This WACV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

SemST: Semantically Consistent Multi-Scale Image Translation via Structure-Texture Alignment

Ganning Zhao [1], Wenhui Cui [1], Suya You [2], C.-C. Jay Kuo [1]
University of Southern California, Los Angeles, California, USA[1]
DEVCOM Army Research Laboratory, Los Angeles, California, USA[2]

Abstract

Unsupervised image-to-image translation learns cross-domain image mapping that transfers input from the source domain to output in the target domain while preserving its semantics. One challenge is that different semantic statistics in source and target domains result in content discrepancy known as semantic distortion. To address this problem, a novel I2I method that maintains semantic consistency in translation is proposed and named SemST in this work. SemST reduces semantic distortion by employing contrastive learning and aligning the structural and textural properties of input and output by maximizing their mutual information. Furthermore, a multi-scale approach is introduced to enhance translation performance, thereby enabling the applicability of SemST to domain adaptation in high-resolution images. Experiments show that SemST effectively mitigates semantic distortion and achieves state-of-the-art performance. Also, the application of SemST to domain adaptation is explored. It is demonstrated by preliminary experiments that SemST can be utilized as a beneficial pre-training for the semantic segmentation task.

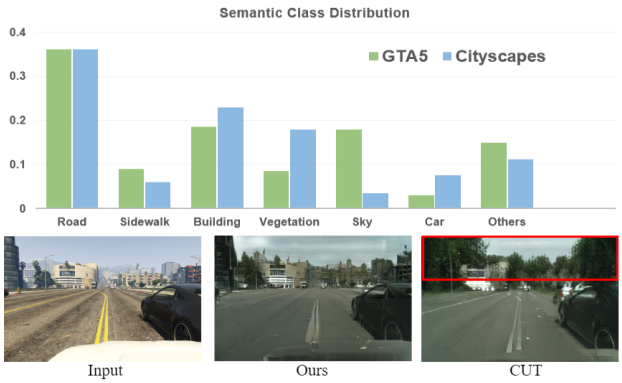

Figure 1. (top): The discrepancy in semantics distributions between GTA5 and Cityscapes. More pixels of sky in GTA5, while more building, vegetation, and car in Cityscapes. This significant difference in class distributions introduces semantic distortion. (bottom): Illustration of semantic distortion. In the GTA5-to-Cityscapes translation task, the sky region is mistakenly transformed to vegetation by CUT [29], due to more vegetation and less sky in Cityscapes. In contrast, our proposed SemST method preserves semantic consistency between input and output.

1. Introduction

The objective of image-to-image (I2I) translation involves learning a mapping from a source domain to a target domain. Specifically, it aims at transforming images of the source style to those of the target style with content consistency. While there is a domain gap, it can be mitigated by aligning the distributions of the source and the target domains. Nevertheless, disparities between class distributions of the source and target domains result in semantic distortion (see Figure 1); namely, different semantics of correspondent regions between input and output. The semantic distortion could potentially impact the efficacy of downstream tasks, such as semantic segmentation or object classification.

Early works [2, 34] employed adversarial training to align distributions in different domains with limited success. Since then, various techniques have been developed to accomplish this task. Bidirectional structures that ensured cycle consistency were proposed in [19,42,46]. However, their strict constraint of bijective projection could result in distortion. Although one-sided image translation [1, 4, 10] offers an alternative, its semantic distortion remains to be a significant problem. Recently, contrastive-learning-based methods were proposed, e.g., [29]. Despite a large amount of effort, such as leveraging more powerful loss functions [5, 41], mining informative positive/negative samples [15, 31, 36], and integrating various methods [45], the capability of the proposed methods in refining synthetic images and/or domain adaptation remains limited and semantic distortion still exists.

Besides image translation, reducing semantic distor-

tion finds applications in unsupervised domain adaptation (UDA). Training deep neural networks (DNNs) in semantic segmentation demands expensive and labor-intensive data labeling. It is desired to train DNNs on source datasets containing existing (or more affordable) annotations and deploy them on unlabeled target datasets. The main challenge in UDA is domain shift, the discrepancy between the source and target domains. Extensive efforts have been exerted to resolve this issue by aligning features between the two domains. Since the domain gap in the image space limits its performance, researchers have recently turned to translating images between domains and then aligning features from images. This new direction is proven advantageous [13, 23, 25–27]. However, current image translation approaches are usually applied to images of low-resolution or downsampled to low-resolution, which inevitably restricts performance in UDA that require high-resolution images. Recent work [14] demonstrates the performance degradation when training UDA on images downsampled to low resolution.

In this work, we propose a novel contrastive-learning-based method that alleviates semantic distortion by ensuring semantic consistency between input and output images. This is achieved by enhancing inter-dependence of structure and texture features between input and output by maximizing their mutual information. In addition, we exploit multi-scale predictions to boost the I2I translation performance by employing global context and local detail information jointly to predict translated images of superior quality, especially for high-resolution images. Hard negative sampling is also applied to reduce semantic distortion by sampling informative negative samples. For brevity, we refer to our method as SemST. Experiments conducted on I2I translation across various datasets demonstrate the state-of-the-art performance of the SemST method. Additionally, utilizing refined synthetic images in different UDA tasks confirms its potential for enhancing the performance of UDA.

2. Related Work

2.1. Unsupervised Image-to-Image Translation

Initial investigations of unsupervised adversarial learning have focused on augmenting the realism of synthetic images while maintaining annotation information [2,34]. They were primarily applied to simple grayscale images such as eyes and hands. Nonetheless, these methods found limited success when applied to more complex datasets.

Extensive efforts have been made to preserve semantics between input and output images. The cycle consistency loss was employed in [19, 35, 42, 46], which assumed a bijective translation function between the source and target domains. They enforced consistency between an input image in the source domain and the reconstructed image, inversely translated from the corresponding target domain image. However, these methods require an additional generator/discriminator pair. Besides, the bijective assumption could introduce distortions [20, 29, 35, 36].

Alternatively, one-sided image translation methods have emerged. They enforced geometry consistency between a source image and its transformed counterpart in the target domain [10] or ensured a strong correlation between matched pairwise distances in individual domains [4, 44]. Furthermore, some research aimed to reduce semantic distortion caused by mismatched semantic statistics by imposing structure consistency [11] or semantically robust loss [17]. However, many challenges still exist, including but not limited to semantic distortion, training instability, and limited applicability to high-resolution images.

2.2. Contrastive Learning

Contrastive Learning (CL) has been applied to image translation, offering a means to learn useful representations by exploring relationships among positive and negative pairs [29]. One idea is to develop suitable loss functions. The InfoNCE loss [28] linked corresponding patches and disassociates others through cross-entropy loss, gaining popularity in a few follow-ups, say, [3, 8, 12, 29]. Improved loss functions were proposed to address issues arising from small batch sizes [5] and alleviate the negative-positive coupling (NPC) effect [41].

Another line of research focuses on hard negative mining. One can employ techniques like using a negative sample generator [36], sampling negatives via the von Mises Fisher distribution [31], or resorting to adversarial training [15].

In addition, some studies aim to mitigate semantic distortion by exploring semantic relations among samples. For instance, one can ensure cross-domain consistency between positive and negative samples in source and target domains [18, 37, 45]. The idea to encode hierarchical semantic structures in the embedding space using the EM algorithm was tried and reported in [22].

2.3. Unsupervised Domain Adaptation

Most work on unsupervised domain adaptation (UDA) has concentrated on feature-level adaptation through adversarial models. Research on image-level translation has received less attention. Recently, it has been shown in [13, 23, 27, 33] that models trained on synthetic images translated from real image domains can enhance performance significantly in the semantic segmentation task. This indicates that, compared with feature-level alignment, the domain gap can be further reduced by image-level alignment. It was also reported in [25, 26] that both image-level and feature-level alignments contribute to performance improvement of domain adaptation.

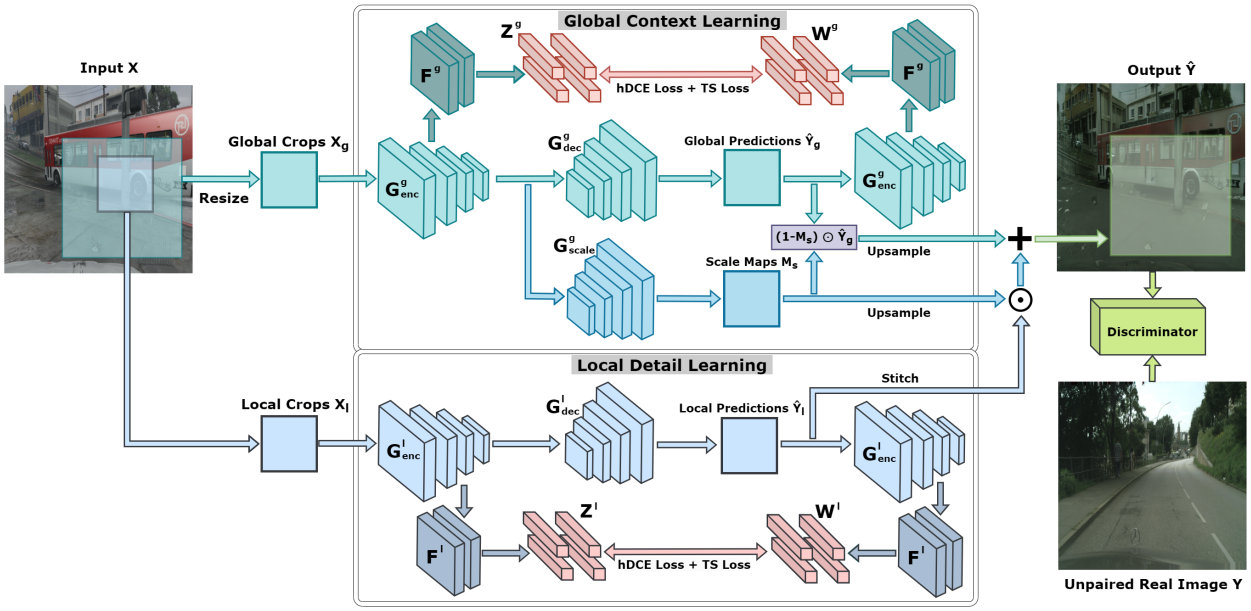

Figure 2. Overview of the proposed SemST method. It consists of a global context and a local detail learning branch. Global and local crops, extracted from input images, are fed separately to global or local learning branch. Generators of the encoder-decoder structure, $G^g_{enc-dec}$ and $G^l_{enc-dec}$, learn global and local predictions, respectively. Predictions are finally fused together with a scale map M_s that assesses the trustworthiness of global or local predictions. In each branch, embeddings of source and target domains are learned via the shared fully connected layers F applied to encoders of input and output, respectively. The hDCE and the TS losses are employed to align semantics within the embeddings in both global and local branches. A discriminator is trained to minimize the domain gap.

3. Proposed SemST Method

3.1. Motivation and System Overview

The distributions of semantic labels are usually different in source and target domains [18, 37, 40], as observed in Figure 1 and prior methods [11, 17], which not only leads to pixels with error semantics but also adversely influences downstream tasks that involve domain adaptation in the pipeline. In practical applications, image semantics are correlated with low-level texture and structure properties. For instance, sky, buildings and vegetation should exhibit similar visual appearance in input and output domains. Thus, we employ the joint texture (i.e., smooth or edge regions) and structure information to maintain semantic consistency between input and output.

The block diagram of the proposed SemST is depicted in Figure 2. We will elaborate on the three components: 1) structure and texture alignment for semantic consistency as indicated in pink; 2) multi-scale prediction as indicated in gray; 3) semantics-aided hard negative sampling.

3.2. Structure and Texture Alignment

To alleviate semantic distortion, we propose a loss function to preserve texture and structure consistency between input and output by maximizing their mutual information. Generally speaking, embeddings in shallow layers of higher resolutions capture the specific texture while embeddings in deeper layers of lower resolutions reflect the generalized structure information as illustrated in Fig. 3. The figure depicts embeddings from shallow to deep layers obtained based on receptive fields of varying sizes, encompassing the small-scale texture information to the large-scale structure information.

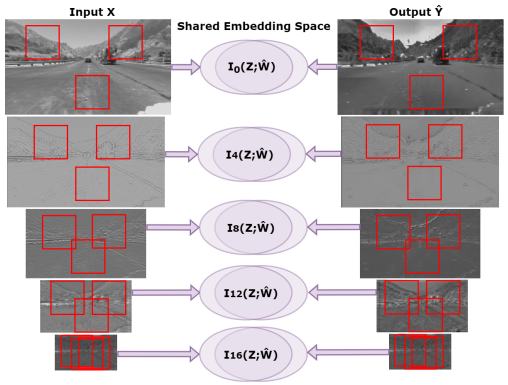

Figure 3. The mutual information, I_x (where x indicates layer indexes), between the input and output embedding spaces is maximized to maintain semantic consistency.

Embeddings in different layers of input and output images are extracted and aligned for semantic consistency. To maintain their consistency, we use mutual information to measure non-linear dependence between input and output embeddings. For efficient backpropagation and robust learning, we adopt the relative Squared-loss Mutual Information (rSMI).

Mathematically, for the embedding, z_i, of the input image and the embedding, $\hat{w}_i$, of the output image, we use Z_i and $\widehat{W}_i$ to denote their respective random variables. The texture-structure consistency (TS) loss is written as

$$L_{TS} = -\frac{1}{N}\sum_{i=1}^{N} \text{rSMI}(Z_i, \widehat{W}_i), \tag{1}$$

where N is the sample number. $\text{rSMI}(Z_i, \widehat{W}_i)$ is computed by the relative Pearson (rPE) divergence [11,39], defined as

$$\text{rSMI}(Z_i, \widehat{W}_i) = D_{rPE}(P_{Z_i} \otimes P_{\widehat{W}_i} || (P_{(Z_i, \widehat{W}_i)}), \tag{2}$$

In practice, $\text{rSMI}(Z_i, \widehat{W}_i)$, is estimated by a linear combination of kernel functions. It is solved by least-squares density-difference estimation. As a result, the mutual information estimator is in form of

$$\widehat{\text{rSMI}}(Z_i, \widehat{W}_i) = 2\hat{\alpha}^T \hat{h} - \hat{\alpha}^T \hat{H} \hat{\alpha} - 1. \tag{3}$$

where $\hat{\alpha}$, $\hat{h}$, and $\hat{H}$ are parameters computed via least-squares density-difference estimation [11].

3.3. Multi-Scale Framework

Most prior art on I2I translation directly manipulated images downsampled to a lower resolution, say, 256×256. However, this process inevitably limits performance due to information loss in the downsampling and subsequent upsampling of images back to their original resolution. Besides, these methods failed to predict smaller objects (e.g., poles and bikes) accurately and object borders with high quality. The performance could be even more compromised when dealing with intricate, high-resolution images containing objects of various scales. Although training random crops could be a solution, it fails to learn scene layout and relationships among objects, thereby introducing errors. To address these challenges, we propose a multi-scale framework that concurrently predicts local crops on a small scale and global crops on a larger scale. The model can learn detailed information with local crops, e.g., small objects and intricate borders, and the context information with global crops, e.g., layout and relationships among objects.

The above idea can be formalized as follows. We randomly crop the large global crops, X_g, from input images and downsample (T) them to size $h_g \times w_g$:

$$X_g = T(X_{ori}[h_{g0} : h_{g1}, w_{g2} : w_{g3}]; h_g, w_g). \tag{4}$$

Furthermore, small local crops X_l of size $h_l \times w_l$ are randomly cropped from X_{ori}:

$$X_l = T(X_{ori}[h_{l0} : h_{l1}, w_{l2} : w_{l3}]; h_l, w_l). \tag{5}$$

Global and local crops are predicted by the generator of the encoder-decoder structure, indicated by $\hat{Y}_g = G^g_{enc-dec}(X_g)$ and $\hat{Y}_l = G^l_{enc-dec}(X_l)$, respectively. Different generators are employed for local and global crops, given the different scales of their content and their requirement for distinct embedding spaces. Notably, this approach allows flexibility in the sizes of local and global crops, which can be equal or different. The overlapping predictions are averaged to increase robustness when stitching images for the subsequent fusion of local and global predictions.

To integrate predictions across different scales effectively, scale attention [6, 7, 14] is used to generate scale maps, $M_s \in [0, 1]$. The scale maps assist in determining which regions of the output should rely more on local or global predictions. For instance, smaller objects and complex structures such as trees and distant objects tend to rely on local predictions. In contrast, simpler regions, such as roads and proximate buildings depend more on global predictions. The final predictions are obtained by the fusion of global and local crops from different scales in the form of

$$\hat{Y} = M_s \odot \hat{Y}_l + (1 - M_s) \odot \hat{Y}_g. \tag{6}$$

3.4. Semantics-aided Hard Negative Sampling

Easy negative samples are uncorrelated with the query sample, diminishing the learning rate from more informative correlated hard negative samples [41]. This is the negative-positive coupling (NPC) effect. The decoupled InfoNCE (DCE) loss is crucial in alleviating the NPC effect.

Here, we adopt the DCE loss by excluding the positive pair from the denominator of InfoNCE. Concurrently, we sample hard negative samples that exhibit semantic correlations with the query sample by the von Mises-Fisher distribution [18,31]. This approach ensures that negative samples and query samples correspond to distinct latent classes, while also maintaining a substantial semantic similarity, quantified through the inner product. As a result, we can express this relationship as

$$z^- \sim q_\beta(z^-), \quad \text{where} \quad q_\beta(z^-) \propto e^{\beta z^T z^-} \cdot p(z^-), \tag{7}$$

where β is a concentration parameter that controls the similarity of hard negative samples with query samples. Combining it with the DCE loss, we obtain the hard Decoupled Contrastive Entropy (hDCE) loss:

$$L_{hDCE} = \mathbb{E}_{(z,\hat{w})}\left[-log\frac{\exp(\hat{w}^T z/\tau)}{N\mathbb{E}_{z^- \sim q_\beta}[\exp(\hat{w}^T z^-/\tau)]}\right], \tag{8}$$

where N is the number of negative patches and τ is a temperature parameter that controls the strength of penalties on hard negative samples. Then, the approximate expectation can be obtained by [31]

$$\begin{aligned}&\mathbb{E}_{z^- \sim q_\beta}[\exp(\hat{w}^T z^-/\tau)] \\ &= \frac{1}{N}\mathbb{E}_{z^- \sim p}[\exp(\hat{w}^T z^-/\tau)\exp(\beta z^T z^-)].\end{aligned} \tag{9}$$

For the implementation, we reweight the negative samples by their correlations with the positive sample, $z^T z^-$.

4. Experiments

To demonstrate the effectiveness of the proposed SemST method, we conduct a series of experiments involving image translation on various datasets. These experiments prove that our method can improve performance by mitigating semantic distortions. Furthermore, we perform testing to confirm that the refined synthetic images can effectively aid the downstream semantic segmentation task and potentially serve as a beneficial pre-training procedure for UDA.

4.1. Experimental Settings

Our implementation is based on the source code of CUT [29]. We substitute the original loss with the TS loss and the hDCE loss proposed in this work. Moreover, we restructure the original network to a multi-scale architecture. The global crop parameter (h_g, w_g) and the local crop parameter (h_l, w_g) are both set to 256 (see Figure 2).

4.2. Image-to-Image Translation

To validate the effectiveness of our SemST method in the image-to-image translation task and its capability to maintain semantic consistency between input and output images, we have extensively tested it on multiple datasets, including paired datasets (e.g., photo to map) and unpaired datasets (GTA to Cityscapes, etc.). Both quantitative results (see Table 1) and qualitative results (see Figure 4 and 5) demonstrate its superior performance. These results are elaborated below.

4.2.1 Simulation to Real: GTA5 → Cityscapes

To prove our model can enhance the realism of synthetic images by converting them into the domain of real-world captured images, we convert the images from GTA5 [30] to Cityscapes [9] domains. We train the model based on GTA5's official training split, comprising 6,202 images with a resolution of 1920×1080. In inference, we refine the first 500 images in the official test split and evaluate the performance by feeding them to FCN-8s [24] pre-trained on Cityscapes by pix2pix [16] to predict semantic label maps. We compute pixel accuracy, class accuracy, and mean IoU by comparing the predicted and ground-truth label maps. Higher scores indicate a similar distribution between output and target images and consistent semantics between input and output images. Thus, such a model can offer refined synthetic images of higher quality and potentially benefit downstream semantic segmentation.

SemST significantly outperforms other benchmarking methods in all metrics, achieving state-of-the-art performance, as shown in Table 1. Exemplary images translated by different methods are visualized in Figure 4. Evidently, SemST attains superior visual quality by preserving the texture and structure of synthetic input images and, consequently, maintaining the semantic information. In contrast, semantic distortions exist in other benchmarking methods. For example, some sky region is converted to vegetation or buildings, which are marked by red bounding boxes.

4.2.2 Parsing → Image

We train our model on the official training split of 3,975 and test on the validation split of 500 images from Cityscapes, which has a resolution of 2048×1024. Specifically, we transform semantic label maps into corresponding images. Similar to the simulation-to-real experiment, we assess different methods using metrics computed on pre-trained FCN-8s [16, 24]. Higher evaluation metrics represent less semantic distortion between input and output.

As shown in Table 1, SemST gives new state-of-the-art performance in all metrics. Figure 4 provides a qualitative comparison with other methods, demonstrating the capability of SemST to produce finer borders and preserve each segmentation region's semantics effectively.

4.2.3 Photo → Maps

We use the Maps dataset [16] to further demonstrate the performance of SemST on the I2I translation task. The dataset contains 2,194 pairs of aerial photo-to-map images, with 1,096 training and 1,098 testing pairs. Following [10], we use RMSE and pixel accuracy with threshold δ ($\delta_1 = 5$ and $\delta_2 = 10$) to evaluate performance, where given ground truth pixel $p_i = (r_i, g_i, b_i)$ and the prediction $\hat{p}_i = (\hat{r}_i, \hat{g}_i, \hat{b}_i)$, pixel accuracy is computed by the indicator function $\sum_{i=1}^{N} \mathbf{I}\{max(|r_i - \hat{r}_i|, |g_i - \hat{g}_i|, |b_i - \hat{b}_i|) < \delta\}$.

Again, Table 1 shows that SemST has the best performance regarding pixel accuracy with δ_1. While CycleGAN produces lower RMSE and pixel accuracy with δ_2 due to its cycle consistency loss, which is particularly beneficial for the paired Maps dataset, SemST still generates the best results among all contrastive-learning-based approaches. This demonstrates the effectiveness of the proposed TS loss function and the multi-scale framework in improving the image translation task, highlighting the robustness and versatility of SemST.

Table 1. Quantitative evaluations of our method and benchmarking methods. The methods with + are reproduced by [11] and ✳ are reproduced by us on a single GPU using the codes provided by the authors. The best results are highlighted in bold.

Methods	GTA5 → Cityscapes			Cityscapes Parsing → Image			Photo → Map		
	pixel acc ↑	class acc ↑	mean IoU ↑	pixel acc ↑	class acc ↑	mean IoU ↑	RMSE ↓	acc%(δ_1) ↑	acc%(δ_2) ↑
DRIT++ [21]	0.423	0.138	0.071	\	\	\	32.12	29.8	52.1
CycleGAN [46]	0.232$^+$	0.127$^+$	0.043$^+$	0.520	0.170	0.110	**26.81**	43.1	**65.6**
GcGAN [10]	0.405$^+$	0.139$^+$	0.068$^+$	0.551	0.197	0.129	27.98	42.8	64.6
CUT [29]	0.546$^+$	0.165$^+$	0.095$^+$	0.695$^+$	0.259$^+$	0.178$^+$	28.48$^+$	40.1$^+$	61.2$^+$
SRUNIT [17]	0.581✳	0.135✳	0.079✳	0.505✳	0.175✳	0.096✳	28.40✳	41.2✳	60.5✳
SRC [18]	0.597✳	0.187✳	0.111✳	0.787✳	0.259✳	0.207✳	27.98✳	41.2✳	61.7✳
VSAIT [35]	0.603✳	0.179✳	0.109✳	0.755✳	0.250✳	0.205✳	\	\	\
CUT+SCC [11]	0.572	0.185	0.110	0.699	0.263	0.182	27.34	39.2	60.5
SSC [45]	0.654	0.186	0.113	0.714	0.263	0.184	27.19	41.8	62.1
Ours	**0.693**	**0.205**	**0.135**	**0.790**	**0.266**	**0.213**	27.15	**45.7**	63.7

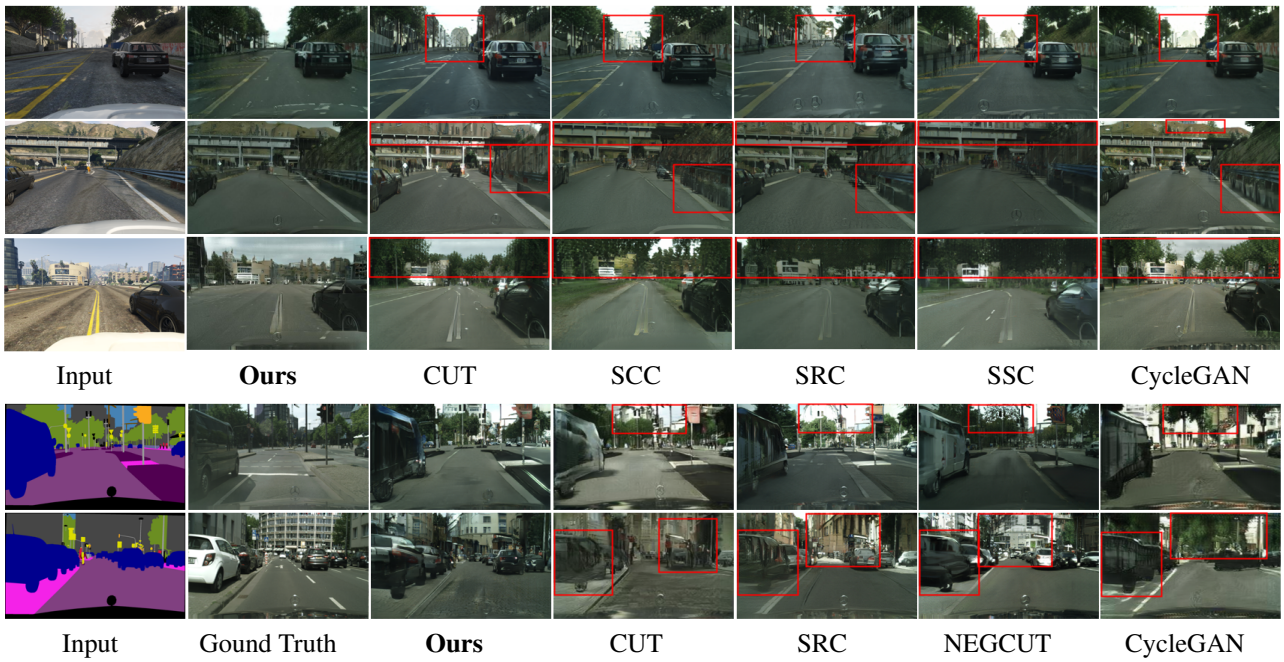

Figure 4. Qualitative visual comparison of images refined by our SemST method and other benchmarking methods on GTA5 → Cityscapes (top) and Parsing → Image (bottom). Our method reduces the semantic distortion and has fewer artifacts highlighted by bounding boxes.

4.2.4 Qualitative Results on Low-resolution Images

We provide more visual results on two popular datasets in Figure 5. They are the Horse → Zebra dataset and the Summer → Winter dataset. The former has unpaired $1,067$ horse images and $1,334$ zebra images. The latter contains $1,231$ summer scenes and 962 winter scenes in Yosemite. There exists a difference in semantic statistics between the two domains for each dataset. Since the resolution of images is small (i.e., 256×256), we use the single-scale method to predict results directly. These experiments demonstrate the effectiveness of our proposed loss function in preserving semantics and the resulting images have better or comparable quality compared to others.

4.3. Enhancing Semantic Segmentation

As discussed in Section 2, training on refined synthetic images can enhance the downstream semantic segmentation task on real-world datasets. Here, we demonstrate images refined by SemST can assist unsupervised domain adaptation (UDA) by incorporating them in the training of domain adaptation networks. Specifically, we train UDA models using images from the source domain and output images obtained by our proposed domain mapper (i.e., refined synthetic images). We compare the IoU scores obtained by different UDA methods and their enhanced variants achieved by incorporating SemST-refined images into the training process in Table 2. The results are discussed below.

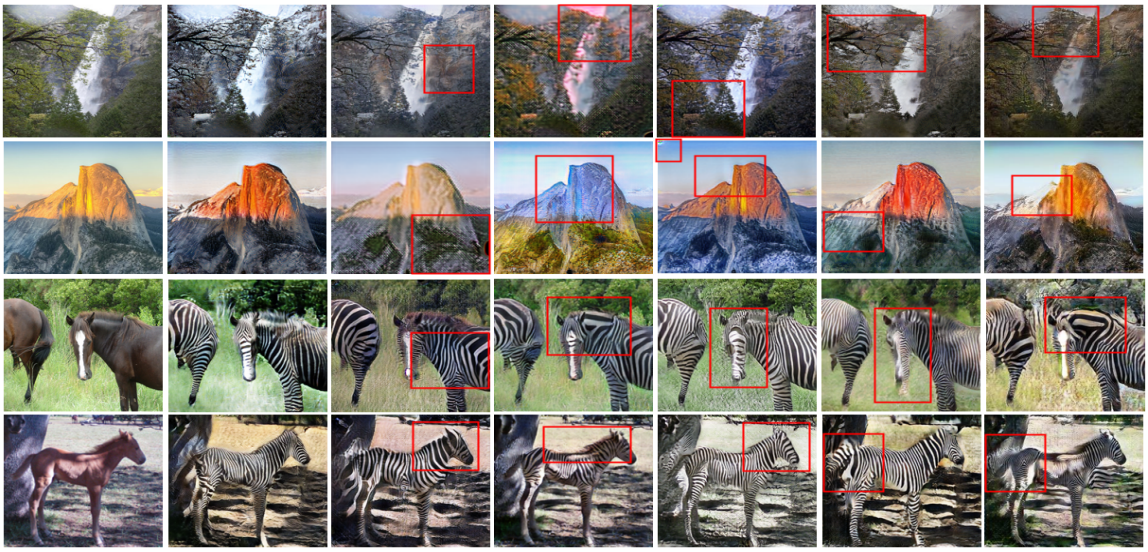

Figure 5. Qualitative visual comparison of images refined by our SemST method versus other benchmarking methods on summer → winter and horse → zebra. In the former, our results realistically cloaked leaves and mountains with snow, exhibiting superior or comparable authentic color representations. In the latter, we generate better or comparable natural color tones and preserve the horse's morphology. Generally, our outcomes contain fewer artifacts.

Table 2. The IoU performance comparison of UDA methods and their enhanced variants by incorporating synthetic images refined by SemST (highlighted in gray shadow) in training. All methods are based on DeepLab-V2 with ResNet-101. Training with images refined by SemST improves UDA, demonstrating the effectiveness of our method in UDA by reducing image-level domain gap.

Method	road	side.	buil.	wall	fence	pole	light	sign	veg.	terr.	sky	pers.	rider	car	truck	bus	train	mbike	bike	mIoU
GTA5 → Cityscapes																				
SePiCo [38]	95.6	69.2	89.0	40.8	38.6	44.3	56.3	64.4	88.3	46.5	88.6	73.1	47.6	90.7	58.9	53.8	5.4	22.4	43.8	58.8
+SemST	95.8	70.2	88.4	45.9	37.2	45.6	53.4	62.1	86.9	39.9	82.3	70.9	47.0	90.5	54.5	60.4	0.1	48.4	62.2	**60.1**
ProDA [43]	87.8	56.0	79.7	46.3	44.8	45.6	53.5	53.5	88.6	45.2	82.1	70.7	39.2	88.8	45.5	59.4	1.0	48.9	56.4	57.5
+SemST	91.8	62.6	83.6	43.5	45.5	47.7	54.2	56.4	88.7	49.2	82.6	70.5	38.6	88.9	47.1	56.4	0.1	47.7	56.5	**58.5**
BDL [23]	91.0	44.7	84.2	34.6	27.6	30.2	36.0	36.0	85.0	43.6	83.0	58.6	31.6	83.3	35.3	49.7	3.3	28.8	35.6	48.5
+SemST	92.6	49.0	85.7	36.4	30.0	32.6	34.4	32.7	84.3	46.2	84.3	57.5	34.9	82.8	42.6	50.7	0.3	36.6	39.5	**50.2**
SYNTHIA → Cityscapes																				
SePiCo [38]	79.2	42.9	85.6	9.9	4.2	38.0	52.5	53.3	80.6	-	81.2	73.7	47.4	86.2	-	63.1	-	48.0	63.2	57.3
+SemST	79.5	45.3	80.0	3.2	1.2	38.3	61.2	54.1	83.4	-	81.3	74.8	49.9	90.3	-	64.3	-	50.9	69.6	**58.0**
ProDA [43]	87.8	45.7	84.6	37.1	0.6	44.0	54.6	37.0	88.1	-	84.4	74.2	24.3	88.2	-	51.1	-	40.5	45.6	55.5
+SemST	86.5	43.2	90.3	37.2	0.1	46.1	53.5	36.2	92.9	-	87.9	80.1	29.1	86.1	-	56.8	-	41.1	48.2	**57.2**

4.3.1 GTA5 → Cityscapes

We translate images from the GTA5 dataset to the domain of the Cityscapes dataset and, then, include the translated images in the training of UDA methods. Experimental results show that training with SemST-refined synthetic images improves mIoU on different UDA methods, which indicates that SemST can be potentially employed as a beneficial pretraining for domain adaptation. Another observation is the effect of class imbalance on semantic segmentation performance. Specifically, the failure in predicting train class results from their low probability in class distribution and different appearances across domains. In contrast, success in road class prediction comes from high probability and similar features across domains.

4.3.2 SYNTHIA → Cityscapes

We experiment on another source dataset called SYNTHIA-RAND-CITYSCAPES. It is a subset of the synthetic urban scene dataset known as SYNTHIA [32]. It contains 9,400 images of resolution 1280×760 and 16 common semantic annotations with Cityscapes. After refining the SYNTHIA dataset to the Cityscapes domain by SemST and subsequently training the domain adaptor with refined images, we observe a performance improvement. Experiments in Table 2 showcase an improvement in the mIoU scores after training on refined images.

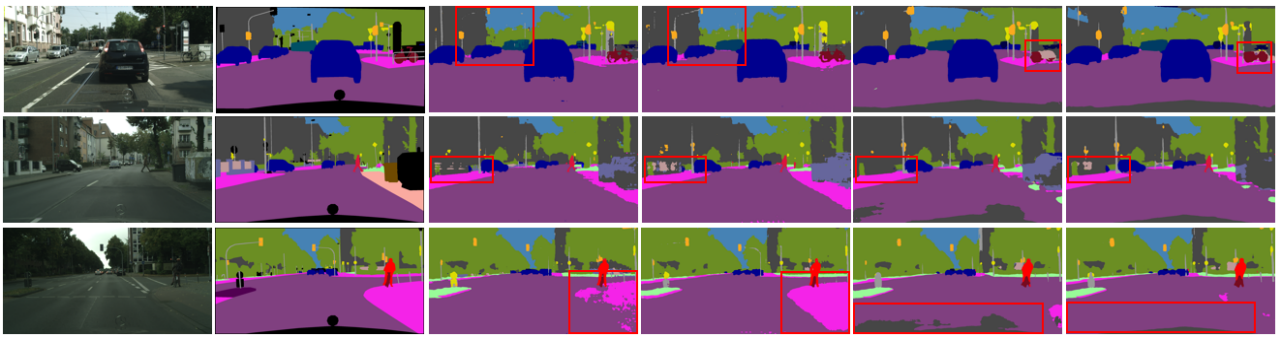

Figure 6. Qualitative visual comparison of results from domain adaptation on GTA5 → Cityscapes using benchmarking methods and those methods trained in combination with SemST-refined images. The latter ones have more accurate label predictions and refined borders, as highlighted within the red bounding boxes.

Figure 7. Qualitative ablation study for GTA5 to Cityscapes. The red bounding boxes indicate artifacts. Removing multi-scale prediction yields inaccurate predictions and blurry results. Eliminating either the hDCE or TS loss introduces more artifacts and semantic distortion. As the weight of TS loss increases, semantic distortion is reduced.

Table 3. Quantitative ablation study demonstrating the contributions of different components

Methods	w/o Components		Weights of TS Loss		
	Multiscale	hDCE Loss	0	1	**2**
pixel acc ↑	0.645	0.624	0.598	0.679	**0.693**
class acc ↑	0.182	0.175	0.169	0.198	**0.205**
mean IoU ↑	0.115	0.113	0.110	0.126	**0.135**

5. Ablation Study

We examine the contribution of each individual component by excluding them and varying the hyperparameters of the TS loss, denoted as λ_{TS}, in the context of the GTA5 to Cityscapes task. The results of our investigations are presented in Table 3 and Figure 7. Notably, removing any component results in decreased performance. Specifically, multi-scale prediction ensures superior performance on high-resolution images by local and global information learning. The hDCE loss alleviates the NPC effect and enables more efficient learning from informative hard negative samples. Furthermore, increasing the value of λ_{TS} enhances performance by mitigating semantic distortion. However, caution is needed in selecting the magnitude of λ_{TS}, as excessively high values would prompt the model to prioritize input-output consistency at the potential expense of neglecting the style information learned from the target domain.

6. Conclusion

A multi-scale image translation method that preserves the semantic consistency between input and output images, called SemST, was presented in this work. The multi-scale framework was used to predict local detail and global context, which improves performance and enables the application to higher-resolution images for UDA. Semantic consistency was achieved by introducing TS loss that aligns semantics between input and output images by maximizing their mutual information in a shared embedding space. Extensive experiments were conducted to demonstrate the state-of-the-art performance of SemST in image translation and its value in facilitating UDA was also validated.

References

[1] Matthew Amodio and Smita Krishnaswamy. Travelgan: Image-to-image translation by transformation vector learning. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 8983–8992, 2019. 1

[2] Charith Atapattu and Banafsheh Rekabdar. Improving the realism of synthetic images through a combination of adversarial and perceptual losses. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. IEEE, 2019. 1, 2

[3] Philip Bachman, R Devon Hjelm, and William Buchwalter. Learning representations by maximizing mutual information across views. *Advances in neural information processing systems*, 32, 2019. 2

[4] Sagie Benaim and Lior Wolf. One-sided unsupervised domain mapping. *Advances in neural information processing systems*, 30, 2017. 1, 2

[5] Junya Chen, Zhe Gan, Xuan Li, Qing Guo, Liqun Chen, Shuyang Gao, Tagyoung Chung, Yi Xu, Belinda Zeng, Wenlian Lu, et al. Simpler, faster, stronger: Breaking the log-k curse on contrastive learners with flatnce. *arXiv preprint arXiv:2107.01152*, 2021. 1, 2

[6] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017. 4

[7] Liang-Chieh Chen, Yi Yang, Jiang Wang, Wei Xu, and Alan L Yuille. Attention to scale: Scale-aware semantic image segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3640–3649, 2016. 4

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

[9] 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

[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/CVF Conference on Computer Vision and Pattern Recognition*, pages 2427–2436, 2019. 1, 2, 5, 6

[11] Jiaxian Guo, Jiachen Li, Huan Fu, Mingming Gong, Kun Zhang, and Dacheng Tao. Alleviating semantics distortion in unsupervised low-level image-to-image translation via structure consistency constraint. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18249–18259, 2022. 2, 3, 4, 6

[12] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 2

[13] 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. 2

[14] Lukas Hoyer, Dengxin Dai, and Luc Van Gool. Hrda: Context-aware high-resolution domain-adaptive semantic segmentation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXX*, pages 372–391. Springer, 2022. 2, 4

[15] Qianjiang Hu, Xiao Wang, Wei Hu, and Guo-Jun Qi. Adco: Adversarial contrast for efficient learning of unsupervised representations from self-trained negative adversaries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1074–1083, 2021. 1, 2

[16] 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. 5

[17] 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, 3, 6

[18] Chanyong Jung, Gihyun Kwon, and Jong Chul Ye. Exploring patch-wise semantic relation for contrastive learning in image-to-image translation tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18260–18269, 2022. 2, 3, 4, 6

[19] Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jung Kwon Lee, and Jiwon Kim. Learning to

discover cross-domain relations with generative adversarial networks. In *International conference on machine learning*, pages 1857–1865. PMLR, 2017. 1, 2

[20] 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. 2

[21] Hsin-Ying Lee, Hung-Yu Tseng, Qi Mao, Jia-Bin Huang, Yu-Ding Lu, Maneesh Singh, and Ming-Hsuan Yang. Drit++: Diverse image-to-image translation via disentangled representations. *International Journal of Computer Vision*, 128:2402–2417, 2020. 6

[22] Junnan Li, Pan Zhou, Caiming Xiong, and Steven CH Hoi. Prototypical contrastive learning of unsupervised representations. *arXiv preprint arXiv:2005.04966*, 2020. 2

[23] Yunsheng Li, Lu Yuan, and Nuno Vasconcelos. Bidirectional learning for domain adaptation of semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6936–6945, 2019. 2, 7

[24] 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. 5

[25] Haoyu Ma, Xiangru Lin, Zifeng Wu, and Yizhou Yu. Coarse-to-fine domain adaptive semantic segmentation with photometric alignment and category-center regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4051–4060, 2021. 2

[26] Haoyu Ma, Xiangru Lin, and Yizhou Yu. I2f: A unified image-to-feature approach for domain adaptive semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 2

[27] Luigi Musto and Andrea Zinelli. Semantically adaptive image-to-image translation for domain adaptation of semantic segmentation. *arXiv preprint arXiv:2009.01166*, 2020. 2

[28] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 2

[29] Taesung Park, Alexei A Efros, Richard Zhang, and Jun-Yan Zhu. Contrastive learning for unpaired image-to-image translation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*, pages 319–345. Springer, 2020. 1, 2, 5, 6

[30] Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 102–118. Springer, 2016. 5

[31] Joshua Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. Contrastive learning with hard negative samples. *arXiv preprint arXiv:2010.04592*, 2020. 1, 2, 4, 5

[32] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M Lopez. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3234–3243, 2016. 7

[33] Tingwei Shen, Ganning Zhao, and Suya You. A study on improving realism of synthetic data for machine learning, 2023. 2

[34] 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. 1, 2

[35] Justin Theiss, Jay Leverett, Daeil Kim, and Aayush Prakash. Unpaired image translation via vector symbolic architectures. In *European Conference on Computer Vision*, pages 17–32. Springer, 2022. 2, 6

[36] Weilun Wang, Wengang Zhou, Jianmin Bao, Dong Chen, and Houqiang Li. Instance-wise hard negative example generation for contrastive learning in unpaired image-to-image translation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14020–14029, 2021. 1, 2

[37] Chen Wei, Huiyu Wang, Wei Shen, and Alan Yuille. Co2: Consistent contrast for unsupervised visual representation learning. *arXiv preprint arXiv:2010.02217*, 2020. 2, 3

[38] Binhui Xie, Shuang Li, Mingjia Li, Chi Harold Liu, Gao Huang, and Guoren Wang. Sepico: Semantic-guided pixel contrast for domain adaptive semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 7

[39] 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

[40] Chuanguang Yang, Zhulin An, Linhang Cai, and Yongjun Xu. Mutual contrastive learning for visual

representation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 3045–3053, 2022. 3

[41] Chun-Hsiao Yeh, Cheng-Yao Hong, Yen-Chi Hsu, Tyng-Luh Liu, Yubei Chen, and Yann LeCun. Decoupled contrastive learning. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVI*, pages 668–684. Springer, 2022. 1, 2, 4

[42] 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, 2

[43] Pan Zhang, Bo Zhang, Ting Zhang, Dong Chen, Yong Wang, and Fang Wen. Prototypical pseudo label denoising and target structure learning for domain adaptive semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12414–12424, 2021. 7

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

[45] Ganning Zhao, Tingwei Shen, Suya You, and C-C Jay Kuo. Unsupervised synthetic image refinement via contrastive learning and consistent semantic and structure constraints. *arXiv preprint arXiv:2304.12591*, 2023. 1, 2, 6

[46] 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, 6