Rethinking the Truly Unsupervised Image-to-Image Translation

Kyungjune Baek*
Yonsei University
bkjbkj12@yonsei.ac.kr

Yunjey Choi
NAVER AI Lab
yunjey.choi@navercorp.com

Youngjung Uh
Yonsei University
yj.uh@yonsei.ac.kr

Jaejun Yoo
UNIST
jaejun.yoo@unist.ac.kr

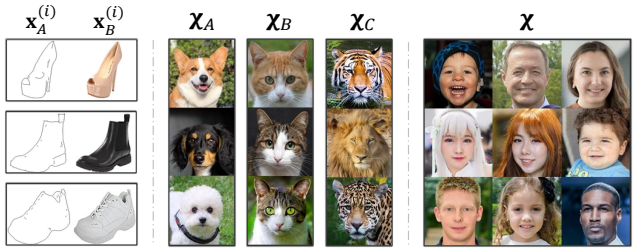
Hyunjung Shim†
Yonsei University
kateshim@yonsei.ac.kr

Abstract

*Every recent image-to-image translation model inherently requires either image-level (*i.e. *input-output pairs) or set-level (*i.e. *domain labels) supervision. However, even set-level supervision can be a severe 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 a truly unsupervised image-to-image translation model (TUNIT) that simultaneously learns to separate image domains and translates input images into the estimated domains. Experimental results show that our model achieves comparable or even better performance than the set-level supervised model trained with full labels, generalizes well on various datasets, and is robust against the choice of hyperparameters (*e.g. *the preset number of pseudo domains). Furthermore, TUNIT can be easily extended to semi-supervised learning with a few labeled data.*

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 [26, 31], many image translation methods have been proposed either using *image-level* supervision (*e.g.* paired data) [14, 12, 39, 33, 28] or using *set-level* supervision (*e.g.* domain labels) [38, 18, 21, 13, 22, 20]. 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*. This assumption can be a serious bottleneck in practice as the number of domains and samples increases. For example, labeling individual samples of a large dataset (*e.g.* FFHQ) is expensive, and the distinction across domains can be vague.



(a) Image-level (b) Set-level (c) Unsupervised

Figure 1: **Levels of supervision.** To perform image-to-image translation, existing methods need either **(a)** a dataset with input-output pairs or, **(b)** a dataset with domain information. Our method is capable of learning mappings among multiple domains using **(c)** a dataset without any supervision.

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 rigorous definition, our goal is to develop an unsupervised translation model given a mixed set of images of many domains (Figure 1). We argue that the unsupervised translation model is valuable in three aspects. First of all, it significantly reduces the effort of data annotation for model training. As a natural byproduct, the unsupervised model can be robust against the noisy labels produced by the manual labeling process. More importantly, it serves as a strong baseline to develop the semi-supervised image translation models. To tackle this problem, we design our model having three sub-modules: 1) clustering the images by approximating the set-level characteristics (*i.e.* domains), 2) encoding the individual content and style of an input image, respectively, and 3) learning a mapping function among the estimated domains.

To this end, we introduce a `guiding network`. The guiding network consists of a shared encoder with two

*Work done during his internship at Clova AI Research.
†Hyunjung Shim is a corresponding author.

branches, where one provides pseudo domain labels and the other encodes images into feature vectors (style codes). We employ a differentiable clustering method based on mutual information maximization for estimating the domain labels and contrastive loss for extracting the style codes. The clustering helps the guiding network to group similar images into the same category. Meanwhile, the contrastive loss helps the model to understand the dissimilarity among images and learn better representations. We find that, by solving two tasks together within the same module, both benefit from each other. Specifically, the clustering can exploit rich representations learned by the contrastive loss and improve the accuracy of estimated domain labels. By taking advantage from the clustering module, the style code can also acknowledge the similarity within the same domain, thereby faithfully reflecting the domain-specific nature.

For both more efficient training and effective learning, we jointly train the guiding network and GAN in an end-to-end manner. This allows the guiding network to understand the recipes of domain-separating attributes based on GAN's feedback, and the generator encourages the style code to contain rich information so as to fool the domain-specific discriminator. Thanks to these internal and external interactions of the guiding network and GAN, our model successfully separates domains and translates images; a truly unsupervised image-to-image translation.

We quantitatively and qualitatively compare the proposed model with the existing set-level supervised models under unsupervised and semi-supervised settings. The experiments on various datasets show that the proposed model outperforms the baselines over all different levels of supervision. Our ablation study shows that the guiding network helps the image translation model to largely improve the performance. Our contributions are summarized as follows:

- We clarify the definition of unsupervised image-to-image translation and to the best of our knowledge, our model is the first to succeed in this task in an end-to-end manner.
- We propose the guiding network to handle the unsupervised translation task and show that the interaction between translation and clustering is helpful for the task.
- The quantitative and qualitative comparisons for the unsupervised translation task on four public datasets show the effectiveness of TUNIT, which clearly outperforms the previous arts.
- TUNIT is insensitive to the hyperparameter (*i.e*. the number of clusters) and serves as a strong baseline for the semi-supervised setting– TUNIT outperforms the current state-of-the-art semi-supervised image translation model.

2. Related work

Image-to-image translation. Since the seminal work of Pix2Pix [14], image-to-image translation models have shown impressive results [38, 21, 18, 12, 19, 4, 13, 22, 36, 5, 34]. Exploiting the cycle consistency constraint or shared latent space assumption, 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 [35]. Not only does this complicates the data collection, but it restricts the methods only applicable to the existing dataset and domains. S^3GAN [24] and Self-conditioned GAN [23] integrated a clustering method and GAN for high-quality generation using the fewer number or none of the labeled data, respectively. Inspired from few shot learning, Liu *et al.* [22] proposed FUNIT that works on previously unseen target classes. However, FUNIT still requires the labels for training. Wang *et al.* [34] utilized the noise-tolerant pseudo labeling scheme to reduce the label cost at the training process. Recently, Bahng et al. [1] partially addressed this by adopting the ImageNet pre-trained classifier for extracting domain information. Unlike the previous methods, we aim to design an image translation model that can be applied without supervision such 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 [7, 6, 11, 15, 29, 8, 2, 32]. Recently, by incorporating contrastive learning into a dictionary learning framework, MoCo [8, 3] achieved outstanding performance in various downstream tasks under reasonable mini-batch size. On the other hand, IIC [15] utilized the mutual information maximization in an unsupervised manner 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. Truly Unsupervised Image-to-Image Translation (TUNIT)

We address the 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. Throughout the paper, we denote K as the actual number of domains in a dataset and $\hat{K}$ as the arbitrarily chosen number of domains to train models.

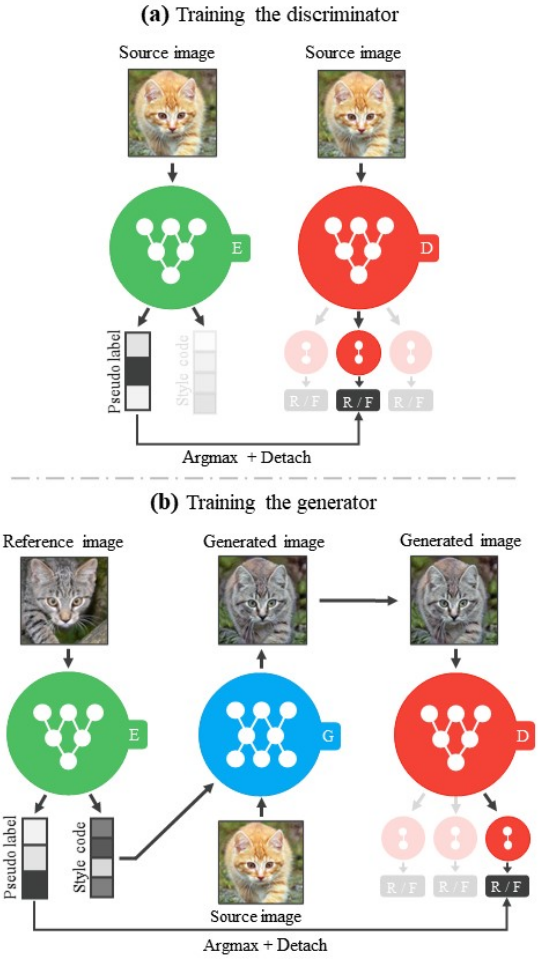


Figure 2: **Overview of our proposed method.** The figure illustrates how our model changes the breed of a cat. **(a)** Our guiding network E estimates the domain and use it to train the multi-task discriminator D. **(b)** Both the style code and the estimated domain of a reference image is used for training the generator G.

3.1. Overview

In our framework, the guiding network (E in Figure 2) plays a central role as an unsupervised domain classifier as well as a style encoder. It guides the translation by feeding the style code of a reference image to the generator and its pseudo domain labels to the discriminator. Using the feedback from the domain-specific discriminator, the generator synthesizes an image of the target domain (*e.g.* breeds) while respecting the style (*e.g.* fur patterns) of the reference image and the content (*e.g.* pose) of the source image.

3.2. Learning to produce domain labels and encode style features

The guiding network E consists of two branches, E_C and E_S, each of which learns to provide domain labels and style codes, respectively. In experiments, we compare our guiding network against straightforward approaches, *i.e.*, K-means on image or feature space.

Unsupervised domain classification. The discriminator requires a target domain label to provide useful gradients to the generator for translating an image into the target domain. Therefore, we adopt the differentiable clustering technique [15] that maximizes the mutual information (MI) between an image $\mathbf{x}$ and its randomly augmented version $\mathbf{x}^+$. The optimum of the mutual information $I(\mathbf{p}, \mathbf{p}^+)$ is reached when the entropy $H(\mathbf{p})$ is maximum and the conditional entropy $H(\mathbf{p}|\mathbf{p}^+)$ is minimum, where $\mathbf{p} = E_C(\mathbf{x})$ represents the softmax output from E_C, indicating a probability vector of $\mathbf{x}$ over $\hat{K}$ domains. Please refer to Section 4.2 for more details about $\hat{K}$. Maximizing MI encourages E_C to assign the same domain label to the pair ($\mathbf{x}$ and $\mathbf{x}^+$) while evenly distributing entire samples to all domains. Formally, E_C maximizes the mutual information:

$$\mathcal{L}_{MI} = I(\mathbf{p}, \mathbf{p}^+) = I(\mathbf{P}) = \sum_{i=1}^{\hat{K}} \sum_{j=1}^{\hat{K}} \mathbf{P}_{ij} \ln \frac{\mathbf{P}_{ij}}{\mathbf{P}_i \mathbf{P}_j}, \quad (1)$$
$$s.t.\ \mathbf{P} = \mathbb{E}_{\mathbf{x}^+ \sim f(\mathbf{x}) | \mathbf{x} \sim p_{data}(\mathbf{x})} [E_C(\mathbf{x}) \cdot E_C(\mathbf{x}^+)^T].$$

Here, f is a composition of random augmentations such as random cropping and affine transformation. $\mathbf{P}_i = \mathbf{P}(\mathbf{p} = i)$ denotes the $\hat{K}$-dimensional marginal probability vector, and $\mathbf{P}_{ij} = \mathbf{P}(\mathbf{p} = \texttt{argmax}(i), \mathbf{p}^+ = \texttt{argmax}(j))$ denotes the joint probability. To provide a deterministic one-hot label to the discriminator, we use the $\texttt{argmax}$ operation (i.e. $y = \texttt{argmax}(E_C(\mathbf{x}))$). Note that the mutual information is one way to implement TUNIT, and any differentiable clustering methods can be adopted such as SCAN [32].

Style encoding and improved domain classification. To perform a reference-guided image translation, the generator needs to understand the style features of the given image. In our framework, E_S encodes an image into a style code $\mathbf{s}$, which is later used to guide the generator for image translation. Here, to learn the style representation, we use the contrastive loss [8]:

$$\mathcal{L}^E_{style} = -\log \frac{\exp(\mathbf{s} \cdot \mathbf{s}^+ / \tau)}{\sum_{i=0}^{N} \exp(\mathbf{s} \cdot \mathbf{s}_i^- / \tau)}, \quad (2)$$

where $\mathbf{x}$ and $\mathbf{x}^+$ denote an image and randomly augmented version of $\mathbf{x}$, respectively, and $\mathbf{s} = E_S(\mathbf{x})$. 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 ($\mathbf{s}$, $\mathbf{s}_i^-$). We adopt a queue to store the negative codes $\mathbf{s}_i^-$ of the previously sampled images as MoCo [8]. By doing so, we can conduct the contrastive learning efficiently without large batch sizes [29].

Interestingly, we find that E_S also helps the unsupervised domain classification task—$(-\mathcal{L}_{MI} + \mathcal{L}^E_{style})$ significantly improves the quality of the clustering, compared to using only $-\mathcal{L}_{MI}$, which is the original IIC [15]. Since E_S shares the embeddings with E_C, imposing the contrastive loss on the style codes improves the representation power

of the shared embeddings. This is especially helpful when samples are complex and diverse, and IIC solely fails to scale up (*e.g.*, AnimalFaces [22]). To evaluate the effect of Eq.(2), we measure the accuracy and the ratio of the inter-variance over the intra-variance (IOI) in terms of the cosine similarity of each clustering result. The clustering result can be more discriminative when the intra-variance and the inter-variance become low and high, respectively. Therefore, the higher IOI indicates a more discriminative clustering result. The table below summarizes the accuracies and IOI on AnimalFaces-10 and Food-10.

	AnimalFaces-10		**Food-10**	
	IIC	IIC + Eq.(2)	IIC	IIC + Eq.(2)
IOI	2.05	3.04	1.34	2.50
Accuracy	68.0%	85.0%	54.2%	86.0%

For both datasets, IIC with Eq.(2) shows a significantly higher accuracy and a higher IOI value. Based on this analysis, we choose to use both $-\mathcal{L}_{MI}$ and $\mathcal{L}^{E}_{style}$ for training the guiding network.

3.3. Learning to translate images

We describe how to perform the unsupervised image-to-image translation under the guidance of our guiding network. For successful translation, the model should provide realistic images containing the visual feature of the target domain. To this end, we adopt three losses for training the generator G: 1) adversarial loss to produce realistic images, 2) style contrastive loss that encourages the model not to ignore the style codes, 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 a variant of conditional discriminator, the multi-task discriminator [25]. It is designed to conduct discrimination for each domain simultaneously. During training, its gradient is calculated only with the loss for estimating the input domain. For the domain label of the input, we utilize the pseudo label from the guiding network. Formally, given the pseudo label $\tilde{y}$ for a reference image $\tilde{\mathbf{x}}$, we train our generator G and multi-task discriminator D via the adversarial loss:

$$\begin{aligned}\mathcal{L}_{adv} = &\mathbb{E}_{\tilde{\mathbf{x}}\sim p_{data}(\mathbf{x})}[\log D_{\tilde{y}}(\tilde{\mathbf{x}})] \\ &+ \mathbb{E}_{\mathbf{x},\tilde{\mathbf{x}}\sim p_{data}(\mathbf{x})}[\log(1 - D_{\tilde{y}}(G(\mathbf{x},\tilde{\mathbf{s}})))],\end{aligned} \tag{3}$$

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

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

$$\mathcal{L}^{G}_{style} = \mathbb{E}_{\mathbf{x},\tilde{\mathbf{x}}\sim p_{data}(\mathbf{x})}\left[-\log \frac{\exp(\mathbf{s}'\cdot\tilde{\mathbf{s}})}{\sum_{i=0}^{N}\exp(\mathbf{s}'\cdot\mathbf{s}_i^{-}/\tau)}\right]. \tag{4}$$

Here, $\mathbf{s}' = E_S(G(\mathbf{x},\tilde{\mathbf{s}}))$ denotes the style code of the translated image $G(\mathbf{x},\tilde{\mathbf{s}})$ and $\mathbf{s}_i^{-}$ denotes the negative style codes, which are from the same queue used in Eq. (2). Besides, the same training scheme of MoCo [8] is applied for the generator training as Eq. (2). This loss guides the generated image $G(\mathbf{x},\tilde{\mathbf{s}})$ to have a style similar to the reference image $\tilde{\mathbf{x}}$ and dissimilar to negative (other) samples. By doing so, we avoid the degenerated solution where the encoder maps all the images to the same style code of the reconstruction loss [5] based on L1 or L2 norm. Eqs. (2) and (4) are based on contrastive loss, but they are used for different purposes. Please refer to Appendix 12. for more discussion.

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

$$\mathcal{L}_{rec} = \mathbb{E}_{\mathbf{x}\sim p_{data}(\mathbf{x})}[\|\mathbf{x} - G(\mathbf{x},\mathbf{s})\|_1]. \tag{5}$$

This objective not only ensures the generator G to preserve domain-invariant characteristics (e.g., pose) of its source 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 $\mathbf{x}$.

Overall objective. Finally, we train the three networks jointly as follows:

$$\begin{aligned}\mathcal{L}_D &= -\mathcal{L}_{adv}, \\ \mathcal{L}_G &= \mathcal{L}_{adv} + \lambda^{G}_{style}\mathcal{L}^{G}_{style} + \lambda_{rec}\mathcal{L}_{rec}, \\ \mathcal{L}_E &= \mathcal{L}_G - \lambda_{MI}\mathcal{L}_{MI} + \lambda^{E}_{style}\mathcal{L}^{E}_{style}\end{aligned} \tag{6}$$

where λ's are hyperparameters. Note that our guiding network E receives feedback from L_G, which is essential for our method. We discuss the effect of feedback to E on performance in Section 4.1.

4. Experiments

We first investigate the effect of each component of TUNIT (Section 4.1). We quantitatively and qualitatively compare the performance on labeled datasets. We show that TUNIT is robust against the choice of hyperparameters (e.g. the preset number of clusters, $\tilde{K}$) and extends well to the semi-supervised scenario (Section 4.2). Lastly, we move on to unlabeled datasets to validate our model in the unsupervised scenario in the wild (Section 4.3).

Datasets. For the labeled datasets, we select ten classes among 149 classes of AnimalFaces and 101 classes of Food-101, which we call AnimalFaces-10 and Food-10, respectively. Here, the labels are used only for the evaluation purpose except for the semi-supervised setting. For the unlabeled datasets, we use AFHQ, FFHQ, and LSUN Car [5, 16, 37], which do not have any or are missing with fine-grained labels. Specifically, AFHQ roughly has three groups (*i.e.*, dog, cat and wild), but each group contains

	Configuration	AnimalFaces-10			Food-10		
		mFID	D & C	Acc.	mFID	D & C	Acc.
A	Baseline FUNIT (supervised)	74.0	0.749 / 0.671	**1.000**	68.4	0.989 / 0.782	**1.000**
B	(A) + Improved G & D (supervised)	**46.2**	**0.896 / 0.732**	**1.000**	**57.6**	**1.284 / 0.857**	**1.000**
C	(B) + K-means on image space	110.7	0.822 / 0.615	0.215	90.7	0.849 / 0.648	0.201
D	(B) + K-means on feature space	76.2	0.770 / 0.597	0.428	64.6	0.968 / 0.808	0.331
E	(B) + Differentiable clustering	73.5	0.940 / 0.588	0.680	64.2	1.038 / 0.819	0.542
F	TUNIT w/ sequential training	**46.0**	1.060 / 0.789	**0.850**	61.1	0.908 / 0.777	**0.860**
G	TUNIT w/ joint training	47.7	**1.039 / 0.805**	0.841	**52.2**	**1.079 / 0.875**	0.848

Table 1: Comparison results. mFID, Density / Coverage (D&C), and classification accuracy (Acc) of each training configuration. Note that the configurations (A) - (B) use ground-truth class labels, while (C) - (G) use pseudo-labels. We **bold** the best results separately for supervised and unsupervised settings. For D& C, we boldify the one which has the best coverage that has a clear maximum.

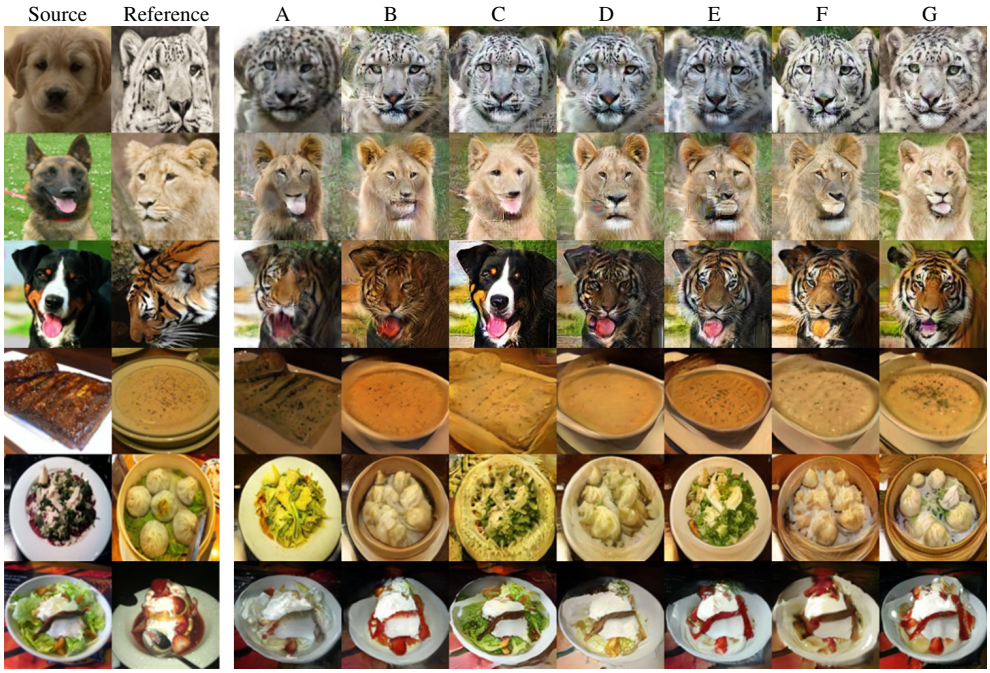
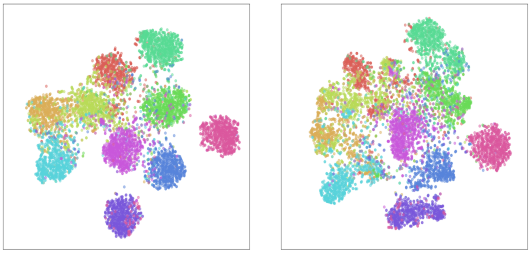


Figure 3: Qualitative comparison of translation results using each configuration in Table 1. Here, B reflects the style feature (*e.g.* species or type of food) of the reference images while A does not. The model C performs much worse than A and B in that it overly adopts the source image, not adequately merging styles and contents from both sides. The model D generates more plausible images than C but fails to reflect the characteristics of the reference images. For example, D on fifth row does not look like *several pieces of dumpling* due to its shape and dish color, meaning that the reference styles are not properly reflected. Similarly, E also fails to generate the dumpling in the fifth row. TUNIT with sequential training F reflects the visual features of each reference on both datasets. However, in terms of visual fidelity, we observe that G consistently outperforms F. Akin to the quantitative results, TUNIT achieves equivalent or even better visual quality than the set-level supervised model A and B.

diverse species and these species labels are not provided. FFHQ and LSUN Car contain various human faces and cars without any labels, respectively.

Evaluation metrics. We report two scores to assess the generated images. First, to provide a general sense of image quality, we use the mean of class-wise Fréchet Inception Distance (mFID) [10]. It can avoid the degenerate case of the original FID, which assigns a good score when the model conveying the source image as is. Additionally, to provide a finer assessment of the generated images, we report Density and Coverage (D&C) [27]. D&C separately evaluates the fidelity and the diversity of the model outputs, which is also known to be robust against outliers and model hyperparameters (e.g. the number of samples used for evaluation). A lower mFID score means better image quality, and D&C scores that are bigger or closer to 1.0 indicate better fidelity and diversity, respectively. Please refer to Appendix 3. for the detailed information.



(a) t-SNE of $\hat{K}$=10 (b) t-SNE of $\hat{K}$=20

$\hat{K}$	AnimalFaces-10		Food-10	
	mFID	D & C	mFID	D & C
1	129.6	0.561 / 0.512	95.1	1.113 / 0.771
4	77.7	0.879 / 0.738	67.4	0.851 / 0.785
7	62.7	1.016 / 0.729	52.7	1.024 / 0.846
10	**47.7**	**1.039 / 0.805**	**52.2**	**1.079 / 0.875**
13	56.8	0.993 / 0.720	54.8	0.970 / 0.845
16	54.1	1.093 / 0.782	54.8	1.029 / 0.857
20	55.4	1.019 / 0.778	57.7	0.937 / 0.846
50	63.8	0.858 / 0.701	60.8	1.067 / 0.837
500	67.2	0.921 / 0.694	63.2	0.986 / 0.826
1000	66.9	0.908 / 0.707	60.7	0.945 / 0.845

Table 2: t-SNE visualization of the model with (a) $\hat{K}$=10 and (b) $\hat{K}$=20 trained on AnimalFaces-10 and quantitative evaluation of our method by varying the number of pseudo domains $\hat{K}$. Each point is colored with the ground-truth labels. As shown in t-SNE visualizations, even if $\hat{K}$ is set to overly larger than the actual number of domains, the guiding network clusters the domains reasonably well. For D& C, we bold the one which has the best coverage that has a clear maximum.

4.1. Comparative Evaluation on Labeled Datasets

Table 1 summarizes the effect of each component of TUNIT and rigorous comparisons with the state-of-the-art supervised method, FUNIT. First, we report the set-level supervised performance of FUNIT and its variant (Table 1). Here, A is the original FUNIT and B denotes the modified FUNIT using our architecture (*e.g.* we do not use PatchGAN discriminator), which brings a large improvement over every score on both datasets. One simple way to extend B to the unsupervised scenario is to add an off-the-shelf clustering method and use its estimated labels instead of the ground truth. We employ K-means clustering on the image space for C, and the pretrained feature space for D. Here, we use ResNet-50 [9] features trained with MoCo v2 [3] on ImageNet. Not surprisingly, because the estimated labels are inaccurate, the overall performance significantly drops. Although using the pretrained features helps a little, not only is it far from the set-level supervised performance but it requires three steps to train the entire model, which complicates the application. This can be partially addressed by employing the differentiable clustering method [15], which trains VGG-11BN [30] with mutual information maximization from scratch that makes E. This reduces the number of training steps from three to two and provides better label

Figure 4: **Cross-domain attribute translation using 0.1% of labeled samples.** As a practical application, we also apply TUNIT to FFHQ with few labels as a form of cross-domain translation. For each attribute, we train TUNIT separately (there are four models.). To this end, we manually label 35 images for each domain – one contains the attribute and another does not contain it. Then, we train TUNIT with 70 labeled samples (0.1% of the dataset) and remaining unlabeled samples. To add or remove the attribute, we use the average style vector of each domain. This result shows that TUNIT can greatly reduce the labeling cost.

estimation, which enables the model to approach the performance of original FUNIT A. However, as seen in the coverage score, the sample diversity is unsatisfactory.

Finally, we build TUNIT by introducing the guiding network and the new objective functions described in Section 3. The changes significantly improve the accuracy on both datasets, particularly achieving similar mFID of the improved set-level supervised model B. Our final model, G matches or outperforms mFID and D&C of B. This is impressive because B utilizes oracles for training while G has no labels. Notably, TUNIT can improve the coverage by 0.073 (7%p) on AnimalFaces-10 than B. We conjecture that TUNIT benefits from the guiding network, which jointly learns the style encoding and clustering with the shared encoder. Because their clustering modules are as powerful as TUNIT, the performance drawback of C, D and E supports that the feedback from style encoding is a key success factor of TUNIT. By comparing F and G, we confirm that they are comparable in terms of clustering and G is more stable in terms of inter-dataset performance. Therefore, we adopt the joint training of style encoder and clustering as our final model (G). In addition, we investigate the effect of joint training between GAN and the guiding network by removing the adversarial loss for training the guiding network. It directly degrades the performance; mFID changes from 47.7 to 63.0 on AnimalFaces-10. It indicates that our training scheme takes an important portion of performance gains. Qualitative results also show the superiority of TU-

Model	AnimalFaces-10								Food-10							
	1%	2%	4%	8%	20%	40%	60%	80%	1%	2%	4%	8%	20%	40%	60%	80%
FUNIT [22]	179.8	174.3	154.3	144.0	124.4	106.4	96.0	79.6	195.7	159.2	141.2	135.2	111.4	85.8	74.8	70.3
SEMIT [34]	68.9	63.0	63.5	60.3	57.3	64.3	66.1	62.8	70.7	66.7	65.3	65.1	65.2	62.8	61.6	64.1
TUNIT (ours)	**42.0**	**42.6**	**43.9**	**46.2**	**42.0**	**42.6**	**43.9**	**46.2**	**53.6**	**56.2**	**52.8**	**53.4**	**53.6**	**56.2**	**52.8**	**53.4**
Model	**AnimalFaces-149**								**Food-101**							
	1%	2%	4%	8%	20%	40%	60%	80%	1%	2%	4%	8%	20%	40%	60%	80%
FUNIT [22]	147.3	133.6	116.9	101.7	102.7	112.4	102.9	100.3	136.8	94.9	80.9	72.6	67.6	66.5	**66.4**	67.4
SEMIT [34]	137.9	123.1	134.0	120.3	113.5	114.5	116.9	116.6	**74.7**	81.7	**71.2**	**72.5**	68.4	68.7	69.8	69.5
TUNIT (ours)	**104.9**	**99.9**	**96.8**	**87.5**	**84.0**	**79.9**	**80.1**	**78.4**	75.0	**74.4**	74.9	75.2	**64.7**	**64.0**	68.2	**66.8**

Table 3: Quantitative evaluation (mFID) when few labels are available during training. We note that TUNIT is the same as G in Table 1

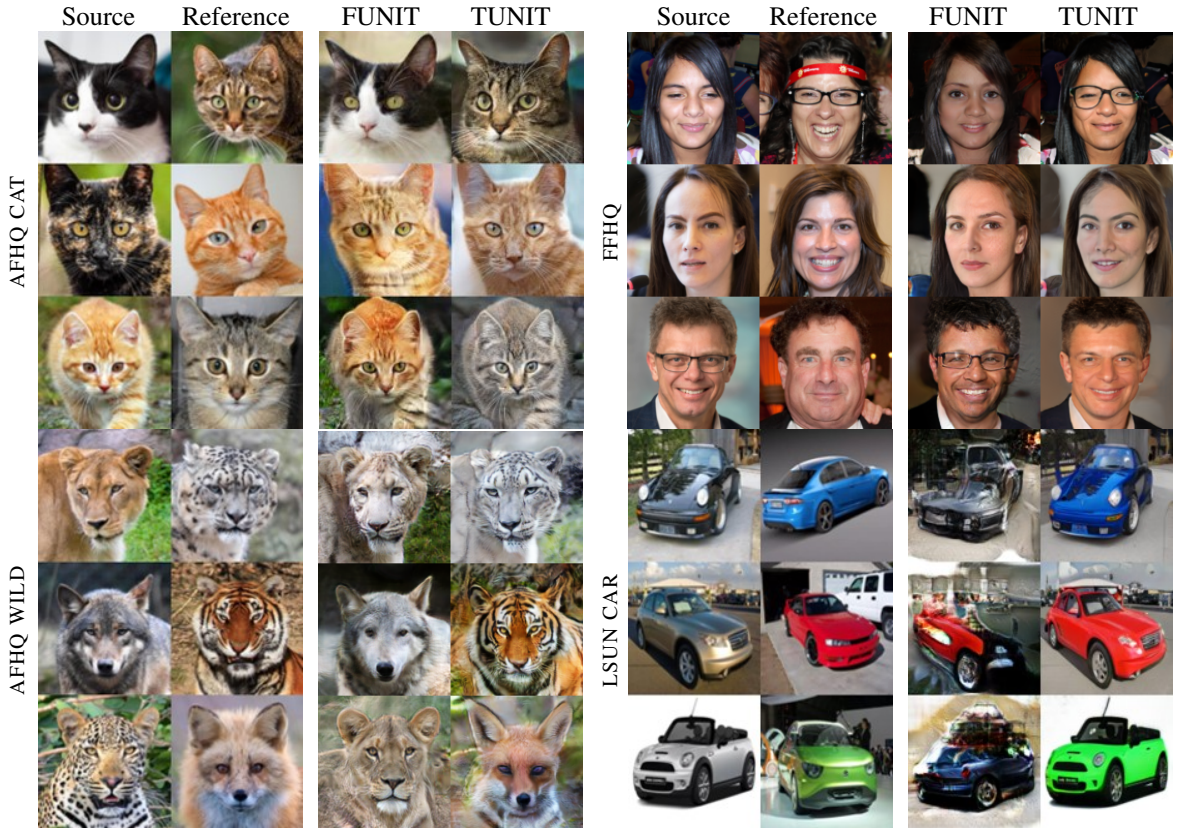


Figure 5: Reference-guided image translation results on unlabeled datasets.

NIT over competitors (Figure 3).

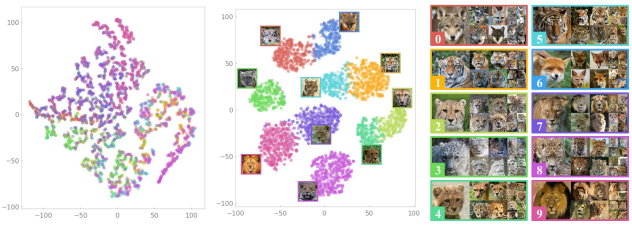
4.2. Analysis on Generalizability

Robustness to various $\hat{K}$'s. When TUNIT conducts clustering for estimating domain labels, the number of clusters $\hat{K}$ can affect the performances. Here, we study the effects on different $\hat{K}$ on the labeled datasets and report them in Table 2. For the qualitative comparison, please refer to Appendix 1. As expected, the model performs best in terms of mFID when $\hat{K}$ equals to the ground truth K (*i.e.* $\hat{K}$=10).

One thing to note here is that TUNIT performs reasonably well for a sufficiently large $\hat{K}$ ($\geq$ 7). More interestingly, even with 100 times larger $\hat{K}$ than the actual number of the domains, TUNIT still works well on both datasets. This trend is also seen in the t-SNE visualization(Table 2).

From this study, we conclude that TUNIT is relatively robust against $\hat{K}$ as long as it is sufficiently large. Thus, in practice, we suggest to use a sufficiently large $\hat{K}$ or to study different $\hat{K}$'s in log scale for finding the optimal model.

With Few labels. We also investigate whether or not TUNIT is effective for a more practical scenario, semi-supervised image translation. To utilize the labels, AnimalFaces and Food datasets are chosen for this experiment. Specifically, AnimalFaces-10 and Food-10 are used for evaluating the models on the small datasets while AnimalFaces-149 and Food-101 are used for assessing the models on the large datasets. We partition the dataset $\mathcal{D}$ into the labeled set $\mathcal{D}_{sup}$ and the unlabeled set $\mathcal{D}_{un}$ with varying ratio $\gamma = |\mathcal{D}_{sup}|/|\mathcal{D}|$. For the semi-supervising setting, we train E_C of TUNIT by an additional cross-



(a) t-SNE of FUNIT (L) / TUNIT (R) (b) Example images

Figure 6: t-SNE visualization of style space trained on AFHQ Wild. Since AFHQ Wild does not have ground-truth labels, each point is colored with the guiding network's prediction. Although we set the number of domains to be larger ($\hat{K} = 10$) than it seems to have, the network practically creates six clusters by closely locating clusters of one species.

entropy loss between the ground truth domain labels and the predicted domain labels on $\mathcal{D}_{sup}$. Besides, the true domain labels for $\mathcal{D}_{sup}$ are utilized for training the domain-specific discriminator. As a counterpart in this scenario, FUNIT [22] and SEMIT [34] are selected because both models can be applied to the semi-supervised image translation (SEMIT achieves the current state-of-the-art performance in the semi-supervised setting). We train the two competitors and TUNIT by changing γ from 0.01 to 0.8 and report the results in Table 3. For the small datasets, the performance of FUNIT significantly degrades as γ decreases. Meanwhile, TUNIT and SEMIT produce relatively similar mFID scores despite γ decreases. Even when SEMIT maintains mFID, TUNIT significantly outperforms SEMIT by 20% of mFID on small datasets. For the large datasets, TUNIT either outperforms or is comparable to the competitors. Especially, on AnimalFaces-149, TUNIT clearly outperforms both competitors. By the experiments on the semi-supervised setting, we conclude that TUNIT can be easily adapted to the semi-supervised image translation with the simple modification (*i.e*. adding the supervised training on the labeled samples), and serve as a strong baseline model. As a result, TUNIT for the semi-supervised setting achieves impressive performance, which is comparable to the state-of-the-art semi-supervised translation method. We provide the qualitative comparison in the appendix.

4.3. Validation on Unlabeled Dataset

Finally, we evaluate TUNIT on the unlabeled datasets (AFHQ, FFHQ and LSUN-Car), having no clear separations of the domains. For AFHQ, we train three individual models for *dog*, *cat* and *wild*. For all experiments, we use FUNIT as a baseline, where all the labels for training are regarded the same as one. We set the number of clusters $\hat{K}$=10 for all the TUNIT models.

Figure 5 demonstrates the results. We observe that the results of TUNIT adequately reflect the style feature of the references such as the textures of cats or cars and the species of the wilds. Although FFHQ has no clear domain distinctions, TUNIT captures the existence of glasses or smile as domains, and then add or remove glasses or smile. However, FUNIT performs much worse than TUNIT in this truly unsupervised scenario. For example, FUNIT outputs the inputs as is (cats and wilds) or insufficiently reflects the species (third row of AFHQ Wild). For FFHQ, despite that FUNIT makes some changes, the changes are not interpreted as meaningful domain translations. For LSUN Car, FUNIT fails to keep the fidelity.

We also visualize the style space of both models to qualitatively assess the quality of the representation. Figure 6 shows the t-SNE maps trained on AFHQ Wild and the examples of each cluster; the sample color corresponds to the box color of representative images. Surprisingly, TUNIT organizes the samples according to the species where it roughly separates the images into six species. Although we set $\hat{K}$ to be overly large, the model represents one species into two domains where those two domains position much closely (*e.g*. tiger). From these results, we confirm that the highly disentangled, meaningful style features can be an important factor in the success of our model. On the other hand, the style features of FUNIT hardly learn meaningful domains so that the model cannot conduct the translation properly as shown in Figure 5. Because of the page limit, we include more results including qualitative comparison and the t-SNE visualization in Appendix 8,9,10 and 11.

5. Conclusion

We argue that an unsupervised image translation should denote a task that does not utilize any kinds of supervision, neither image-level (*i.e*. paired) nor set-level (*i.e*. unpaired). In this regime, most of the previous studies fall into the set-level supervised framework, using the domain information at a minimum. In this paper, we proposed TUNIT, a truly unsupervised image translation method. By exploiting synergies between clustering and representation learning, TUNIT finds pseudo labels and style codes so that it can translate images without using any external information. The experimental results show that TUNIT can successfully perform an unsupervised image translation while being robust against hyperparameter changes (*e.g*., the preset number of clusters, $\hat{K}$). Our model is easily extended to the semi-supervised setting, providing comparable results to the state-of-the-art semi-supervised method. These highlight TUNIT has great potential in practical applications.

Acknowledgements. All experiments were conducted on NAVER Smart ML (NSML) [17] platform. This research was supported by the NRF Korea funded by the MSIT (NRF-2019R1A2C2006123), the IITP grant funded by the MSIT (2020-0-01361, YONSEI University, 2020-0-01336, Artificial Intelligence graduate school support (UNIST)), and the Korea Medical Device Development Fund grant (Project Number: 202011D06).

References

[1] H. Bahng, S. Chung, S. Yoo, and J. Choo. Exploring unlabeled faces for novel attribute discovery. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5820–5829, 2020. 2

[2] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 2

[3] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 2, 6

[4] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul 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] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 2, 4

[6] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. In *International Conference on Learning Representations*, 2018. 2

[7] Raia Hadsell, Sumit Chopra, and Yann 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

[8] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020. 2, 3, 4

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6

[10] 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 *Advances in neural information processing systems*, pages 6626–6637, 2017. 5

[11] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. In *International Conference on Learning Representations*, 2019. 2

[12] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *International conference on machine learning*, pages 1989–1998, 2018. 1, 2

[13] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 172–189, 2018. 1, 2

[14] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei 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

[15] Xu Ji, João F Henriques, and Andrea 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, 6

[16] Tero Karras, Samuli Laine, and Timo 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. 4

[17] Hanjoo Kim, Minkyu Kim, Dongjoo Seo, Jinwoong Kim, Heungseok Park, Soeun Park, Hyunwoo Jo, KyungHyun Kim, Youngil Yang, Youngkwan Kim, et al. NSML: Meet the MLaaS platform with a real-world case study. *arXiv preprint arXiv:1810.09957*, 2018. 8

[18] Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jung Kwon Lee, and Jiwon 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. 1, 2

[19] Orest Kupyn, Volodymyr Budzan, Mykola Mykhailych, Dmytro Mishkin, and Jiří 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

[20] Hsin-Ying Lee, Hung-Yu Tseng, Qi Mao, Jia-Bin Huang, Yu-Ding Lu, Maneesh Singh, and Ming-Hsuan Yang. Drit++: Diverse image-to-image translation via disentangled representations. *International Journal of Computer Vision*, pages 1–16, 2020. 1

[21] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. In *Advances in neural information processing systems*, pages 700–708, 2017. 1, 2

[22] Ming-Yu Liu, Xun Huang, Arun Mallya, Tero Karras, Timo Aila, Jaakko Lehtinen, and Jan 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, 7, 8

[23] Steven Liu, Tongzhou Wang, David Bau, Jun-Yan Zhu, and Antonio Torralba. Diverse image generation via self-conditioned gans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[24] Mario Lučić, Michael Tschannen, Marvin Ritter, Xiaohua Zhai, Olivier Bachem, and Sylvain Gelly. High-fidelity image generation with fewer labels. In *International Conference on Machine Learning*, pages 4183–4192, 2019. 2

[25] Lars Mescheder, Sebastian Nowozin, and Andreas Geiger. Which training methods for gans do actually converge? In *ICML*, 2018. 4

[26] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014. 1

[27] Muhammad Ferjad Naeem, Seong Joon Oh, Youngjung Uh, Yunjey Choi, and Jaejun Yoo. Reliable fidelity and diversity metrics for generative models. 2020. 5

[28] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *CVPR*, 2019. 1

[29] Nikunj Saunshi, Orestis Plevrakis, Sanjeev Arora, Mikhail Khodak, and Hrishikesh Khandeparkar. A theoretical analysis of contrastive unsupervised representation learning. In *International Conference on Machine Learning*, pages 5628–5637, 2019. 2, 3

[30] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015. 6

[31] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. In *Advances in neural information processing systems*, pages 3483–3491, 2015. 1

[32] Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, Marc Proesmans, and Luc Van Gool. Scan: Learning to classify images without labels. In *European Conference on Computer Vision (ECCV)*, 2020. 2, 3

[33] 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. 1

[34] Yaxing Wang, Salman Khan, Abel Gonzalez-Garcia, Joost van de Weijer, and Fahad Shahbaz Khan. Semi-supervised learning for few-shot image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4453–4462, 2020. 2, 7, 8

[35] I Zeki Yalniz, Hervé Jégou, Kan Chen, Manohar Paluri, and Dhruv Mahajan. Billion-scale semi-supervised learning for image classification. *arXiv preprint arXiv:1905.00546*, 2019. 2

[36] Dingdong Yang, Seunghoon Hong, Yunseok Jang, Tiangchen Zhao, and Honglak Lee. Diversity-sensitive conditional generative adversarial networks. In *International Conference on Learning Representations*, 2019. 2

[37] Fisher Yu, Yinda Zhang, Shuran Song, Ari Seff, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015. 4

[38] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei 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

[39] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. Toward multimodal image-to-image translation. In *Advances in neural information processing systems*, pages 465–476, 2017. 1