Semi-supervised Learning for Few-shot Image-to-Image Translation

Yaxing Wang[1]*, Salman Khan[2], Abel Gonzalez-Garcia[1], Joost van de Weijer[1], Fahad Shahbaz Khan[2,3]
[1] Computer Vision Center, Universitat Autònoma de Barcelona, Spain
[2] Inception Institute of Artificial Intelligence, UAE [3] CVL, Linköping University, Sweden
{yaxing,agonzalez,joost}@cvc.uab.es, salman.khan@inceptioniai.org, fahad.khan@liu.se

Abstract

In the last few years, unpaired image-to-image translation has witnessed remarkable progress. Although the latest methods are able to generate realistic images, they crucially rely on a large number of labeled images. Recently, some methods have tackled the challenging setting of few-shot image-to-image translation, reducing the labeled data requirements for the target domain during inference. In this work, we go one step further and reduce the amount of required labeled data also from the source domain during training. To do so, we propose applying semi-supervised learning via a noise-tolerant pseudo-labeling procedure. We also apply a cycle consistency constraint to further exploit the information from unlabeled images, either from the same dataset or external. Additionally, we propose several structural modifications to facilitate the image translation task under these circumstances. Our semi-supervised method for few-shot image translation, called SEMIT, *achieves excellent results on four different datasets using as little as 10% of the source labels, and matches the performance of the main fully-supervised competitor using only 20% labeled data. Our code and models are made public at: https://github.com/yaxingwang/SEMIT.*

1. Introduction

Image-to-image (I2I) translations are an integral part of many computer vision tasks. They include transformations between different modalities (*e.g.*, from RGB to depth [27]), between domains (*e.g.*, horses to zebras [46]) or editing operations (*e.g.*, artistic style transfer [13]). Benefiting from large amounts of labeled images, I2I translation has obtained great improvements on both paired [8, 15, 19, 40, 47] and unpaired image translation [2, 7, 22, 42, 44, 46]. Recent research trends address relevant limitations of earlier approaches, namely diversity and scalability. Current methods [1, 18, 25] improve over the single-sample limitation of deterministic models by generating diverse translations given an input image. The scalability problem has also been successfully alleviated [9, 33, 34, 39], enabling translations across several domains using a single model. Nonetheless, these approaches still suffer from two issues. First, the target domain is required to contain the same categories or attributes as the source domain at test time, therefore failing to scale to unseen categories (see Fig. 1(a)). Second, they highly rely upon having access to vast quantities of labeled data (Fig. 1(a, b)) at train time. Such labels provide useful information during the training process and play a key role in some settings (*e.g.* scalable I2I translation).

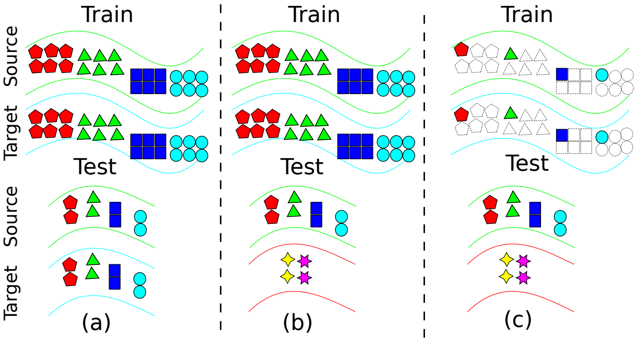


Figure 1. Comparison between unpaired I2I translation scenarios. Each colored symbol indicates a different image label, and dashed symbols represent unlabeled data. (a) *Standard* [9, 18, 46]: target classes are the same as source classes and all are seen during training. (b) *Few-shot* [28]: actual target classes are different from source classes and are unseen during training. Only a few examples of the unseen target classes are available at test time. For training, source classes act temporarily as target classes. (c) *Few-shot semi-supervised* (Ours): same as few-shot, but the source domain has only a limited amount of labeled data at train time.

Recently, several works have studied I2I translation given a few images of the *target class* (as in Fig. 1(b)). Benaim and Wolf [3] approach one-shot I2I translation by first training a variational autoencoder for the seen domain and then adapting those layers related to the unseen domain. ZstGAN [26] introduces zero-shot I2I translation, employing the annotated attributes of unseen categories instead

[*] Work done as an intern at Inception Institute of Artificial Intelligence

of the labeled images. FUNIT [28] proposes few-shot I2I translation in a multi-class setting. These models, however, need to be trained using large amounts of hand-annotated ground-truth labels for images of the source domain (Fig. 1 (b)). Labeling large-scale datasets is costly and time consuming, making those methods less applicable in practice. In this paper, we overcome this limitation and explore a novel setting, introduced in Fig. 1(c). Our focus is few-shot I2I translation in which only limited labeled data is available *from the source classes during training*.

We propose using semi-supervised learning to reduce the requirement of labeled source images and effectively use unlabeled data. More concretely, we assign pseudo-labels to the unlabeled images based on an initial small set of labeled images. These pseudo-labels provide soft supervision to train an image translation model from source images to unseen target domains. Since this mechanism can potentially introduce noisy labels, we employ a pseudo-labeling technique that is highly robust to noisy labels. In order to further leverage the unlabeled images from the dataset (or even external images), we use a cycle consistency constraint [46]. Such a cycle constraint has generally been used to guarantee the content preservation in unpaired I2I translation [22, 44, 46, 28], but we propose here also using it to exploit the information contained in unlabeled images.

Additionally, we introduce further structural constraints to facilitate the I2I translation task under this challenging setting. First, we consider the recent Octave Convolution (OctConv) operation [6], which disentangles the latent representations into high and low frequency components and has achieved outstanding results for some discriminative tasks [6]. Since I2I translation mainly focuses on altering high-frequency information, such a disentanglement could help focalize the learning process. For this reason, we propose a novel application of OctConv for I2I translation, making us the first to use it for a generative task. Second, we apply an effective entropy regulation procedure to make the latent representation even more domain-invariant than in previous approaches [18, 25, 28]. This leads to better generalization to target data. Notably, these techniques are rather generic and can be easily incorporated in many current I2I translation methods to make the task easier when there is only limited data available.

Experiments on four datasets demonstrate that the proposed method, named *SEMIT*, consistently improves the performance of I2I translation using only 10% to 20% of the labels in the data. Our main contributions are:

- We are the first to approach few-shot I2I translation in a semi-supervised setting, reducing the amount of required labeled data for *both* source and target domains.
- We propose several crucial modifications to facilitate this challenging setting. Our modifications can be easily adapted to other image generation architectures.
- We extensively study the properties of the proposed approaches on a variety of I2I translation tasks and achieve significant performance improvements.

2. Related work

Semi-supervised learning. The methods in this category employ a small set of labeled images and a large set of unlabeled data to learn a general data representation. Several works have explored applying semi-supervised learning to Generative Adversarial Networks (GANs). For example, [31, 36] merge the discriminator and classifier into a single network. The generated samples are used as unlabeled samples to train the ladder network [31]. Springenberg [37] explored training a classifier in a semi-supervised, adversarial manner. Similarly, Li *et al.* [10] proposed Triple-GAN that plays minimax game with a generator, a discriminator and a classifier. Other works [11, 12] either learn two-way conditional distributions of both the labels and the images, or add a new network to predict missing labels. Recently, Lucic *et al.* [29] proposed bottom-up and top-down methods to generate high resolution images with fewer labels. To the best of our knowledge, no previous work addresses I2I translation to generate highly realistic images in a semi-supervised manner.

Zero/few-shot I2I translation. Several recent works used GANs for I2I translation with few test samples. Lin *et al*. proposed zero-shot I2I translation, ZstGAN [26]. They trained a model that separately learns domain-specific and domain-invariant features using pairs of images and captions. Benaim and Wolf [3] instead considered one image of the target domain as an exemplar to guide image translation. Recently, FUNIT [28] learned a model that performs I2I translation between seen classes during training and scales to unseen classes during inference. These methods, however, rely on vast quantity of labeled source domain images for training. In this work, we match their performance using only a small subset of the source domain labels.

3. Proposed Approach: SEMIT

Problem setting. Our goal is to design an unpaired I2I translation model that can be trained with minimal supervision (Fig. 1 (c)). Importantly, in the few-shot setting the target classes are unseen during training and their few examples are made available only during the inference stage. In contrast to previous state-of-the-art [28], which trains on a large number of labeled samples of the source domain (some of which act as 'target' during training), we assume only limited labeled examples of the source classes are available for training. The remaining images of the source classes are available as unlabeled examples.

Suppose we have a training set $\mathcal{D}$ with N samples. One portion of the dataset is labeled, $\mathcal{D}_l = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^{N_l}$, where

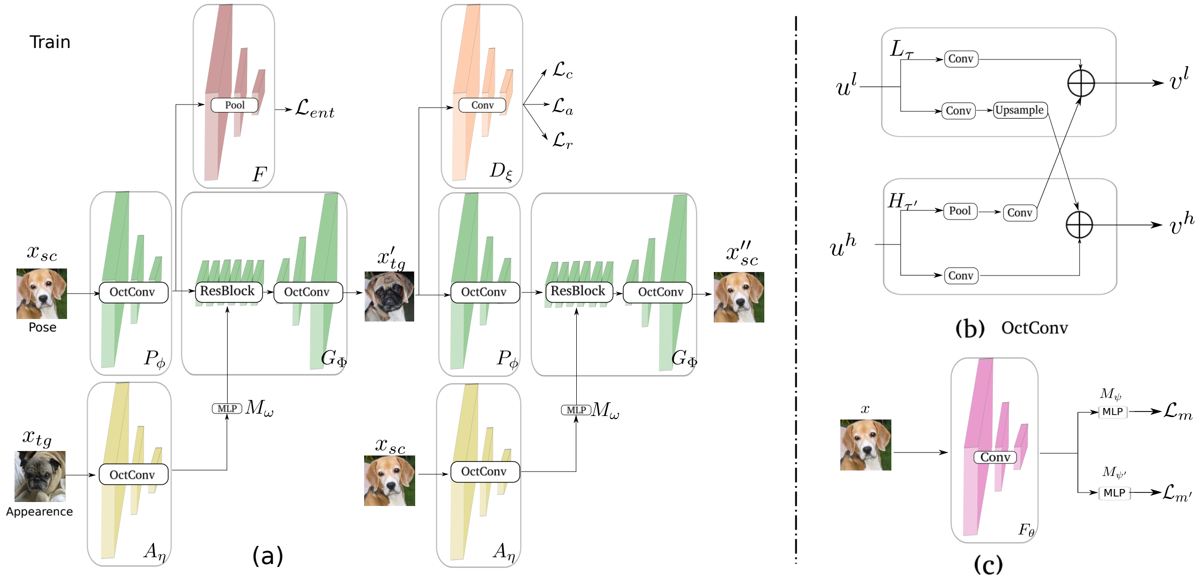


Figure 2. Model architecture for training. (a) The proposed approach is composed of two main parts: Discriminator D_ξ and the set of Pose encoder P_ϕ, Appearance encoder A_η, Generator G_Φ, Multilayer perceptron M_ω and feature regulator F. (b) The OctConv operation contains high-frequency block ($H_{\tau'}$) and low-frequency block (L_τ). (c) Noise-tolerant Pseudo-labeling architecture.

$\boldsymbol{x}_i \in \mathbb{R}^D$ denotes an image, $\boldsymbol{y}_i \in \{0,1\}^C : \mathbf{1}^\top \boldsymbol{y}_i = 1$ denotes a one-hot encoded label and C is the total number of classes. We consider a relatively larger unlabeled set, $\mathcal{D}_u = \{\boldsymbol{x}_i\}_{i=1}^{N_u}$, that is available for semi-supervised learning. Overall, the total number of images are $N = N_u + N_l$.

We initially conduct semi-supervised learning, where we learn a classifier to assign pseudo-labels $\tilde{\boldsymbol{y}}$ to the unlabeled data, generating a set $\tilde{\mathcal{D}} = \{(\boldsymbol{x}_i, \tilde{\boldsymbol{y}}_i)\}_{i=1}^{N}$, where $\tilde{\boldsymbol{y}}_i = \boldsymbol{y}_i$ for $\boldsymbol{x}_i \in \mathcal{D}_l$ i.e., for a sample whose ground-truth label is available. The pseudo-labels predicted by the model form a soft label-space, i.e., $\tilde{\boldsymbol{y}}_i \in [0,1]^C : \mathbf{1}^\top \tilde{\boldsymbol{y}}_i = 1$. Then, our method performs unsupervised multi-domain I2I translation on the set $\tilde{\mathcal{D}}$ with few labeled images and a large unlabeled set. The dual-mode training procedure is explained below.

3.1. Noise-tolerant Pseudo-labeling

The assigned pseudo-labels are used to train the I2I translator network in the next stage. Therefore, the labeling approach must avoid generating false predictions while being able to tolerate noise in the label space. To achieve these requisites, we develop a Noise-tolerant Pseudo-Labeling (NTPL) approach that is trained progressively with a soft-labeling scheme to avoid the noise accumulation problem.

As illustrated in Fig. 2 (c), our pseudo-labeling scheme consists of a feature extractor F_θ and a couple of classification heads, M_ψ and $M'_{\psi'}$. The semi-supervised labeling model is designed to suffice the following principles, (a) decision consolidation and (b) high-confidence sampling for a noise-tolerant pseudo-labeling. Firstly, the two classification heads are used to assess the uncertainty for a given unlabeled sample, i.e., a pseudo-label is considered valid only if both the classifier outputs agree with each other. Secondly, we add the pseudo-labels to the training set only if both classifier confidences are above a set threshold. Each classification head is trained using a loss $\mathcal{L}_c$ that is based on the probabilistic end-to-end noise correction framework of [23]. The overall classifier loss function is the sum of losses for classification heads M_ψ and $M'_{\psi'}$,

$$\mathcal{L}_c = \mathcal{L}_m + \mathcal{L}_{m'}. \tag{1}$$

For both classification heads M_ψ and $M'_{\psi'}$, the loss function consists of three components: (i) *Compatibility loss*, which tries to match the label distribution with the pseudo-label; (ii) *Classification loss*, which corrects the noise in labels; and (iii) *Entropy regulation loss*, which forces the network to peak at one category rather than being flat (i.e., confusing many classes). Below, we explain the loss components for $\mathcal{L}_m$ and the formulation for loss $\mathcal{L}_{m'}$ is analogous.

Compatibility loss. The compatibility loss encourages the model to make predictions that are consistent with the ground-truth or pseudo-labels. Since in many cases, the current estimates of labels are correct, this loss function avoids estimated labels far away from the assigned labels,

$$\mathcal{L}_{cmp} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C} \tilde{y}_{ij} \log(y^h_{ij}), \tag{2}$$

where $\boldsymbol{y}^h = \text{softmax}(\boldsymbol{y}')$ is the underlying label distribution for noisy labels and $\boldsymbol{y}'$ can be updated by back-propagation during training. The tunable variable $\boldsymbol{y}'$ is initialized with $\boldsymbol{y}' = K\tilde{\boldsymbol{y}}$, where K is a large scalar (1000).

Classification loss. We follow the operand-flipped KL-divergence formulation from [23], which was shown to improve robustness against noisy labels. This loss is given by,

$$\mathcal{L}_{cls} = \frac{1}{n}\sum_{i=1}^{N} \mathrm{KL}(M_\psi(F_\theta(\boldsymbol{x}_i))) \| \boldsymbol{y}_i^h). \tag{3}$$

Entropy regulation loss. Confused models tend to output less confident predictions that are equally distributed over several object categories. The entropy regulation loss forces the estimated output distribution to be focused on one class,

$$\mathcal{L}_{ent} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C} M_\psi(F_\theta(\boldsymbol{x}_i))_j \log\left(M_\psi(F_\theta(\boldsymbol{x}_i))_j\right). \tag{4}$$

The full loss of M_ψ is given by,

$$\mathcal{L}_m = \tau_{cls}\mathcal{L}_{cls} + \tau_{cmp}\mathcal{L}_{cmp} + \tau_{ent}\mathcal{L}_{ent}, \tag{5}$$

where τ_{cls}, τ_{cmp} and τ_{ent} are the hyper-parameters.

Training procedure. Our semi-supervised training procedure includes both labeled and pseudo-labeled examples. Therefore, we must select reliable pseudo-labels. Similar to existing work [35], we perform the following procedure to reach this goal. Initially, we train the model (Fig. 2 (c)) with only cleanly labeled images i.e., without any pseudo-labeled images. After the sub-nets converge, we estimate the pseudo-label for each unlabeled image $\boldsymbol{x}_i \in \mathcal{D}_u$. We define $\boldsymbol{y}_i^m$ and $\boldsymbol{y}_i^{m'}$ as the predictions of M_ψ and $M'_{\psi'}$ branches, respectively. Then, ℓ_i^m and $\ell_i^{m'}$ are the classes which have the maximum estimated probability in $\boldsymbol{y}_i^m$ and $\boldsymbol{y}_i^{m'}$. We set two requirements to obtain the pseudo-label. First, we ensure that both the predictions agree i.e., $\ell_i^m = \ell_i^{m'}$. At the same time, the labeling network must be highly confident about the prediction i.e., the maximum probability exceeds a threshold value (0.95). When both requirements are fulfilled, we assign the pseudo-label $\tilde{\boldsymbol{y}}_i$ for an unlabeled image $\boldsymbol{x}_i$. We combine both the cleanly labeled image-set and pseudo-labeled image-set to form our new training set, which is used to train the labeling network (Fig. 2 (c)). This process progressively adds reliable pseudo-labels in the training set. Besides, this cycle gradually reduces the error in the pseudo-labels for unlabeled samples. We repeat this process 100 times (Sec. 5.1).

3.2. Unpaired Image-to-Image Translation

In this work, we perform unpaired I2I translation with only few labeled examples during training. Using the pseudo-labels provided by NTPL, we now describe the actual training of the I2I translation model.

Method overview. As illustrated in Fig. 2 (a), our model architecture consists of six sub-networks: Pose encoder P_ϕ, Appearance encoder A_η, Generator G_Φ, Multilayer perceptron M_ω, feature regulator F, and Discriminator D_ξ, where indices denote the parameters of each sub-net. Let $\boldsymbol{x}_{sc} \in \mathcal{X}$ be the input source image which provides *pose* information, and $\boldsymbol{x}_{tg} \in \mathcal{X}$ the target image which contributes *appearance*, with corresponding labels $\ell_{sc} \in \{1, \dots, C\}$ for the source and $\ell_{tg} \in \{1, \dots, C\}$ for the target. We use the pose extractor and the appearance extractor to encode the source and target images, generating $P_\phi(\boldsymbol{x}_{sc})$ and $A_\eta(\boldsymbol{x}_{tg})$, respectively. The appearance information $A_\eta(\boldsymbol{x}_{tg})$ is mapped to the input parameters of the Adaptive Instance Normalization (AdaIN) layers [18] (scale and shift) by the multilayer perceptron M_ω. The generator G_Φ takes both the output of pose extractor $P_\phi(\boldsymbol{x}_{sc})$ and the AdaIN parameters output by the multilayer perceptron $M_\omega(A_\eta(\boldsymbol{x}_{tg}))$ as its input, and generates a translated output $\boldsymbol{x}'_{tg} = G_\Phi(P_\phi(\boldsymbol{x}_{sc}), M_\omega(A_\eta(\boldsymbol{x}_{tg})))$. We expect G_Φ to output a target-like image in terms of appearance, which should be classified as the corresponding label ℓ_{tg}.

Additionally, we generate another two images, $\boldsymbol{x}'_{sc}$ and $\boldsymbol{x}''_{sc}$, that will be used in the reconstruction loss (Eq. (7)). The former is used to enforce content preservation [28], and we generate it by using the source image $\boldsymbol{x}_{sc}$ as input for both the pose extractor P_ϕ and the appearance extractor A_η, i.e. $\boldsymbol{x}'_{sc} = G_\Phi(P_\phi(\boldsymbol{x}_{sc}), M_\omega(A_\eta(\boldsymbol{x}_{sc})))$[1]. On the other hand, we generate $\boldsymbol{x}''_{sc}$ by transforming the generated target image $\boldsymbol{x}'_{tg}$ back into the source domain of $\boldsymbol{x}_{sc}$. We achieve this by considering $\boldsymbol{x}_{sc}$ as the target appearance image, that is, $\boldsymbol{x}''_{sc} = G_\Phi(P_\phi(\boldsymbol{x}'_{tg}), M_\omega(A_\eta(\boldsymbol{x}_{sc})))$. This is inspired by CycleGAN [46] and using it for few-shot I2I translation is a novel application. The forward-backward transformation allows us to take advantage of unlabeled data since cycle consistency constraints do not require label supervision.

In order to enforce the pose features to be more class-invariant, we include an *entropy regulation* loss akin to Eq. (4). More concretely, we process input pose features via feature regulator F, which contains a stack of average pooling layers (hence, it does not add any parameters). The output $F(P_\phi(\boldsymbol{x}_{sc}))$ is then entropy-regulated via $\mathcal{L}_{ent}$, forcing the pose features to be sparse and focused on the overall spatial layout rather than domain-specific patterns.

A key component of our generative approach is the discriminator sub-net. We design the discriminator to output three terms: $D_\xi(\boldsymbol{x}) \rightarrow \left\{D^c_{\xi'}(\boldsymbol{x}), D^a_{\xi''}(\boldsymbol{x}), F_\Xi(\boldsymbol{x})\right\}$. Both $D^c_{\xi'}(\boldsymbol{x})$ and $D^a_{\xi''}(\boldsymbol{x})$ are probability distributions. The goal of $D^c_{\xi'}(\boldsymbol{x})$ is to classify the generated images into their correct target class and thus guide the generator to synthesize target-specific images. We use $D^a_{\xi''}(\boldsymbol{x})$ to distinguish between real and synthesized (fake) images of the target class. On the other hand, $F_\Xi(\boldsymbol{x})$ is a feature map. Similar to previous works [4, 18, 28], $F_\Xi(\boldsymbol{x})$ aims to match the appearance of translated image $\boldsymbol{x}'_{tg}$ to the input $\boldsymbol{x}_{tg}$.

The overall loss is a multi-task objective comprising (a) *adversarial loss* that optimizes the game between the generator and the discriminator, i.e. $\{P_\phi, A_\eta, M_\omega, G_\Phi\}$ seek to minimize while discriminator $D^a_{\xi''}$ seeks to max-

[1] Not shown in Fig. 2 for clarity.

Datasets	Animals [28]	Birds [38]	Flowers [30]	Foods [20]
#classes train	119	444	85	224
#classes test	30	111	17	32
#images	117,574	48,527	8.189	31,395

Table 1. Datasets used in the experiments

imize it; (b) *classification loss* that ensures that sub-nets $\{P_\phi, A_\eta, M_\omega, G_\Phi\}$ map source images $\boldsymbol{x}_{sc}$ to target-like images; (c) *entropy regularization loss* that enforces the pose feature to be class-invariant; and (d) *reconstruction loss* that strengthens the connection between the translated images and the target image $\boldsymbol{x}_{tg}$, and guarantees the translated images reserve the pose of the input source image $\boldsymbol{x}_{sc}$.

Adversarial loss. We require $D^a_{\xi''}$ to address multiple adversarial classification tasks simultaneously, as in [28]. Specifically, given output $D^a_{\xi''} \in \mathbb{R}^C$, we locate the ℓ_{th} class response, where $\ell_i \in \{1, \ldots C\}$ is the category of input image to discriminator. Using the response for ℓ_{th} class, we compute the adversarial loss and back-propagate gradients. For example, when updating D_ξ, $\ell_{th} = \ell_{sc}$; when updating $\{P_\phi, A_\eta, M_\omega, G_\Phi\}$, $\ell_{th} \in \{\ell_{sc}, \ell_{tg}\}$. We employ the following adversarial objective [16],

$$\begin{aligned} \mathcal{L}_a \;=\; & \mathop{\mathbb{E}}_{\boldsymbol{x}_{sc}\sim\mathcal{X}}\left[\log D^a_{\xi''}(\boldsymbol{x}_{sc})_{\ell_{sc}}\right] \\ + & \mathop{\mathbb{E}}_{\boldsymbol{x}_{sc,tg}\sim\mathcal{X}}\left[\log\left(1 - D^a_{\xi''}(\boldsymbol{x}'_{tg})_{\ell_{tg}}\right)\right]. \end{aligned} \tag{6}$$

Classification loss. Inspired by [32], we use an auxiliary classifier in our GAN model to generate target-specific images. However, in our case the labels may be noisy for the pseudo-labeled images. For this reason, we employ here the noise-tolerant approach introduced in Sec. 3.1 and use the single-head loss (Eq. (5)) as loss function $\mathcal{L}_c$.

Reconstruction loss. For successful I2I translation, we would like that the translated images keep the pose of the source image $\boldsymbol{x}_{sc}$ while applying the appearance of the target image $\boldsymbol{x}_{tg}$. We use the generated images $\boldsymbol{x}'_{sc}$ and $\boldsymbol{x}''_{sc}$ and the features $F_\Xi(\boldsymbol{x})$ output by the discriminator to achieve these goals via the following reconstruction loss,

$$\begin{aligned} \mathcal{L}_r = & \,\mathbb{E}_{\boldsymbol{x}_{sc}\sim\mathcal{X}, \boldsymbol{x}'_{sc}\sim\mathcal{X}'}\left[\|\boldsymbol{x}_{sc} - \boldsymbol{x}'_{sc}\|_1\right] \\ & + \mathbb{E}_{\boldsymbol{x}_{sc}\sim\mathcal{X}, \boldsymbol{x}''_{sc}\sim\mathcal{X}''}\left[\|\boldsymbol{x}_{sc} - \boldsymbol{x}''_{sc}\|_1\right] \\ & + \mathbb{E}_{\boldsymbol{x}_{sc}\sim\mathcal{X}, \boldsymbol{x}'_{sc}\sim\mathcal{X}'}\left[\|F_\Xi(\boldsymbol{x}_{sc}) - F_\Xi(\boldsymbol{x}'_{sc})\|_1\right] \\ & + \mathbb{E}_{\boldsymbol{x}_{tg}\sim\mathcal{X}, \boldsymbol{x}'_{tg}\sim\mathcal{X}'}\left[\left\|F_\Xi(\boldsymbol{x}_{tg}) - F_\Xi(\boldsymbol{x}'_{tg})\right\|_1\right]. \end{aligned} \tag{7}$$

Full Objective. The final loss function of our model is:

$$\min_{P_\phi, A_\eta, M_\omega, G_\Phi} \max_{D_\xi} \lambda_a \mathcal{L}_a + \lambda_c \mathcal{L}_c + \lambda_r \mathcal{L}_r + \lambda_e \mathcal{L}_{ent}, \tag{8}$$

where λ_a, λ_c, λ_r and λ_e are re-weighting hyper-parameters.

3.3. Octave network

An important aspect of our generator model is the Octave Convolution (OctConv) operator [6]. This operator has not been studied before for generative tasks. Specifically, OctConv aims to separate low and high-frequency feature maps. Since image translation mainly focuses on altering high-frequency information, such disentanglement can help with the learning. Furthermore, the low-frequency processing branch in OctConv layers has a wider receptive field that is useful to learn better context for the encoders. Let $\boldsymbol{u} = \{\boldsymbol{u}^h, \boldsymbol{u}^l\}$ and $\boldsymbol{v} = \{\boldsymbol{v}^h, \boldsymbol{v}^l\}$ be the inputs and outputs of OctConv layer, respectively. As illustrated in Fig. 2 (b), the forward pass is defined as,

$$\boldsymbol{v}^l = L_\tau(\boldsymbol{u}^h, \boldsymbol{u}^l), \quad \boldsymbol{v}^h = H_{\tau'}(\boldsymbol{u}^h, \boldsymbol{u}^l), \tag{9}$$

where, L_τ and $H_{\tau'}$ are the high and low-frequency processing blocks with parameters τ and τ', respectively. The complete architecture of the OctConv layer used in our work is shown in Fig. 2 (b). We explore suitable proportions of low-frequency and high-frequency channels for networks P_ϕ, A_η, and G_Φ in Sec. 5.1. For the discriminator D_ξ, we empirically found the OctConv does not improve performance.

4. Experimental setup

Datasets. We consider four datasets for evaluation, namely *Animals* [28], *Birds* [38], *Flowers* [30], and *Foods* [20] (see Table 1 for details). We follow FUNIT's inference procedure [28] and randomly sample 25,000 source images from the training set and translate them to each target domain (not seen during training). We consider the 1, 5, and 20-shot settings for the target set. For efficiency reasons, in the ablation study we use the same smaller subset of 69 Animals categories used in [28], which we refer to as *Animals-69*.

Evaluation metrics. We consider the following three metrics. Among them, two are commonly used *Inception Score (IS)* [36] and *Fréchet Inception Distance (FID)* [17]. Moreover, we use *Translation Accuracy* [28] to evaluate whether a model is able to generate images of the target class. Intuitively, we measure translation accuracy by the Top1 and Top5 accuracies of two classifiers: *all* and *test*. The former is trained on both source and target classes, while the latter is trained using only target classes.

Baselines. We compare against the following baselines (see Suppl. Mat. (Sec. 3) for training details). *CycleGAN* [46] uses two pairs of domain-specific encoders and decoders, trained to optimize both an adversarial loss and the cycle consistency. *StarGAN* [9] performs scalable image translation for all classes by inputting the label to the generator. *MUNIT* [18] disentangles the latent representation into the content space shared between two classes, and the class-specific style space. *FUNIT* [28] is the first few-shot I2I translation method.

Variants. We explore a wide variety of configurations for our approach, including: semi-supervised learning (S), OctConv (O), entropy regulation (E), and cycle consistency (C).

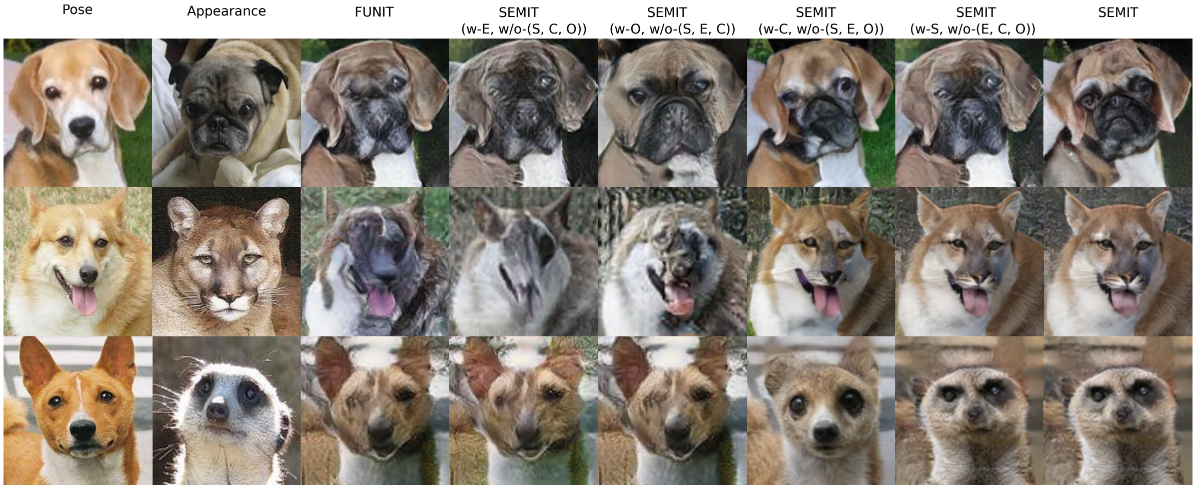


Figure 3. Comparison between FUNIT [28] and variants of our proposed method. For example, *SEMIT (w-E, w/o-(S, C, O))* indicates the model trained with only entropy regulation. More examples are in Suppl. Mat. (Sec. 1).

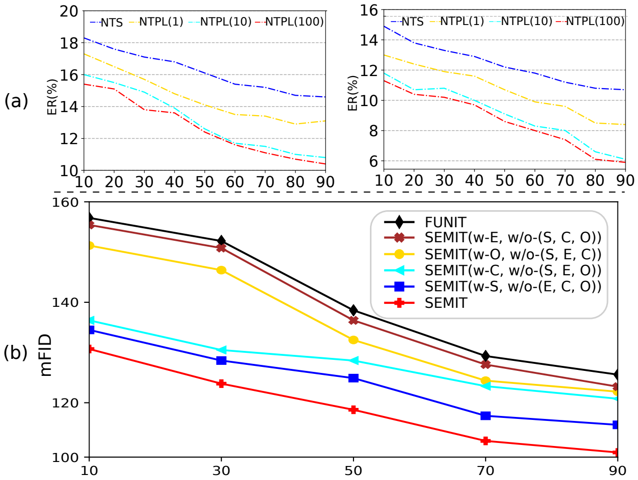


Figure 4. (a) Ablation study on classification for (left) Animals-69 and (right) Birds, measured by Error Rate (ER). (b) Ablation study of the variants of our method for one-shot on Animals-69. The x-axis shows the percentage of the labeled data used for training.

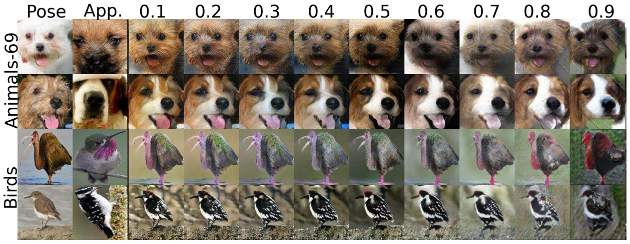
We denote them by *SEMIT* followed by the present (w) and absent (w/o) components, *e.g*. *SEMIT*(w-O, w/o-(S, E, C)) refers to model with OctConv and without semi-supervised learning, entropy regulation or cycle consistency.

5. Experiments

5.1. Ablation study

Here, we evaluate the effect of each independent contribution to SEMIT and their combinations. Full experimental configurations are in Suppl. Mat. (Sec. 4).

Noise-tolerant Pseudo-labeling. As an alternative to our NTPL, we consider the state-of-the-art approach for fine-grained recognition NTS [43], as it outperforms other fine-grained methods [24, 5, 14] on our datasets. We adopt NTS's configuration for Animals-69 and Birds and divide the datasets into train set (90%) and test set (10%). In order to study the effect of NTPL for limited labeled data, we randomly divide the train set into labeled data and unlabeled data, for which we ignore the available labels. All models are evaluated on the test set. To confirm that the iterative process in Sec. 3.1 leads to better performance, we consider three NTPL variants depending on the number of times that we repeat this process. NTPL (100) uses the standard 100 iterations to progressively add unlabeled data into the train set, whereas NTPL (10) uses 10 and NTPL (1) uses a single iteration. We report results in terms of the Error Rate (ER) in Fig. 4 (a). We can see how for both NTPL and NTS, the performance is significantly lower for regimes with less labeled data. With 10% of labeled data, NTS obtains a higher error than NTPL (100), *e.g*. for Animals-69: 18.3% vs. 15.2%. The training times for each variant are as follows NTS: 28.2min, NTPL (1): 36.7min, NTPL (10): 91.2min, NTPL (100): 436min. Note that each model NTPL (k) is initialized with the previous model, NTPL ($k-1$). For any given percentage of labeled data, our NTPL-based training clearly obtains superior performance, confirming that NTPL contributes to predicting better labels for unlabeled



Dataset	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Animals	130.3	129.8	128.4	128.6	**127.1**	128.5	128.6	129.4	130.9
Birds	118.9	116.4	113.4	113.6	**112.7**	114.6	115.2	119.7	135.4

Figure 5. Qualitative (top) and quantitative (bottom) results for several ratios of high/low frequency channels in OctConv. Results correspond to one-shot I2I translation on Animals-69 and Birds with *90%* labeled data. More examples in Suppl. Mat. (Sec. 2).

	Setting	Top1-all	Top5-all	Top1-test	Top5-test	IS-all	IS-test	mFID
100%	CycleGAN-20	28.97	47.88	38.32	71.82	10.48	7.43	197.13
	MUNIT-20	38.61	62.94	53.90	84.00	10.20	7.59	158.93
	StarGAN-20	24.71	48.92	35.23	73.75	8.57	6.21	198.07
	FUNIT-1	17.07	54.11	46.72	82.36	22.18	10.04	93.03
	FUNIT-5	33.29	78.19	68.68	96.05	22.56	13.33	70.24
	FUNIT-20	39.10	84.39	73.69	97.96	22.54	14.82	66.14
	SEMIT-1	29.42	65.51	62.47	90.29	24.48	13.87	75.87
	SEMIT-5	35.48	78.96	71.23	94.86	25.63	15.68	68.32
	SEMIT-20	**45.70**	**88.5**	**74.86**	**99.51**	**26.23**	**16.31**	**49.84**
20%	FUNIT-1	12.01	30.59	29.86	55.44	19.23	4.59	139.7
	FUNIT-5	15.25	36.48	36.47	66.58	21.12	6.16	128.3
	FUNIT-20	16.95	41.43	42.61	68.92	21.48	6.78	117.4
	SEMIT-1	26.71	69.48	65.48	85.49	23.52	12.63	92.21
	SEMIT-5	39.56	78.34	71.81	96.25	24.01	14.17	69.28
	SEMIT-20	**44.25**	**85.60**	**73.80**	**98.62**	**24.67**	**15.04**	**65.21**
10%	FUNIT-1	10.21	28.41	27.42	49.54	17.24	4.05	156.8
	FUNIT-5	13.04	35.62	31.21	61.70	19.12	4.87	138.8
	FUNIT-20	14.84	39.64	37.52	65.84	19.64	5.53	127.8
	SEMIT-1	16.25	51.55	39.71	81.47	22.58	8.61	99.42
	SEMIT-5	29.40	76.14	62.72	92.13	22.98	13.24	78.46
	SEMIT-20	**39.02**	**82.90**	**69.70**	**95.40**	**23.43**	**14.07**	**69.40**

Table 2. Performance comparison with baselines on Animals [28].

data and improves the robustness against noisy labels.

OctConv layer. Fig. 5 (top) presents qualitative results on the Animals-69 and Birds datasets (one-shot, 90% labeled data) for varying proportions of channels devoted to high or low frequencies (Sec. 3.3). Changing this value has a clear effect on how our method generates images. As reported in Fig. 5 (bottom), we find using OctConv with half the channels for each frequency (0.5) obtains the best performance. For the rest of the paper, we set this value to 0.5. We conclude that OctConv facilitates the I2I translation task by disentangling the feature space into frequencies.

Other SEMIT variants. Fig. 4 (b) presents a comparison between several variants of SEMIT and FUNIT [28] in terms of mean FID (mFID) for various percentages of labeled training data. Adding either Entropy regulation (SEMIT (w-E, w/o-(S, C, O)) or OctConv layers (SEMIT (w-O, w/o-(S, E, C)) improves the performance of I2I translation compared to FUNIT [28] at all levels of labeled data. We attribute this to the architectural advantage and enhanced optimization granted by our contributions to the I2I translation task in general. Next, adding either cycle consistency or semi-supervised learning achieves a further boost in performance. The improvement is remarkably substantial for low percentages of labeled data (10%-30%), which is our main focus. This shows how such techniques, especially semi-supervised learning, can truly exploit the information in unlabeled data and thus relax the labeled data requirements. Finally, the complete SEMIT obtains the best mFID score, indicating that our method successfully performs I2I translation even with much fewer labeled images. Similar conclusions can be drawn from the qualitative examples in Fig. 3, where SEMIT successfully transfers the appearance of the given target to the input pose image.

	Setting	Top1-all	Top5-all	Top1-test	Top5-test	IS-all	IS-test	mFID
100%	CycleGAN-20	9.24	22.37	19.46	42.56	25.28	7.11	215.30
	MUNIT-20	23.12	41.41	38.76	62.71	24.76	9.66	198.55
	StarGAN-20	5.38	16.02	13.95	33.96	18.94	5.24	260.04
	FUNIT-1	11.17	34.38	30.86	60.19	67.17	17.16	113.53
	FUNIT-5	20.24	51.61	45.40	75.75	74.81	22.37	99.72
	FUNIT-20	23.50	56.37	49.81	1.286	76.42	24.00	97.94
	SEMIT-1	15.64	42.85	43.7.62	72.41	69.63	20.12	105.82
	SEMIT-5	23.57	55.96	49.42	80.41	78.42	24.98	90.48
	SEMIT-20	**28.15**	**62.41**	**54.62**	**83.32**	**82.64**	**27.51**	**83.56**
20%	FUNIT-1	6.21	20.31	15.34	28.45	29.23	8.23	184.4
	FUNIT-5	10.25	22.34	22.75	43.24	43.62	12.53	168.6
	FUNIT-20	11.76	28.51	26.47	46.38	58.40	15.75	145.1
	SEMIT-1	13.58	48.16	43.97	64.27	59.29	16.48	109.84
	SEMIT-5	19.23	53.25	50.34	73.16	67.84	22.27	98.38
	SEMIT-20	**21.49**	**57.55**	**52.34**	**76.41**	**72.31**	**23.44**	**95.41**
10%	FUNIT-1	6.04	19.34	12.51	38.84	32.62	7.47	203.3
	FUNIT-5	8.82	22.52	19.85	42.53	38.59	9.53	175.7
	FUNIT-20	10.98	26.41	22.48	48.36	41.37	13.85	154.9
	SEMIT-1	11.21	37.14	35.14	59.41	48.48	12.57	128.4
	SEMIT-5	13.54	43.63	40.24	68.75	59.84	17.58	119.4
	SEMIT-20	**15.41**	**48.36**	**42.51**	**71.49**	**65.42**	**19.87**	**109.8**

Table 3. Performance comparison with baselines on Birds [38].

5.2. Results for models trained on a single dataset

Tables 2 and 3 report results for all baselines and our method on Animals [28] and Birds [38], under three percentages of labeled source images: 10%, 20%, and 100%. We use the 20-shot setting as default for all baselines but also explore 1-shot and 5-shot settings for FUNIT [28] and our method. All the baselines that are not specialized for few-shot translation (i.e. CycleGAN [46], MUNIT [47], and StarGAN [9]) suffer a significant disadvantage in the few-shot scenario, obtaining inferior results even with 100% of labeled images. However, both FUNIT and SEMIT perform significantly better, and SEMIT achieves the best results for all metrics under all settings. Importantly, SEMIT trained with only 20% of ground-truth labels (*e.g.* mFID of 65.21 for Animals) is comparable to FUNIT with 100% labeled data (mFID 66.14), clearly indicating that the proposed method effectively performs I2I translation with ×5 less labeled data. Finally, our method achieves competitive performance even with only 10% available labeled data. We also provide many-shot case in Suppl. Mat. (Sec. 5)

Fig. 6 shows example images generated by FUNIT and SEMIT using 10% labeled data. On Animals, Birds, and Food, FUNIT manages to generate somewhat adequate target-specific images. Nonetheless, under closer inspection, the images look blurry and unrealistic, since FUNIT fails to acquire enough guidance for generation without exploiting the information present in unlabeled data. Besides, it completely fails to synthesize target-specific images of Flowers, possibly due to the smaller number of images per class in this dataset. SEMIT, however, successfully synthesizes convincing target-specific images for all datasets, including the challenging Flowers dataset. These results again support our conclusion: SEMIT effectively applies the target appearance onto the given pose image despite using much less labeled data.

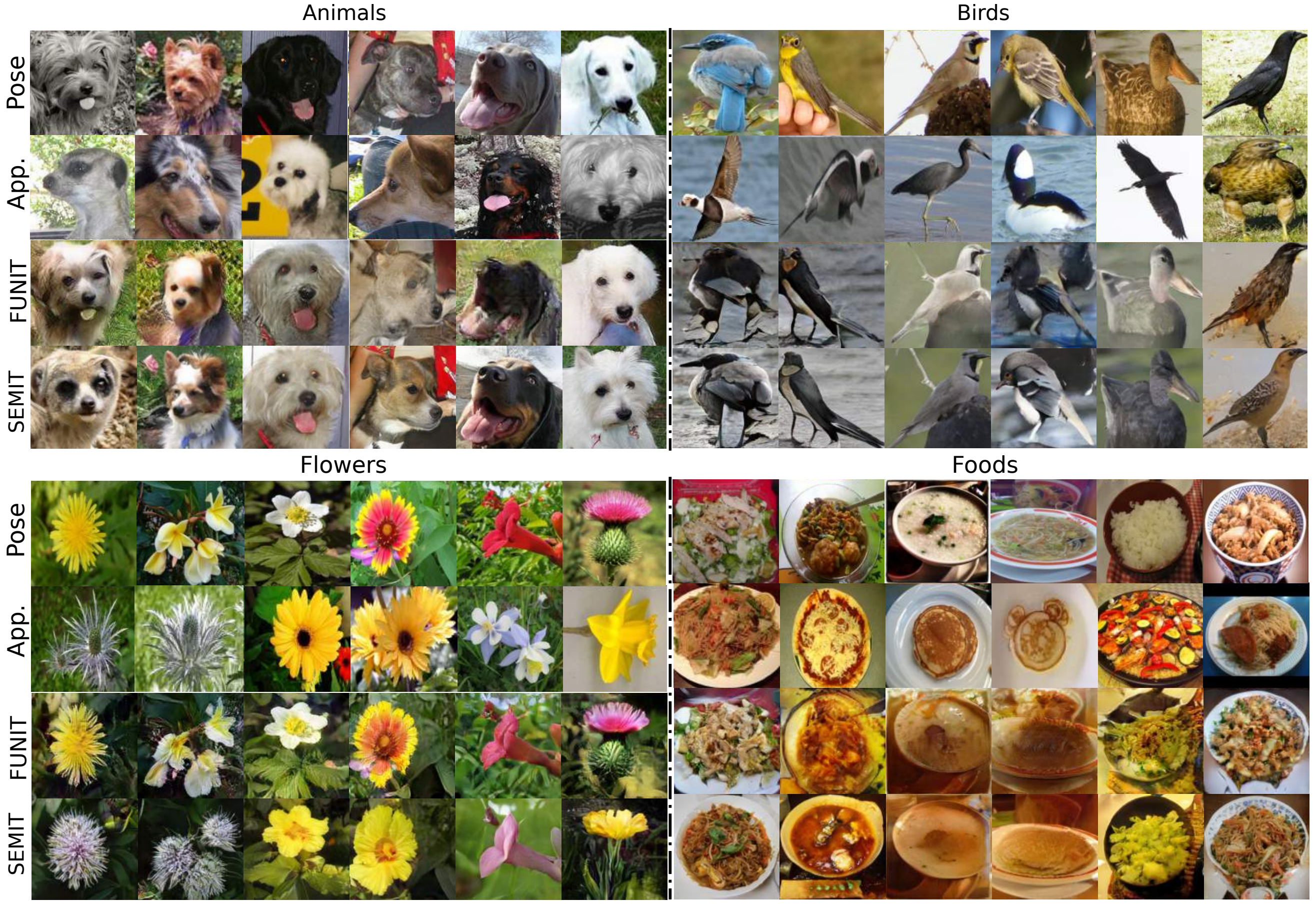


Figure 6. Qualitative comparison between our method and FUNIT [28] on the four datasets. More examples are in Suppl. Mat. (Sec. 6).

5.3. Results for models trained on multiple datasets

We investigate whether SEMIT can learn from multiple datasets simultaneously. For this, we merge an additional 20,000 unlabeled animal faces (from [25, 45, 21] or retrieved via search engine) into the Animals dataset, which we call *Animals*++. We also combine 6,033 unlabeled bird images from CUB-200-2011 [41] into Birds and name it *Birds*++. We term our model trained on the original dataset as Ours (SNG) and the model trained using the expanded versions as Ours (JNT). We experiment using 10% labeled data from the original datasets. Note, we do not apply the classification loss (Eq. 1) for the newly added images, as the external data might include classes not in the source set. Fig. 7 shows results which illustrate how Ours (SNG) achieves successful target-specific I2I translation, but Ours (JNT) exhibits even higher visual quality. This is because Ours (JNT) can leverage the additional low-level information (color, texture, etc.) provided by the additional data. We provide quantitative results in Suppl. Mat. (Sec. 8).

6. Conclusions

We proposed semi-supervised learning to perform few-shot unpaired I2I translation with fewer image labels for the source domain. Moreover, we employ a cycle consistency constraint to exploit the information in unlabeled data, as well as several generic modifications to make the I2I translation task easier. Our method achieves excellent results on several datasets while requiring only a fraction of the labels.

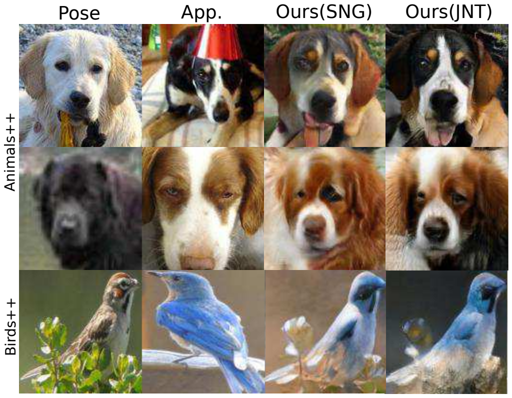


Figure 7. Results of our method on a single dataset (SNG) and joint datasets (JNT). More examples are in Suppl. Mat. (Sec. 7).

Acknowledgements. We thank the Spanish project TIN2016-79717-R and also its CERCA Program of the Generalitat de Catalunya.

References

[1] Yazeed Alharbi, Neil Smith, and Peter Wonka. Latent filter scaling for multimodal unsupervised image-to-image translation. In *CVPR*, 2019.

[2] Matthew Amodio and Smita Krishnaswamy. Travelgan: Image-to-image translation by transformation vector learning. In *CVPR*, June 2019.

[3] Sagie Benaim and Lior Wolf. One-shot unsupervised cross domain translation. In *NIPS*, 2018.

[4] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *NIPS*, pages 2172–2180, 2016.

[5] Yue Chen, Yalong Bai, Wei Zhang, and Tao Mei. Destruction and construction learning for fine-grained image recognition. In *CVPR*, 2019.

[6] Yunpeng Chen, Haoqi Fang, Bing Xu, Zhicheng Yan, Yannis Kalantidis, Marcus Rohrbach, Shuicheng Yan, and Jiashi Feng. Drop an octave: Reducing spatial redundancy in convolutional neural networks with octave convolution. *arXiv preprint arXiv:1904.05049*, 2019.

[7] Ying-Cong Chen, Xiaogang Xu, Zhuotao Tian, and Jiaya Jia. Homomorphic latent space interpolation for unpaired image-to-image translation. In *CVPR*, pages 2408–2416, 2019.

[8] Wonwoong Cho, Sungha Choi, David Keetae Park, Inkyu Shin, and Jaegul Choo. Image-to-image translation via group-wise deep whitening-and-coloring transformation. In *CVPR*, June 2019.

[9] 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 *CVPR*, June 2018.

[10] LI Chongxuan, Taufik Xu, Jun Zhu, and Bo Zhang. Triple generative adversarial nets. In *NIPS*, pages 4088–4098, 2017.

[11] Zhijie Deng, Hao Zhang, Xiaodan Liang, Luona Yang, Shizhen Xu, Jun Zhu, and Eric P Xing. Structured generative adversarial networks. In *NIPS*, 2017.

[12] Zhe Gan, Liqun Chen, Weiyao Wang, Yuchen Pu, Yizhe Zhang, Hao Liu, Chunyuan Li, and Lawrence Carin. Triangle generative adversarial networks. In *Advances in Neural Information Processing Systems*, pages 5247–5256, 2017.

[13] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *CVPR*, pages 2414–2423, 2016.

[14] Weifeng Ge, Xiangru Lin, and Yizhou Yu. Weakly supervised complementary parts models for fine-grained image classification from the bottom up. In *CVPR*, pages 3034–3043, 2019.

[15] Abel Gonzalez-Garcia, Joost van de Weijer, and Yoshua Bengio. Image-to-image translation for cross-domain disentanglement. In *NIPS*, pages 1294–1305, 2018.

[16] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*, pages 2672–2680, 2014.

[17] 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 *NIPS*, pages 6626–6637, 2017.

[18] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *ECCV*, pages 172–189, 2018.

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

[20] Yoshiyuki Kawano and Keiji Yanai. Automatic expansion of a food image dataset leveraging existing categories with domain adaptation. In *ECCV*, pages 3–17. Springer, 2014.

[21] Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Li Fei-Fei. Novel dataset for fine-grained image categorization. In *First Workshop on Fine-Grained Visual Categorization, CVPR*, 2011.

[22] Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jungkwon Lee, and Jiwon Kim. Learning to discover cross-domain relations with generative adversarial networks. *ICML*, 2017.

[23] Yi Kun and Wu Jianxin. Probabilistic End-to-end Noise Correction for Learning with Noisy Labels. In *CVPR*, 2019.

[24] Michael Lam, Behrooz Mahasseni, and Sinisa Todorovic. Fine-grained recognition as hsnet search for informative image parts. In *CVPR*, July 2017.

[25] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Kumar Singh, and Ming-Hsuan Yang. Diverse image-to-image translation via disentangled representations. In *ECCV*, 2018.

[26] Jianxin Lin, Yingce Xia, Sen Liu, Tao Qin, and Zhibo Chen. Zstgan: An adversarial approach for unsupervised zero-shot image-to-image translation. *arXiv preprint arXiv:1906.00184*, 2019.

[27] Fayao Liu, Chunhua Shen, Guosheng Lin, and Ian Reid. Learning depth from single monocular images using deep convolutional neural fields. *IEEE Trans. on PAMI*, 38(10):2024–2039, 2016.

[28] 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.

[29] Mario Lucic, Michael Tschannen, Marvin Ritter, Xiaohua Zhai, Olivier Bachem, and Sylvain Gelly. High-fidelity image generation with fewer labels. *ICML*, 2019.

[30] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *ICVGIP*, pages 722–729. IEEE, 2008.

[31] Augustus Odena. Semi-supervised learning with generative adversarial networks. *arXiv preprint arXiv:1606.01583*, 2016.

[32] Augustus Odena, Christopher Olah, and Jonathon Shlens. Conditional image synthesis with auxiliary classifier gans. In *ICML*, pages 2642–2651. JMLR. org, 2017.

[33] Guim Perarnau, Joost Van De Weijer, Bogdan Raducanu, and Jose M Álvarez. Invertible conditional gans for image editing. *Advances in neural information processing systems Workshop on Adversarial Training*, 2016.

[34] Andrés Romero, Pablo Arbeláez, Luc Van Gool, and Radu Timofte. Smit: Stochastic multi-label image-to-image translation. *arXiv preprint arXiv:1812.03704*, 2019.

[35] Kuniaki Saito, Yoshitaka Ushiku, and Tatsuya Harada. Asymmetric tri-training for unsupervised domain adaptation. *ICML*, 2017.

[36] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *NIPS*, pages 2234–2242, 2016.

[37] Jost Tobias Springenberg. Unsupervised and semi-supervised learning with categorical generative adversarial networks. *ICLR*, 2016.

[38] Grant Van Horn, Steve Branson, Ryan Farrell, Scott Haber, Jessie Barry, Panos Ipeirotis, Pietro Perona, and Serge Belongie. Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection. In *CVPR*, pages 595–604, 2015.

[39] Yaxing Wang, Abel Gonzalez-Garcia, Joost van de Weijer, and Luis Herranz. SDIT: Scalable and diverse cross-domain image translation. In *ACM MM*, 2019.

[40] Yaxing Wang, Joost van de Weijer, and Luis Herranz. Mix and match networks: encoder-decoder alignment for zero-pair image translation. In *CVPR*, pages 5467–5476, 2018.

[41] Peter Welinder, Steve Branson, Takeshi Mita, Catherine Wah, Florian Schroff, Serge Belongie, and Pietro Perona. Caltech-ucsd birds 200. 2010.

[42] Wayne Wu, Kaidi Cao, Cheng Li, Chen Qian, and Chen Change Loy. Transgaga: Geometry-aware unsupervised image-to-image translation. In *CVPR*, June 2019.

[43] Ze Yang, Tiange Luo, Dong Wang, Zhiqiang Hu, Jun Gao, and Liwei Wang. Learning to navigate for fine-grained classification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 420–435, 2018.

[44] Zili Yi, Hao Zhang, Ping Tan Gong, et al. Dualgan: Unsupervised dual learning for image-to-image translation. In *ICCV*, 2017.

[45] Weiwei Zhang, Jian Sun, and Xiaoou Tang. Cat head detection-how to effectively exploit shape and texture features. In *ECCV*, pages 802–816. Springer, 2008.

[46] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, 2017.

[47] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. Toward multimodal image-to-image translation. In *NIPS*, pages 465–476, 2017.