Rethinking the Truly Unsupervised Image-to-Image Translation

Kyungjune Baek[1]* Yunjey Choi[2] Youngjung Uh[2] Jaejun Yoo[3] Hyunjung Shim[1]

[1]Yonsei University [2]Clova AI Research, NAVER Corp. [3]EPFL

Figure 1. Our model is able to conduct image-to-image translation without any supervision. The output images are generated with the source image and the the average style code of each estimated domain. The breed of the output cat changes according to the domain while preserving the pose of the source image.

Abstract

Every recent image-to-image translation model uses either image-level (i.e. input-output pairs) or set-level (i.e. domain labels) supervision at minimum. However, even the set-level supervision can be a serious bottleneck for data collection in practice. In this paper, we tackle image-to-image translation in a fully unsupervised setting, i.e., neither paired images nor domain labels. To this end, we propose the truly unsupervised image-to-image translation method (TUNIT) that simultaneously learns to separate image domains via an information-theoretic approach and generate corresponding images using the estimated domain labels. Experimental results on various datasets show that the proposed method successfully separates domains and translates images across those domains. In addition, our model outperforms existing set-level supervised methods under a semi-supervised setting, where a subset of domain labels is provided. The source code is available at https://github.com/clovaai/tunit.

* Work done during his internship at Clova AI Research.

1. Introduction

Given an image of one domain, image-to-image translation is a task to generate the plausible images of the other domains. Based on the success of conditional generative models [33, 41], many image translation methods have been proposed either by using image-level supervision (*e.g.* paired data) [17, 43, 35] or by using set-level supervision (*e.g.* domain labels) [48, 28, 25, 29]. Though the latter approach is generally called 'unsupervised' as a counterpart of the former, it actually assumes that the domain labels are given *a priori*, which can be a serious bottleneck in practical applications as the number of domains and samples becomes large. For example, labeling individual samples of a large dataset, such as FFHQ [20], is very costly, and the distinction across domains can often be ambiguous.

In this work, we first clarify that unsupervised image-to-image translation should strictly denote the task *without any supervision* neither paired images nor domain labels. Under this definition, our goal is to develop an unsupervised translation model given a mixed set of images of many domains (Figure 2c). We tackle this problem by formulating three sub-problems: 1) distinguishing the set-level characteristics

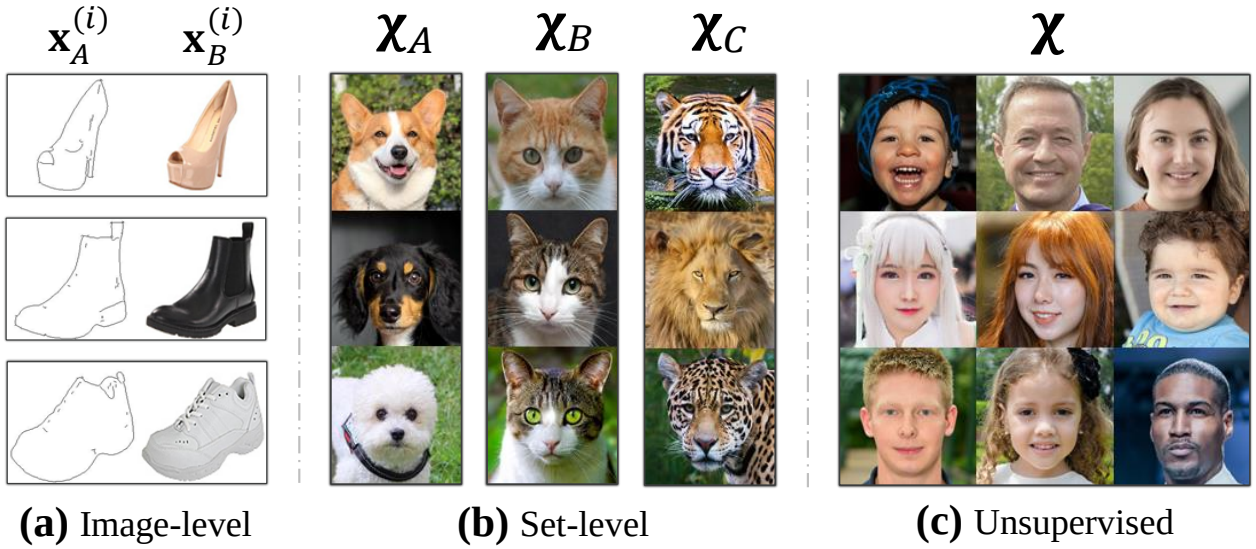


Figure 2. **Levels of supervision.** The previous methods conduct image-to-image translation relying on either **(a)** image-level or **(b)** set-level supervision. Our proposed method can perform the task using **(c)** a dataset without any supervision.

of images (*i.e*. domains), 2) encoding the individual content and style of an input image, and 3) learning a mapping function among the estimated domains.

To this end, we introduce a `guiding network` that has two branches of providing pseudo domain labels and encoding style features, which are later used in the discriminator and the generator, respectively. More specifically, for estimating domain labels, we employ an unsupervised approach [18] that maximizes the mutual information (MI) between the domain assignments of one image and its augmented version (*e.g*. random cropping, horizontal flip). This helps the guiding network to group similar images together while evenly separating their categories. For embedding style codes, we employ a contrastive loss [9, 10], which leads the model to further understand the dissimilarity between images, resulting in better representation learning. By participating in the image translation process, the guiding network can also exploit gradients from the generator and the discriminator. The guiding network now understands the recipes of domain-separating attributes because the generator wants the style code to contain enough information to fool the domain-specific discriminator, and vice versa. Thanks to this interaction between the guiding network and the adversarial networks, our model successfully separate domains and translate images.

Our experimental analysis results show that, by exploiting the synergy between two tasks, the guiding network helps the image translation model to largely improve the generation performance, and vice versa. We quantitatively and qualitatively compare our model with the existing set-level supervised methods under a semi-supervised setting, where only a subset of images has the domain labels. The experiments on various datasets show that the proposed model outperforms the previous methods across all different levels of supervision. To the best of our knowledge, our model is the first that successfully conducts the truly unsupervised image-to-image translation task in an end-to-end manner.

2. Related work

Image-to-image translation. Since the seminal work of Pix2Pix [17] , image-to-image translation models have shown impressive results [48, 28, 25, 2, 29, 22, 15, 44, 24, 4, 16, 31, 46]. Exploiting the cycle consistency constraint, these methods were able to train the model with a set-level supervision (domains) solely. However, acquiring domain information can be a huge burden in practical applications where a large amount of data are gathered from several mixed domains, *e.g.*, web images [45]. Not only does this complicates the data collection, but it restricts the methods only applicable to the existing dataset and domains. Inspired from few-shot learning, Liu *et al.* [29] proposed FUNIT that works on previously unseen target classes. On the other hand, InfoGAN [3] utilizes an information-theoretic approach enabling GAN to learn meaningful representations under unsupervised manner and S^3GAN [30] has proposed to integrate a clustering method and GANs so that it can conduct high quality generation task using the fewer number of the labeled data. However, FUNIT requires the labels for training, while S^3GAN needs a subset of labeled data for the best performance. Recently, Bahng et al. [1] has partially addressed this by adopting the pre-trained classifier for extracting domain information. Unlike the previous methods, we aim to design an image-to-image translation model that can be applied without any supervision such

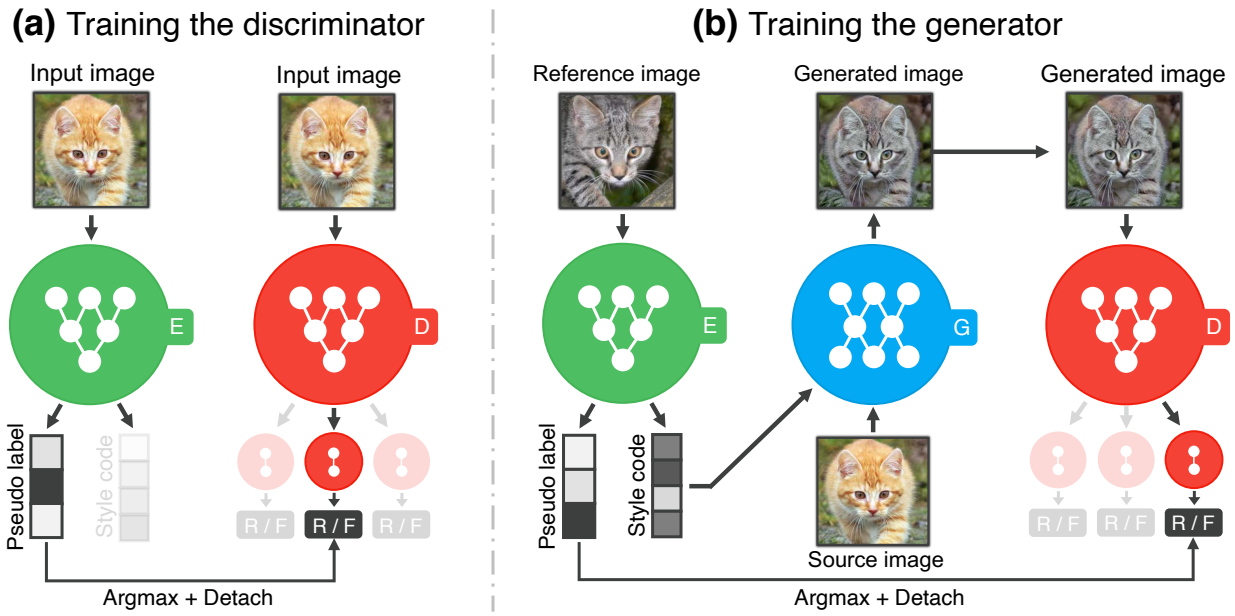


Figure 3. **Overview of our proposed method.** The figure illustrates how our model changes the breed of the cat. **(a)** An estimated domain from our guiding network E is used to train the multi-task discriminator D. **(b)** E provides the generator G with the style code of a reference image and the estimated domain is again used for GAN training.

as a pre-trained network or supervision on both the train and the test datasets.

Unsupervised representation learning and clustering. Unsupervised representation learning aims to extract meaningful features for downstream tasks without any human supervision. To this end, many researchers have proposed to utilize the information that can be acquired from the data itself [18, 10, 6, 34, 8, 14]. Recently, by incorporating the contrastive learning into a dictionary learning framework, MoCo [10] has achieved the state-of-the-art performance in various downstream tasks under reasonable mini-batch size. On the other hand, IIC [18] has utilized the mutual information maximization in a self-supervised way so that the network clusters images while assigning the images evenly. Though IIC provided a principled way to perform unsupervised clustering, the method fails to scale up when combined with a difficult downstream task such as image-to-image translation. By taking the best of both worlds, we aim to solve unsupervised image-to-image translation.

3. Method

We consider an unsupervised image-to-image translation problem, where we have images $\mathcal{X}$ from K domains ($K \geq 2$) without domain labels y. Here, K is an unknown property of the dataset.

We suggest a module that integrates both a domain classifier and a style encoder, which we call a guiding network. It guides the translation by feeding the style code to the generator and the pseudo domain labels to the discriminator. Using the feedback from the discriminator, the generator synthesizes images of the target domains (*e.g.* breeds) while respecting styles (*e.g.* fur patterns) of reference images and maintaining the content (*e.g.* pose) of source images (Figure 3).

3.1. Learning to produce domain labels and encode style features

In our framework, the guiding network E plays a central role as an unsupervised domain classifier as well as a style encoder. Our guiding network E consists of two branches of E_{class} and E_{style}, each of which learns to provide domain labels and style codes, respectively.

Unsupervised domain classification. We employ the unsupervised clustering approach [18] to automatically produce a domain label of a given image. Specifically, we maximize the mutual information (MI) between the domain assignments of an image $\mathbf{x}$ and those of its randomly augmented version $\mathbf{x}^+$ (*e.g.* random cropping, horizontal flip):

$$I(\mathbf{p}, \mathbf{p}^+) = H(\mathbf{p}) - H(\mathbf{p}|\mathbf{p}^+), \tag{1}$$

where $\mathbf{p} = E_{class}(\mathbf{x})$ represents the softmax output of the guiding network E that is the probability vector of $\mathbf{x}$ over K domains. Here, we set K as an arbitrarily large number. Note that the optimum of Eq. (1) is reached as the entropy $H(\mathbf{p})$ is maximum and the conditional entropy $H(\mathbf{p}|\mathbf{p}^+)$ is minimum. By maximizing the MI, the network is encouraged to distribute all the samples as evenly as possible over K domains while confidently classifying the paired samples

$(\mathbf{x}, \mathbf{x}^+)$ as the same domain. The joint probability matrix for calculating the MI is given by $K \times K$ matrix $\mathbf{P}$:

$$\mathbf{P} = \mathbb{E}_{\mathbf{x}^+ \sim f(\mathbf{x}) | \mathbf{x} \sim p_{data}(\mathbf{x})}[E_{class}(\mathbf{x}) \cdot E_{class}(\mathbf{x}^+)^T], \quad (2)$$

where f is a composition of random augmentations such as random cropping and affine transformation.

With the joint probability matrix, our guiding network is trained by directly maximizing the MI via the following objective function:

$$\mathcal{L}_{MI} = I(\mathbf{p}, \mathbf{p}^+) = I(\mathbf{P}) = \sum_{i=1}^{K} \sum_{j=1}^{K} \mathbf{P}_{ij} \ln \frac{\mathbf{P}_{ij}}{\mathbf{P}_i \mathbf{P}_j}, \quad (3)$$

where $\mathbf{P}_i = \mathbf{P}(\mathbf{p} = i)$ denotes the K-dimensional marginal probability vector, and $\mathbf{P}_{ij} = \mathbf{P}(\mathbf{p} = i, \mathbf{p}^+ = j)$ denotes the joint probability (Please refer to [18] for more details). To provide a deterministic one-hot label to the discriminator, we find the index of highest probability of $\mathbf{p}$ using the `argmax` operation (*i.e.* $y = \texttt{argmax}(E_{class}(\mathbf{x}))$).

Improving domain classification. Though maximizing $\mathcal{L}_{MI}$ provides a way to automatically generate the domain labels for input images, it fails to scale up when the resolution of images becomes higher than 64×64 or samples become complex and diverse (*e.g.* AnimalFaces [29]). We overcome this by adding an auxiliary branch E_{style} to the guiding network E and imposing the contrastive loss [9]:

$$\mathcal{L}^E_{style} = -\log \frac{\exp(\mathbf{s} \cdot \mathbf{s}^+ / \tau)}{\sum_{i=0}^{N} \exp(\mathbf{s} \cdot \mathbf{s}_i^- / \tau)}, \quad (4)$$

where $\mathbf{s} = E_{style}(\mathbf{x})$ is the style code of $\mathbf{x}$. This $(N+1)$-way classification enables E to utilize not only the similarity of the positive pair $(\mathbf{s}, \mathbf{s}^+)$ but also the dissimilarity of the negative pairs $(\mathbf{s}, \mathbf{s}_i^-)$, where the negative style codes $\mathbf{s}_i^-$ are stored into a queue using previously sampled images (Please refer to [10] for more details). We observe that adding this objective significantly improves unsupervised classification accuracy on AnimalFaces from 50.6% to 84.1% compared to the previous overclustering approach [18].

3.2. Image-to-image translation with the domain guidance

In this subsection, we describe how to perform the unsupervised image-to-image translation under the guidance of our guiding network. For successful translation, the translation model should provide the realistic images containing the visual feature of the target domain. To this end, we adopt three losses 1) adversarial loss to produce realistic images, 2) style contrastive loss that encourages the model not to ignore the style code, 3) image reconstruction loss for preserving the domain-invariant features. We explain each loss and the overall objective for each network.

Adversarial loss. For adversarial training, we adopt the multi-task discriminator [32]. It is designed to conduct discrimination for each domain simultaneously. However, its gradient is calculated only with the loss for estimating the domain of the input image. That is, the discriminator outputs a binary vector whose length is the number of domains (K). Then, the network is trained only with the gradient related to the prediction of the input's domain. For the domain label of the input image, we utilize the pseudo label from the guiding network. Formally, given the pseudo labels y and $\widetilde{y}$ for a source image $\mathbf{x}$ and a reference image $\widetilde{\mathbf{x}}$ respectively, we train our generator G and multi-task discriminator D via the adversarial loss:

$$\begin{aligned}\mathcal{L}_{adv} = &\mathbb{E}_{\mathbf{x} \sim p_{data}(\mathbf{x})}[\log D_y(\mathbf{x}) \\ &+ \mathbb{E}_{\mathbf{x}, \widetilde{\mathbf{x}} \sim p_{data}(\mathbf{x})}[\log(1 - D_{\widetilde{y}}(G(\mathbf{x}, \widetilde{\mathbf{s}})))],\end{aligned} \quad (5)$$

where $D_y(\cdot)$ denotes the logit from the domain-specific (y) discriminator, and $\widetilde{\mathbf{s}} = E_{style}(\widetilde{\mathbf{x}})$ denotes a target style code of the reference image $\widetilde{\mathbf{x}}$. The generator G learns to translate $\mathbf{x}$ to the target domain $\widetilde{y}$ while reflecting the style code $\widetilde{\mathbf{s}}$.

Style constrastive loss. In order to prevent degenerate situation where the generator ignores the given style code $\widetilde{\mathbf{s}}$ and synthesize a random image of the domain $\widetilde{y}$, we impose a style contrastive loss to the generator:

$$\mathcal{L}^G_{style} = \mathbb{E}_{\mathbf{x}, \widetilde{\mathbf{x}} \sim p_{data}(\mathbf{x})} \left[-\log \frac{\exp(\mathbf{s}' \cdot \widetilde{\mathbf{s}})}{\sum_{i=0}^{N} \exp(\mathbf{s}' \cdot \mathbf{s}_i^- / \tau)} \right]. \quad (6)$$

Here, $\mathbf{s}' = E_{style}(G(\mathbf{x}, \widetilde{\mathbf{s}}))$ denotes the style code of the translated image $G(\mathbf{x}, \widetilde{\mathbf{s}})$ and $\mathbf{s}_i^-$ denotes the negative style codes, which are from the same queue used in Eq. (4). The above loss guides the generated image $G(\mathbf{x}, \widetilde{\mathbf{s}})$ to have a style similar to the reference image $\widetilde{\mathbf{x}}$ and dissimilar to random negative samples. By doing so, we can avoid the degenerated solution where the encoder maps all the images to the same style code of the reconstruction loss [5, 16] based on L1 or L2 norm.

Image reconstruction loss. To ensure that the generator G can reconstruct the source image $\mathbf{x}$ when given with its original style $\mathbf{s} = E_{style}(\mathbf{x})$, we impose an image reconstruction loss:

$$\mathcal{L}_{rec} = \mathbb{E}_{\mathbf{x} \sim p_{data}(\mathbf{x})}[\|\mathbf{x} - G(\mathbf{x}, \mathbf{s})\|_1]. \quad (7)$$

This objective not only ensures the generator G to preserve domain-invariant characteristics (*e.g.*, pose) of its input image $\mathbf{x}$, but also helps to learn the style representation of the guiding network E by extracting the original style $\mathbf{s}$ of the source image $\mathbf{x}$.

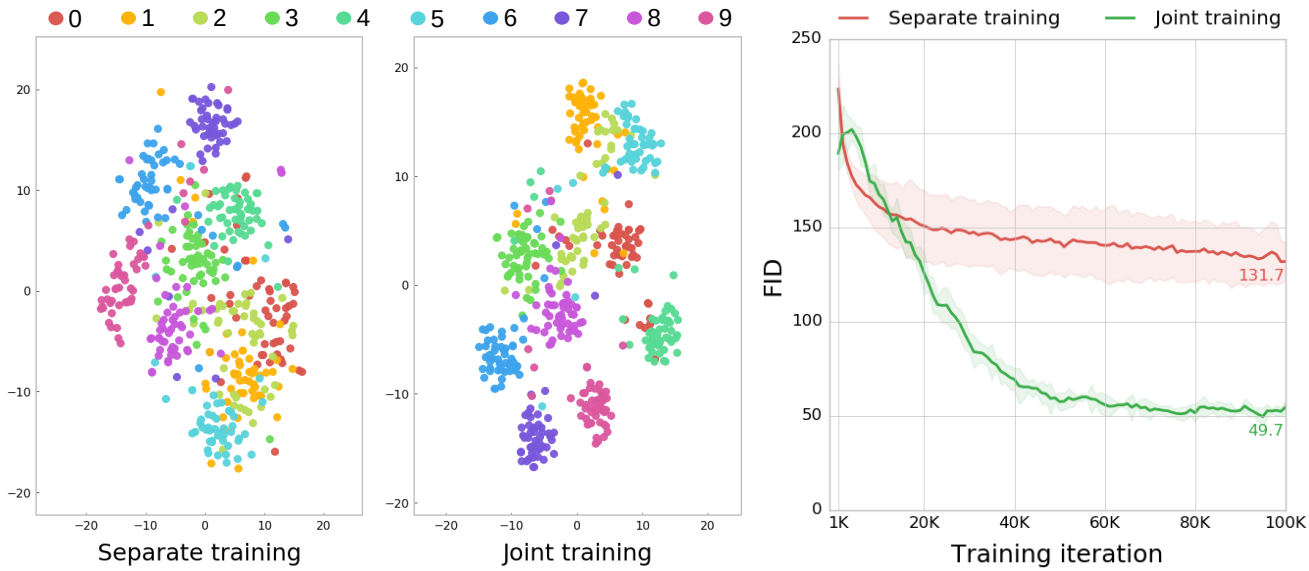


Figure 4. **Comparison of separate and joint training. (Left)** t-SNE visualization of style codes extracted by our guiding network. The ground truth domains of all test images in AnimalFaces-10 are represented in different colors. **(Right)** FID curves over training iterations. Joint training significantly outperforms separate training where the guiding network cannot receive feedback from the translation loss.

Overall objective. Finally, we train our discriminator, generator, and guiding network jointly as follows:

$$\begin{aligned} \mathcal{L}_D &= -\mathcal{L}_{adv}, \\ \mathcal{L}_G &= \mathcal{L}_{adv} + \lambda^G_{style}\mathcal{L}^G_{style} + \lambda_{rec}\mathcal{L}_{rec}, \\ \mathcal{L}_E &= \mathcal{L}_G - \lambda_{MI}\mathcal{L}_{MI} + \lambda^E_{style}\mathcal{L}^E_{style} \end{aligned} \tag{8}$$

where λ's are hyperparameters. Note that our guiding network E receives feedback from the translation loss L_G, which is essential for our method. We analyze and discuss the effect of backpropagating L_G to E on performance in Section 4.1. For training details, please see Appendix A.

4. Experiments

In this section, we present three experiments: analyzing the effect of the proposed objective functions and training strategy (Section 4.1), an unsupervised image-to-image translation on three unlabeled datasets (Section 4.2), and comparison to the state-of-the-art (SOTA) techniques under a semi-supervised setting (Section 4.3). For quantitative evaluation, we use Fréchet Inception Distance (FID) [12]. Because FID cannot penalize the case when the model conveys the source image as is for the output image, we calculate the mean of class-wise FIDs (mFID). For the details, please see Appendix B

Baselines. We use FUNIT [29] and MSGAN [31], which are the state of the art multi-domain and cross-domain image-to-image translation models, respectively. We observe that vanilla FUNIT fails on AnimalFaces-10 (mFID >150 for all ratios). As we believe that comparing the failing model is meaningless and unfair, we try our best to get a reasonable result by adjusting FUNIT (e.g. changing discriminator architecture). Under the semi-supervised scenario, we compare our method with both baselines in two training schemes. One is a naïve scheme that utilizes only the subset of the labeled dataset for entire training. The other is to train a simple domain classifier using the labeled data and utilize its domain output of the unlabeled data so that GANs can be trained with the whole images (Section

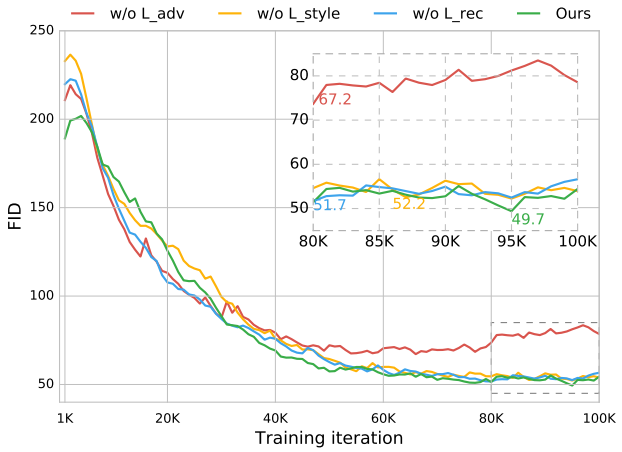


Figure 5. **Effects of translation loss on AnimalFaces-10.** During joint training, the generator is trained with entire translation loss (L_{adv}, L_{style}, and L_{rec}), but the guiding network is not received feedback from one of three losses. The FID score significantly increases when the guiding network is unable to receive feedback from the adversarial loss L_{adv}. Inset shows the zoomed-in final iterations.

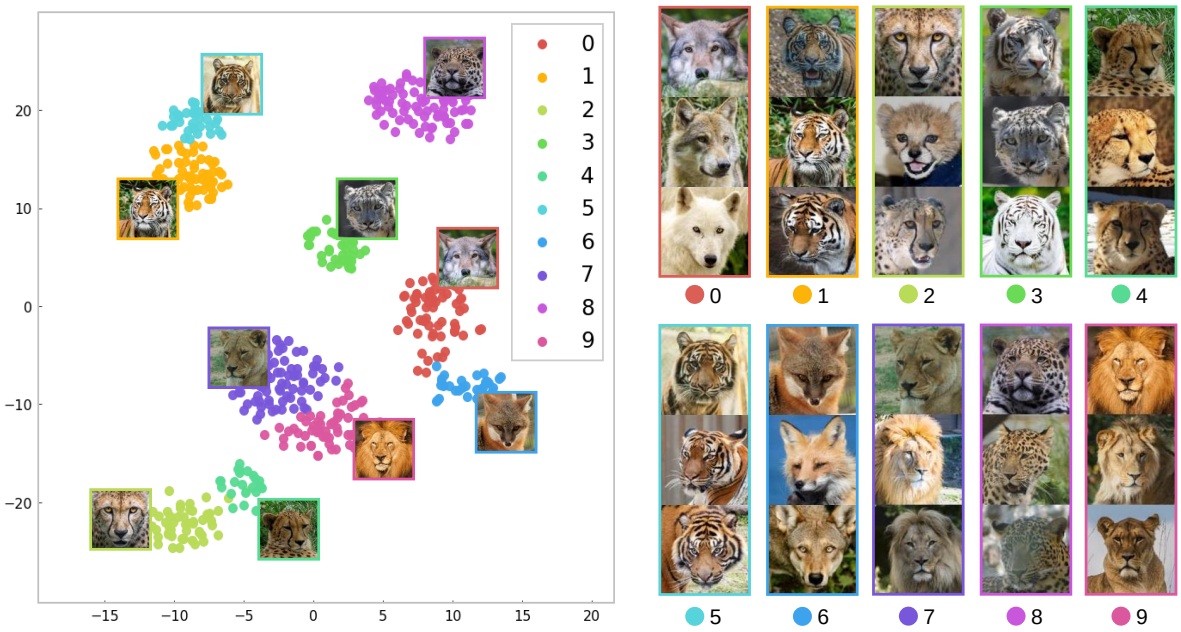


Figure 6. **t-SNE visualization of the style space** of our guiding network trained on AFHQ wild. Since AFHQ wild does not have ground-truth domain labels, each data point is colored with the guiding network's prediction. Although we set the number of domains to be quite large ($K = 10$), the network separates one species into two domains, which are so closely located that the model successfully creates six clusters.

4.3).

Datasets. For unsupervised translation, we evaluate our method on AFHQ, FFHQ, and LSUN Car datasets [5, 20, 47]. Note that these datasets do not have fine-grained set-level labels. AFHQ consists of three roughly labeled domains (*i.e.* cat, dog and wild), where each domain contains diverse species. FFHQ and LSUN Car contain various kinds of human faces and cars, respectively. For comparing with previous methods, we conduct the semi-supervised scenario by varying the supervision levels. Here, we use AnimalFaces [29] (multi-domain) and Summer2winter (cross-domain). Among total 149 animal classes from AnimalFaces, ten classes are chosen arbitrarily for training and testing. We call the selected dataset as AnimalFaces-10.

4.1. Effect of proposed strategy

In this section, we investigate the effect of our training strategy, which simultaneously performs representation learning as well as training the translation networks. Though one can easily think of separately training the guiding network and GANs, we show that this significantly degrades the overall performance. For this analysis, we choose AnimalFaces-10 and compare two training strategies with the proposed model; 1) joint training and 2) separate training. More specifically, the former is to train all the networks in an end-to-end manner as described in Section 3, and the latter is to first train the guiding network with $\mathcal{L}_{MI}$ and $\mathcal{L}^{C}_{style}$ for 100k iterations and then train the generator and the discriminator using the outputs of the frozen guiding network as their inputs. Note that for the separate training, the guiding network does not receive feedback from the translation loss $\mathcal{L}_G$ in Eq. (8).

To analyze the effects of different training strategies, we plot the class-wise FID over training iteration and provide t-SNE visualization in Fig. 4. From the FID comparison, the separate training strategy records much higher average FID score with higher standard deviation than the joint training strategy. This clearly shows that the joint training is more effective in terms of the image quality and the stable performance. To understand the performance gain by the joint training strategy, we visualize the style space of the guiding network via t-SNE. Here, the accuracy of the guiding network for each strategy is 83.6% and 70.4%, respectively. When the model is trained in the end-to-end manner, the clusters become more compact and clearly separated. Meanwhile, the separate training strategy leads the feature space to be entangled with significant overlaps. From this observation, we conclude that our joint training strategy induces more compact and disentangled clusters, thereby the generator can benefit from the style codes representing meaningful domain features, which eventually leads to the better translation model. Under the joint training strategy, we study the effect of each component ($\mathcal{L}_{adv}$, $\mathcal{L}_{rec}$, $\mathcal{L}^{G}_{style}$) in $\mathcal{L}_G$ for the guiding network on the translation performance. We train the model while excluding one of

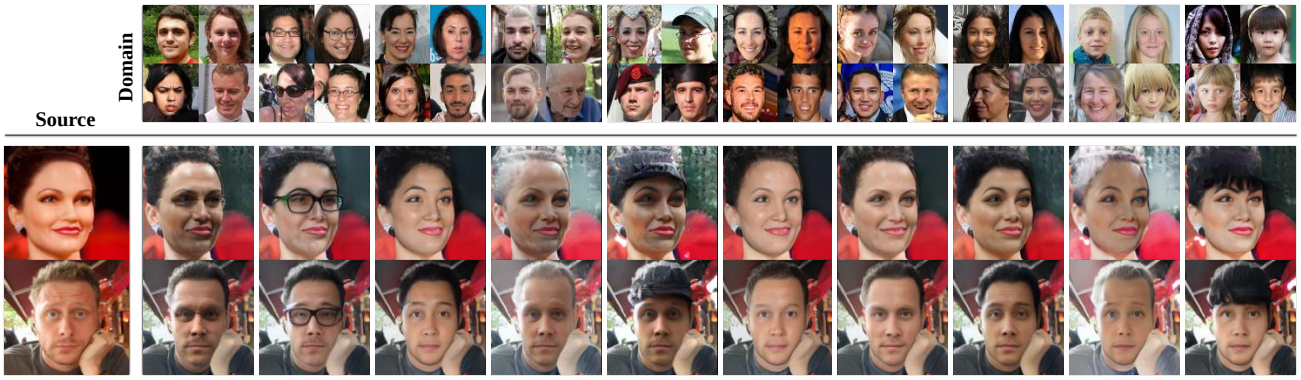
(a) AFHQ Cat.

(b) AFHQ Dog.

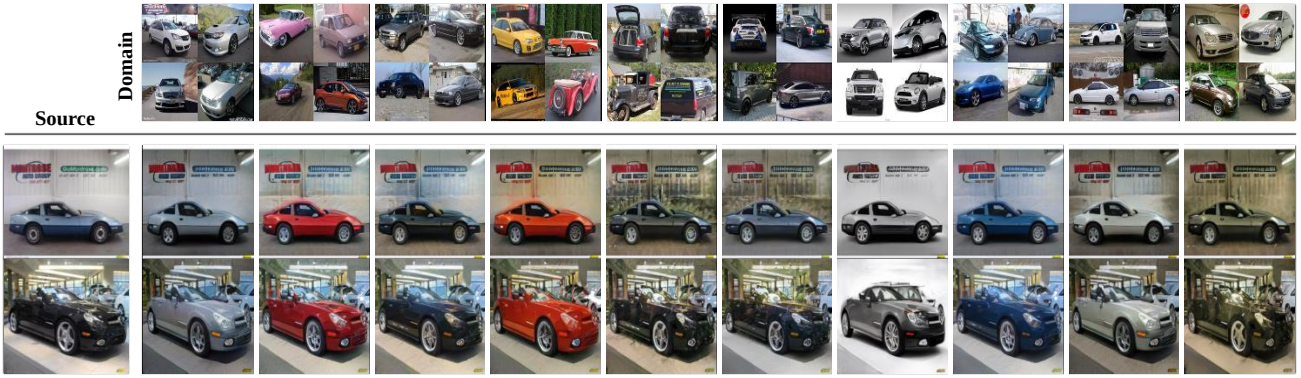
(c) AFHQ Wild.

Figure 7. **Unsupervised image-to-image translation results on AFHQ.** We set the number of domains (K) to ten in all cases. The top row shows representative images of ten domains estimated by our guiding network. Each source image is translated using the average style codes for each domain in test dataset. We note that all images are uncurated. The t-SNE visualization for wild can be found in Fig. 6.

the losses one at a time when training the guiding network. Fig. 5 shows the FID curves over training iterations for each setting. Removing $\mathcal{L}_{rec}$ or $\mathcal{L}^G_{style}$ does not show a meaningful impact on the translation performance. However, removing $\mathcal{L}_{adv}$ drops the translation performance significantly. We conjecture that it is because the adversarial loss cares not only how realistic the output images are but also their domains. From the experiment, we verify that the adversarial loss imposed to the guiding network enhances the overall translation performance of our model. From the study on the training strategy, we confirm that the interaction between the guiding network and GANs indeed enhances the translation performance. Therefore, in the following experiments, we adopt the joint training scheme and update the guiding network with full $\mathcal{L}_G$.

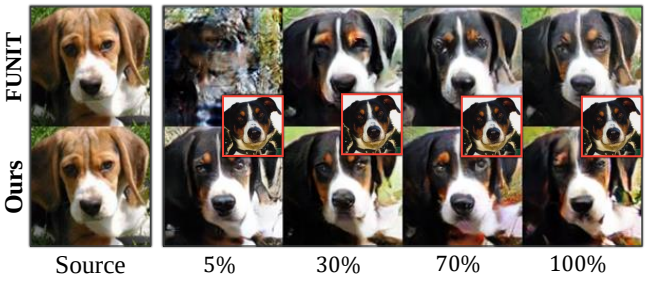


(a) FFHQ

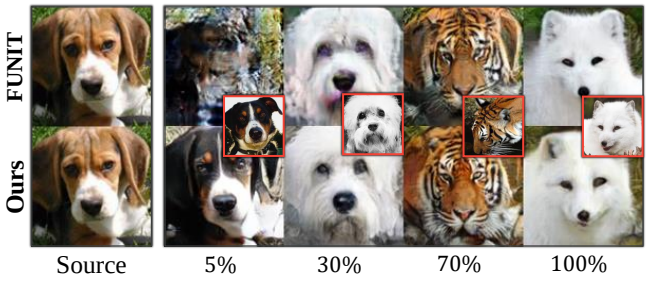


(b) LSUN Car

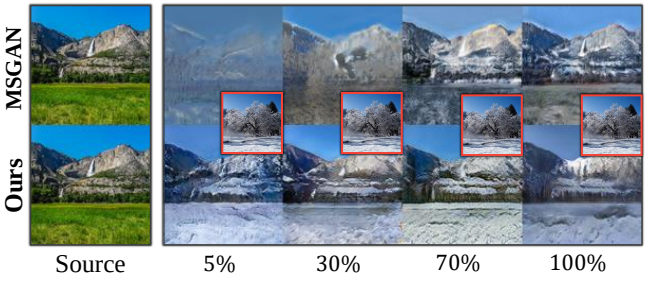
Figure 8. **Unsupervised image-to-image translation results on FFHQ and LSUN Car.** The experimental settings are the same as in Fig. 7.



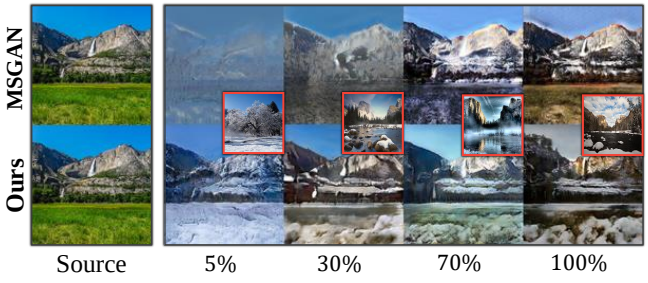
(a) Same references on AnimalFaces-10



(b) Different references on AnimalFaces-10

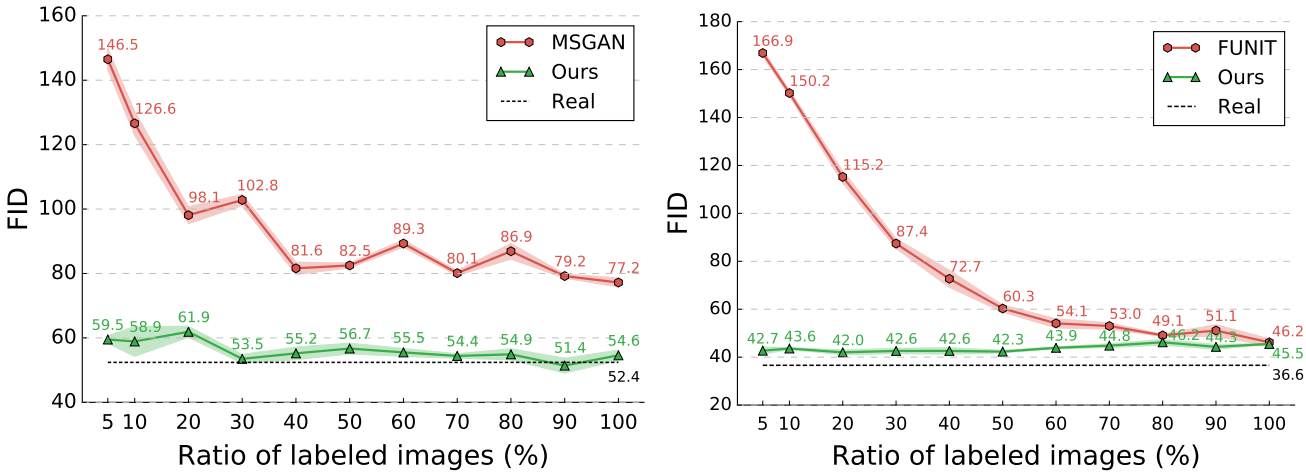


(c) Same references on Summer2winter



(d) Different references on Summer2winter

Figure 9. **Qualitative comparison for varying ratios of labeled images.** Red box indicates a reference image and the value under each image indicates the ratio of $\mathcal{D}_{sup}$.



(a) Summer2winter (b) AnimalFaces-10

Figure 10. **FID curves for varying ratios of labeled images under naïve scheme.** The dashed line indicates the expected lower bound (Real), which is calculated by dividing the training data in half. Our method is able to generate high-fidelity images using only 5% of the labeled data and outperforms the baselines in all ratios.

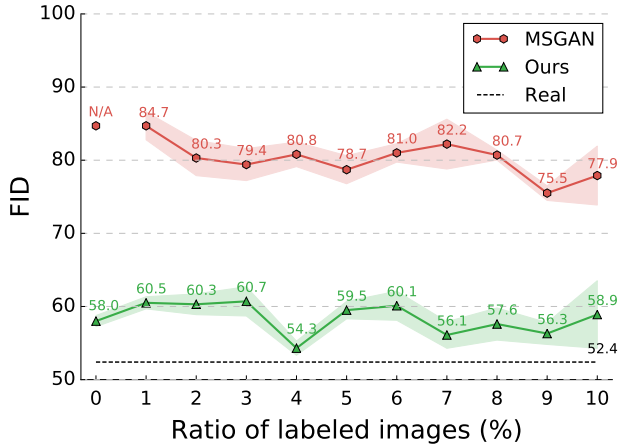
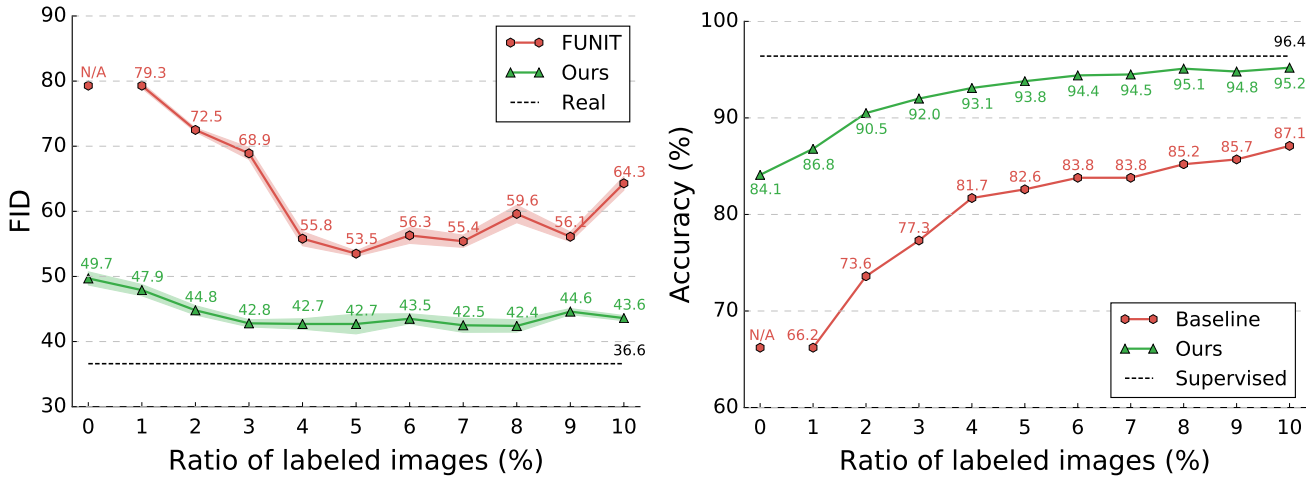


Figure 11. **FID curves on Summer2winter under alternative scheme.** Plots are the FID mean and standard deviation across five runs. Even if we introduce an auxiliary classifier to make MSGAN stronger, our method significantly outperforms MSGAN in all ratios. The dashed line indicates the FID value calculated by dividing the training data in half.

4.2. Image-to-image translation without any labels

To verify that the proposed method is able to handle unsupervised image-to-image translation, we evaluate our model on AFHQ, FFHQ and LSUN Car datasets. Only for AFHQ, we utilize the rough labels; dog, cat, and wild through training three sets individually, which means that we train one model per each domain of AFHQ. Because none of previous techniques consider a truly unsupervised image-to-image translation problem, we solely evaluate the proposed method on these datasets. For all datasets, we do not know the exact number of classes. Therefore, we arbitrarily set the number of domains (K) to 10 assuming these are sufficiently large to cover the hidden classes. Interestingly, we find that our method shows consistent performance across different K's. Because the datasets do not contain fine labels, we cannot conduct quantitative evaluation for this experiment. Instead, we visualize the style space from the guiding network to qualitatively assess the quality of representation learning. Fig. 6 shows the t-SNE result of the guiding network trained on AFHQ *wild* class and the example images from each domain. Surprisingly, the guiding network organizes the samples according to the species where it roughly separates the AFHQ wild images into seven species. Although we set the number of domains to be overly large, the network represents one species into two domains where those two domains position much closely. Thanks to this highly disentangled style space, our model can successfully handle the unsupervised image-to-image translation problem. The qualitative image translation results are demonstrated in Fig. 7. Each image is synthesized with the source image and the average style code of all test images in each domain. Four representative images of each domain are shown above. We consistently observe that each output successfully reflects the visual feature of each domain (*i.e.* the fur pattern and the color) and the visual feature of its species. The results from FFHQ and LSUN Car are shown in Fig. 8. Though it is not clear how to define the 'domains' in FFHQ, the network successfully separates the images into the visually distinctive categories such as glasses, hair color, and bang. For LSUN Car, the color of output images seamlessly follows the color of the reference images while preserving the type of the source



(a) FID (b) Accuracy

Figure 12. **(a) FID curves on AnimalFaces-10 under alternative scheme.** Even if we introduce an auxiliary classifier to make FUNIT stronger, our method outperforms FUNIT in all ratios. **(b) Classification accuracy on AnimalFaces-10.** Our guiding network produces much more accurate domain labels compared to the baseline classfier. The dashed line indicates the accuracy when the baseline classifier utilizes the entire labels for training. We note that IIC clustering achieves 50.4% accuracy and FUNIT with IIC achieves 112.2 of FID.

images. Please refer to the supplementary material for t-SNE visualizations, reference-guided translation results for all the datasets, and images representing each domain for FFHQ and LSUN Car.

4.3. Image-to-image translation with fewer labels

In this section, we compare our model to the state-of-the-art translation models trained in two schemes under the semi-supervised learning setting. We partition the dataset $\mathcal{D}$ into the labeled set $\mathcal{D}_{sup}$ and the unlabeled set $\mathcal{D}_{un}$ with varying ratio $\gamma = |\mathcal{D}_{sup}|/|\mathcal{D}|$. The first scheme is to train the models naïvely using only $\mathcal{D}_{sup}$. The second scheme is proposed to address the unfairness of the number of usable samples for training translation models. Here, an auxiliary classifier is trained on $\mathcal{D}_{sup}$ and is employed to produce pseudo-labels for $\mathcal{D}_{un}$. Using these labels, the translation model can be trained on the entire $\mathcal{D}$. Unlike these competitors, our model always utilize the entire dataset $\mathcal{D}$ by estimating pseudo labels via the guiding network. In the semi-supervised setting, the proposed model exploits an additional cross-entropy loss between the ground truth and the probability vector on $\mathcal{D}_{sup}$ for training E_{class}. By employing powerful few-shot or semi-supervised learning algorithms to improve the classifier, one might be able to improve the image-to-image translation model with fewer labels [40, 7]. Though this is an interesting direction, we leave it as a future research topic.

Naïve scheme. Fig. 10 (a) and (b) demonstrate the quantitative results using the class-wise FID on Summer2winter and AnimalFaces-10, respectively. As the ratio (γ) decreases, the performance of baseline models significantly degrades while the proposed model maintains FID around 60 and 45 regardless of γ. More importantly, the proposed model trained with only 5% outperforms MSGAN with all the labels (FID scores of 59.5 vs 77.2 on Summer2winter) and is comparable to FUNIT (FID scores of 42.7 vs 46.2 on AnimalFaces-10) with the entire dataset. Fig. 9 shows the qualitative comparisons between our results and the baselines trained with naïve scheme. We generate the translated images from the same reference but with the different γ and from the different references with the different γ. The results from FUNIT and MSGAN have poor quality for 5% labels and contain artifacts for even 30% labels.

Alternative scheme. The baseline trained with naïve scheme does not fully utilize the training samples as it simply disregards $\mathcal{D}_{un}$. To better exploit the entire training samples, we train an auxiliary classifier with $\mathcal{D}_{sup}$ from scratch to produce pseudo labels for $\mathcal{D}_{un}$. We vary the ratio of labeled data (γ) from 0% to 10% (0% is only for our method). We use VGG11-BN network [38] for both datasets. We note that the architecture of the guiding network is also VGG11-BN network, therefore this is fair comparison. Fig. 11 plots FID versus γ. Similar to the naïve scheme, the proposed model significantly outperforms MSGAN across all γ's. Interestingly, the performance of MSGAN using the alternative training is no longer sensitive to the changes in γ. Fig. 12 shows both the classification accuracy and FID scores on AnimalFaces-10. The accuracy of auxiliary classifier improves as the number of training samples (with labels) increases. The translation quality naturally improves upon the higher classification accuracy. Although the classification accuracy of the ratio 8% case

for FUNIT and 1% case for our model are similar (around 86%), the translation performance shows a noticeable gap (59.6 for FUNIT and 47.9 for our model). It implies that the accuracy is not the only factor for governing the translation performance. Based on our extensive comparisons and evaluations, we show that the proposed model is effective for the semi-supervised scenario and outperforms the baselines with the significant improvements.

5. Conclusion

We argue that the unsupervised image-to-image translation should not utilize any supervision, such as image-level (*i.e.* paired) or set-level (*i.e.* unpaired) supervision. Under this rigorous regime, many previous studies fall into the supervised framework that uses the domain information at minimum. In this paper, for the first time, we proposed an effective model to handle the truly unsupervised image-to-image translation. To this end, we suggested a guiding network that performs unsupervised representation learning for providing pseudo domain labels and the image translation tasks. The experimental results showed that the guiding network indeed exploits the synergy between two tasks, and the proposed model successfully conducts the unsupervised-image-to-image translation. Under the semi-supervised learning scenario, our model showed consistently better performance than the other state-of-the-art image translation models, regardless of the various ratio of labeled samples.

Acknowledgements

All the experiments are conducted on NAVER Smart Machine Learning (NSML) [21]. We thank Jung-Woo Ha for the feedback.

References

[1] H. Bahng, S. Chung, S. Yoo, and J. Choo. Exploring unlabeled faces for novel attribute discovery, 2019. 2

[2] S. Benaim and L. Wolf. One-shot unsupervised cross domain translation. In *Advances in Neural Information Processing Systems*, pages 2104–2114, 2018. 2

[3] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in neural information processing systems*, pages 2172–2180, 2016. 2

[4] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8789–8797, 2018. 2

[5] Y. Choi, Y. Uh, J. Yoo, and J.-W. Ha. Stargan v2: Diverse image synthesis for multiple domains. *arXiv preprint arXiv:1912.01865*, 2019. 4, 6

[6] A. Dosovitskiy, J. T. Springenberg, M. Riedmiller, and T. Brox. Discriminative unsupervised feature learning with convolutional neural networks. In *Advances in neural information processing systems*, pages 766–774, 2014. 3

[7] S. Gidaris, A. Bursuc, N. Komodakis, P. Pérez, and M. Cord. Boosting few-shot visual learning with self-supervision. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8059–8068, 2019. 10

[8] S. Gidaris, P. Singh, and N. Komodakis. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*, 2018. 3

[9] R. Hadsell, S. Chopra, and Y. LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742. IEEE, 2006. 2, 4

[10] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick. Momentum contrast for unsupervised visual representation learning. *arXiv preprint arXiv:1911.05722*, 2019. 2, 3, 4

[11] K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015. 13

[12] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in neural information processing systems*, pages 6626–6637, 2017. 5, 13

[13] G. Hinton, N. Srivastava, and K. Swersky. Neural networks for machine learning lecture 6a overview of mini-batch gradient descent. 13

[14] R. D. Hjelm, A. Fedorov, S. Lavoie-Marchildon, K. Grewal, P. Bachman, A. Trischler, and Y. Bengio. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*, 2018. 3

[15] J. Hoffman, E. Tzeng, T. Park, J.-Y. Zhu, P. Isola, K. Saenko, A. A. Efros, and T. Darrell. Cycada: Cycle-consistent adversarial domain adaptation. *arXiv preprint arXiv:1711.03213*, 2017. 2

[16] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz. Multimodal unsupervised image-to-image translation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 172–189, 2018. 2, 4

[17] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017. 1, 2

[18] X. Ji, J. F. Henriques, and A. Vedaldi. Invariant information clustering for unsupervised image classification and segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9865–9874, 2019. 2, 3, 4

[19] T. Karras, T. Aila, S. Laine, and J. Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017. 13

[20] T. Karras, S. Laine, and T. Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019. 1, 6

[21] H. Kim, M. Kim, D. Seo, J. Kim, H. Park, S. Park, H. Jo, K. Kim, Y. Yang, Y. Kim, et al. Nsml: Meet the mlaas platform with a real-world case study. *arXiv preprint arXiv:1810.09957*, 2018. 11

[22] T. Kim, M. Cha, H. Kim, J. K. Lee, and J. Kim. Learning to discover cross-domain relations with generative adversarial networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1857–1865. JMLR. org, 2017. 2

[23] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 13

[24] O. Kupyn, V. Budzan, M. Mykhailych, D. Mishkin, and J. Matas. Deblurgan: Blind motion deblurring using conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8183–8192, 2018. 2

[25] H.-Y. Lee, H.-Y. Tseng, J.-B. Huang, M. Singh, and M.-H. Yang. Diverse image-to-image translation via disentangled representations. In *Proceedings of the European conference on computer vision (ECCV)*, pages 35–51, 2018. 1, 2

[26] J. H. Lim and J. C. Ye. Geometric gan. *arXiv preprint arXiv:1705.02894*, 2017. 13

[27] M. Lin, Q. Chen, and S. Yan. Network in network. *arXiv preprint arXiv:1312.4400*, 2013. 14

[28] M.-Y. Liu, T. Breuel, and J. Kautz. Unsupervised image-to-image translation networks. In *Advances in neural information processing systems*, pages 700–708, 2017. 1, 2

[29] M.-Y. Liu, X. Huang, A. Mallya, T. Karras, T. Aila, J. Lehtinen, and J. Kautz. Few-shot unsupervised image-to-image translation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 10551–10560, 2019. 1, 2, 4, 5, 6

[30] M. Lucic, M. Tschannen, M. Ritter, X. Zhai, O. Bachem, and S. Gelly. High-fidelity image generation with fewer labels. *arXiv preprint arXiv:1903.02271*, 2019. 2

[31] Q. Mao, H.-Y. Lee, H.-Y. Tseng, S. Ma, and M.-H. Yang. Mode seeking generative adversarial networks for diverse image synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1429–1437, 2019. 2, 5

[32] L. Mescheder, S. Nowozin, and A. Geiger. Which training methods for gans do actually converge? In *ICML*, 2018. 4, 13

[33] M. Mirza and S. Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014. 1

[34] A. v. d. Oord, Y. Li, and O. Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 3

[35] T. Park, M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu. Semantic image synthesis with spatially-adaptive normalization. In *CVPR*, 2019. 1

[36] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer. Automatic differentiation in pytorch. 2017. 13

[37] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved techniques for training gans. In *Advances in neural information processing systems*, pages 2234–2242, 2016. 13

[38] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 10

[39] S. Singh and S. Krishnan. Filter response normalization layer: Eliminating batch dependence in the training of deep neural networks. *arXiv preprint arXiv:1911.09737*, 2019. 14

[40] K. Sohn, D. Berthelot, C.-L. Li, Z. Zhang, N. Carlini, E. D. Cubuk, A. Kurakin, H. Zhang, and C. Raffel. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *arXiv preprint arXiv:2001.07685*, 2020. 10

[41] K. Sohn, H. Lee, and X. Yan. Learning structured output representation using deep conditional generative models. In *Advances in neural information processing systems*, pages 3483–3491, 2015. 1

[42] D. Tran, R. Ranganath, and D. M. Blei. Deep and hierarchical implicit models. *arXiv preprint arXiv:1702.08896*, 7:3, 2017. 13

[43] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *CVPR*, 2018. 1

[44] J. M. Wolterink, A. M. Dinkla, M. H. Savenije, P. R. Seevinck, C. A. van den Berg, and I. Išgum. Deep mr to ct synthesis using unpaired data. In *International workshop on simulation and synthesis in medical imaging*, pages 14–23. Springer, 2017. 2

[45] I. Z. Yalniz, H. Jégou, K. Chen, M. Paluri, and D. Mahajan. Billion-scale semi-supervised learning for image classification. *arXiv preprint arXiv:1905.00546*, 2019. 2

[46] D. Yang, S. Hong, Y. Jang, T. Zhao, and H. Lee. Diversity-sensitive conditional generative adversarial networks. *arXiv preprint arXiv:1901.09024*, 2019. 2

[47] F. Yu, Y. Zhang, S. Song, A. Seff, and J. Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015. 6

[48] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017. 1, 2

A. Training details

We train the guiding network for the first 35K iterations while freezing the update from both the generator and the discriminator. Then, we train the whole framework 100K more iterations for training GANs. The batch size is set to 32 and 16 for 128×128 and 256×256 images, respectively. Training takes about 36 hours on a single Tesla V100 GPU with our implementation using PyTorch[36]. We use Adam [23] optimizer with $\beta_1 = 0.9, \beta_2 = 0.99$ for the guiding network, and RMSprop [13] optimizer with $\alpha = 0.99$ for the generator and the discriminator. All learning rates are set to 0.0001 with a weight decay 0.0001. We adopt hinge version adversarial loss [26, 42] with R_1 regularization [32] using $\gamma = 10$ (Eq. 5). We set $\lambda_{\text{rec}} = 0.1, \lambda^G_{\text{style}} = 0.01, \lambda^E_{\text{style}} = 1$, and $\lambda_{\text{MI}} = 5$ in Eq. 8 for all experiments. When the guiding network is simultaneously trained with the generator, we decrease λ^E_{style} and λ_{MI} to 0.1 and 0.5, respectively. For evaluation, we use the exponential moving average over the parameters [19] of the guiding network and the generator. We initialize the weights of convolution layers with He initialization [11], all biases to zero, and weights of linear layers from $N(0, 0.01)$ with zero biases. The source code is available at https://github.com/clovaai/tunit.

B. Evaluation protocol

Dataset with ground truth labels (*e.g.* AnimalFaces-10 and Summer2Winter)**.** For evaluation, we use Fréchet Inception Distance (FID) [12], which measures the discrepancy between real and generated distributions. We report the mean value of the class-wise FIDs (mFID), each of which is calculated using specific target domain. We set the number of generated samples to be equal to that of training samples in a target domain. For AnimalFaces-10, we choose the source images from all domains except the target domain and translate them using five reference images randomly sampled from the target domainâĂŹs test set. In the experiments of semi-supervised translation, we report the average of the top-5 FIDs over 100K iterations.

Dataset without any labels (*e.g.* AFHQ, FFHQ, LSUN Car)**.** Neither inception score [37] nor FID can penalize the generator just reconstructing input images; The model that doesnâĂŹt translate at all achieves a fairly good FID of 26.9 on AFHQ (Appendix G). In addition, mFID based on pseudo-labels is prone to wrong classification even with a good translation. Thus, we believe that the evaluation criteria should be fixed to ground truth (GT). Since GT is not available for some datasets, considering the above issues, we provide various qualitative results (+t-SNE) instead.

C. Architecture details

For the guiding network, we use `VGG11` before the linear layers followed by the average pooling operation as the shared part and append two branches E_{class} and E_{style}. The branches are one linear layer with K and 128 dimensional outputs, respectively. The detailed information of the generator, the guiding network and the discriminator architectures are provided in Table 1, Table 2 and Table 3.

LAYER	RESAMPLE	NORM	OUTPUT SHAPE
Image $\mathbf{x}$	-	-	$128 \times 128 \times 3$
Conv7×7	-	IN	$128 \times 128 \times ch$
Conv4×4	Stride 2	IN	$64 \times 64 \times 2ch$
Conv4×4	Stride 2	IN	$32 \times 32 \times 4ch$
Conv4×4	Stride 2	IN	$16 \times 16 \times 8ch$
ResBlk	-	IN	$16 \times 16 \times 8ch$
ResBlk	-	IN	$16 \times 16 \times 8ch$
ResBlk	-	AdaIN	$16 \times 16 \times 8ch$
ResBlk	-	AdaIN	$16 \times 16 \times 8ch$
Conv5×5	Upsample	AdaIN	$32 \times 32 \times 4ch$
Conv5×5	Upsample	AdaIN	$64 \times 64 \times 2ch$
Conv5×5	Upsample	AdaIN	$128 \times 128 \times ch$
Conv7×7	-	-	$128 \times 128 \times 3$

Table 1. Generator architecture. âĂIJchâĂİ represents the channel multiplier that is set to 64. IN and AdaIN indicate instance normalization and adaptive instance normalization, respectively.

LAYER	RESAMPLE	NORM	OUTPUT SHAPE
Image $\mathbf{x}$	-	-	$128 \times 128 \times 3$
Conv3×3	MaxPool	BN	$64 \times 64 \times ch$
Conv3×3	MaxPool	BN	$32 \times 32 \times 2ch$
Conv3×3	-	BN	$32 \times 32 \times 4ch$
Conv3×3	MaxPool	BN	$16 \times 16 \times 4ch$
Conv3×3	-	BN	$16 \times 16 \times 8ch$
Conv3×3	MaxPool	BN	$8 \times 8 \times 8ch$
Conv3×3	-	BN	$8 \times 8 \times 8ch$
Conv3×3	MaxPool	BN	$4 \times 4 \times 8ch$
GAP	-	-	$1 \times 1 \times 8ch$
FC	-	-	128
FC	-	-	K

Table 2. Guiding network architecture. âĂIJchâĂİ represents the channel multiplier that is set to 64. The architecture is based on VGG11-BN. GAP and FC denote global average polling [27] and fully connected layer, respectively.

LAYER	RESAMPLE	NORM	OUTPUT SHAPE
Image $\mathbf{x}$	-	-	$128 \times 128 \times 3$
Conv3×3	-	-	$128 \times 128 \times ch$
ResBlk	-	FRN	$128 \times 128 \times ch$
ResBlk	AvgPool	FRN	$64 \times 64 \times 2ch$
ResBlk	-	FRN	$64 \times 64 \times 2ch$
ResBlk	AvgPool	FRN	$32 \times 32 \times 4ch$
ResBlk	-	FRN	$32 \times 32 \times 4ch$
ResBlk	AvgPool	FRN	$16 \times 16 \times 8ch$
ResBlk	-	FRN	$16 \times 16 \times 8ch$
ResBlk	AvgPool	FRN	$8 \times 8 \times 16ch$
ResBlk	-	FRN	$8 \times 8 \times 16ch$
ResBlk	AvgPool	FRN	$4 \times 4 \times 16ch$
LReLU	-	-	$4 \times 4 \times 16ch$
Conv4×4	-	-	$1 \times 1 \times 16ch$
LReLU	-	-	$1 \times 1 \times 16ch$
Conv1×1	-	-	K

Table 3. Discriminator architecture. âĂIJchâĂİ and K represent the channel multiplier that is set to 64 and the number of clusters, respectively. FRN indicates filter response normalization [39].

D. t-SNE visualization & cluster example images

D.1. AFHQ Cat

Figure 13. **t-SNE visualization and representative images of each domain.**

D.2. AFHQ Dog

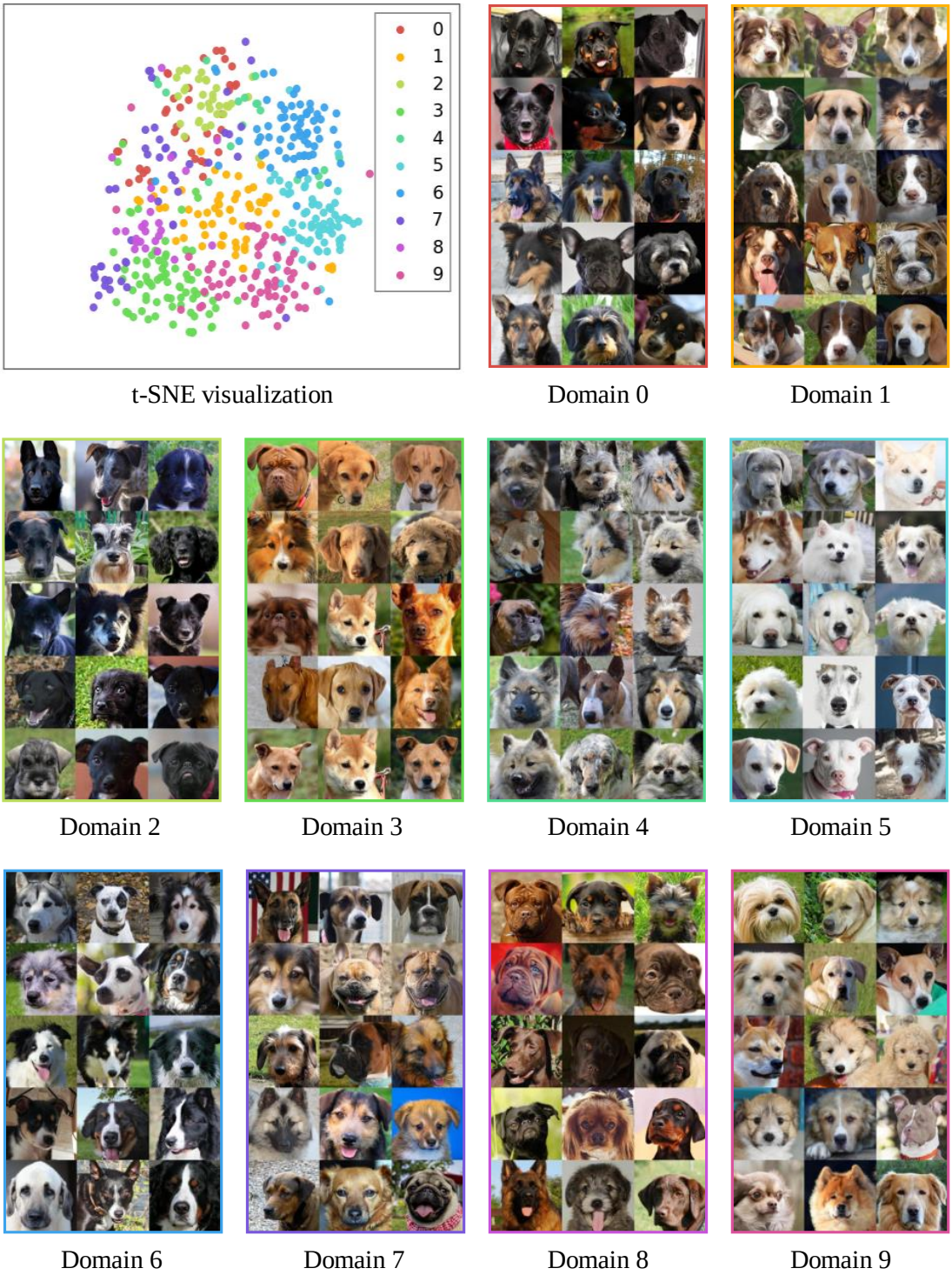


Figure 14. **t-SNE visualization and representative images of each domain.**

D.3. FFHQ

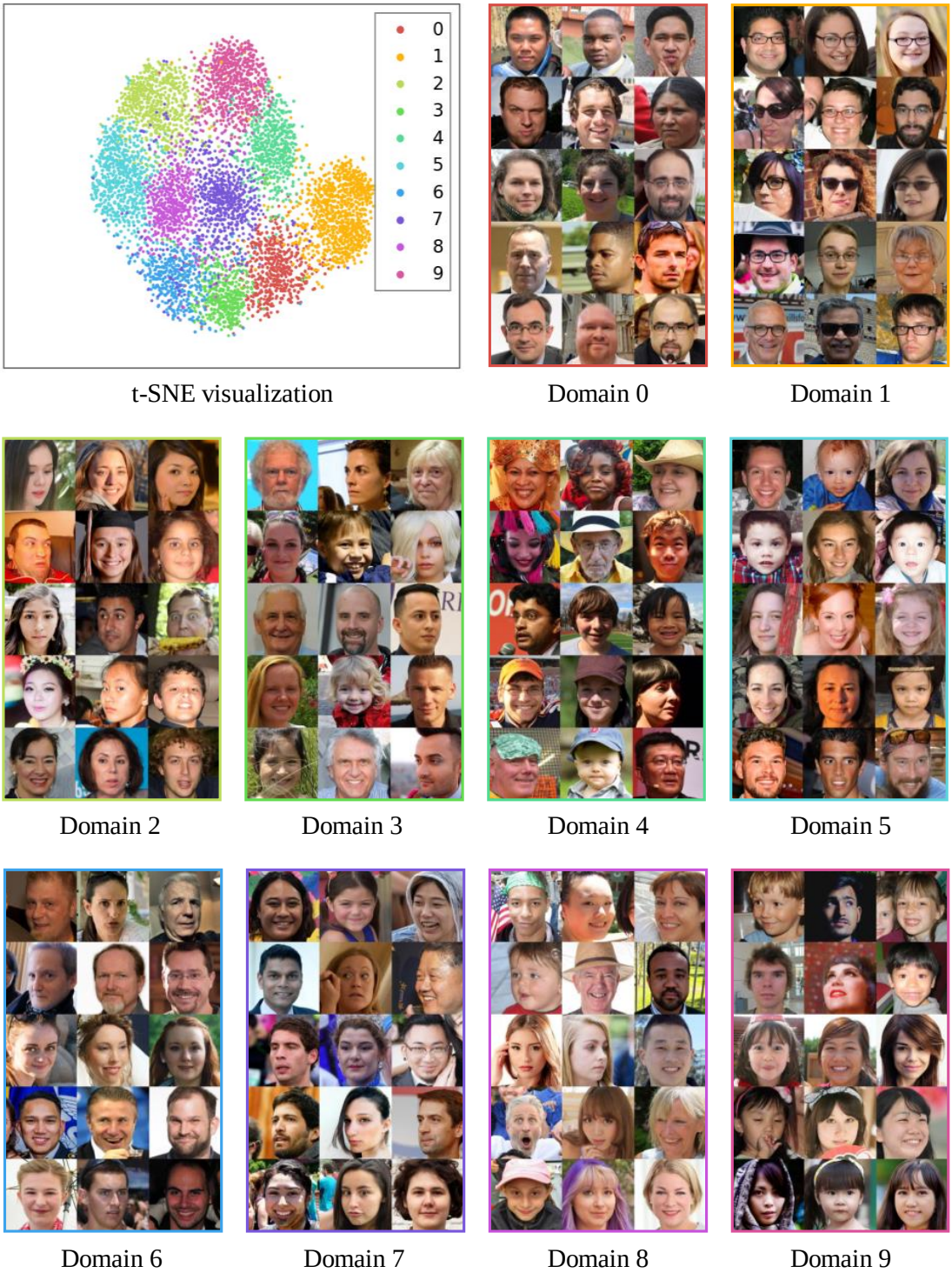


Figure 15. **t-SNE visualization and representative images of each domain.**

D.4. LSUN Car

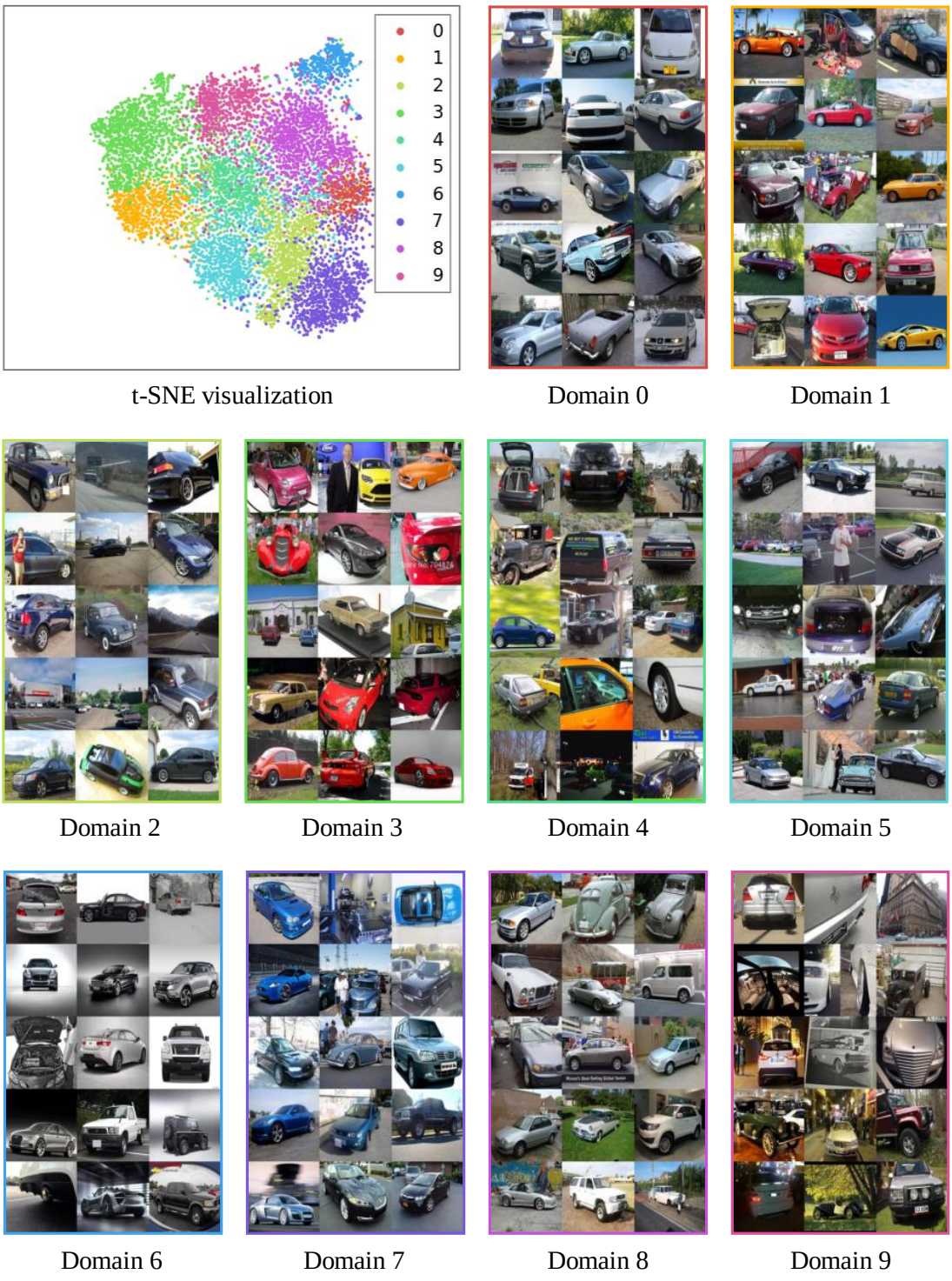
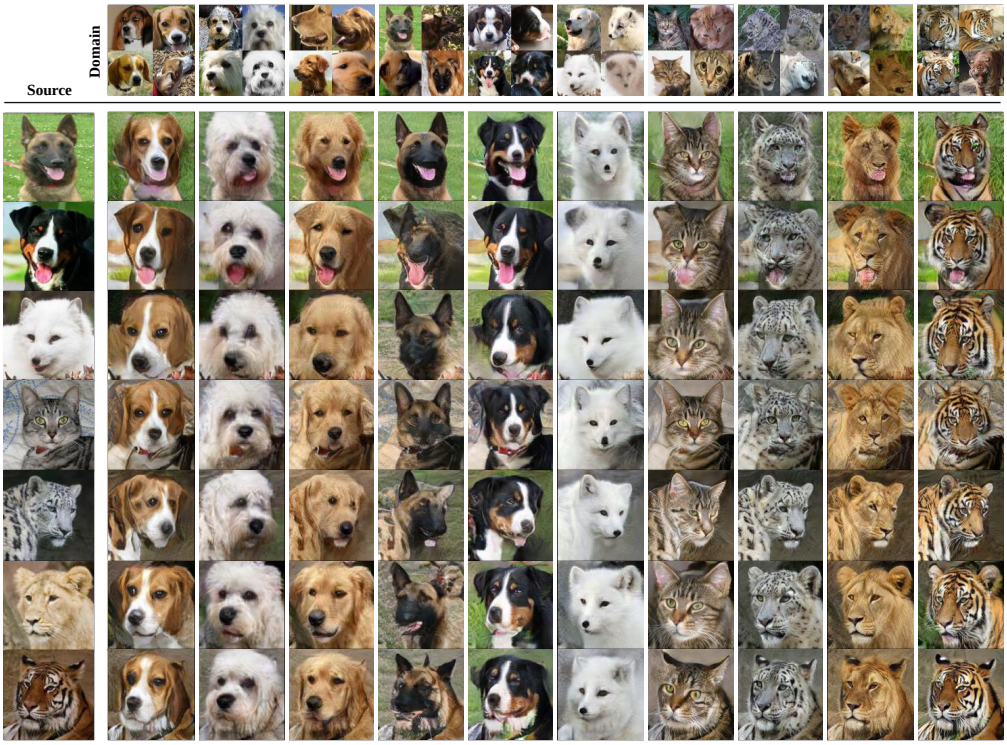


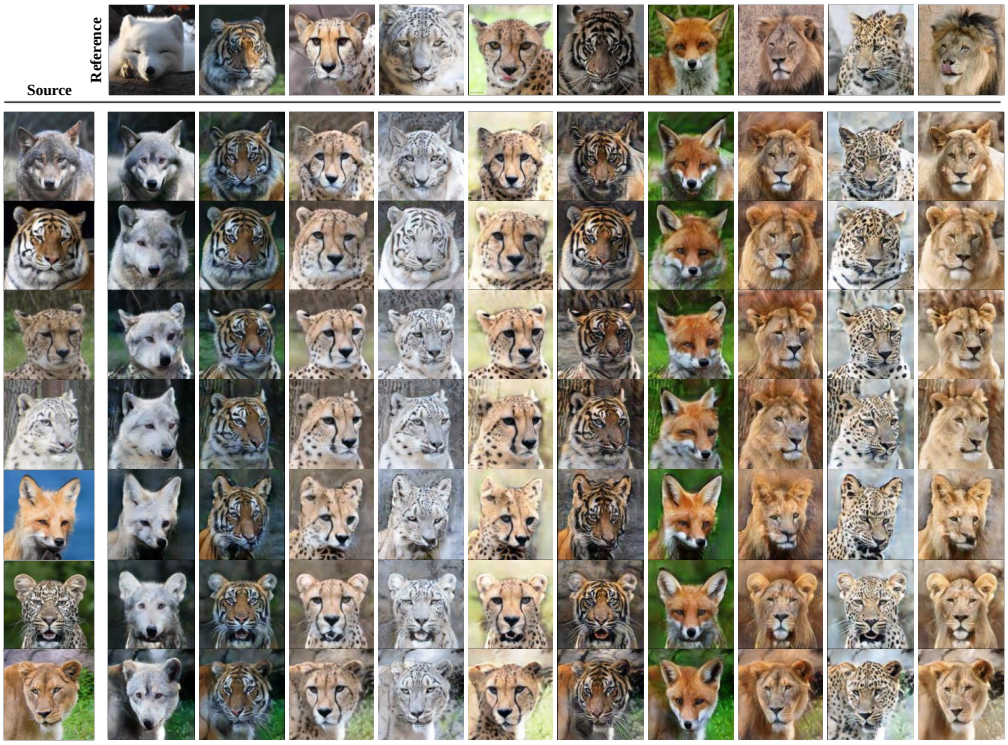
Figure 16. **t-SNE visualization and representative images of each domain.**

E. AFHQ, LSUN Car, FFHQ, AnimalFaces-10, and S2W

E.1. AnimalFaces-10



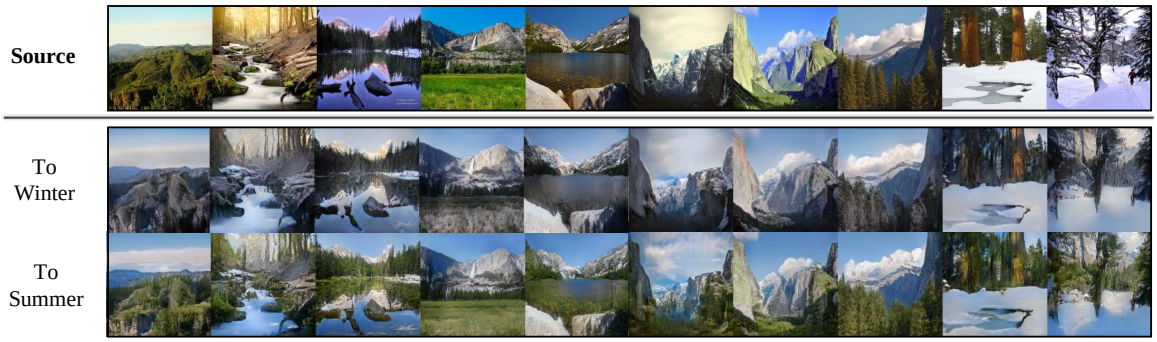
(a) Results guided by average style vectors



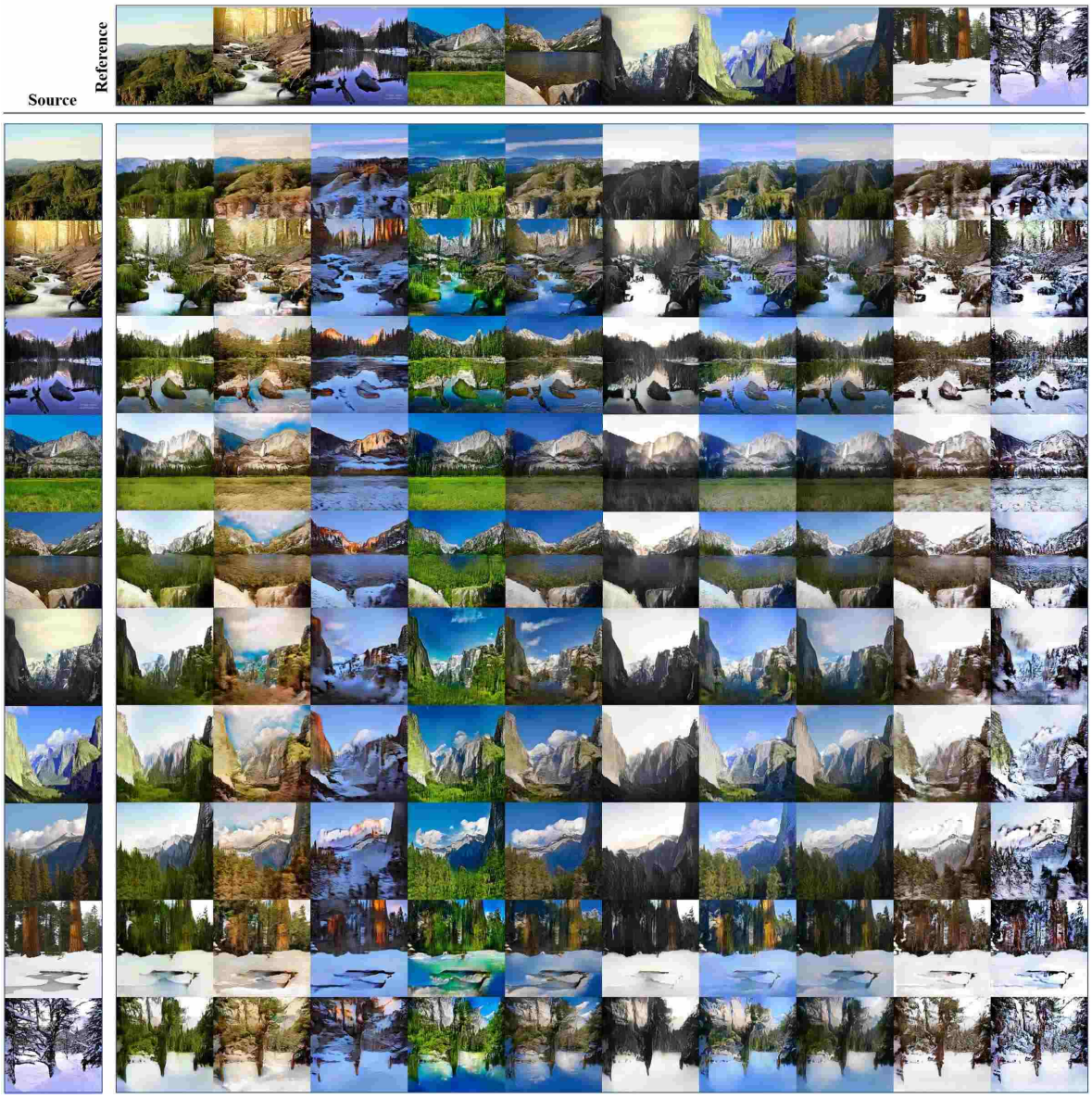
(b) Results guided by reference images

Figure 17. AnimalFaces-10, unsupervised image-to-image translation results.

E.2. Summer2Winter (S2W)



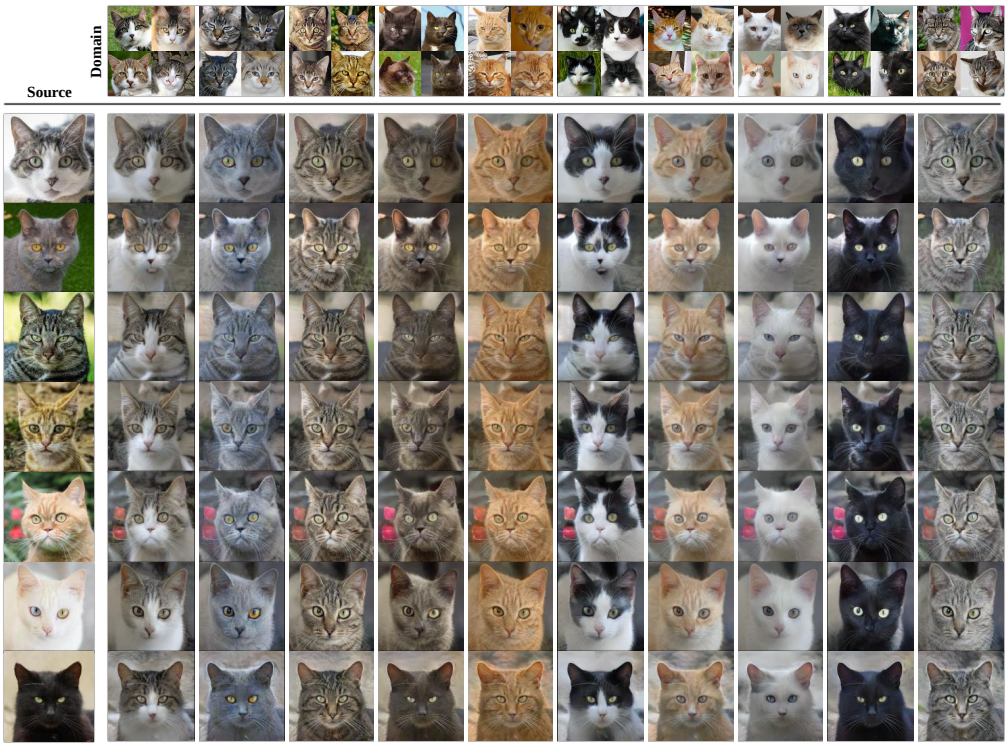
(a) Results guided by the average style code of each domain



(b) Results guided by reference images

Figure 18. Summer2Winter (S2W), unsupervised image-to-image translation results.

E.3. AFHQ Cat

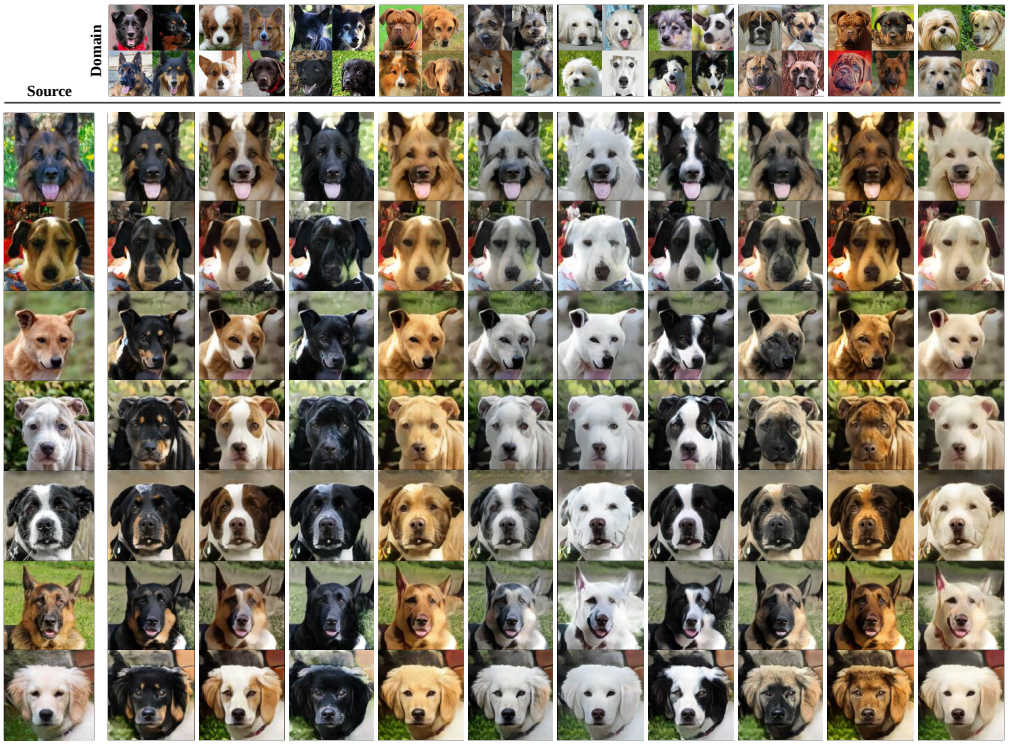


(a) Results guided by the average style code of each domain

(b) Results guided by reference images

Figure 19. AFHQ Cat, unsupervised image-to-image translation results.

E.4. AFHQ Dog

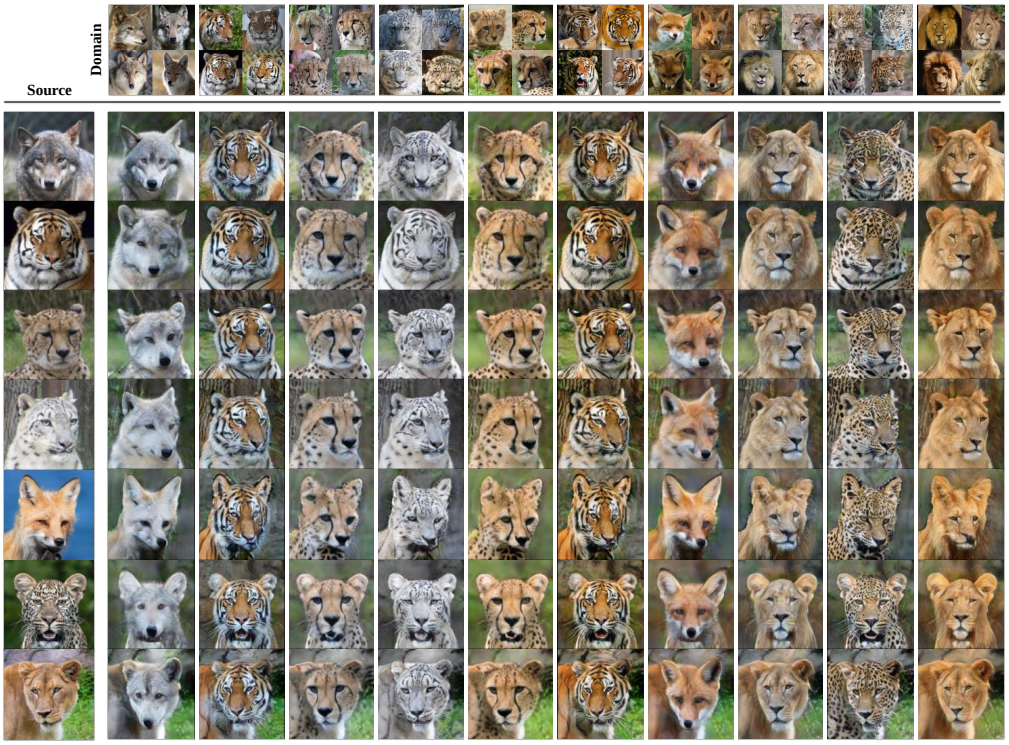


(a) Results guided by the average style code of each domain

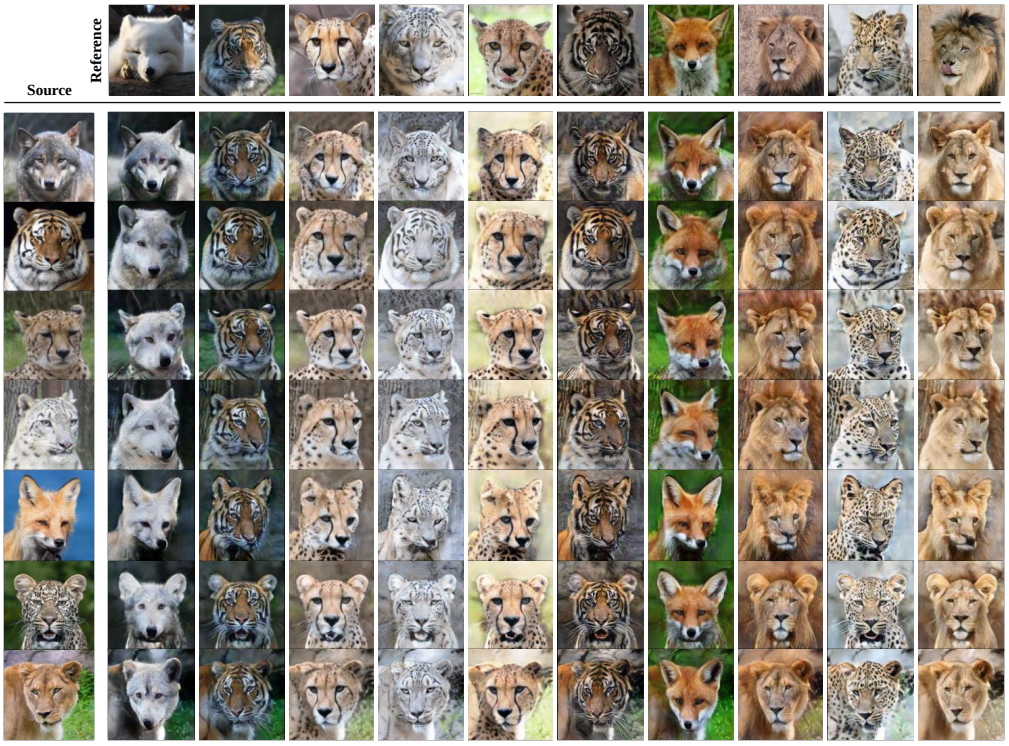
(b) Results guided by reference images

Figure 20. AFHQ Dog, unsupervised image-to-image translation results.

E.5. AFHQ Wild



(a) Results guided by the average style code of each domain

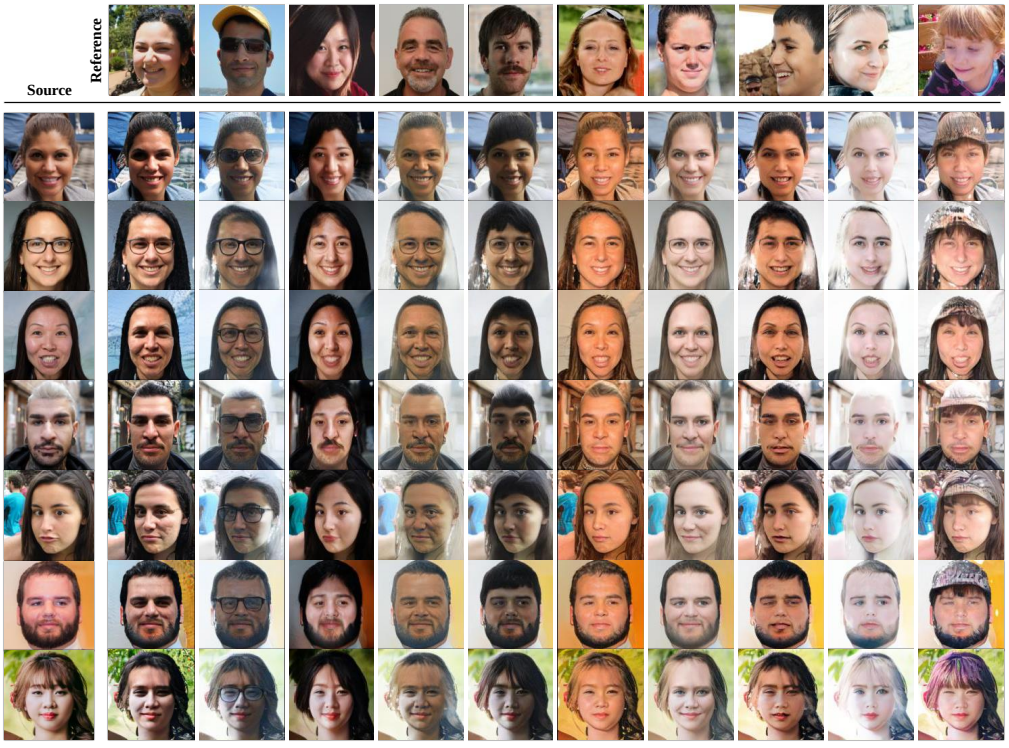


(b) Results guided by reference images

Figure 21. AFHQ Wild, unsupervised image-to-image translation results.

E.6. FFHQ

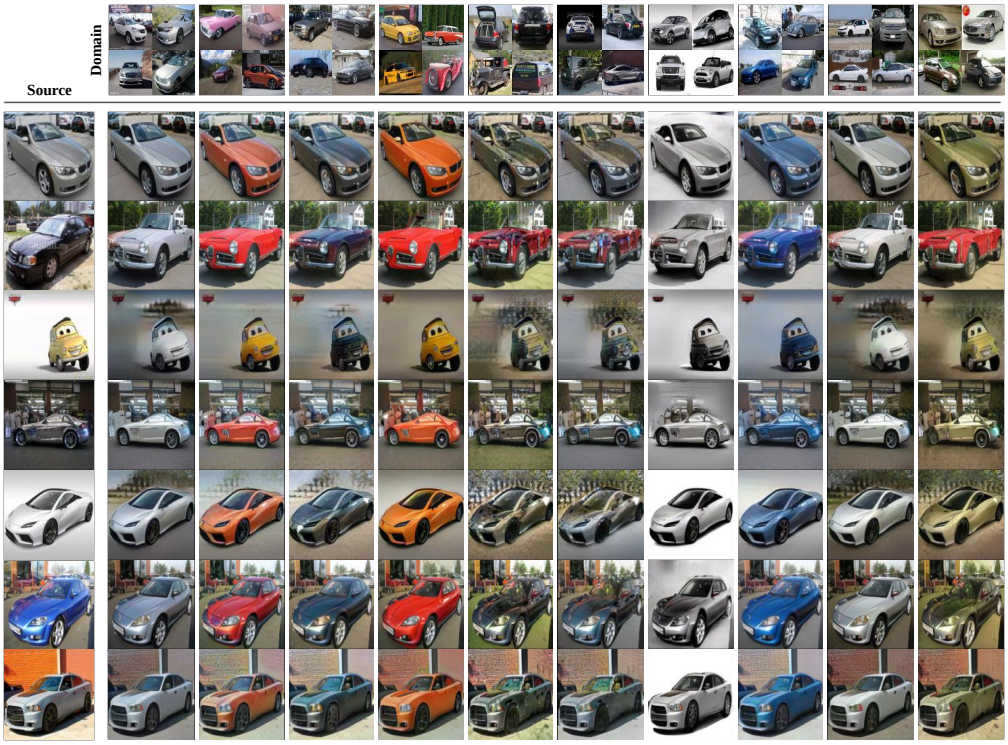
(a) Results guided by the average style code of each domain



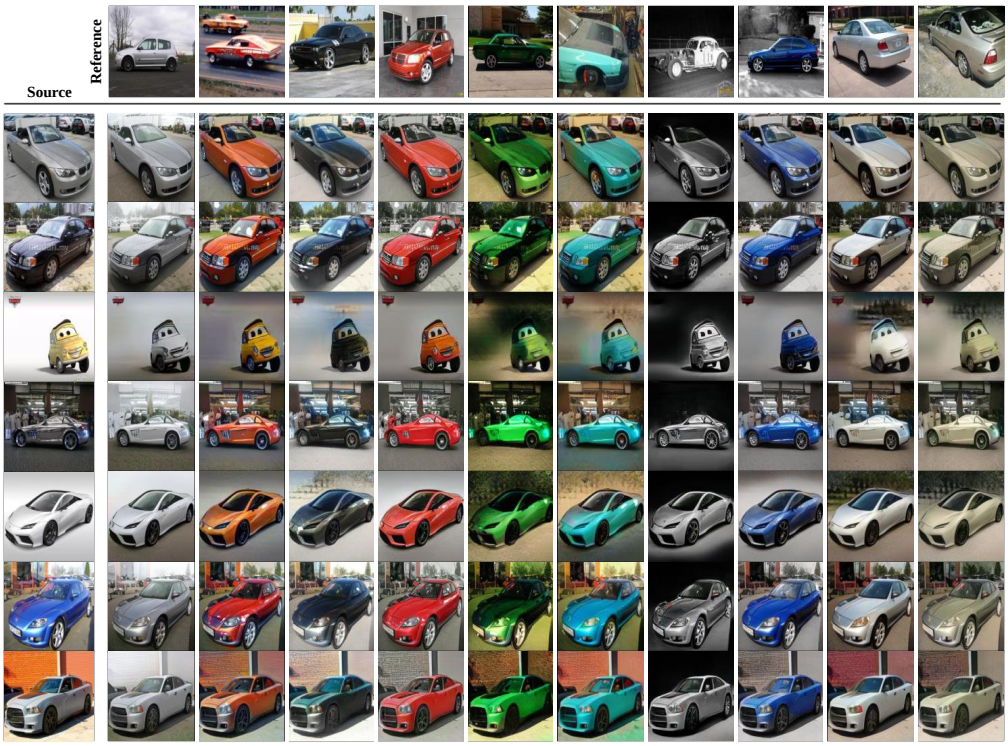
(b) Results guided by reference images

Figure 22. FFHQ, unsupervised image-to-image translation results.

E.7. LSUN Car



(a) Results guided by the average style code of each domain



(b) Results guided by reference images

Figure 23. LSUN Car, unsupervised image-to-image translation results.

F. Effects on different K on AnimalFaces-10

As stated in the main text, except for the case of Summer2winter where K is set to 2, we have set K as an arbitrarily large number (*e.g.* 10) for all experiments. Here, we empirically demonstrate that our method is quite robust to several K values. Fig. 24 shows the FID values and qualitative results for varying K values on AnimalFaces-10. For evaluation, we calculated the average of class-wise FID values as mentioned in Section B, using the actual domain information ($K = 10$). Note that we did not use the actual K information during the training phase.

As can be seen in Fig. 24, our model performs robustly when K is set to large enough ($K \geq 10$). We observe that the model performs best when the K value we set matches the ground-truth K value (*i.e.* $K = 10$). Interestingly, when K is set to 1, the model fails to completely change the breeds of the input image and only changes its texture.

Figure 24. AnimalFace, model performance with varying K's. The value under K is FID. Note that the FID is the loweset (49.7) when $K = 10$ (the number of actual domains).

G. Failure case: $K = 1$

To examine the failure case of the proposed model, we train our model on AFHQ wild and set $K = 1$. In this extreme case, the generator can not translate images between domains. The output images are almost the same as the source images. We conjecture that it is because the discriminator can not penalize the domain features of the output images so that the generator and the guiding network do not receive the feedback related to the domain. We note that the model that always reconstructs the input image achieves FID of 26.9. Therefore, we do not adopt FID for evaluating models trained on unlabeled dataset.

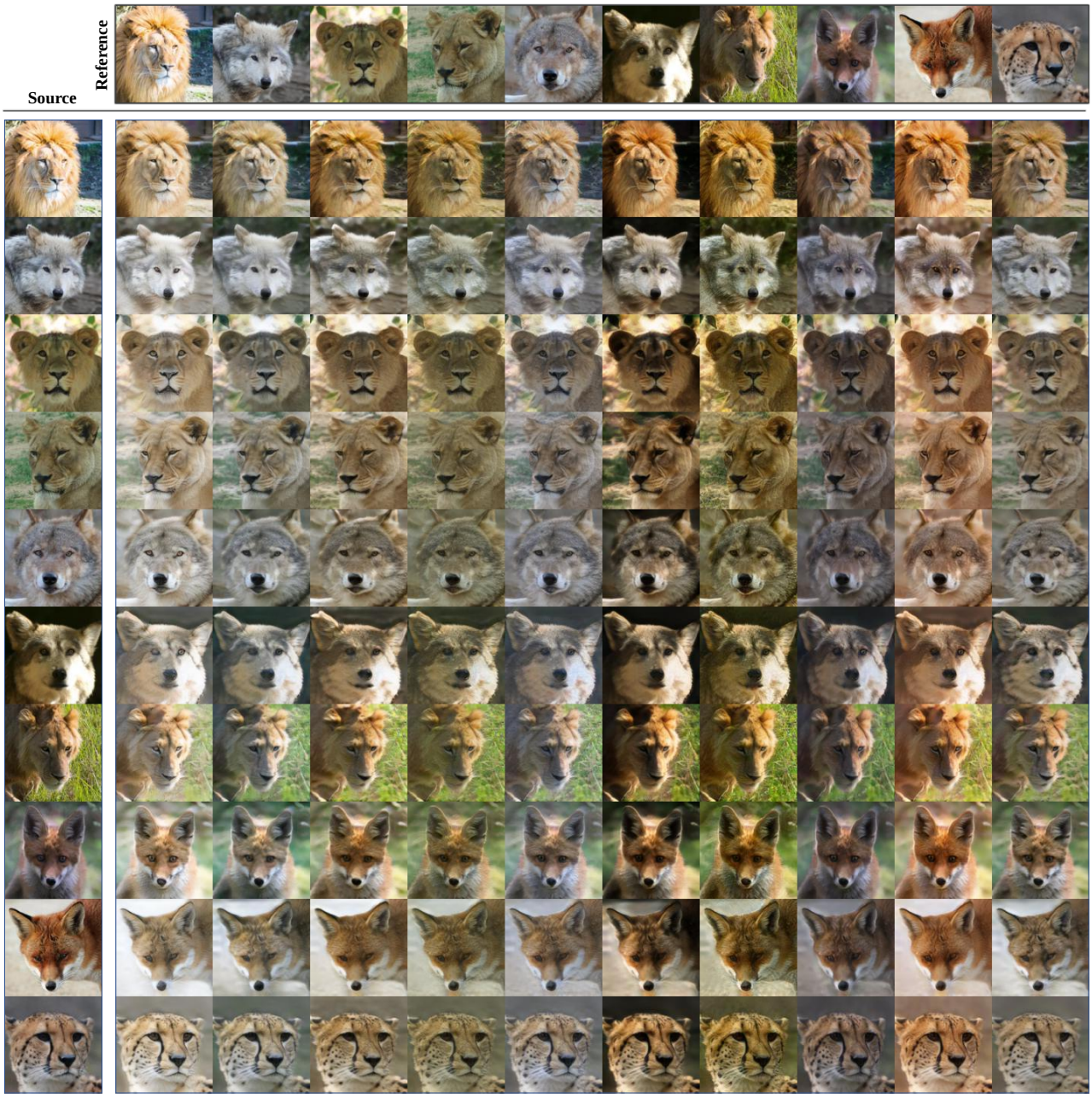


Figure 25. Failure case. Only the overall colors of the reference is translated.

H. Scalability: AnimalFaces-50

To test the scalability of the method, we train the models on AnimalFaces-50 that consists of 50 classes chosen arbitrarily for training and testing. We train the models under Naïve scheme. The benefits from the proposed model still exist.

Table 4. **FID scores for varying ratios of labeled images under Naïve scheme.**

Model	0%	20%	40%	60%	80%	100%
Ours	121.3	53.5	49.0	50.4	53.0	59.9
FUNIT	N/A	125.0	82.9	68.8	56.6	56.4

I. Style code interpolation

Please refer the video in the attachment. The outputs are generated with the model trained under unsupervised manner. The dataset is AFHQ wild.