Semantically Robust Unpaired Image Translation for Data with Unmatched Semantics Statistics

Zhiwei Jia[1]* Bodi Yuan[2]* Kangkang Wang[2] Hong Wu[2]
David Clifford[2] Zhiqiang Yuan[2] Hao Su[1]
[1]UC San Diego {zjia,haosu}@eng.ucsd.edu
[2]X {bodiyuan,kangkang,wuh,davidclifford,zyuan}@google.com

Abstract

Many applications of unpaired image-to-image translation require the input contents to be preserved semantically during translations. Unaware of the inherently unmatched semantics distributions between source and target domains, existing distribution matching methods (i.e., GAN-based) can give undesired solutions. In particular, although producing visually reasonable outputs, the learned models usually flip the semantics of the inputs. To tackle this without using extra supervisions, we propose to enforce the translated outputs to be semantically invariant w.r.t. small perceptual variations of the inputs, a property we call "semantic robustness". By optimizing a robustness loss w.r.t. multi-scale feature space perturbations of the inputs, our method effectively reduces semantics flipping and produces translations that outperform existing methods both quantitatively and qualitatively.

1. Introduction

Recently, unpaired image-to-image translation [11] has been very popular in the computer vision community. Due to its general assumptions on the inputs (unlabelled images collected from different domains) and the easy accessibility of training data (does not use paired images), it is widely used in fields such as image manipulations, style transfer, domain adaptation, data augmentation, etc. [13, 58, 43, 22, 28, 9, 30, 41, 45, 2]. On the other hand, unpaired image translation remains a very challenging task owing to its unsupervised learning nature. Without paired images that specify the exact domain mapping, one has to rely on visual cues to perform distribution matching (i.e., via GANs [16]). Existing GAN-based methods all rely on the adversarial loss that aims to optimally align image statistics between translation images and the target ones (in the marginal sense). *However, what if the two distributions should not be the same?* In fact, the underlying distributions of semantics from the two domains are usually different, let alone the image distribution of translated images and target one. We call this the unmatched semantics statistics problem, which is under-explored yet both critical and common for unpaired image translation tasks.

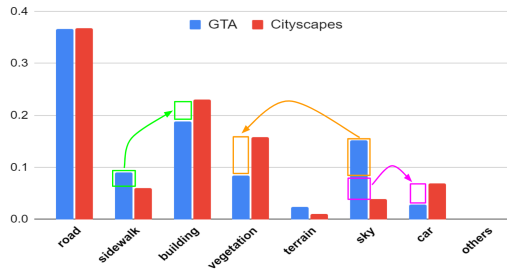

Figure 1: The class distributions in GTA vs. Cityscapes. During unpaired image translation, the generator has to flip the inputs' semantics to match the target distributions. Instances from over-represented semantic classes in the source domain (e.g., sky) can be flipped to those from underrepresented classes (e.g., vegetation).

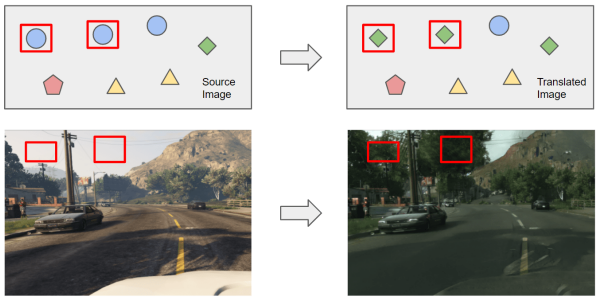

Figure 2: (**top**) Conceptually, forcing the distribution of translated images to match the target one causes semantics (the different colored shapes) of the input images to get flipped. (**bottom**) An example of semantics flipping (highlighted in red boxes) from the GTA to Cityscapes task.

*Co-first authors; work partly done during an internship at X. Code available at https://github.com/SeanJia/SRUNIT.

Similar to the language translation, the semantics of an image should be preserved during translations. For instance, in the GTA to Cityscapes dataset [12], while trees look different across domains, their identity/semantics remains the same. In the Horse to Zebra task [11], a horse or zebra remains in the class Equus, not turning into a shack. Consider the translation as a two-stage process: firstly project an image from one domain to the shared semantics space, and then project it to the other domain. When the source and target images are projected to the same semantic space and have different distributions in that space, we say the data have *unmatched semantics statistics*. Unpaired data from different domains generally have unmatched semantics statistics, unless they are very carefully constructed. For example, in the Horse to Zebra dataset, there are more zebras than horses; in the GTA to Cityscapes dataset, more trees in Cityscapes than in GTA (see Fig. 1). Given unmatched semantics statistics, forcibly matching distributions between the translated and the target images can only give spurious solutions, where semantics get flipped only to match the target semantics statistics (see Fig. 2 for an example). In Sec. 5, we demonstrate that semantics flipping is a critical and common issue in various GAN-based unpaired image translation frameworks.

There are a few direct attempts at preserving the semantics during translations and thus reducing flipping. However, they either require extra supervision or pre-trained models [22, 51] or are too restrictive (dataset-specific) and prone to artifacts [7, 60, 56]. In this paper, we propose to tackle this problem by encouraging that, during image translations, perceptually similar contents should be mapped to contents with high semantic similarity. We call this property of the mapping as *semantic robustness*. In essence, semantic robustness ensures a consistent mapping that prevents the semantics of the inputs from being flipped easily. Specifically, based on the recently proposed framework CUT [40], we propose a semantic robustness loss w.r.t. multi-scale feature space perturbations of the input images. We call our method SRUNIT (Semantically Robust Unpaired Image Translation) and empirically demonstrate its effectiveness in reducing semantics flipping. SRUNIT outperforms existing GAN-based methods both qualitatively and quantitatively on several common datasets.

2. Related Work

Unpaired Image-to-image Translation Although lacking pixel-wise supervision, advances have been made in unpaired image-to-image translation by utilizing Generative Adversarial Networks (GANs) [16]. The central idea is to minimize the statistical difference (measured using the discriminators) between generated and target images by updating the generators. These methods can roughly be sorted into two-sided methods such as [32, 62, 29] and one-sided counterparts such as [24, 34, 35].

Preserving Semantics in Image Translation More recently, efforts have been made in preserving the semantic content of the source images during the unpaired image translation. There are several existing approaches. Cycle consistency [62] is proposed to enforce bijective mappings between domains so that semantic information will not be lost during translation. Geometry consistency [15] enforces equivariance of the generators regarding geometric transformations. DistanceGAN [7] and HarmonicGAN [60] encourage visual similarities within the source domain to be reflected in the target domain. A spectral constraint based on the Fourier transform of the input images is proposed in [56]. Attention-based methods [37, 52] are used to preserve the background during the translations. Moreover, multiple work [51, 8, 46, 33, 40] adopt the idea that input and output images should be similar, measured by a function either pre-defined or learned contrastively.

Robustness & Generalization of DNNs Semantic robustness and semantics flipping discussed in our paper is related to both adversarial robustness and generalization ability. Some work has tackled the adversarial robustness of GANs [11, 6, 54]. And some [59, 53, 3] has explored their generalization properties. In a broader context, both adversarial attack and defence have been extensively studied [48, 19, 23, 50, 49, 42, 17, 1], and many recent advances have been made for understanding the generalizability of DNNs [38, 47, 55, 5, 61, 14, 4, 27, 26].

3. Semantic Robustness

Many applications of unpaired image translation (style transfer, domain adaptation, data augmentation [22, 28, 2]) require the semantics of the inputs to be preserved during translations. In this section, we will discuss the semantics flipping issue and the concept of semantic robustness.

3.1. Unmatched Semantics Statistics

Most existing approaches for unpaired image-to-image translation do not explicitly study the mismatched distributions of the semantics across source and target domains. This prevalent phenomena in unpaired translation tasks usually incur serious artifacts (see Fig. 2 bottom row). To begin with, let us define the terms. When translating images from one domain to the other, it is natural to assume an intermediate semantics space where resides the information to be preserved during translations. When converting images from a domain to the shared semantics space, we refer to the resulting distribution in this space as semantics distribution. Due to the nature of unpaired image translation tasks where direct supervision of the paired relations is missing, we should assume that the unpaired data

from different domains have different semantics distributions (i.e., the unmatched semantics statistics). Most widely available datasets fall into this category (e.g., see Fig. 1 for unmatched semantics statistics between GTA [44] and Cityscapes [12]). A few exceptions are those originally constructed for paired translation (e.g., Maps to Photos [25]).

3.2. The Semantics Flipping Issue

We argue that provided the unmatched semantics statistics between source and target domains, an inherent problem for GAN-based unpaired image translation frameworks is the semantics flipping issue.

The central idea of GAN-based methods is to match the image statistics between translated images and target images as much as possible. Multiple metrics for evaluating the translation performance follow this principle, i.e. they measure some sort of statistical distance between generated and target images (FID, MMD, etc. [21, 18]). This is indeed problematic, as the generated and the target distribution should not be the same, provided that the source and target domains have discrepancies in semantics statistics. We observe that the learned translation models by existing methods usually are undesired solutions (e.g., Fig. 2 bottom row), which, although producing visually reasonable outputs, systematically flip contents into other semantics. This is because only through semantics flipping can the generators produce images that match the statistics of the target domain (see Fig. 1 as an illustration of this process).

3.3. Limitations of Existing Approaches

Most existing unpaired image translation frameworks do not explicitly tackle and, in fact, suffer from the semantics flipping issue (empirically demonstrated in Sec. 5). For two-sided domain mapping methods, cycle consistency [62] are the most popular technique which suggests using bijective (and therefore information-preserving) mappings. However, as pointed out in [11], CycleGAN can learn to hide information in plain sight such that semantics flipping still occurs while the information is preserved during the translations. One-sided domain mapping approaches directly pose constraints on the generators to preserve meaningful information. GcGAN [15] proposes geometry consistency to enforce that the translation functions are equivariant w.r.t. common geometric transformations. However, spurious solutions with semantics flipping can also be equivariant as such. Another line of work is to enforce some sort of relations between input and output images (or image patches), for example, by perceptual similarity or statistical dependency [33, 40]. Since these methods have their correspondence learned unsupervisedly (e.g., contrastively), its inaccuracy can lead to spurious enforcement with more semantics flipping (or artifacts). Alternatively, [56] uses a spectral constraint to maintain semantics. The approach might fail in general and was only shown to be successful in translation tasks across visually similar domains. Although methods with ground truth perceptual similarity can reduce semantics flipping [22, 51], they require extra supervision or pre-trained models that are not available for a general unpaired image translation task.

3.4. Semantic Robustness to the Rescue

Other than directly enforcing relations between input and output images, we propose to encourage that small perceptual variation of the input images (or patches) should not change the semantics of the corresponding transformed images (or patches). We call this property of the generators as *semantic robustness*. Notice that the perceptual similarity between images (or patches) refers to the distance measured in the feature space (e.g., CNN features of the images), rather than in the raw pixel space. We argue that increasing the semantic robustness of the generators can effectively reduce semantics flipping during translations. Intuitively, an input image (or patch) should have its semantics invariant under small perceptual perturbations, and thus, the semantics of the corresponding translated image (or patch) should also be invariant. Remember that semantics flipping happens as the generator is forced to match the target statistics by transforming semantically over-represented contents from the source domain to less represented ones (see Fig. 1). *Semantic robustness encourages a consistent translation such that contents of the same semantics are not transformed into contents of several different semantics.* As a result, it prevents forceful distribution matching and mitigates the flipping issue.

How do we obtain the semantics from images in the first place? Without relying on extra supervision or pre-trained models, contrastive learning approaches (e.g., [40]) can learn to extract features that are domain-invariant, which we consider as the semantics of the inputs. One might find it intuitive to directly enforce that the translation should not change these "semantics" of the input to reduce semantics flipping. However, this direct approach does not work well (see our ablation study in Sec. 6). Interestingly, these extracted semantics can be used to effectively reduce flipping by instead enforcing semantic robustness (i.e., the semantics of the translated image be invariant to perceptual variations of the inputs). This is partly because the latter indirect constraint is a "soft" version of the former direct constraint and is more robust w.r.t. the inaccuracy of the extracted semantics which is contrastively learned.

4. Method

4.1. Preliminary: CUT

The goal of unpaired image-to-image translation is to learn functions between two domains X and Y given train-

ing samples $\{x_i\}_{i=1}^N$, $\{y_j\}_{j=1}^M$ sampled from $p_X(x)$ and $p_Y(y)$. Recently, several one-sided methods were proposed, which essentially learn the generator $G : X \to Y$ and the discriminator D_Y that aims to distinguish between images $\{x\}$ and translated images $\{F(y)\}$. Commonly, the training objective consists of multiple pieces. The first one is the adversarial losses [16], Eqn. 1, for matching the marginal distribution of generated images to that of the target images.

$$\mathcal{L}_{\text{GAN}}(G, D_Y, X, Y) = \mathbb{E}_{y \sim p_Y(y)} \left[\log D_Y(y)\right] + \mathbb{E}_{x \sim p_X(x)} \left[\log\left(1 - D_Y(G(x))\right)\right] \quad (1)$$

The second part is usually a loss constraining the generator G to perverse desired contents during translations. For instance, the recent state-of-the-art method Contrastive Unpaired Translation (CUT) [40] tries to maximize the mutual information between the input and generated output via contrastive learning. It utilizes InfoNCE loss [39] to learn an embedding that associates corresponding patches (of input and translated images) to each other while disassociating them if otherwise. By doing so, it learns encoders that extract domain-invariant features of the input images at multiple scales. At each scale, the feature (an $\mathbb{R}^{256}$ vector) at one position from the input image is denoted as "query" v; the corresponding feature in the translated image is denoted the "positives" v^+; the features at N other locations from the input images are the "negatives" v^-. Formally, the contrastive loss is set up as an $(N+1)$–way classification as below (where τ is the temperature).

$$\ell\left(\boldsymbol{v}, \boldsymbol{v}^+, \boldsymbol{v}^-\right) = -\log\left[\frac{\exp\left(\boldsymbol{v} \cdot \boldsymbol{v}^+ / \tau\right)}{\exp\left(\boldsymbol{v} \cdot \boldsymbol{v}^+ / \tau\right) + \sum_{n=1}^{N} \exp\left(\boldsymbol{v} \cdot \boldsymbol{v}_n^- / \tau\right)}\right] \quad (2)$$

Although encouraging semantic correspondence between input and output images, CUT still suffers from semantics flipping when the two domains have unmatched semantics statistics. This is because the contrastively learned semantics are not accurate enough to ensure successful correspondence enforcement across domains. Nevertheless, when combined with other techniques to improve semantic robustness, these semantics extractors can be used to successfully reduce semantics flipping.

4.2. Semantically Robust Unpaired Image Translation (SRUNIT)

Our method is based on CUT [40] and the semantic robustness we proposed in Sec. 3.4. As illustrated in Fig. 3, in CUT, K layers (denoted $\{G_k\}$), including the input layer G_1 (an identity function), are selected from the first half of the generator G. Together with the rest of the network, denoted G_{K+1}, we have $G = G_{K+1} \circ G_K \circ \cdots \circ G_1$. We further define $\mathcal{G}_i^j = G_j \circ \cdots \circ G_{i+1}$ as a forward propagation via $(j-i)$ components of G. For instance, $G(x) = \mathcal{G}_1^{K+1}(x)$. During CUT training, at each scale $k \in \{1 \ldots K\}$ (just like in CUT, there are K scales in total), a feature extractor F_k that consumes the output of G_k (a layer in the generator) is learned by Eqn. 2, where v, v^- and v^+ are the outputs of F_k. The optimization of Eqn. 2 encourages the feature $F_k(\mathcal{G}_1^k(G(x)))$ to remain close to $F_k(\mathcal{G}_1^k(x))$.

We consider the perceptual variations mentioned in our semantic robustness concept as the random perturbations in the output space of G_k, and consider outputs of F_k as the semantics (by Eqn. 2 F_k extracts domain-invariant features). Then, we propose to improve semantic robustness by minimizing the loss $\mathcal{L}_{robust} = \frac{1}{K}\sum_{k=1}^{K} \mathcal{L}_k$, where

$$\mathcal{L}_k = \mathbb{E}_x \Big[\frac{1}{||\tau_k||_2} \Big\| F_k(\mathcal{G}_1^k(x)) - F_k(\mathcal{G}_1^k(\mathcal{G}_k^{K+1}(\mathcal{G}_1^k(x) + \tau_k))) \Big\|_2\Big] \quad (3)$$

The τ_k refers to some random perturbation. As shown in Fig. 3, $\mathcal{L}_k$ measures the distance between extracted semantics at scale k from the input image and that of the corresponding translated image under feature space perturbation to the input. We can see that minimizing $\mathcal{L}_k$ indirectly enforces semantic robustness, which is the condition that "transformed images should have their semantics invariant to small feature space variations of the inputs". Formally, this condition can be measured by

$$\mathbb{E}_x \Big[\frac{1}{||\tau_k||_2} \Big\| F_k(\mathcal{G}_1^k(G(x))) - F_k(\mathcal{G}_1^k(\mathcal{G}_k^{K+1}(\mathcal{G}_1^k(x) + \tau_k))) \Big\|_2\Big] \quad (4)$$

$\mathcal{L}_k$ and Eqn. 4 are closely related by the triangle inequality since, by contrastive learning (Eqn. 2), we have $F_k(\mathcal{G}_1^k(G(x)))$ remain close to $F_k(\mathcal{G}_1^k(x))$. In fact, our approach produces better translation results than directly optimizing Eqn. 4, because the latter can harm the diversity of the translation (the mode collapse issue). Optimizing $\mathcal{L}_k$ can be seen as an adaptive version of optimizing Eqn. 4, adjusted by the distance between $F_k(\mathcal{G}_1^k(G(x)))$ and $F_k(\mathcal{G}_1^k(x))$. See Sec. 6 for empirical evidence and see Appendix for a detailed discussion.

One might ask a question: why not directly minimize the distance between $F_k(\mathcal{G}_1^k(G(x)))$ and $F_k(\mathcal{G}_1^k(x))$ to enforce semantics-preserving translations. This will make duplicate efforts, similar to the contrastive loss (Eqn. 2) used in CUT. We show in the ablation study (Sec. 6) that doing so can actually hurt the performance.

Moreover, we adopt the patch-based approach so that x refers to the input image patches. In practice, one can choose to only include a random subset of $\{\mathcal{L}_k\}$ in each training iteration to reduce the computational complexity of optimizing $\mathcal{L}_{robust}$. See Sec. 5.3 for more details.

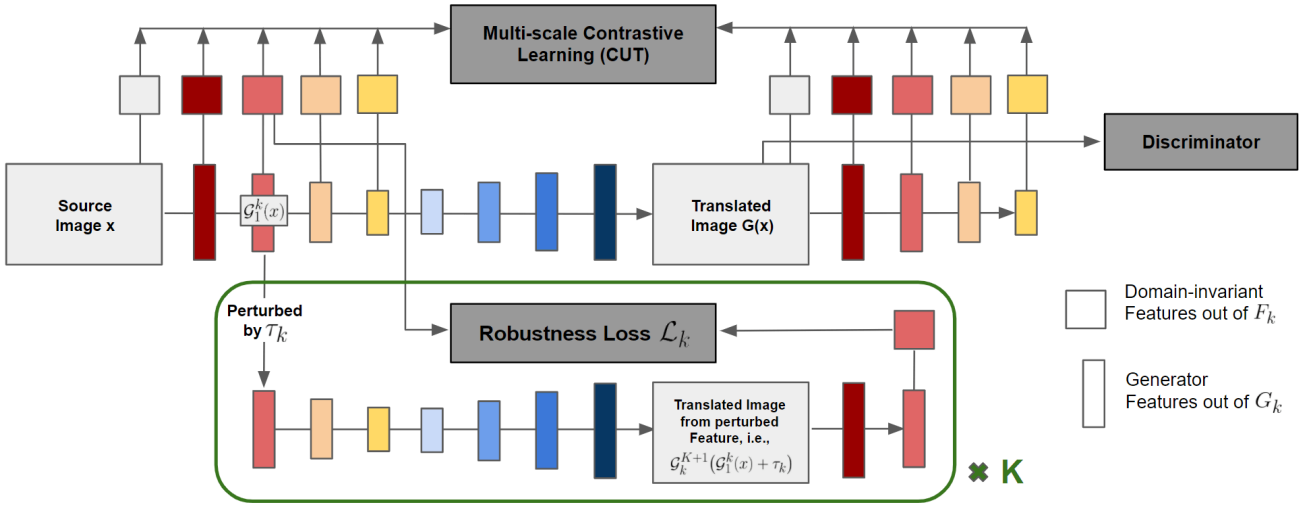

Figure 3: Our method improves semantic robustness by making the semantics of translated output invariant to small feature space variations of the inputs. The contents in the green box exemplify the data-flow for $\mathcal{L}_k$ at one specific scale k. There are in total K such losses; each corresponds to one of the K selected layers in the generator, respectively.

4.3. Advantages Over Distance-Preserving Methods

Our semantic robustness approach is advantageous to the distance-preserving approach (e.g., DistanceGAN & HarmonicGAN [7, 60]), which aims to improve the semantic consistency of the translations by maintaining the distance between different parts of the same sample during the transformations. See Sec. 6 for the empirical results.

Firstly, assuming no access to extra supervision or pre-trained models, the distance used in this line of work is based on image pixels, for instance, the (standardized) L1 distance between raw pixels or color histograms. Compared to the CNN feature-based perceptual similarity used in the concept of semantic robustness, pixel space similarity metrics are far more sensitive to geometric transformations, change of lighting conditions, etc.; thus, they usually fail to capture the essential information in the images. The authors of HarmonicGAN provide an option to use CNN features from pre-trained models to measure visual distance, which is domain-specific and requires prior knowledge, making this approach restrictive. Ours utilizes the contrastively learned features, being both universal and effective.

Secondly, the underlying principle of the distance-preserving approach is usually violated, i.e., the visual distance between contents in the source domain become changed when translated to the target domain. For instance, in the Label to Image task from Cityscapes [12] (i.e., translating semantic labels as inputs to street-view images), two identical image patches of the same semantics should not be mapped to visually identical outputs. Forcefully maintaining the visual distance as such leads to serious artifacts and hurts the diversity of the translations. Instead, our method encourages the outputs to be semantically the same, which can still be diverse and high-quality.

Thirdly, our method focuses on improving the robustness of the translations from the source domain w.r.t. feature space perturbations in all directions. Whereas, distance-preserving methods do so only w.r.t. pixel space perturbations in the direction of other source images (or in practice, in the direction of other patches from the same source image). This makes our approach much more efficient and effective in reducing semantics flipping.

5. Experiments

In this section, we demonstrate how our method (denoted SRUNIT) effectively reduces semantics flipping and produces translations that outperform existing approaches both quantitatively and qualitatively for several popular unpaired image translation tasks. In specific, we compare SRUNIT with CycleGAN [62], GcGAN [15], DRIT [31] and CUT [40]. Notice that some popular datasets are designed for paired translation tasks and are not very realistic for unpaired image translation (they admit perfectly aligned semantics statistics across the domains). Therefore, we subsample them (which aggravates the flipping issue) so that the setup becomes more realistic.

5.1. Quantitative Evaluation

It is critical to choose the right metrics to quantitatively evaluate the translation performance, as the focus of this paper is on reducing semantics flipping. Popular metrics such as FID and MMD [21, 18] ignore the unmatched semantics statistics nature of unpaired image translation datasets (see the discussion in Sec. 3.2); consequently, they are not suitable and can be even misleading here. Instead, we

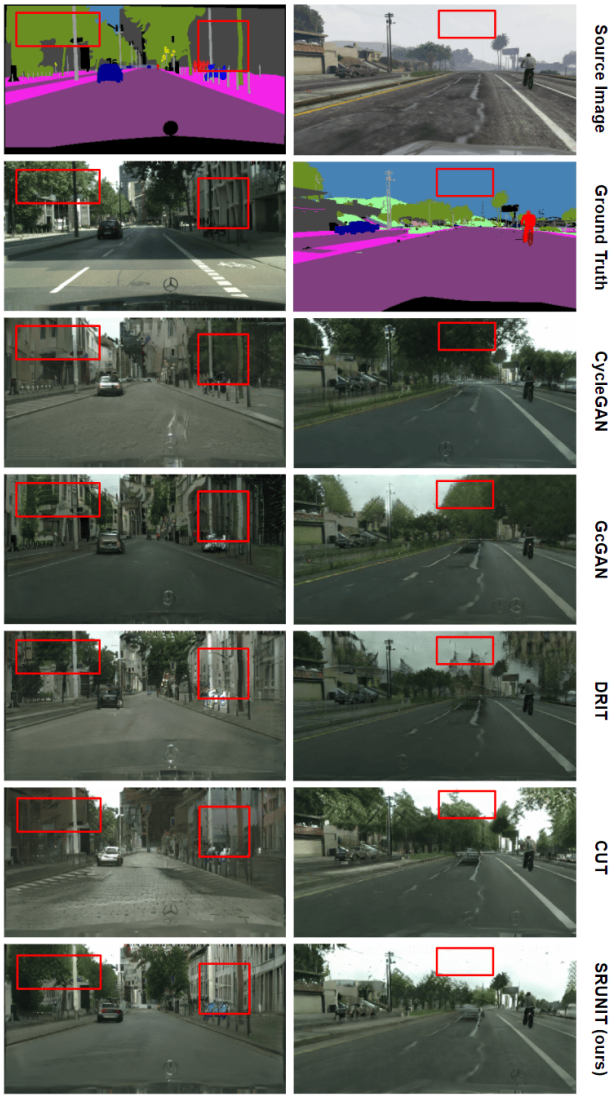

Figure 4: Visual results of the Label to Image and GTA to Cityscapes tasks (first and second column, respectively). Row 2 of column 2 shows the ground truth mask as no such ground truth image exists. Although not solved perfectly, semantics flipping is effectively reduced by our method (some improvements are highlighted in red boxes).

use datasets where (partial) information of the ground truth translated results are available and use corresponding metrics for evaluation. Some datasets (e.g., Aerial Photo to Google Map) directly provide ground truth correspondence, easing the evaluation of the translation quality. Others (Label to Image, GTA to Cityscapes, etc.) do not have such "ground truth" translation. On these datasets, we follow the common practice to compute metrics based on pre-trained models [11, 15, 40]. The intuition is that the more accurate the models (trained on source images) can classify the target images, the better these generated images are [25].

5.1.1 Cityscapes Label → Image

Cityscapes [12] is a real-world image dataset popular for benchmarking semantic segmentation and image translation. The dataset is originally constructed for paired translation. To ensure a reasonable level of unmatched semantics statistics between the two domains, we sub-sample around 1500 images from the RGB semantic label images and 1500 from the street-view images according to K-means clustering results based on histograms of the semantic labels. Each image is resized to 512×256 and during training, we randomly crop the 256×256 patches. As a result, the two domains have unmatched semantics statistics. We use the 500 validation set in Cityscapes for evaluation (whose ground truth semantic labels are provided). We use three metrics (as in [40]) to provide a comprehensive evaluation of the translation quality. They are the mean pixel accuracy, class accuracy (i.e., class-weighted pixel accuracy) and mean IoU (default metric for semantic segmentation). These metrics are computed by using a light-weight publicly available DeepLab V3 [10] model pre-trained on the Cityscapes semantic segmentation task (refer to the Appendix for details). We notice that, for these datasets, there are no standard pre-trained models for evaluation across existing work (e.g., CUT uses DRN [57] and CycleGAN & GcGAN use FCN [36]). We choose DeepLab V3 because its pre-trained models are public available and it is in general a better model for semantic segmentation. Table 1 and Fig. 4 show that SRUNIT produces results better than existing methods by a large margin. All detailed information is in the Appendix.

5.1.2 GTA → Cityscapes

GTA5 [44] is another popular dataset of 24966 synthesized images from the game Grand Theft Auto 5. We set aside 500 images from GTA5 for evaluation and use the remaining ones together with all the 2975 Cityscapes images (from Cityscapes' fine-labeled training set) for training. Similar to the Label to Image task, we resize all images to 512×256 and randomly crop 256×256 patches for training. The two datasets have quite different semantics statistics (as displayed in Fig. 1). Again, we use the DeepLab model to compute the three metrics. The quantitative results in Table 1 and qualitative results in Fig. 4 demonstrate the effectiveness of our proposed SRUNIT.

5.1.3 Google Map → Aerial Photo

The Google Maps dataset [25] contains in total 2194 (map, aerial photo) pairs of images around New York City and is widely used in both paired and unpaired image translation [25, 62, 15]. The dataset is split into 1096 pairs and 1098 pairs for training and test sets, respectively. Since it is original constructed for paired image translation, we sub-sample

	Label → Image			GTA → Cityscapes			Map → Photo			Photo → Map		
method	pxAcc	clsAcc	mIoU	pxAcc	clsAcc	mIoU	Dist	Acc(δ_1)	Acc(δ_2)	Dist	Acc(δ_1)	Acc(δ_2)
CycleGAN	66.36	27.24	21.31	66.33	32.53	**23.84**	70.16	28.67	43.88	23.02	16.11	32.65
GcGAN	65.30	27.78	21.41	65.62	32.38	22.64	71.47	28.87	43.48	23.62	15.00	30.65
DRIT	72.74	28.13	22.06	64.28	32.17	20.99	70.87	28.97	43.56	24.19	13.94	29.01
CUT	75.09	29.70	23.43	64.59	32.19	20.35	70.28	28.86	44.07	23.44	16.25	31.34
SRUNIT (ours)	**80.70**	**33.95**	**27.23**	**67.21**	**32.97**	22.69	**68.55**	**30.41**	**45.91**	**23.00**	**17.67**	**32.78**

Table 1: Average pixel prediction accuracy (pxAcc), average class prediction accuracy (clsAcc), Mean IoU (mIoU), Average L2 distance (Dist) and pixel accuracy with threshold (Acc) measured for the tasks. The best entries are highlighted in bold.

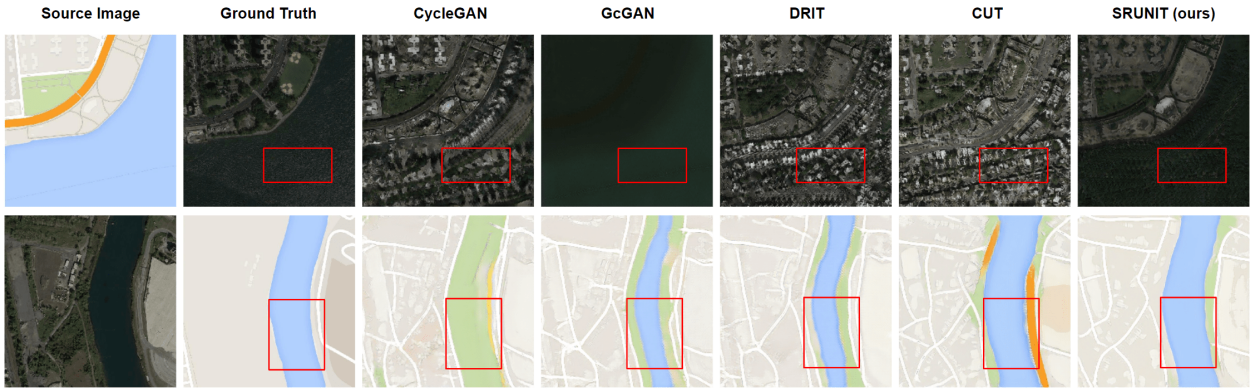

Figure 5: Visually in the Photo to Map tasks, our method effectively reduces semantics flipping (highlighted in red boxes).

map images and aerial photos from the training set (around 600 from each domain) so that there is a reasonable amount of difference between the semantics statistics from the two sets. We do so by K-means clustering the color histogram of the images (see the Appendix for details). We use all the 1098 test set pairs for evaluation. Images are resized to 256×256 to accommodate all methods in our comparison. We measure the quality of translation by average pixel L2 distance (Dist) and pixel accuracy (%), following [15], where given a ground truth pixel $p_i = (r_i, g_i, b_i)$ and the prediction $p'_i = (r'_i, g'_i, b'_i)$, the accuracy of p'_i is computed as 1 if $\max(|r_i - r'_i|, |g_i - g'_i|, |b_i - b'_i|) < \delta$ and 0 otherwise: We use $\delta = 30, 50$ as the domain of aerial photos has large diversity. Table 1 and Fig. 5 demonstrate the clear advantages of our method over existing ones.

5.1.4 Aerial Photo → Google Map

The evaluation protocol is similar as above, except that we use smaller $\delta = 3, 5$ since the domain of google maps has much less diversity than the other domain. Again, Table 1 and Fig. 5 show the advantage of our approach.

5.2. Qualitative Evaluation

Besides the aforementioned tasks, we show more visual results on the following three popular datasets (all training images resized to 256×256). In Fig, 6, we demonstrate that SRUNIT produce images of better or comparable quality compared to others. Due to the lack of ground truth translation information, it is not possible to quantitatively measure how well our model reduces semantics flipping.

Horse → Zebra A famous dataset consists of 1067 and 1334 training images for horses and zebras, respectively (link). The two domains have different semantics statistics.

Summer → Winter A dataset of photos of Yosemite constructed by authors of CycleGAN. The training set consists of 1231 summer images and 962 winter images (link). Again, the two domains have different semantics statistics.

Day → Night A dataset of outdoor scenes used in [25, 15]. While the original dataset consists of paired images, we sub-sample 1418 day images and 391 night images so that the semantics statistics are different (details in Appendix).

5.3. Implementation details

We follow CUT [40] for the choice of network architecture and the training setup (learning rate, number of epochs, etc.). The τ_k used in Eqn. 3 is sampled independently as a vector for each coordinate of the feature maps produced by G_k. We first project standard multivariate Gaussian random variable into the unit sphere and then resize it with its magnitude sampled uniformly in $[10^{-7}, T]$, where the default T we choose is 0.1 (we fine-tune it in $[0.01, 0.2]$). We set the default coefficient of the loss term $\mathcal{L}_{robust}$ as 10^{-4}, and fine-tune it in $[10^{-5}, 10^{-3}]$. By default we compute

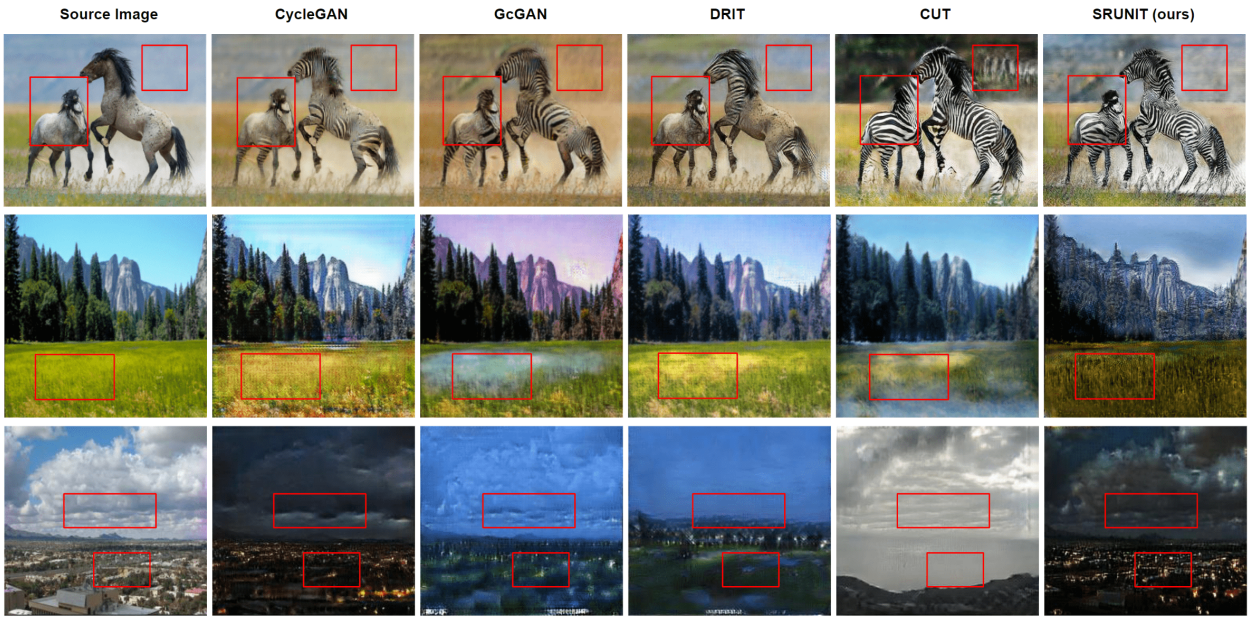

Figure 6: Visual comparisons on the three datasets (Horse to Zebra, Summer to Winter, Day to Night). Semantics flipping (or, in general, areas where our method improves over others) are highlighted in red boxes. See Appendix for more samples.

$\{\mathcal{L}_k\}$ (see Eqn. 3) for all 5 feature extractors $\{F_k\}$ used in the CUT paper. We fine-tune it by leaving one $\mathcal{L}_k$ out for each variant. We find all these hyper-parameters relatively robust. Optimizing $\mathcal{L}_{robust}$ might lead to training instability at the beginning of the adversarial training; thus we only apply it after finishing $\frac{1}{4}$ of the total epochs. See Appendix for full implementation details.

6. Ablation Study

Here we justify the design choice of our proposed semantic robustness loss $\mathcal{L}_{robust}$. We perform all the following experiments on the Label to Image dataset as it is a relatively challenging task. We use CUT as the backbone (as used in SRUNIT). The results are shown in Table 2.

Firstly, we show our method's advantage over the distance preserving approach (see also Sec. 4.3 for a discussion). We train a model denoted E1 by adding the self-distance constraints from DistanceGAN [7] to the CUT backbone and E2 by adding a patch-based distance-preserving constraint in the style of HarmonicGAN [60] (not exactly the same, though). To verify the necessity of using feature extractor F_k in $\mathcal{L}_{robust}$, we train E3 by removing the function calls to F_k in Eqn. 3. To show that Eqn. 3 is a better proxy of Eqn. 4 (as discussed in Sec. 4.2), we train E4 by using Eqn. 4 instead of Eqn. 3 when optimizing the $\mathcal{L}_{robust}$. We train a model E5 to illustrate that directly minimizing the distance between semantics (extracted by F_k) of the input and that of the output does not work (also see Sec. 4.2 for a discussion). We further show that applying constraints on the discriminator instead of the generator is not a better way to improve the semantic robustness of the model. We do so by training a model E6 with Lipschitz penalty [20] on the discriminator in the spirit of WGAN [20]. We provide full details in the Appendix.

	E1	E2	E3	E4	E5	E6	SRUNIT
pixAcc	74.46	75.31	75.42	76.38	74.86	76.25	**80.70**
clsAcc	29.84	30.13	30.43	31.13	29.79	30.92	**33.95**
mIoU	23.52	23.86	23.89	24.71	23.22	23.51	**27.23**

Table 2: Ablation studies (using CUT as backbone) on the Label to Image task in defense of our choice in SRUNIT.

7. Conclusion

In this paper, we tackle the semantic flipping problem in unpaired image translation which is critical for many of its applications. We argue that the inherently unmatched semantics distributions across different domains should be responded to by improving the semantic robustness of the generators. We do so by proposing a semantic robustness loss that enforces the semantics of the translated images to be invariant to the perceptual perturbations (specifically the multi-scale feature space perturbations) of the inputs. Quantitative and qualitative evaluations on multiple datasets suggest that our approach can effectively reduce semantics flipping that existing GAN-based methods suffer from.

References

[1] Naveed Akhtar and Ajmal Mian. Threat of adversarial attacks on deep learning in computer vision: A survey. *IEEE Access*, 6:14410–14430, 2018. 2

[2] Antreas Antoniou, Amos Storkey, and Harrison Edwards. Data augmentation generative adversarial networks. *arXiv preprint arXiv:1711.04340*, 2017. 1, 2

[3] Sanjeev Arora, Rong Ge, Yingyu Liang, Tengyu Ma, and Yi Zhang. Generalization and equilibrium in generative adversarial nets (gans). *arXiv preprint arXiv:1703.00573*, 2017. 2

[4] Sanjeev Arora, Rong Ge, Behnam Neyshabur, and Yi Zhang. Stronger generalization bounds for deep nets via a compression approach. In *International Conference on Machine Learning*, pages 254–263, 2018. 2

[5] Peter L Bartlett, Dylan J Foster, and Matus J Telgarsky. Spectrally-normalized margin bounds for neural networks. In *Advances in Neural Information Processing Systems*, pages 6240–6249, 2017. 2

[6] Dina Bashkirova, Ben Usman, and Kate Saenko. Adversarial self-defense for cycle-consistent gans. In *Advances in Neural Information Processing Systems*, pages 637–647, 2019. 2

[7] Sagie Benaim and Lior Wolf. One-sided unsupervised domain mapping. In *Advances in neural information processing systems*, pages 752–762, 2017. 2, 5, 8

[8] Konstantinos Bousmalis, Nathan Silberman, David Dohan, Dumitru Erhan, and Dilip Krishnan. Unsupervised pixel-level domain adaptation with generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3722–3731, 2017. 2

[9] Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, and Dumitru Erhan. Domain separation networks. In *Advances in neural information processing systems*, pages 343–351, 2016. 1

[10] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 6

[11] Casey Chu, Andrey Zhmoginov, and Mark Sandler. Cyclegan, a master of steganography. *arXiv preprint arXiv:1712.02950*, 2017. 1, 2, 3, 6

[12] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 2, 3, 5, 6

[13] Chris Donahue, Julian McAuley, and Miller Puckette. Adversarial audio synthesis. *arXiv preprint arXiv:1802.04208*, 2018. 1

[14] Gintare Karolina Dziugaite and Daniel M Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. *arXiv preprint arXiv:1703.11008*, 2017. 2

[15] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, Kun Zhang, and Dacheng Tao. Geometry-consistent generative adversarial networks for one-sided unsupervised domain mapping. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2427–2436, 2019. 2, 3, 5, 6, 7

[16] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. 1, 2, 4

[17] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014. 2

[18] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012. 3, 5

[19] Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg. Badnets: Identifying vulnerabilities in the machine learning model supply chain. *arXiv preprint arXiv:1708.06733*, 2017. 2

[20] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. In *Advances in neural information processing systems*, pages 5767–5777, 2017. 8

[21] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *arXiv preprint arXiv:1706.08500*, 2017. 3, 5

[22] 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. PMLR, 2018. 1, 2, 3

[23] Shanjiaoyang Huang, Weiqi Peng, Zhiwei Jia, and Zhuowen Tu. One-pixel signature: Characterizing cnn models for backdoor detection. *arXiv preprint arXiv:2008.07711*, 2020. 2

[24] 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. 2

[25] 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. 3, 6, 7

[26] Zhiwei Jia and Hao Su. Information-theoretic local minima characterization and regularization. *arXiv preprint arXiv:1911.08192*, 2019. 2

[27] Yiding Jiang, Dilip Krishnan, Hossein Mobahi, and Samy Bengio. Predicting the generalization gap in deep networks with margin distributions. In *International Conference on Learning Representations*, 2019. 2

[28] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016. 1, 2

[29] Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jung Kwon Lee, and Jiwon Kim. Learning to discover cross-domain relations with generative adversarial networks. *arXiv preprint arXiv:1703.05192*, 2017. 2

[30] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken,

Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017. 1

[31] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Diverse image-to-image translation via disentangled representations. In *Proceedings of the European conference on computer vision (ECCV)*, pages 35–51, 2018. 5

[32] Minjun Li, Haozhi Huang, Lin Ma, Wei Liu, Tong Zhang, and Yugang Jiang. Unsupervised image-to-image translation with stacked cycle-consistent adversarial networks. In *Proceedings of the European conference on computer vision (ECCV)*, pages 184–199, 2018. 2

[33] Xiaodan Liang, Hao Zhang, and Eric P Xing. Generative semantic manipulation with contrasting gan. *arXiv preprint arXiv:1708.00315*, 2017. 2, 3

[34] 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. 2

[35] Ming-Yu Liu and Oncel Tuzel. Coupled generative adversarial networks. In *Advances in neural information processing systems*, pages 469–477, 2016. 2

[36] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 6

[37] Youssef A Mejjati, Christian Richardt, James Tompkin, Darren Cosker, and Kwang In Kim. Unsupervised attention-guided image to image translation. *arXiv preprint arXiv:1806.02311*, 2018. 2

[38] Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. Norm-based capacity control in neural networks. In *Conference on Learning Theory*, pages 1376–1401, 2015. 2

[39] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 4

[40] Taesung Park, Alexei A Efros, Richard Zhang, and Jun-Yan Zhu. Contrastive learning for unpaired image-to-image translation. In *European Conference on Computer Vision*, pages 319–345. Springer, 2020. 2, 3, 4, 5, 6, 7

[41] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2536–2544, 2016. 1

[42] Aaditya Prakash, Nick Moran, Solomon Garber, Antonella DiLillo, and James Storer. Deflecting adversarial attacks with pixel deflection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8571–8580, 2018. 2

[43] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. *arXiv preprint arXiv:1605.05396*, 2016. 1

[44] Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *European conference on computer vision*, pages 102–118. Springer, 2016. 3, 6

[45] Pouya Samangouei, Maya Kabkab, and Rama Chellappa. Defense-gan: Protecting classifiers against adversarial attacks using generative models. *arXiv preprint arXiv:1805.06605*, 2018. 1

[46] Ashish Shrivastava, Tomas Pfister, Oncel Tuzel, Joshua Susskind, Wenda Wang, and Russell Webb. Learning from simulated and unsupervised images through adversarial training. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2107–2116, 2017. 2

[47] Jure Sokolić, Raja Giryes, Guillermo Sapiro, and Miguel RD Rodrigues. Robust large margin deep neural networks. *IEEE Transactions on Signal Processing*, 65(16):4265–4280, 2017. 2

[48] Jiawei Su, Danilo Vasconcellos Vargas, and Kouichi Sakurai. One pixel attack for fooling deep neural networks. *IEEE Transactions on Evolutionary Computation*, 23(5):828–841, 2019. 2

[49] Bo Sun, Nian-hsuan Tsai, Fangchen Liu, Ronald Yu, and Hao Su. Adversarial defense by stratified convolutional sparse coding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11447–11456, 2019. 2

[50] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013. 2

[51] Yaniv Taigman, Adam Polyak, and Lior Wolf. Unsupervised cross-domain image generation. *arXiv preprint arXiv:1611.02200*, 2016. 2, 3

[52] Hao Tang, Hong Liu, Dan Xu, Philip HS Torr, and Nicu Sebe. Attentiongan: Unpaired image-to-image translation using attention-guided generative adversarial networks. *arXiv preprint arXiv:1911.11897*, 2019. 2

[53] Hoang Thanh-Tung, Truyen Tran, and Svetha Venkatesh. Improving generalization and stability of generative adversarial networks. *arXiv preprint arXiv:1902.03984*, 2019. 2

[54] Kiran K Thekumparampil, Ashish Khetan, Zinan Lin, and Sewoong Oh. Robustness of conditional gans to noisy labels. In *Advances in neural information processing systems*, pages 10271–10282, 2018. 2

[55] Huan Xu and Shie Mannor. Robustness and generalization. *Machine learning*, 86(3):391–423, 2012. 2

[56] Yanchao Yang, Dong Lao, Ganesh Sundaramoorthi, and Stefano Soatto. Phase consistent ecological domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9011–9020, 2020. 2, 3

[57] Fisher Yu, Vladlen Koltun, and Thomas Funkhouser. Dilated residual networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 472–480, 2017. 6

[58] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE*

international conference on computer vision, pages 5907–5915, 2017. 1

[59] Pengchuan Zhang, Qiang Liu, Dengyong Zhou, Tao Xu, and Xiaodong He. On the discrimination-generalization tradeoff in gans. *arXiv preprint arXiv:1711.02771*, 2017. 2

[60] Rui Zhang, Tomas Pfister, and Jia Li. Harmonic unpaired image-to-image translation. *arXiv preprint arXiv:1902.09727*, 2019. 2, 5, 8

[61] Wenda Zhou, Victor Veitch, Morgane Austern, Ryan P. Adams, and Peter Orbanz. Non-vacuous generalization bounds at the imagenet scale: a PAC-bayesian compression approach. In *International Conference on Learning Representations*, 2019. 2

[62] 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. 2, 3, 5, 6