR*, 2014. 2

[34] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *4th International IEEE Workshop on 3D Representation and Recognition (3dRR-13)*, Sydney, Australia, 2013. 5

[35] A. Krizhevsky, I. Sutskever, and G. Hinton. ImageNet classification with deep convolutional neural networks. In *NeurIPS*, 2012. 11

[36] Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. Photo-realistic single image super-resolution using a generative adversarial network. In *CVPR*, 2017. 2

[37] Bowen Li, Xiaojuan Qi, Thomas Lukasiewicz, and Philip H. S. Torr. ManiGAN: Text-guided image manipulation. In *CVPR*, 2020. 2, 4, 5, 6

[38] Huan Ling, Karsten Kreis, Daiqing Li, Seung Wook Kim, Antonio Torralba, and Sanja Fidler. EditGAN: High-precision semantic image editing. *arXiv preprint*, arXiv:2111.03186, 2021. 2

[39] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *CVPR*, 2019. 2

[40] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of StyleGAN imagery. In *ICCV*, 2021. 2, 4, 5, 6

[41] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *arXiv preprint*, arXiv:2103.00020, 2021. 2, 3, 7

[42] Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the latent space of GANs for semantic face editing. In *CVPR*, 2020. 1, 2

[43] Jing Shi, Ning Xu, Trung Bui, Franck Dernoncourt, Zheng Wen, and Chenliang Xu. A benchmark and baseline for language-driven image editing. In *ACCV*, 2020. 2

[44] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015. 3

[45] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, 2016. 11

[46] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Herve Jegou. Training data-efficient image transformers & distillation through attention. In *ICML*, 2021. 4, 11

[47] Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or. Designing an encoder for StyleGAN image manipulation. *ACM Transactions on Graphics (TOG)*, 40(4):1–14, 2021. 1, 5, 7, 11, 15

[48] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint*, arXiv:1807.03748, 2018. 2

[49] A. van den Oord, O. Vinyals, and K. Kavukcuoglu. Neural discrete representation learning. In *NeurIPS*, 2017. 2

[50] Andrey Voynov and Artem Babenko. Unsupervised discovery of interpretable directions in the GAN latent space. In *ICML*, 2020. 2

[51] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional GANs. In *CVPR*, 2018. 2

[52] Yaxing Wang, Luis Herranz, and Joost van de Weijer. Mix and match networks: Cross-modal alignment for zero-pair image-to-image translation. *IJCV*, 128(12):2849–2872, 2020. 2

[53] Jiahui Yu, Xin Li, Jing Yu Koh, Han Zhang, Ruoming Pang, James Qin, Alexander Ku, Yuanzhong Xu, and Yonghui Wu Jason Baldridge. Vector-quantized image modeling with improved VQGAN. *arXiv preprint*, arXiv:2110.04627, 2021. 2

[54] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S. Huang. Generative image inpainting with contextual attention. In *CVPR*, 2018. 2

[55] Xiaoming Yu, Yuanqi Chen, Thomas Li, Shan Liu, and Ge Li. Multi-mapping image-to-image translation via learning disentanglement. *NeurIPS*, 2019. 2

[56] R. Zhang, P. Isola, A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 3, 5, 11

[57] Jiapeng Zhu, Yujun Shen, Deli Zhao, and Bolei Zhou. In-domain GAN inversion for real image editing. In *ECCV*, 2020. 1

[58] J.-Y. Zhu, T. Park, P. Isola, and A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, 2017. 2

[59] Peiye Zhuang, Oluwasanmi O Koyejo, and Alex Schwing. Enjoy your editing: Controllable GANs for image editing via latent space navigation. In *ICLR*, 2021. 1, 2

A. Appendix

A.1. Assets

We provide a list of the assets used in our work (datasets, code, and models) in Table 2 (links) and Table 3 (licences).

A.2. Datasets

To design transformation queries from ImageNet classes, we have grouped classes into clusters by semantic similarity, upon manual inspection of the WordNet hierarchy of classes. The resulting clusters are shown in Table 8. This process resulted in 273 classes gathered in 47 clusters. We have not included all ImageNet classes because (i) we wanted to reduce the large number of dog breed classes, and (ii) a lot of classes were "standalone classes" with no natural target for transformation among the other classes. The clusters are then grouped into 13 bigger "groups" that are used solely for visualization in Figure 6 of the main paper.

A.3. Evaluation metrics

LPIPS. As recommanded by the authors of [56], we use the AlexNet [35] backbone to compute LPIPS distance, when we use it as an evaluation metric. To avoid using the same metric for optimization, we compute the LPIPS in the perceptual regularization term $\mathcal{L}_{perc}$, see Eq. (3) of the main paper, using the VGG16 network [35].

The LPIPS distance is computed at an image resolution of 256, for both evaluation and optimization. In the main paper, all LPIPS scores have been multiplied by 100 for readibility.

(C)SFID. The FID metric [22] measures the distance between the distributions of the real images and generated images in the feature space of an InceptionV3 classifier [45].

More formally, let μ^r and σ^r be the mean and standard deviation of inception features for the real images, and μ^s and σ^s for the synthetic images. The Simplified FID [31] is computed as

$$SFID(\alpha) = \|\mu^r - \mu^s\|^2 + \alpha\|\sigma^r - \sigma^s\|^2. \quad (6)$$

It does not take into account the off-diagonal terms in the feature covariance matrix to avoid numerical instability.

The Conditional Simplified FID (CSFID) is computed in the same manner but for each target class separately, and then averaging the resulting scores: With μ_c^r and σ_c^r the mean and standard deviation of inception features for the real images belonging to class c, and μ_c^s and σ_c^s for the synthetic images, we have

$$CSFID(\alpha) = \frac{1}{|C|}\sum_c \|\mu_c^r - \mu_c^s\|^2 + \frac{\alpha}{|C|}\sum_c \|\sigma_c^r - \sigma_c^s\|^2. \quad (7)$$

We have noticed that the distance on standard deviations was not very discriminative: since we are modifying images and not generating images from scratch, we already have a lot of diversity in the generated images. Experimentally, using $\alpha > 0$ mostly consisted in adding a bias term in this metric, therefore we chose to use $\alpha = 0$ in the (C)SFID scores.

Since the images we transform are extracted from the ImageNet validation set, we use the ImageNet training set as our reference distribution to compute the (C)SFID scores. As for LPIPS, the (C)SFID scores are computed at an image resolution of 256.

Accuracy. We use a DeiT classifier [46] trained on ImageNet, which takes images of size 384×384. Smaller images are upsampled before being passed to the classifier.

A.4. Details on the multimodal encoder

For data augmentations, we use a random horizontal flipping and a random rotation between -10 and 10 degrees, followed by cropping the image (keeping at least 80% of the input image) with aspect ratio between 0.9 and 1.1. For the CLIP-based multimodal encoders, we have considered all CLIP networks freely available, listed in Table 4.

Backbone	Params.	Latent dim.
RN50	38M	512
RN50x4	87M	640
ViT-B/32	88M	512
ViT-B/16	86M	512
RN50x16	167M	768

Table 4. Visual backbones used for the multimodal encoder. Our default configuration only includes the ViT-B/32, the RN50 and the RN50x4.

A.5. Additional qualitative results

In Figure 12, we show qualitative results when we replace the VQGAN image encoder with other GAN-based encoders. VQGAN has a native encoder and decoder, and thus the initial latent vector is obtained directly. For StyleGAN2 [30], we use the e4e encoder [47] followed by an additional 1,000 steps of LPIPS minimization. For the IC-GAN [6] model, we use the BigGAN [4] backbone as generator. IC-GAN is naturally conditioned on the SwaV embedding [5] of the input image; for added robustness we sample 1,000 latent points and choose the one yielding smallest LPIPS distance with respect to the input image.

Figure 13 shows intermediate transformation results with FlexIT for 0, 8, 16, 32 and 160 optimization steps. The result after zero optimization steps shows the effect of autoencoding the input image, without changing the latent representation. Figure 14 show representative failure cases for our method, due to either the regularization method or the multimodal embedding space. Finally, in Figure **??** we

Asset Name	Link
ImageNet	https://www.image-net.org
Cars	https://ai.stanford.edu/ jkrause/cars/car_dataset.html
LPIPS	https://github.com/richzhang/PerceptualSimilarity
FID	https://github.com/mseitzer/pytorch-fid
DeiT	https://github.com/facebookresearch/deit
CLIP	https://github.com/openai/CLIP
VQGAN	https://github.com/CompVis/taming-transformers
IC-GAN	https://github.com/facebookresearch/ic_gan
StyleGAN2	https://github.com/justinpinkney/awesome-pretrained-stylegan2
e4e	https://github.com/omertov/encoder4editing

Table 2. List of asset links.

Asset Name	Asset type	License
ImageNet	Images	https://www.image-net.org/download.php
Cars	Images	https://ai.stanford.edu/ jkrause/cars/car_dataset.html
LPIPS	Code and Models	BSD-2-Clause License
FID	Code and Models	Apache-2.0 License
DeiT	Code and Models	Apache License 2.0
CLIP	Code and Models	MIT License
VQGAN	Code and Models	MIT License
IC-GAN	Code and Models	Attribution-NonCommercial 4.0 International
StyleGAN2	Code and Models	https://github.com/justinpinkney/awesome-pretrained-stylegan2
e4e	Code	MIT License

Table 3. List of asset licenses.

show additional color transformation results on the Cars30k dataset.

A.6. Ablation results

In Table 5, we show ablation experiments for all FlexIT parameters; this includes the CSFID scores of the hyperparameter configurations reported in Figure 11 of the main paper.

In Table 6, we show ablations for combining multiple CLIP networks and using multiple data augmentations in the multimodal encoder. This includes the CSFID scores reported in Figure 9 of the main paper; we also report the runtime needed for each algorithm.

A.7. Results on ManiGAN evaluation setup

Evaluating text-driven image editing requires (1) a list of sensible transformation queries, and (2) a method for evaluating the quality and accuracy of the generated result. The evaluation protocol in ManiGAN consists in (1) choosing *random* COCO captions/image pairs and thus leading to noisy transformations and (2) calculating the image-text similarity score which was used as a loss term during their training, leading to bias in the final scores. In the main paper, we compare different methods using our novel evaluation protocol, which was carefully designed to avoid these pitfalls. Nonetheless, we show in Tab. 7 that even with the ManiGAN protocol, FlexIT improves upon the ManiGAN scores by a large margin.

	IS↑	SIM↑	DIFF↓	MP↑
ManiGAN	14.96	0.087	0.216	0.068
FlexIT	**18.19**	**0.177**	**0.146**	**0.151**

Table 7. ManiGAN evaluation on random edits from COCO.

	Acc.↑	LPIPS↓	CSFID↓	SFID↓
$\lambda_I = 0$	**64.8**	27.6	65.4	12.3
$\lambda_I = 0.1$	60.6	25.9	57.8	8.3
$\lambda_I = 0.2$	52.6	24.6	**55.9**	6.4
$\lambda_I = 0.3$	45.8	23.5	56.3	5.5
$\lambda_I = 0.4$	38.6	**22.6**	58.6	**5.0**
$\lambda_S = 0.0$	34.3	**23.8**	60.2	**4.8**
$\lambda_S = 0.2$	45.9	24.0	57.3	5.5
$\lambda_S = 0.4$	52.6	24.6	**55.9**	6.4
$\lambda_S = 0.5$	56.2	25.0	56.5	7.1
$\lambda_S = 0.8$	**60.0**	26.5	65.5	11.7
$\lambda_z = 0.0$	**59.4**	26.5	56.1	7.1
$\lambda_z = 0.05$	52.6	24.6	**55.9**	6.4
$\lambda_z = 0.1$	51.0	**23.3**	56.7	**6.3**
$\lambda_p = 0.05$	**66.2**	28.8	**56.0**	7.9
$\lambda_p = 0.1$	59.1	26.4	**56.0**	7.2
$\lambda_p = 0.15$	52.6	24.6	**55.9**	6.4
$\lambda_p = 0.2$	47.9	**23.3**	57.5	**6.3**
ℓ_1	**54.2**	**24.6**	56.3	6.5
ℓ_2	52.4	**24.5**	**55.9**	6.8
$\ell_{2,1}$	52.6	**24.6**	**55.9**	**6.4**
$lr = 0.025$	47.6	**22.5**	58.3	**6.0**
$lr = 0.5$	52.6	24.6	55.9	6.4
$lr = 0.1$	**60.4**	27.6	**54.8**	7.2
resolution 256	53.8	24.8	56.8	7.2
resolution 288	**52.6**	24.6	**55.9**	**6.4**
resolution 320	54.3	**24.0**	57.4	7.3

Table 5. FlexIT ablation results. lr is the learning rate. Lines corresponding to our default configuration are marked in light grey. The norms ℓ_1, ℓ_2, and $\ell_{2,1}$ refer to the distance used for regularization in the VQGAN latent space. Best values for each metric are shown in bold inside each group of parameter values.

networks	d	Acc.↑	LPIPS↓	CSFID↓	SFID↓	sec. /im
ViT-B/32	0	9.4	21.8	92.7	7.4	27s
ViT-B/32	1	37.5	26.4	76.5	11.1	27s
ViT-B/32	8	35.1	25.4	76.9	10.7	33s
ViT-B/32	32	35.5	25.0	77.7	10.8	53s
RN50x4	0	13.4	23.8	91.6	11.8	35s
RN50x4	1	32.5	27.4	80.2	13.7	35s
RN50x4	8	31.0	25.2	77.3	12.3	53s
RN50x4	32	27.0	24.2	79.1	11.7	122s
2 nets	0	23.0	22.8	80.7	9.5	39s
2 nets	1	50.6	26.4	63.2	8.9	39s
2 nets	8	47.8	24.9	62.7	8.4	64s
2 nets	32	47.4	24.2	62.9	8.1	160s
3 nets	0	30.4	22.5	72.2	8.3	45s
3 nets	1	54.9	26.0	56.7	6.7	45s
3 nets	8	52.6	24.6	55.9	6.4	75s
3 nets	32	51.7	24.0	56.7	6.7	190s
5 nets	0	39.6	22.4	66.8	7.7	70s
5 nets	1	60.3	25.5	51.9	5.5	70s
5 nets	8	60.1	23.9	52.1	5.4	176s
5 nets	32	52.0	22.8	52.7	5.2	560s

Table 6. Ablation results for the multimodal encoder components. d is the number of augmentations. $d = 0$ means that the encoder takes the unchanged image as input; For $d = 1$, the encoder takes only one (augmented image), which explains why the edit time is the same as $d = 0$. When considering n CLIP networks, we take the first n elements in the following list: RN50x4, ViT-B/32, RN50, ViT-B/16, RN50x16. Our default configuration is marked in light grey. Last column gives computation time per image in seconds.

Group	Cluster	Classes
bird	bird of prey	bald eagle, kite, great grey owl
bird	finch	indigo bunting, goldfinch, house finch, junco
bird	grouse	black grouse, prairie chicken, ptarmigan, ruffed grouse
bird	seabird	king penguin, albatross, pelican, European gallinule, black swan
bird	wading bird	goose, oystercatcher, little blue heron, black stork, bustard, flamingo, spoonbill
container	bag	backpack, plastic bag, purse
container	food container	water jug, beer bottle, water bottle, wine bottle, coffee mug, vase, coffeepot, teapot, measuring cup, cocktail shaker
device	electronics	cassette player, cellular telephone, computer keyboard, desktop computer, dial telephone, hard disc, iPod, laptop
device	measuring	analog clock, digital clock, wall clock, stopwatch, digital watch, odometer, barometer
dog	hound	English foxhound, Italian greyhound, Afghan hound, basset, beagle, otterhound
dog	sporting dog	English springer, cocker spaniel, golden retriever, Irish setter
dog	terrier	American Staffordshire terrier, wire-haired fox terrier, standard schnauzer, Border terrier, Irish terrier, Yorkshire terrier
dog	toy dog	papillon, Chihuahua, Japanese spaniel, Shih-Tzu, toy terrier
dog	working dog	collie, German shepherd, Rottweiler, miniature pinscher, French bulldog, Siberian husky, boxer, Eskimo dog
edible	edible fruit	Granny Smith, strawberry, lemon, orange, banana, custard apple, fig, pineapple, pomegranate
edible	sandwich	cheeseburger, hotdog, bagel
edible	vegetable	bell pepper, broccoli, cauliflower, spaghetti squash, zucchini, butternut squash, artichoke, cardoon, cucumber
fungus	fungus	bolete, coral fungus, earthstar, gyromitra, hen-of-the-woods, stinkhorn
insect	beetle	ground beetle, ladybug, leaf beetle, long-horned beetle, tiger beetle, weevil
insect	butterfly	monarch, admiral, cabbage butterfly, lycaenid, ringlet, sulphur butterfly
insect	spider	black widow, garden spider, tarantula, wolf spider, scorpion
mammal	bear	American black bear, brown bear, ice bear, sloth bear, giant panda, lesser panda
mammal	bovid	ox, ibex, bighorn, gazelle, impala, water buffalo, ram, bison
mammal	canine	Arctic fox, grey fox, red fox, African hunting dog, dingo, coyote, red wolf, timber wolf, white wolf, hyena
mammal	equine	sorrel, zebra
mammal	feline	Persian cat, tabby, cheetah, jaguar, leopard, lion, snow leopard, tiger
mammal	great ape	chimpanzee, gorilla, orangutan
mammal	monkey	capuchin, spider monkey, squirrel monkey, baboon, guenon, macaque
music. instr.	percussion	chime, drum, gong, maraca, marimba, steel drum
music. instr.	stringed	cello, violin, acoustic guitar, electric guitar, banjo
music. instr.	wind	bassoon, oboe, sax, flute, cornet, French horn, trombone
object	ball	golf ball, ping-pong ball, rugby ball, soccer ball, tennis ball
object	handtool	hammer, plane, plunger, screwdriver, shovel
object	headdress	bathing cap, shower cap, bonnet, cowboy hat, sombrero, football helmet
reptile	amphibian	bullfrog, tree frog, axolotl, spotted salamander, common newt, eft, European fire salamander
reptile	snake	rock python, boa constrictor, green mamba, Indian cobra, diamondback, sidewinder, horned viper, king snake, green snake, thunder snake
reptile	turtle	box turtle, mud turtle, terrapin
sea life	aqu. mammal	killer whale, grey whale, sea lion, dugong
sea life	bony fish	goldfish, tench, eel, anemone fish, lionfish, gar, sturgeon
sea-life	crab	American lobster, Dungeness crab, fiddler crab, king crab, rock crab, crayfish, hermit crab, isopod
sea life	shark	great white shark, tiger shark, hammerhead
vehicle	bicycle	motor scooter, tricycle, unicycle, mountain bike, moped
vehicle	boat	speedboat, lifeboat, canoe, fireboat, gondola
vehicle	car	ambulance, beach wagon, cab, convertible, jeep, limousine, minivan, sports car
vehicle	locomotive	electric locomotive, steam locomotive
vehicle	sailing vessel	catamaran, trimaran, schooner
vehicle	truck	minivan, police van, fire engine, garbage truck, pickup, tow truck, trailer truck, school bus

Table 8. Groups and clusters of the ImageNet classes used to define the transformation queries.

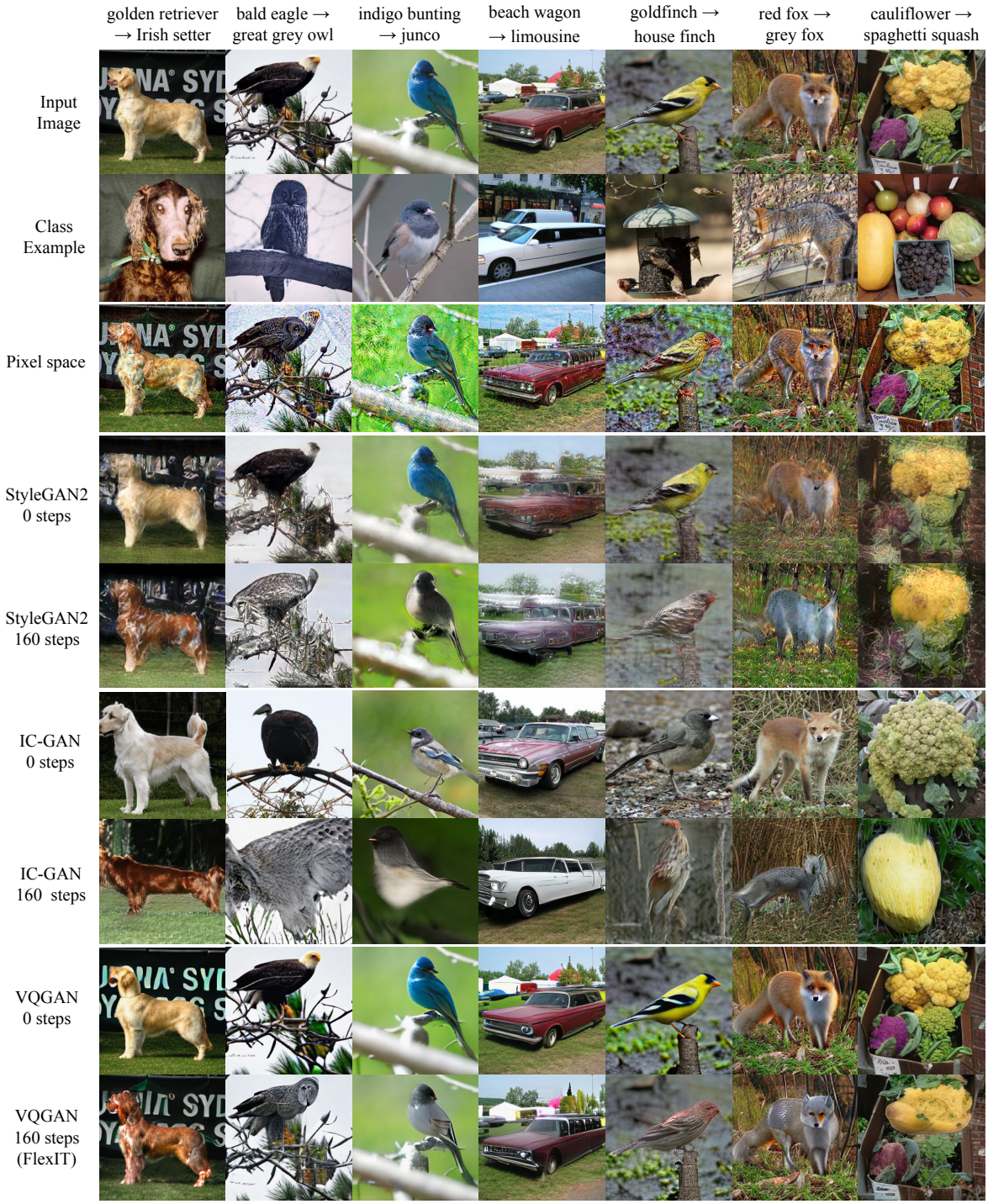

Figure 12. Transformation examples with various backbones for the image latent space. For each latent space, we show the initial image decoded from the initial point z_0, and the resulting image after 160 optimization steps. The three latent spaces differ substantially in their encoding images (0 steps). The IC-GAN latent space provides natural images that are far away from the input image due to the limited generator capacity in conjunction with the smaller latent space size (2560 dim.). StyleGAN2 images preserve the input image appearance thanks to the larger size of its latent space $\mathcal{W}+$ (8192), however images contain many unnatural artifacts due to the challenges of embedding images in this latent space [47]. The VQGAN latent space leads to the best reconstruction results. After 160 steps of optimization, the images generated with StyleGAN2 still have the same unnatural artifacts, and images generated with IC-GAN remain natural but far from the input images. VQGAN, which we use in FlexIT, achieves good edits while preserving the overall image appearance. The pixel-space method introduces high-frequency artifacts, without substantially modifying the high-level semantic image content, resembling adversarial examples for image classification.

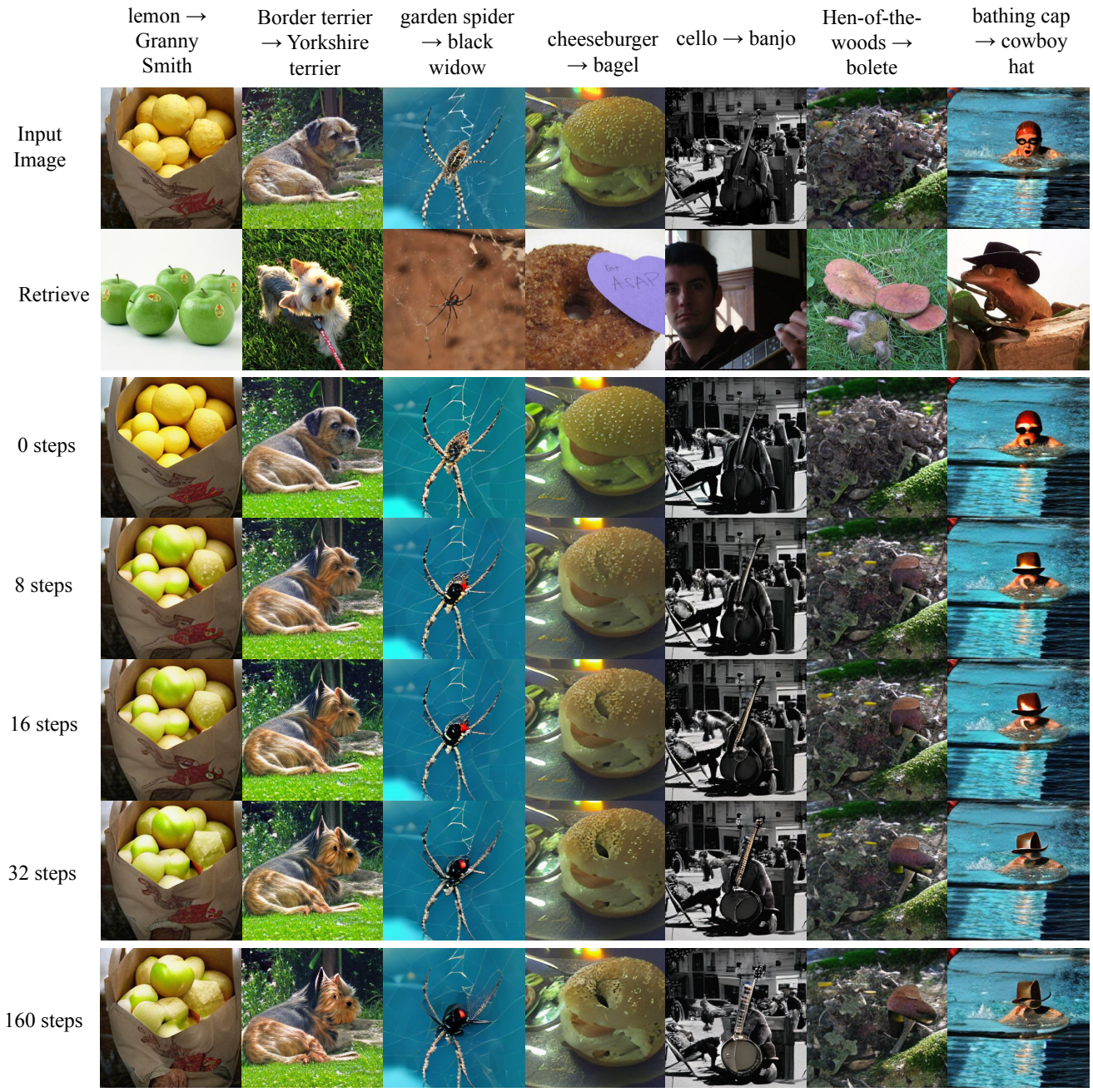

Figure 13. Intermediate transformation results obtained with FlexIT. Note that most edits only require 32 steps to be completed; some edits benefit from longer optimization schemes, such as the spider and the banjo.

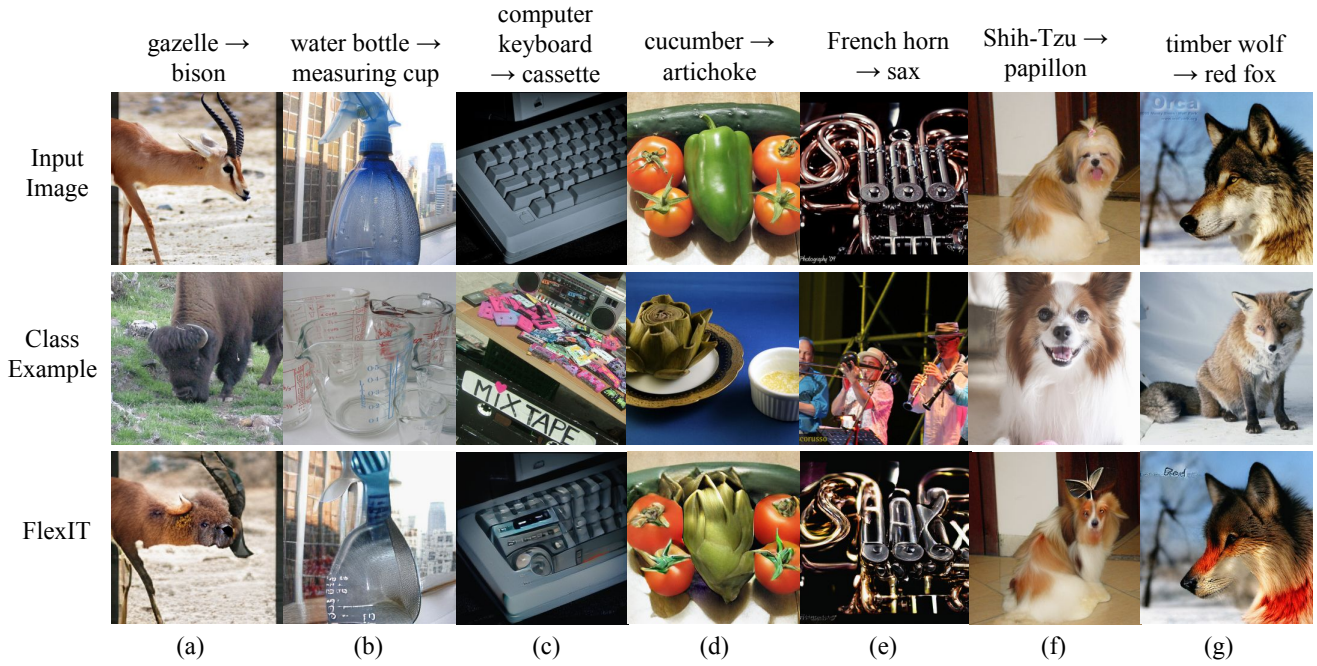

Figure 14. Representative failure cases of FlexIT. The first three columns show examples where the regularization with respect to the initial image was too strong. (a): FlexIT added bison-like texture but fails to change the shape convincingly. (b): markings have been added to the bottle, but without changing its shape to that of a measuring cup. (c): only a part of the input object was changed. (d): the bell pepper rather than the cucumber was transformed, probably because the former is more centered, and has a better initial shape. Columns (e)–(g) show failure cases related to the CLIP embedding space. (e): we observe an interesting text synthesis behaviour where the letters of the target class "sax" have been written in the image. This is related to the OCR capabilities of CLIP. (f): a butterfly is synthesized on the head of the dog (CLIP optimized for both the dog breed papillon and the insect papillon). (g): an unrealistic image is produced by adding saturated red to the image.