Alleviating Semantics Distortion in Unsupervised Low-Level Image-to-Image Translation via Structure Consistency Constraint

Jiaxian Guo[1] Jiachen Li[2] Huan Fu[1]
Mingming Gong[3] Kun Zhang[4,6] Dacheng Tao[1,5]
[1] The University of Sydney [2] Shanghai Jiao Tong University [3] The University of Melbourne
[4] Carnegie Mellon University [5] JD Explore Academy
[6] Mohamed bin Zayed University of Artificial Intelligence
jguo5934@uni.sydney.edu.au lijc0804@sjtu.edu.cn hufu6371@uni.sydney.edu.au
mingming.gong@unimelb.edu.au kunz1@cmu.edu dacheng.tao@gmail.com

Abstract

Unsupervised image-to-image (I2I) translation aims to learn a domain mapping function that can preserve the semantics of the input images without paired data. However, because the underlying semantics distributions in the source and target domains are often mismatched, current distribution matching-based methods may distort the semantics when matching distributions, resulting in the inconsistency between the input and translated images, which is known as the semantics distortion problem. In this paper, we focus on the low-level I2I translation, where the structure of images is highly related to their semantics. To alleviate semantic distortions in such translation tasks without paired supervision, we propose a novel I2I translation constraint, called Structure Consistency Constraint (SCC), to promote the consistency of image structures by reducing the randomness of color transformation in the translation process. To facilitate estimation and maximization of SCC, we propose an approximate representation of mutual information called relative Squared-loss Mutual Information (rSMI) that enjoys efficient analytic solutions. Our SCC can be easily incorporated into most existing translation models. Quantitative and qualitative comparisons on a range of low-level I2I translation tasks show that translation models with SCC outperform the original models by a significant margin with little additional computational and memory costs.

1. Introduction

Image-to-image translation, or domain mapping, aims to translate an image in the source domain $\mathcal{X}$ properly to the target domain $\mathcal{Y}$. It has been applied to various vision tasks [13, 46, 49, 59, 65]. Early works [18, 34, 44] considered supervised image-to-image (I2I) translation on paired datasets, and methods based on conditional generative adversarial networks can generate high-quality translations [18, 44, 60]. However, since paired data are often unavailable or expensive to obtain, unsupervised I2I translation has attracted intense attention in recent years [3, 17, 25, 26, 32, 43, 69, 73].

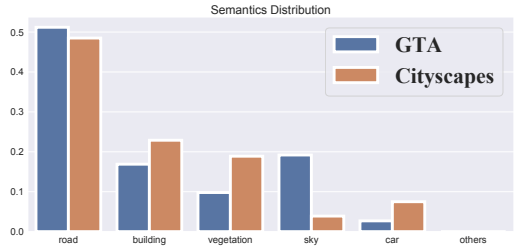


Figure 1. Class distributions in GTA and Cityscapes. We can see that the ratio of the sky in GTA is significantly higher than it in Cityscapes, and thus the distribution matching based method has to translate the sky to vegetation/building to align the distributions.

Benefiting from generative adversarial networks (GANs) [14], many works aim to perform unsupervised I2I translation by finding G_{XY} such that the translated images and target domain images have similar distributions, *i.e.*, $P_{G_{XY}(X)} \approx P_Y$. Due to an infinite number of functions that can satisfy the adversarial loss, GAN alone could learn a function far away from the true one. To remedy this issue, various constraints have been placed on the learned mapping function. For instance, the well-known cycle-consistency [26, 69, 73] enforces the translation function G_{XY} to be bijective. DistanceGAN [3] preserves the pairwise distances in the source images. GcGAN [10] forces the function to be smooth w.r.t. certain geometric transformations of input images. DRIT++ [32] and MUNIT [17] learn disentangled representations by embedding images onto a domain-invariant content space and a domain-specific attribute space and the mapping function can be then derived from representation learning components.

The above methods perform well when the two domains

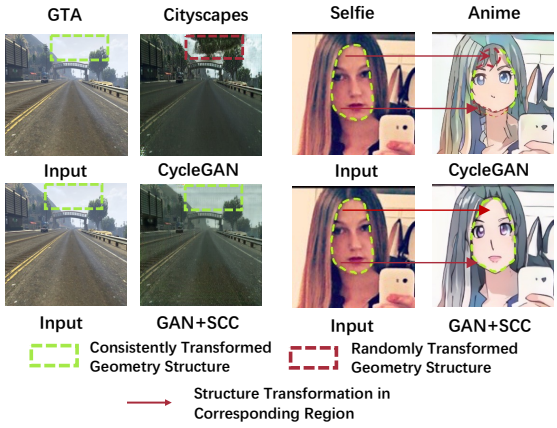


Figure 2. The illustration about the inconsistent geometry structure translation causes the semantic-distortion problem in unsupervised low-level image translation. Visually, we can see that the geometry structures of the sky and human face are distorted during translation in CycleGAN, which causes the semantical distortion *e.g.*, sky to vegetation, a face without fringe to face with fringe.

differ only in style information. However, in most unpaired datasets, not only style but also the underlying semantic distributions differ across source and target datasets [19]. Taking GTA to Cityscapes as an example, we perform the class statistics of GTA and Cityscapes, and the results are given as Figure 1. It can be seen that the class distributions in GTA are different from that in Cityscapes, *e.g.*, the proportion of sky in the GTA is significantly higher than that in Cityscapes, while the proportion of vegetation in GTA is lower than that in Cityscapes. Figure 2 also shows an example in selfie→anime translation, where the ratio of human faces with bangs in the Anime dataset is significantly higher than that in the Selfie dataset. In these cases, previous GAN-based methods *e.g.*, CycleGAN [73], which aims to align the distribution between domain *i.e.*, $P_{G_{XY}(X)} \approx P_Y$, may translate sky to building/vegetation in GTA2cityscape or automatically add the bangs on the human face in selfie2anime for the sake of aligning distribution (Figure 2), resulting in a semantic mismatch between input and translated images *i.e.*, *semantics distortion* problem.

It is hard to solve the *semantics distortion* problem in a universal way [19] when the given source and target dataset have unmatched semantics distributions because the characterization of semantics may vary from task to task. This lack of universally best choice is usually formalized in what is called the "No-Free Lunch" theorem [30, 63, 64], indicating that there is no single I2I algorithm that can perform better than all the other algorithms on all I2I applications. As such, we need to use suitable inductive bias [1, 24] to guide the translation model to preserve the related content according to the specific requirements of different I2I applications. For example, in high-level I2I image translation tasks, the pose/location of an object may be regarded as the semantics, but the type of object (e.g. cat→human face) is the style information that should be translated, and thus [65] introduces the pose bias to preserve pose structure properly during translation.

In this paper, we consider a widely applicable low-level image translation problem [5], which is fundamental in a wide range of computer vision applications, such as domain adaptation [16], segmentation [73], and simulation-to-real [45]. In low-level I2I, the difference between domains arises from the low-level information *e.g.*, resolution, illumination, color rather than geometry variation, while the structure (e.g. the shapes of objects) in images is most invariant across the source and target domains, *i.e.*, the semantics of an image is highly related to its structure (*shape of objects*). Therefore, the semantic distortion can be regarded as the change of structures in the translated images, as illustrated in Figure 2. Motivated by this, a natural solution to alleviate semantic distortion in this translation task would be to preserve the structure of source images.

To guarantee the consistency of image structure between source and translated images, we propose an I2I translation constraint, called *Structure Consistency Constraint* (SCC) . We observe that the pixel values before and after translation are usually highly correlated if the image structure is preserved (Figure 3). Based on this observation, we propose a mutual information (MI)-based dependency measure that models the nonlinear relationships between pixel values in the source and translated images. To efficiently estimate MI between pixel values, we propose the so-called relative Squared-Loss Mutual Information (rSMI) which can be estimated in an analytic form. By maximizing rSMI together with the GAN loss, our approach can significantly reduce the semantic distortion by better preserving image structures. In experiments, to show the effectiveness and compatibility of our *structure consistency constraint*, we incorporate it into the GAN framework and other existing image translation methods (*e.g.*, CycleGAN, CUT [43]). The quantitative and qualitative comparisons with existing I2I methods on several low-level tradatasets demonstrate that models with SCC outperform the corresponding baselines by a significant margin at only little computational and memory costs [1].

2. Methodology

Unsupervised I2I translation aims to find a mapping function G_{XY} between two domains $\mathcal{X}$ and $\mathcal{Y}$ given unpaired samples $\{x_i\}_{i=1}^N$ and $\{y_j\}_{j=1}^M$ drawn from the marginal distributions P_X and P_Y, respectively. To alleviate *semantics distortion* problem in low-level I2I translation, we directly promote the structure consistency of the source and translated images because the image structure is highly related to its semantics in this task. In the following, we first present our motivation of placing the MI-based structure consis-

[1]Codes are available at https://github.com/CR-Gjx/SCC

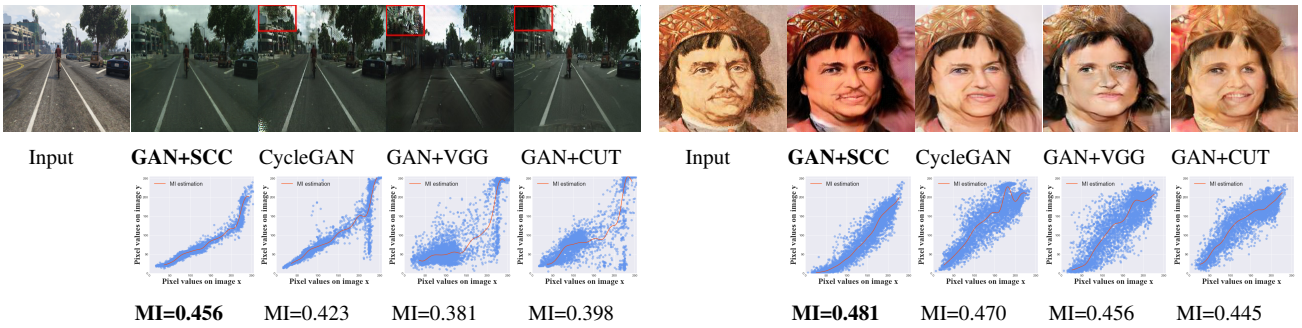


Figure 3. Unsupervised image translation examples on GTA → Cityscapes. Portrait → Photo. The top row is the translated results by each method. The bottom row is the scatter plot of the pixel values in the input image x and its corresponding pixel value in the translated image $\hat{y}$, which shows the non-linear dependency of pixel values in two images. Obviously, the stronger the dependency between pixel values in the input image (X-axis) and the translated images (Y-axis), the better the geometry structure of the input image is maintained. MI stands for the mutual information estimated by our rSMI method. Specifically, the VGG refers to the Contextual loss [39] of VGG features.

tency constraint (SCC), and then give the details about SCC, which aims to reduce the randomness of color transform in the translation process and thus promote the consistency of geometry structure between source and translated images.

2.1. Motivation

As illustrated in Figure 3, 5 (a), and 7, advanced methods, *e.g.*, CycleGAN, CUT [43], Contexual loss [39], U-GAT-IT [25], MUNIT [17], may change the geometry structure of input images and potentially cause the semantics mismatch between input and translated images. Therefore, it is essential to enforce a constraint such that we can ensure the learned function G_{XY} change the image style with minimal structure distortion. Our work is the first to explore such constraints for unsupervised image-to-image translation.

As we know, geometric structures in an images are often outlined by colors. So, if we hope to presereve the geometry structure during translation, we would expect the color translation to be consistent between the input and output images. For example, the green leaf in summer should be translated to yellow in autumn, but we do not expect it to be translated into a colorful one, otherwise, we cannot identify it as a leaf. Based on this observation, we plot the corresponding pixel values of images before and after translation at the bottom row of Figure 3. We can see that if the pixel values in the translated image (Y-axis) are more dependent on the pixel values (X-axis) in the input images, more structures will be preserved. Obviously, previous methods (*e.g.*, CycleGAN, CUT, Contextual loss of VGG feature) fail to translate color within a geometry structure consistently, and such randomness of the color transformations result in the distortion of geometry structure and semantics. Therefore, reducing the randomness of color transformation is an effective way to alleviate the *semantic-distortion* problem in I2I translation.

Motivated by the analysis, we develop the *structure consistency constraint* (SCC) as a general and effective constraint to preserve the pixel-level structure during the translation process. SCC exploits mutual information to model the non-linear dependencies of pixel values between the input and translated images, thus reducing the randomness of color transformation in the translation. As illustrated in Figure 4, our SCC is enforced into the input and translated images and thus allows one-sided unsupervised domain mapping, *i.e.*, G_{XY} can be trained independently from G_{YX}. Applying our SCC to a vanilla GAN, the pixel values before and after translation have stronger dependency (higher MI), and the model therefore better preserves the geometric structures as shown in Figure 3, thus reducing semantic distortion in low-level I2I translation. In the following, we present the details of our approach.

2.2. Approximate Representation of Mutual Information

For a source domain image $x_i \in \mathcal{X}$ and its translation $\hat{y}_i = G_{XY}(x_i)$, we denote V^{x_i} and $V^{\hat{y}_i}$ as the random variables for pixels in x_i and $\hat{y}_i$, respectively. Thus, pixels in x_i, i.e., $\{v_j^{x_i}\}_{j=1}^{M}$, can be regarded as data sampled from $P_{V^{x_i}}$, and the pixels in $\hat{y}_i$, i.e., $\{v_j^{\hat{y}_i}\}_{j=1}^{M}$, can be considered as data sampled from $P_{V^{\hat{y}_i}}$, where M is the number of pixels of the image. Formally, the mutual information between V^{x_i} and $V^{\hat{y}_i}$ is

$$MI(V^{x_i}, V^{\hat{y}_i}) = \mathbb{E}_{(v^{x_i}, v^{\hat{y}_i}) \sim P_{(V^{x_i}, V^{\hat{y}_i})}} \left(\log \frac{P_{(V^{x_i}, V^{\hat{y}_i})}}{P_{V^{x_i}} \otimes P_{V^{\hat{y}_i}}} \right) \tag{1}$$

where $P_{(V^{x_i}, V^{\hat{y}_i})}$ is the joint distribution of V^{x_i} and $V^{\hat{y}_i}$, $P_{V^{x_i}} \otimes P_{V^{\hat{y}_i}}$ is the product of two marginal distributions $P_{V^{x_i}}$ and $P_{V^{\hat{y}_i}}$. Because V^{x_i} and $V^{\hat{y}_i}$ are low-dimensional, a straightforward way to estimate (1) is to estimate the distributions P based on the histogram of the images. Next, we will introduce how we estimate the mutual information between pixels from two domain images and backpropagate it to optimize parameters in the translation network.

To enable efficient backpropagation, we propose the relative Squared-loss Mutual Information (rSMI), which is an extension of the well-known Squared-loss Mutual Information (SMI) [54] and can be estimated analytically. For con-

ventional presentation, we denote $P_{V^{x_i}} \otimes P_{V^{\hat{y}_i}}$ as S_i $P_{(V^{x_i},V^{\hat{y}_i})}$ as Q_i. Then, the SMI based on Pearson Divergence [53] between $P_{V^{x_i}}$ and $P_{V^{\hat{y}_i}}$ is expressed

$$\begin{aligned} SMI(V^{x_i}, V^{\hat{y}_i}) &= D_{PE}(P_{V^{x_i}} \otimes P_{V^{\hat{y}_i}} || P_{(V^{x_i},V^{\hat{y}_i})}) \\ &= D_{PE}(S_i || Q_i) \\ &= \mathbb{E}_{Q_i}[(\frac{S_i}{Q_i} - 1)^2]. \end{aligned}$$

Because $\frac{S_i}{Q_i}$ is unbounded, $SMI(V^{x_i}, V^{\hat{y}_i})$ can be int causing numeric instability in the backpropagation. We use the relative Pearson(rPE) Divergence [67] to alle the problem:

$$D_{rPE}(S_i \,||\, Q_i) = D_{PE}(S_i \,||\, \beta S_i + (1-\beta)Q_i).$$

Here, we introduce the mixture distribution $\beta S_i + (1-\beta)Q_i$, $\beta \in (0,1)$, to replace Q_i. Benefiting from the modification, the density ratio will be bounded to $[0, \frac{1}{\beta}]$. Thus, the proposed rSMI between V^{x_i} and $V^{\hat{y}_i}$ can be written as:

$$\begin{aligned} rSMI(V^{x_i}, V^{\hat{y}_i}) &= D_{rPE}(P_{V^{x_i}} \otimes P_{V^{\hat{y}_i}} || P_{(V^{x_i},V^{\hat{y}_i})}) \\ &= \mathbb{E}_{\beta S_i + (1-\beta)Q_i}[(\frac{S_i}{\beta S_i + (1-\beta)Q_i} - 1)^2] \end{aligned} \tag{4}$$

To estimate the $rSMI(V^{x_i}, V^{\hat{y}_i})$, we directly estimate the density ratio using a linear combination of kernel functions of $\{v_j^{x_i}\}_{j=1}^M$ and $\{v_j^{\hat{y}_i}\}_{j=1}^M$:

$$\begin{aligned} \frac{S_i}{\beta S_i + (1-\beta)Q_i} &= \omega_\alpha(v^{x_i}, v^{\hat{y}_i}) \\ &= \alpha^T \phi(v^{x_i}, v^{\hat{y}_i}) \end{aligned} \tag{5}$$

where $\phi \in \mathbb{R}^m$ is the kernel function, $\alpha \in \mathbb{R}^m$ is the parameter vector we need to solve, and m is the number of kernels. Referring to the least-squares density-difference estimation [52], the solved optimal solution of $\hat{\alpha}$ is (the derivation is given in the appendix A.1):

$$\begin{aligned} \hat{\alpha} &= (\hat{H} + \lambda R)^{-1}\hat{h}, \\ \hat{H} &= \frac{1-\beta}{n}(K \circ L)(K \circ L)^T + \frac{\beta}{n^2}(KK^T) \circ (LL^T), \\ \hat{h} &= \frac{1}{n^2}(K1_n) \circ (L1_n) \end{aligned} \tag{6}$$

where R is a positive semi-definite regularization matrix, n is the sample number, 1_n is the n-dimensional vector filled by ones, and K and L are two $m \times n$ matrices composed by kernel functions, and the Hadamard product of K and L is used to define ϕ, that is $\phi(v^{x_i}, v^{\hat{y}_i}) = K(v^{x_i}) \circ L(v^{\hat{y}_i})$. Finally, an appropriate mutual information estimator of with smaller bias is expressed as:

$$\widehat{rSMI}(V^{x_i}, V^{\hat{y}_i}) = 2\hat{\alpha}^T\hat{h} - \hat{\alpha}^T\hat{H}\hat{\alpha} - 1. \tag{7}$$

Note that, the computation of $\widehat{rSMI}(V^{x_i}, V^{\hat{y}_i})$ is resource friendly, as it can be solved analytically. Thus, the parameters in the translation neural network can be efficiently updated by backprogation.

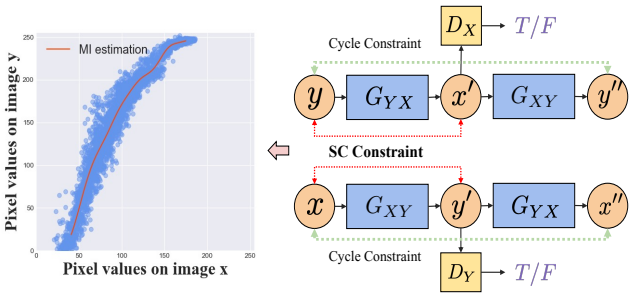


Figure 4. An illustration of structure consistency constraint. The left figure shows that the pixel value in the input image x and its corresponding pixel value in the translated image $\hat{y}$ have strong non-linear dependencies, so we add the structure consistency constraint to model the dependencies of pixel values in two domain images.

2.3. Full Objective

Following the analysis above, our structure consistency constraint (SCC) for I2I translation using mutual information can be expressed as:

$$\mathcal{L}_{SCC} = \frac{1}{N}\sum_{i=1}^{N} \widehat{rSMI}(V^{x_i}, V^{G_{XY}(x_i)}), \tag{8}$$

where N is the number of samples, and $G_{XY}(x_i) = \hat{y}_i$. We directly maximize $\mathcal{L}_{SCC}$ to guarantee more local geometric structures of images being invariant in the translation process. By combining SCC with the standard adversarial loss, the image geometry will be preserved while its style is changed. As a result, one-sided unsupervised domain mapping can be targeted. The full objective will take the form:

$$\begin{aligned} &\min_{G_{XY}} \max_{D_Y} \mathcal{L}_{GAN+SCC}(G_{XY}, D_Y) \\ &\qquad = \mathcal{L}_{GAN}(G_{XY}, D_Y) - \lambda_{SCC}\mathcal{L}_{SCC}(G_{XY}), \end{aligned} \tag{9}$$

where $\mathcal{L}_{gan}$ is the adversarial loss [14], which introduced a discriminator D_Y, to encourage the distribution of output matches the distributions of target domain images, *i.e*, $P_{G_{XY}(X)} \approx P_Y$. In addition, to guarantee the distribution consistency in the pixel level, we use a GAN based on the 1×1 convolution. The objective function is as follows:

$$\begin{aligned} \mathcal{L}_{GAN}(G_{XY}, D_Y) &= \mathbb{E}_{y\sim P_Y}[\log D_Y(y)] \\ &\quad + \mathbb{E}_{x\sim P_X}[\log(1 - D_Y(G_{XY}(x))). \end{aligned} \tag{10}$$

In Equation 9, λ_{SCC} is a hyperparameter to weight $\mathcal{L}_{gan}$ and $\mathcal{L}_{SCC}$ in the training procedure. The proposed SCC can easily be integrated into various I2I translation frameworks, *e.g.,* CycleGAN [73] and CUT [43], by replacing the loss $\mathcal{L}_{gan}$ with the losses in these methods.

3. Experiments

In this section, we perform quantitative experiments on three typical unsupervised low-level image translation benchmarks: Digits Translation, Unsupervised Segmentation and

(a) GAN (b) GAN + SCC (c) CycleGAN (d) CycleGAN + SCC

Figure 5. Qualitative comparisons on SVHN→MNIST. From Figure (a) and (b), we can see that the GAN method has no collapse solution by combining with our SCC. Also, the semantics distortion problem in CycleGAN is alleviated after incorporating with SCC.

Table 1. Classification accuracy for digits experiments.

	Translated Images as Test set			Translated Images as Training set		
Method	S → M	M → M-M	M-M → M	S → M	M → M-M	M-M → M
GAN alone	21.3±9.5	54.6±40.5	80.3±3.5	28.6±10.8	45.7±31.2	95.5±0.4
+ SCC	37.3±1.2	96.3±0.2	90.9±0.5	47.9±2.3	86.2±1.9	96.0±0.1
CycleGAN	26.1±8.1	95.3±0.4	84.7±2.5	31.6±5.6	83.8±3.0	95.9±0.4
+ SCC	38.0±0.5	**96.7±0.1**	91.5±0.3	47.4±2.0	87.7±2.1	**96.1±0.2**
GcGAN-*rot*	32.5±2.0	95.0±0.6	85.9±0.8	40.9±6.5	84.6±2.8	96.0±0.1
+ SCC	36.5±1.3	96.4±0.3	91.8 ±1.0	47.5±1.2	89.5±0.6	96.1±0.1
GcGAN -*vf*	33.3±4.2	95.2±0.4	84.5±1.5	31.6±5.6	83.8±3.0	95.9±0.4
+ SCC	37.0±0.8	96.6±0.3	91.8±0.8	49.5±4.9	87.8±2.3	96.0±0.1
Cyc + rot + SCC	39.0±0.5	96.5±0.3	91.8±1.0	50.5±1.8	**89.8±0.5**	**96.1±0.1**
Cyc + vf + SCC	**44.6±6.8**	**96.7±0.3**	**92.0±0.8**	**51.3±5.4**	89.0±0.8	**96.1±0.1**

Image Generation (*e.g.*, Cityscapes [7]), and Simulation-to-Real (*e.g.*, Maps [18] and GTA2cityscapes [45]). Because these benchmarks have the true label of the translation images, we can quantitatively evaluate whether the translation model causes the semantics distortion problem or not. Further, to qualitatively evaluate the translation quality of our method, we also perform experiments on Selfie → Anime, Portrait → Photo, Horse → Zebra datasets.

Effectiveness and Compatibility We couple our structure consistency constraint (SCC) with the vanilla GAN to show its effectiveness, and incorporate SCC with some popular methods such as CycleGAN [73], GcGAN [10], and U-GAT-IT [25] to show its compatibility. Then we make qualitative and quantitative comparisons with the recent published unsupervised I2I translation methods *e.g.*, CycleGAN [73], GcGAN [10], CoGAN [35], SimGAN [48], BiGAN [8] , DistanceGAN [3], CUT [43]), the VGG-based Contextual loss [39], the VGG-based Content loss [12], L1 loss of VGG feature [39], DRIT++ [32], UNIT [33], MUNIT [17], AGGAN [58], and U-GAT-IT [25]. Specifically, the current baselines have their own advantages and disadvantages: some baselines perform well on one task but perform poorly on other tasks. For example, some style transfer methods do not perform well on unsupervised image segmentation. As such, following the current literature, we compare our methods with SOTA methods for each application.

Sensitivity We perform the sensitivity analysis by varying the hyper-parameter λ_{SCC} on GTA2cityscapes.

In the appendix, we investigate the influence of our SCC on the generation diversity A.2.2 and training stability A.2.3. We examine all the experiments three times and report the average scores to reduce random errors.

For the implementation of the mutual information estimator presented in section 2.2, we set the hyperparameter β to 0.5 (more analysis about other values of β are given at the appendix A.2.1), and utilize nine Gaussian kernels for both input images x and translated images $\hat{y}$. Then we apply our SCC to all the baselines and keep other experimental details including hyper-parameters, networks in baselines the same. Due to page limit, we provide more experimental details and qualitative results in the Appendix A.6 and A.7, respectively.

3.1. Quantitative Evaluation

3.1.1 Digits Translation

We examine three digit I2I translation tasks: SVHN→MNIST, MNIST-M→MNIST and MNIST→MNIST-M [2]. The models are trained on the training split with images size 32 × 32, and λ_{SCC} is set to 20. We adopt the classification accuracy as the evaluation metric, and design two evaluation methods: (1) we train a classifier on the target dataset's training split. The fake images translated from the source dataset's test images are used to compute the classification accuracy. This evaluation method can only measure the quality of translated images. (2) a classifier is trained on the translated images from the source dataset's training images, and test the performance of this classifier on the target dataset's test split. This evaluation method can measure both the quality

[2] refer to S→M, M-M→M and M→M-M

Table 2. Quantitative scores on GTA → Citycapes,Citycapes parsing → image and Photo → Map. The scores with * are reproduced on a single GPU using the codes provided by the authors. More qualitative results are given at the Appendix A.7.2.

Methods	GTA → Citycapes			Citycapes parsing → image			Photo → Map		
	pixel acc ↑	class acc ↑	mean IoU ↑	pixel acc↑	class acc↑	mean IoU↑	RMSE ↓	acc%(δ_1) ↑	acc%(δ_2) ↑
CoGAN	\	\	\	0.40	0.10	0.06	\	\	\
BiGAN/ALI	\	\	\	0.19	0.06	0.02	\	\	\
SimGAN	\	\	\	0.20	0.10	0.04	\	\	\
DistanceGAN	\	\	\	0.53	0.19	0.11	\	\	\
GAN + VGG	0.216	0.098	0.041	0.551	0.199	0.133	34.38	28.1	48.8
DRIT++	0.423	0.138	0.071	\	\	\	32.12	29.8	52.1
GAN *	0.382	0.137	0.068	0.437	0.161	0.098	33.22	19.3	42.0
+ SCC	0.487	0.148	0.089	0.642	0.215	0.155	28.91	38.6	61.8
GcGAN-rot *	0.405	0.139	0.068	0.551	0.197	0.129	27.98	42.8	64.6
+ SCC	0.445	0.162	0.080	0.651	0.228	0.162	**26.55**	**44.7**	**66.5**
CycleGAN *	0.232	0.127	0.043	0.52	0.17	0.11	26.81	43.1	65.6
+ SCC	0.386	0.161	0.076	0.571	0.192	0.134	26.61	44.7	66.2
CUT *	0.546	0.165	0.095	0.695	0.259	0.178	28.48	40.1	61.2
+ SCC	**0.572**	**0.185**	**0.11**	**0.699**	**0.263**	**0.182**	27.34	39.2	60.5

Figure 6. Unsupervised image translation examples on GTA → Cityscapes. The generated examples clearly show that our SCC can alleviate the semantic distortion problem *e.g.*, sky to tree/building in mainstream translation models. More examples are given at Appendix A.7

and diversity of translation images, but it is unstable [3].

We conduct each experiment five times to reduce the randomness of GAN-based approaches. The scores are reported in Table 1. Generally, by incorporating our SCC, all the baselines show promising improvements in both accuracy and stability, especially for the challenging task S→M. Some qualitative results are shown in Figure 5. More details and results are given in Appendix A.6.1 and A.7.1, respectively.

3.1.2 Segmentation in Cityscapes

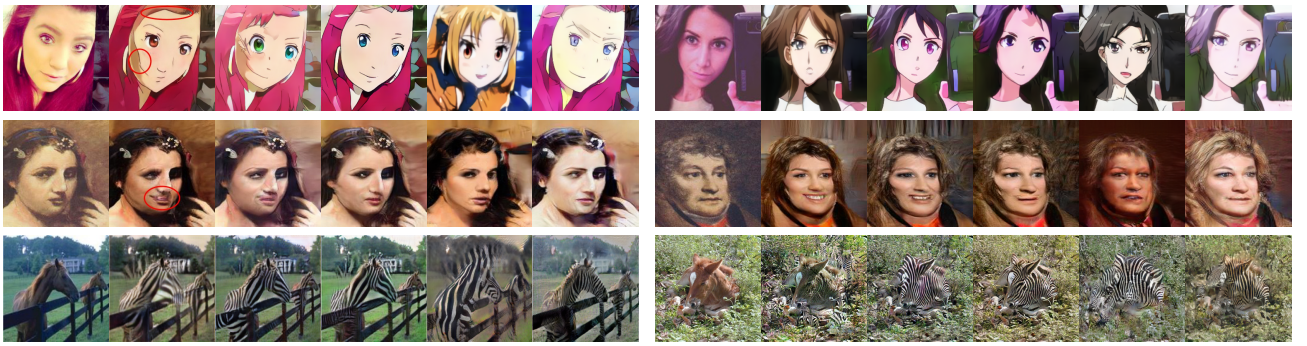
Following [10, 73], we train the models using the unaligned 3975 images of Cityscapes [7] with 128×128 resolution. We evaluate the domain mappers using FCN scores and scene parsing metrics as previously done in [73]. Specifically, for parsing→image, we use the pre-trained FCN-8s [36] provided by pix2pix [18] to predict segmentation label maps from translated images, then compare them with true labels using parsing metrics including pixel accuracy, class accuracy, and mean IoU. We do not report the score of DRIT++, because its network size is too big to perform experiments with 128×128 resolution, resulting in the unfair comparison with other methods, but the results of other datasets can still show the superiority of our method over DRIT++.

As reported in Table 2, the results of all the image translation methods are improved if further constrained by our SCC, which shows the effectiveness of our method on reducing the semantics distortion problem. In particular, GcGAN coupled with SCC yields a promising improvement compared with GcGAN in the parsing → image task.

3.1.3 Maps

The Maps dataset [18] contains 2194 aerial photo-map image pairs, with 1096 pairs for training and 1098 pairs for evaluation. For evaluation, we employ the metrics including RMSE and pixel accuracy with threshold δ ($\delta_1 = 5$ and $\delta_2 = 10$) suggested by GcGAN [10]. All images are resized to 256×256 resolution. Following [10, 73], the network details are similar to the details of Cityscape, but the generator contains 9 res-blocks for images with 256×256 resolution.

[3] Domain adaptation. has access to the labels of source domain images while I2I translation does not.



Input GAN+VGG CycleGAN Cycle+SCC U(light) U(light)+SCC Input GAN+VGG CycleGAN Cycle+SCC U(light) U(light)+SCC

Figure 7. Qualitative results on Selfie → Anime, Portrait → Photo, Horse → Zebra datasets. More qualitative results are given in A.7.3. We can see that the no matter personal identification or horse shape is better preserved by the translation model empowered by our SCC.

The scores are reported in Table 2. Compared with the vanilla GAN, our SCC can significantly improve translation accuracy to 38.6% and 61.8% from 19.3% and 42.0% with the threshold of δ_1 and δ_2, respectively. Moreover, integrating our SC constraint into CycleGAN and GcGAN can generate better translations than both individual ones. This further demonstrates the compatibility of our SCC. Qualitative results are shown in A.7.1.

3.1.4 Simulation to Real: GTA to Cityscapes

To evaluate the effectiveness of our SCC on simulation to real tasks, we use the GTA [45] to cityscapes datasets. Specifically, we use the official training split of GTA dataset the training dataset. All images are resized to 256×256 resolution during training. In the test process, we translate the first 500 images in the GTA test set to the cityscapes style, and use the pre-trained FCN-8s [36] provided by pix2pix [18] to predict the segmentation label maps from translated images, and calculate the scores with the true label in the GTA.

The results are give as Table 2, and the sample translated images are given as Figure 6. Our SCC can consistently alleviate the semantic distortion problem in GTA2cityscape task, as Figure 6 shows, all other translation models tend to translate sky to vegetation to align the distribution, but the translation model with SCC can maintain sky during translation, and thus we can consistently improve the segmentation score when coupling SCC with other models.

3.2. Qualitative Evaluation

We implement the qualitative evaluation on anime2selfie [25], horse2zebra [73], photo2portrait [31]. We choose CycleGAN, GcGAN, AGGAN, DRIT, UNIT, MUNIT, and CUT as baselines. All images are resized to 256×256 resolution. More experimental details are given in A.4.4.

Following [25], we use KID score [4] as the evaluation metric. The results are reported in Appendix A.3.1 because the pages are limited, and we can see that the method coupled with our SCC can even achieve better results than those methods with larger model sizes. As the qualitative results are shown in Figure 7, after adding our SCC, the translated images retain more geometric structure than the original images, and are consistent with the style of the target images. Specifically, the light version of U-GAT-IT with our SCC can achieve better performance than the full version of U-GAT-IT, even with a half size of parameters. Then we conducted a user study, in which 180 participants were asked to choose the best-translated image given the domain names *e.g.*, selfie → anime, exemplar images in the source and target domains, and the corresponding translated images from different methods. The results shown in Table 3 demonstrate that most users choose the outputs of our method, which shows that preserving the structure of the image can significantly improve the appearance attraction of the translated images. More qualitative results are given in appendix A.7.4 .

Table 3. The results of User Study: the percentage of users prefer a particular model. To avoid the concern of cherry-picking, qualitative results of U-GAT-IT and our results are used in the user study. Sample images are given in Appendix A.7.3.

	hor2zeb	sel2ani	pho2por	Paramaters
Cyc+Gc+SCC	**33.20**	**47.85**	**56.89**	45.2MB
U-GAT-IT	32.22	37.22	19.00	134.0MB
MUNIT	1.25	1.67	8.44	46.6MB
DRIT	5.28	2.94	3.00	65.0MB
CycleGAN	28.05	10.32	12.67	28.3MB

3.3. Sensitivity Analysis

We study the influence of SCC by performing experiments with different λ_{SCC}. As shown in Table 4 and Figure 8, the performance of translation models are all improved to some extent after incorporating our SCC. However, when λ_{SCC} becomes too large, the improvement with our SCC is limited as the model focuses on reducing geometry distortion and ignores the style information learned from GAN. More examples are given in Appendix A.7.5. A practical strategy of choosing λ_{SCC} is to find the largest λ_{SCC} with normal style information using binary search. Specifically,

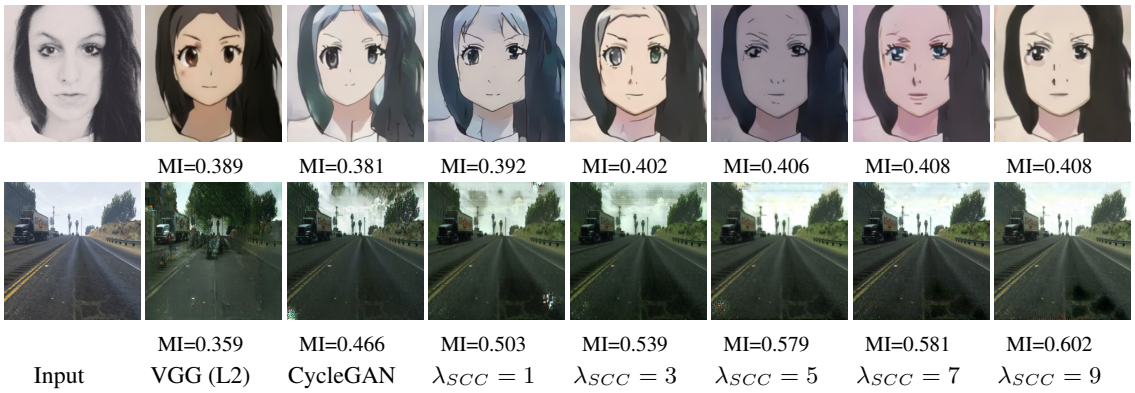


Figure 8. Sensitivity analysis examples on Selfie → Anime and GTA → Cityscapes. Obviously, the semantics distortion problem in CycleGAN is alleviated after incorporating with our SCC.

the first value of λ_{SCC} can be set to 5, which can promote the structure consistency of most translation models.

Table 4. The segmentation scores for different λ_{SCC} of the model CycleGAN + SCC in the datasets GTA2cityscapes.

λ_{SCC}	0	1	3	5	7	9
pixel acc ↑	0.232	0.292	0.322	0.360	0.382	0.386
class acc ↑	0.127	0.136	0.143	0.160	0.160	0.161
mean IoU ↑	0.0432	0.055	0.059	0.070	0.075	0.076

4. Related Work

Unsupervised Image-to-Image Translation. Although unsupervised image-to-image (I2I) translation has obtained some promising progress in recent years, several works study it from an optimization perspective. Specifically, Cyclic consistency based GAN, *e.g.,* CycleGAN [73], DualGAN [69] and DiscoGAN [26], is a general approach for this problem. DistanceGAN [3] and GcGAN [10] further introduced distance and geometry transformation consistency to constraint the search space of mapping functions. Instead of exploiting general constraints for the subject, more works developed novel frameworks to investigate special settings of unsupervised I2I translation. Several other works [6, 17, 31, 32, 47] mapped the content and style information of images into disentangled spaces for multi-modal translations. However, we find that the complex neural networks and many hyper-parameters make the optimization process unstable [25]. [9, 12, 20, 23, 39] tried to reduce the perceptual loss or content loss based on a pre-trained VGG model to reduce the content of two domain image, which is computationally cost and cannot be easily adapted to the data on hand. Moreover, [5, 40, 48, 56, 57, 60, 68, 71] use the attention-based/ pretrained model or pre-define functions to preserve the semantics during translation. SRUNIT [19] promote the robustness of feature translation, but SRUNIT is mainly incorporated into CUT [43]. However, how to preserve the semantics via low-level information is under explored.

Mutual Information (MI). Mutual information is the measure of dependency between two random variables, and it is widely used in machine learning and particularly suitable for canonical tasks, *e.g.*, multi-modalities images registration [37, 38, 74]. Since computing MI is difficult [42], researchers have taken much effort to improve the estimation of MI. For example, early works studied Non-parametric models based on Kernel Density Estimator (KDE) [21, 22, 29, 50, 51], K-nearest Neighbor Method (KNN) [27, 28], and likelihood-ratio estimator [55] for MI estimation. Subsequent works improved the performance in more complicated cases such as discrete-continuous mixtures [11, 41], segmentation [66, 70, 72] and continue learning [61, 62]. Recently, MINE [2, 15] showed that the mutual information between high dimensional continuous random variables can be estimated by gradient descent over neural networks.

5. Conclusion

In this paper, we propose the structure consistency constraint (SCC) to improve the structure consistency in pixel-wise level for unsupervised image-to-image translation. To enable efficient estimation of our constraint, we propose an expression of mutual information called relative Squared-loss Mutual Information(rSMI) with an analytical estimation method. We evaluate our model quantitatively in a wide range of applications. The experimental results demonstrate that SCC can achieve high-quality translation to maintain images' geometry in the original domain.

6. Acknowledge

Mr Jiaxian Guo is supported in part by Australian Research Council Projects FL-170100117 and IH-180100002. Dr Mingming Gong is supported by Australian Research Council Project DE210101624. Prof Kun Zhang would like to acknowledge the support by the National Institutes of Health under Contract R01HL159805, by the NSF-Convergence Accelerator Track-D award #2134901, and by the United States Air Force under Contract No. FA8650-17-C7715.

References

[1] Peter W Battaglia, Jessica B Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinicius Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, et al. Relational inductive biases, deep learning, and graph networks. *arXiv preprint arXiv:1806.01261*, 2018. 2

[2] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeswar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and R Devon Hjelm. Mine: mutual information neural estimation. *arXiv preprint arXiv:1801.04062*, 2018. 8

[3] Sagie Benaim and Lior Wolf. One-sided unsupervised domain mapping. In *Advances in neural information processing systems*, pages 752–762, 2017. 1, 5, 8

[4] Mikołaj Bińkowski, Dougal J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*, 2018. 7

[5] 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, 8

[6] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8789–8797, 2018. 8

[7] 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. 5, 6

[8] Jeff Donahue, Philipp Krähenbühl, and Trevor Darrell. Adversarial feature learning. *arXiv preprint arXiv:1605.09782*, 2016. 5

[9] Alexey Dosovitskiy and Thomas Brox. Generating images with perceptual similarity metrics based on deep networks. *arXiv preprint arXiv:1602.02644*, 2016. 8

[10] 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. 1, 5, 6, 8

[11] Weihao Gao, Sreeram Kannan, Sewoong Oh, and Pramod Viswanath. Estimating mutual information for discrete-continuous mixtures. In *Advances in neural information processing systems*, pages 5986–5997, 2017. 8

[12] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2414–2423, 2016. 5, 8

[13] Arnab Ghosh, Richard Zhang, Puneet K Dokania, Oliver Wang, Alexei A Efros, Philip HS Torr, and Eli Shechtman. Interactive sketch & fill: Multiclass sketch-to-image translation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1171–1180, 2019. 1

[14] 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, 4

[15] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*, 2018. 8

[16] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei A Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. *arXiv preprint arXiv:1711.03213*, 2017. 2

[17] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 172–189, 2018. 1, 3, 5, 8

[18] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017. 1, 5, 6, 7

[19] Zhiwei Jia, Bodi Yuan, Kangkang Wang, Hong Wu, David Clifford, Zhiqiang Yuan, and Hao Su. Semantically robust unpaired image translation for data with unmatched semantics statistics. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14273–14283, 2021. 2, 8

[20] 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. 8

[21] Erik Jonsson and Michael Felsberg. Soft histograms for belief propagation. 2006. 8

[22] Kirthevasan Kandasamy, Akshay Krishnamurthy, Barnabas Poczos, Larry Wasserman, et al. Nonparametric von mises estimators for entropies, divergences and mutual informations. In *Advances in Neural Information Processing Systems*, pages 397–405, 2015. 8

[23] Oren Katzir, Dani Lischinski, and Daniel Cohen-Or. Cross-domain cascaded deep feature translation. *arXiv*, pages arXiv–1906, 2019. 8

[24] Kenji Kawaguchi, Leslie Pack Kaelbling, and Yoshua Bengio. Generalization in deep learning. *arXiv preprint arXiv:1710.05468*, 2017. 2

[25] Junho Kim, Minjae Kim, Hyeonwoo Kang, and Kwanghee Lee. U-gat-it: unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation. *arXiv preprint arXiv:1907.10830*, 2019. 1, 3, 5, 7, 8

[26] Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jung Kwon Lee, and Jiwon Kim. Learning to discover cross-domain relations with generative adversarial networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1857–1865. JMLR. org, 2017. 1, 8

[27] LF Kozachenko and Nikolai N Leonenko. Sample estimate of the entropy of a random vector. *Problemy Peredachi Informatsii*, 23(2):9–16, 1987. 8

[28] Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. Estimating mutual information. *Physical review E*, 69(6):066138, 2004. 8

[29] Akshay Krishnamurthy, Kirthevasan Kandasamy, Barnabas Poczos, and Larry Wasserman. Nonparametric estimation of renyi divergence and friends. In *International Conference on Machine Learning*, pages 919–927, 2014. 8

[30] Tor Lattimore and Marcus Hutter. No free lunch versus occam's razor in supervised learning. In *Algorithmic Probability and Friends. Bayesian Prediction and Artificial Intelligence*, pages 223–235. Springer, 2013. 2

[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. 7, 8

[32] Hsin-Ying Lee, Hung-Yu Tseng, Qi Mao, Jia-Bin Huang, Yu-Ding Lu, Maneesh Singh, and Ming-Hsuan Yang. Drit++: Diverse image-to-image translation via disentangled representations. *arXiv preprint arXiv:1905.01270*, 2019. 1, 5, 8

[33] 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. 5

[34] Ming-Yu Liu, Xun Huang, Arun Mallya, Tero Karras, Timo Aila, Jaakko Lehtinen, and Jan Kautz. Few-shot unsupervised image-to-image translation. *arXiv preprint arXiv:1905.01723*, 2019. 1

[35] Ming-Yu Liu and Oncel Tuzel. Coupled generative adversarial networks. In *Advances in neural information processing systems*, pages 469–477, 2016. 5

[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, 7

[37] Hongxia Luan, Feihu Qi, Zhong Xue, Liya Chen, and Dinggang Shen. Multimodality image registration by maximization of quantitative–qualitative measure of mutual information. *Pattern Recognition*, 41(1):285–298, 2008. 8

[38] Frederik Maes, Andre Collignon, Dirk Vandermeulen, Guy Marchal, and Paul Suetens. Multimodality image registration by maximization of mutual information. *IEEE transactions on Medical Imaging*, 16(2):187–198, 1997. 8

[39] Roey Mechrez, Itamar Talmi, and Lihi Zelnik-Manor. The contextual loss for image transformation with non-aligned data. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 768–783, 2018. 3, 5, 8

[40] 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. 8

[41] Kevin R Moon, Kumar Sricharan, and Alfred O Hero. Ensemble estimation of mutual information. In *2017 IEEE International Symposium on Information Theory (ISIT)*, pages 3030–3034. IEEE, 2017. 8

[42] Liam Paninski. Estimation of entropy and mutual information. *Neural computation*, 15(6):1191–1253, 2003. 8

[43] Taesung Park, Alexei A Efros, Richard Zhang, and Jun-Yan Zhu. Contrastive learning for unpaired image-to-image translation. *arXiv preprint arXiv:2007.15651*, 2020. 1, 2, 3, 4, 5, 8

[44] 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

[45] 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. 2, 5, 7

[46] Matan Sela, Elad Richardson, and Ron Kimmel. Unrestricted facial geometry reconstruction using image-to-image translation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1576–1585, 2017. 1

[47] Zhiqiang Shen, Mingyang Huang, Jianping Shi, Xiangyang Xue, and Thomas Huang. Towards instance-level image-to-image translation. *arXiv preprint arXiv:1905.01744*, 2019. 8

[48] 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. 5, 8

[49] Md Mahfuzur Rahman Siddiquee, Zongwei Zhou, Nima Tajbakhsh, Ruibin Feng, Michael B Gotway, Yoshua Bengio, and Jianming Liang. Learning fixed points in generative adversarial networks: From image-to-image translation to disease detection and localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 191–200, 2019. 1

[50] Shashank Singh and Barnabás Póczos. Exponential concentration of a density functional estimator. In *Advances in Neural Information Processing Systems*, pages 3032–3040, 2014. 8

[51] Shashank Singh and Barnabás Póczos. Generalized exponential concentration inequality for rényi divergence estimation. In *International Conference on Machine Learning*, pages 333–341, 2014. 8

[52] Masashi Sugiyama, Takafumi Kanamori, Taiji Suzuki, Marthinus Christoffel du Plessis, Song Liu, and Ichiro Takeuchi. Density-difference estimation. *Neural Computation*, 25(10):2734–2775, 2013. 4

[53] Masashi Sugiyama, Taiji Suzuki, and Takafumi Kanamori. *Distribution Comparison*, page 140–162. Cambridge University Press, 2012. 4

[54] Taiji Suzuki, Masashi Sugiyama, Takafumi Kanamori, and Jun Sese. Mutual information estimation reveals global associations between stimuli and biological processes. *BMC bioinformatics*, 10(1):S52, 2009. 3

[55] Taiji Suzuki, Masashi Sugiyama, Jun Sese, and Takafumi Kanamori. Approximating mutual information by maximum likelihood density ratio estimation. In *New challenges for feature selection in data mining and knowledge discovery*, pages 5–20, 2008. 8

[56] Yaniv Taigman, Adam Polyak, and Lior Wolf. Unsupervised cross-domain image generation. *arXiv preprint arXiv:1611.02200*, 2016. 8

[57] Hao Tang, Hong Liu, Dan Xu, Philip HS Torr, and Nicu Sebe. Attentiongan: Unpaired image-to-image translation using attention-guided generative adversarial networks. *IEEE Transactions on Neural Networks and Learning Systems*, 2021. 8

[58] Hao Tang, Dan Xu, Nicu Sebe, and Yan Yan. Attention-guided generative adversarial networks for unsupervised image-to-image translation. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2019. 5

[59] Matteo Tomei, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Art2real: Unfolding the reality of artworks via semantically-aware image-to-image translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5849–5859, 2019. 1

[60] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8798–8807, 2018. 1, 8

[61] Zhen Wang, Liu Liu, Yiqun Duan, and Dacheng Tao. Continual learning through retrieval and imagination. In *AAAI Conference on Artificial Intelligence*, 2022. 8

[62] Zhen Wang, Liu Liu, and Dacheng Tao. Deep streaming label learning. In *International Conference on Machine Learning (ICML)*, pages 378–387, 2020. 8

[63] Darrell Whitley and Jean Paul Watson. Complexity theory and the no free lunch theorem. *Search methodologies*, pages 317–339, 2005. 2

[64] David H Wolpert and William G Macready. No free lunch theorems for optimization. *IEEE transactions on evolutionary computation*, 1(1):67–82, 1997. 2

[65] Wayne Wu, Kaidi Cao, Cheng Li, Chen Qian, and Chen Change Loy. Transgaga: Geometry-aware unsupervised image-to-image translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8012–8021, 2019. 1, 2

[66] Yufei Xu, Qiming Zhang, Jing Zhang, and Dacheng Tao. Vitae: Vision transformer advanced by exploring intrinsic inductive bias. *Advances in Neural Information Processing Systems*, 34, 2021. 8

[67] Makoto Yamada, Taiji Suzuki, Takafumi Kanamori, Hirotaka Hachiya, and Masashi Sugiyama. Relative density-ratio estimation for robust distribution comparison. *Neural computation*, 25(5):1324–1370, 2013. 4

[68] 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. 8

[69] Zili Yi, Hao Zhang, Ping Tan, and Minglun Gong. Dualgan: Unsupervised dual learning for image-to-image translation. In *Proceedings of the IEEE international conference on computer vision*, pages 2849–2857, 2017. 1, 8

[70] Qiming Zhang, Yufei Xu, Jing Zhang, and Dacheng Tao. Vitaev2: Vision transformer advanced by exploring inductive bias for image recognition and beyond. *arXiv preprint arXiv:2202.10108*, 2022. 8

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

[72] Shuai Zhao, Yang Wang, Zheng Yang, and Deng Cai. Region mutual information loss for semantic segmentation. In *Advances in Neural Information Processing Systems*, pages 11115–11125, 2019. 8

[73] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017. 1, 2, 4, 5, 6, 7, 8

[74] Barbara Zitova and Jan Flusser. Image registration methods: a survey. *Image and vision computing*, 21(11):977–1000, 2003. 8