Supplementary Materials for Semantically Robust Unpaired Image Translation for Data with Unmatched Semantics Statistics

A. Eqn. 3 vs. Eqn. 4 in the definition of $\mathcal{L}_{robust}$

In the main paper at Sec. 4.2, we briefly discuss the relation between the two versions (namely, Eqn. 3 & 4) of $\mathcal{L}_k$ used in our proposed semantic robustness loss $\mathcal{L}_{robust}$. Here we include more details. To begin with, let us list them below as Eqn. a and b, respectively.

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

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

The difference between the two is the extra $G(\cdot)$ inside the L2 norm of the Eqn. b. Remind that by the notations in the main paper we have $G(x) = \mathcal{G}_k^{K+1}(\mathcal{G}_1^k(x))$. Optimizing Eqn. b directly reflects our definition of semantic robustness in Sec. 3.4, i.e., the transformed image $G(x)$ should have their semantics computed by $F_k \circ \mathcal{G}_k$ invariant to small perturbations τ_k in the feature space (i.e. the output space of G_k) of the input x. On the other hand, optimizing Eqn. a indirectly minimizes Eqn. b, since the contrastive loss (Eqn. 2 in Sec. 4,1 in the main paper) minimizes $||F_k(\mathcal{G}_1^k(x)) - F_k(\mathcal{G}_1^k(G(x)))||_2$ due to all the outputs of F_k constained to be on the same unit sphere (please refer to the CUT paper [8] for details.) In specific, denote $A_k(x) = F_k(\mathcal{G}_1^k(G(x)))$, $B_k(x) = F_k(\mathcal{G}_1^k(x))$ and $C_k(x) = F_k(\mathcal{G}_1^k(\mathcal{G}_k^{K+1}(\mathcal{G}_1^k(x)+\tau_k)))$, we have

$$\begin{aligned}
\mathcal{L}'_k &= \mathbb{E}_x\Big[\frac{1}{||\tau_k||_2}||A_k(x) - C_k(x)||_2\Big] \\
&= \mathbb{E}_x\Big[\frac{1}{||\tau_k||_2}||A_k(x) - B_k(x) + B_k(x) - C_k(x)||_2\Big] \\
&\leq \mathbb{E}_x\Big[\frac{1}{||\tau_k||_2}\big(||A_k(x) - B_k(x)||_2 + ||B_k(x) - C_k(x)||_2\big)\Big] \\
&= \mathbb{E}_x||A_k(x) - B_k(x)||_2 + \mathbb{E}_x\Big[\frac{1}{||\tau_k||_2}||B_k(x) - C_k(x)||_2\Big] \\
&= \mathbb{E}_x||A_k(x) - B_k(x)||_2 + \mathcal{L}_k
\end{aligned}$$

In short, we have

$$\mathcal{L}'_k \leq \mathcal{L}_k + \mathbb{E}_x||A_k(x) - B_k(x)||_2$$

Since $\frac{1}{K}\sum_{k=1}^K \mathbb{E}_x||A_k(x) - B_k(x)||_2$ is minimized by the contrastive loss, our proposal to minimize $\mathcal{L}_{robust} = \frac{1}{K}\sum_{k=1}^K \mathcal{L}_k$ effectively minimizes an upper bound of $\frac{1}{K}\sum_{k=1}^K \mathcal{L}'_k$.

We argue that optimizing such an upper bound (the adaptive version) is better than directly minimizing $\mathcal{L}'_k$ since otherwise the diversity of the translation can be harmed. As a result, it might fail to produce complex visual patterns and create artifacts instead. We numerically compare the performances in the ablation studies (Sec. 6 in the main paper). We also show a visual comparison in Fig. 1.

B. Evaluation on Label to Image and GTA to Cityscapes datasets

When computing the three metrics on the Label to Image and the GTA to Cityscapes tasks, we use a light-weight DeepLab V3 [2] model pre-trained on the Cityscapes semantic segmentation task to evaluate the segmentation masks of the translated images. In specific, we choose the `mobilenetv3_small_cityscapes_trainfine` model, publicly available at TensorFlow Model Page.

For both tasks and all 3 metrics, we perform the evaluation on the 500 (finely annotated) validation set of the Cityscapes dataset which consists of 19 classes. We ignore all unlabelled pixels when computing the metrics. We follow the common evaluation protocol as in [8]. For pxAcc (the average pixel accuracy), we compute the average pixel accuracy of the segmentation masks predicted by the DeepLab model on the translated images. For clsAcc (the class accuracy), we compute the average pixel accuracy per semantic class and then compute the mean of those across all 19 classes. For mIoU (mean IoU), we compute the average IoU per class and then the mean of them.

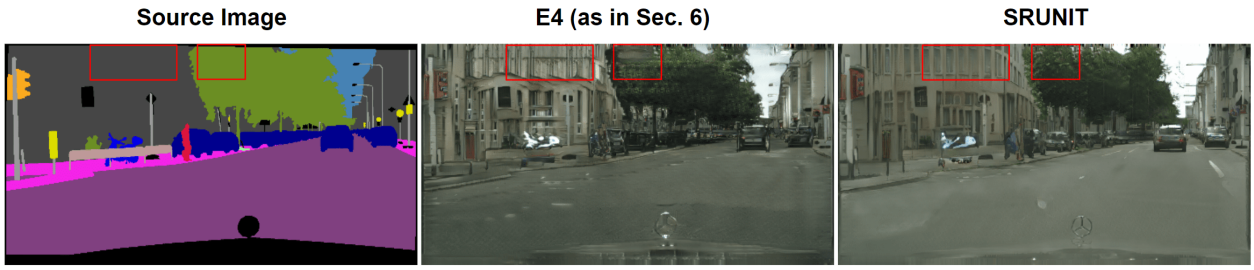

Figure 1. Visual comparisons of the Label to Image task for E4 (a model trained by using $\mathcal{L}'_k$ in optimizing $\mathcal{L}_{robust}$) vs. SRUNIT (by using $\mathcal{L}_k$ instead). The former does harm to the diversity of the translations. The red boxes highlight area where there are artifacts.

C. Dataset Construction with Unmatched Semantics Statistics

Our paper focuses on unpaired image-to-image translation where data from two domains inherently have different semantics distributions. One example where we quantitatively demonstrate this discrepancy is in Fig. 1 (from the main paper) for the GTA to Cityscapes task. Since originally designed for paired image translations, the Label to Image and the Google Map to Aerial Photo datasets in our experiments are sub-sampled to ensure a reasonable amount of difference in their semantics statistics (briefly mentioned in Sec. 5.1.1, 5.1.3 & 5.1.4).

In the Label to Image task from Cityscapes, as the original dataset is paired, we characterize each pair of images (the RGB semantic mask from the source domain and the street-view image from the target domain) by its histogram of the semantic classes (a vector $\in \mathbb{R}^{19}$, where the i^{th} entry represents the ratio of the pixels belonging to the i^{th} class in the image). Then we use the K-means ($K = 2$) algorithm to cluster the 2975 images from the training set in Cityscapes according to these histograms (i.e., vectors). We use all the source images from one cluster and all the target images from the other cluster as the sub-sampled unpaired data. The resulting two clusters are of roughly the same size with different semantics statistics shown in Fig. 2 (top).

In the Google Map to Aerial Photo task (and vice versa), similarly, since the dataset is paired, we can characterize each out of the 1096 training pairs by the color histogram of its google map image. In specific, we convert all google map image from RGB to gray-scale and apply bucketing to the pixel intensities (each bucket includes consecutively 5 out of the 256 total values) so that each histogram becomes a vector $\in \mathbb{R}^{51}$. We apply K-means (again $K = 2$) clustering on these histograms. The resulting clusters are very different in size due to the long tail distribution in the Google Map dataset. To deal with this, we first obtain the two histogram centroids from K-means. Then, from highest to lowest, we rank all pairs of images by the ratio of the distance between its histogram to one centroid over that between its histogram to the other centroid. We then assign the top half of the pairs to one cluster and the rest to the other. Fig. 2 (bottom) shows the resulting distributions, where there is a large difference between the two. We use all Google maps from cluster I together with 10% of randomly sampled Google map images from cluster II (similarly all aerial photos from cluster II and 10% from I) to ensure the reasonable amount of difference in semantics statistic between the two domains.

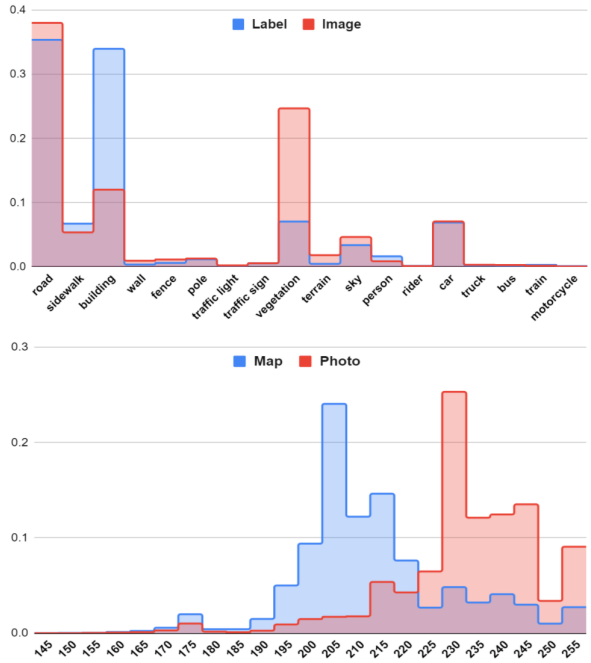

Figure 2. (**top**) Semantics distributions of the Label and Image data, each from one of the two clusters. (**bottom**) Semantics distributions of the Google Maps and Aerial Photos from the two clusters. As mentioned in Appendix C, the latter has a very large semantics discrepancy and thus we mix 10% of each cluster when constructing our sub-sampled dataset.

D. Mismatched Semantics Statistics, Semantics Flipping & Semantic Robustness

In Sec. 3.2 of the main paper, we argue that it is inherently common in unpaired image translation to have mismatched semantics statistics across domains, and as a result, the semantics flipping usually occurs in the spurious solutions obtained by existing GAN-based methods. We claim in Sec. 3.4 that our proposed semantics robustness can effectively mitigate the semantics flipping problems by, to some extent, offsetting the negative effects of the difference in semantics statistics.

To give more insights, we construct training data with better-aligned semantics statistics between the source and target domains by adding more data to the training set from the opposite cluster. For instance, in the Label to Image task where all source images are from cluster I and target ones from cluster II, we add X% of source images randomly sampled from cluster II and X% of target images from cluster I to form a new set of training data. We compare several CUT [8] models (the backbone of our proposed method SRUNIT) trained with data of different levels of discrepancy in semantics statistics. The results (illustrated in Fig. 3) indicate that (1) better-matched semantics statistics of the training data lead to less semantics flipping; (2) our method's improvement over its baseline in reducing semantics flipping is substantial. As the results of SRUNIT are visually comparable to the baseline with 50% more training data (which significantly reduces the difference in semantics statistics across the domains).

Figure 3. The three columns correspond to the experiments on Label to Image, Map to Photo and Photo to Map, respectively. CUT + X% data indicates a CUT [8] model trained with X% more data sampled from the opposite cluster (to mitigate the mismatched semantics statistics problem). We shows that SRUNIT is rather effective in reducing semantics flipping caused by the different semantics statistics. It produces comparable or better translation results than the CUT baseline trained on data with much more "matched" semantics statistics. The red boxes highlight the areas where SRUNIT have improvements over CUT.

E. More Visual Results

We show additional visual results for the 7 unpaired image-to-image translation tasks performed in the main paper. They are Label to Image and GTA to Cityscapes (see Fig. 4), Map to Photo and Photo to Map (see Fig. 5), Horse to Zebra, Summer to Winter and Day to Night (see Fig. 6), respectively.

F. Additional Implementation Details

This section includes more implementation details about our method SRUNIT. Please also refer to Sec. 5.3 of the main paper. We follow CUT [8] for the choice of network architecture and the training setup. In specific, we use the least square loss [7], a ResNet-based generator [5] with 9 residual blocks, and a patch-based discriminator [4]. We keep an image buffer of size 50 to update the discriminator for better training stability [9]. We use the default hyperparameters for relevant loss terms in CUT, including selecting the same $K = 5$ layers in the generator to compute the contrastive loss. We add our proposed loss term $\beta * \mathcal{L}_{robust} = \frac{\beta}{K} \sum_{k=1}^{K} \mathcal{L}_k$ using the same K layers (by default) with the coefficient $\beta = 10^{-4}$ (by default). We fine-tune K to be 4 or 5 by leaving one of the $\{mathcal L_k\}$ out each time. We use Adam optimizers [6] to train our model for 400 epochs with an initial learning rate of 0.0002 and a linear decay for the last 50% of epochs (the same with CUT). The exception is that for the GTA to Cityscapes task, due to the large quantity of the training data, we only train for 20 epochs in total. Moreover, we adopt the patch-based approach as in CUT to compute $\mathcal{L}_{robust}$; i.e., in each training iteration we randomly sample 256 patches from each of the K layers to compute $\mathcal{L}_k$ (assuming batch size is 1). This can significantly reduce the computation complexity in optimizing $\mathcal{L}_{robust}$.

G. Details about the Ablation Studies

For all models (E1 to E6) in the ablation studies section (Sec. 6 in the main paper), we use CUT as the backbone (the same as our proposed method SRUNIT).

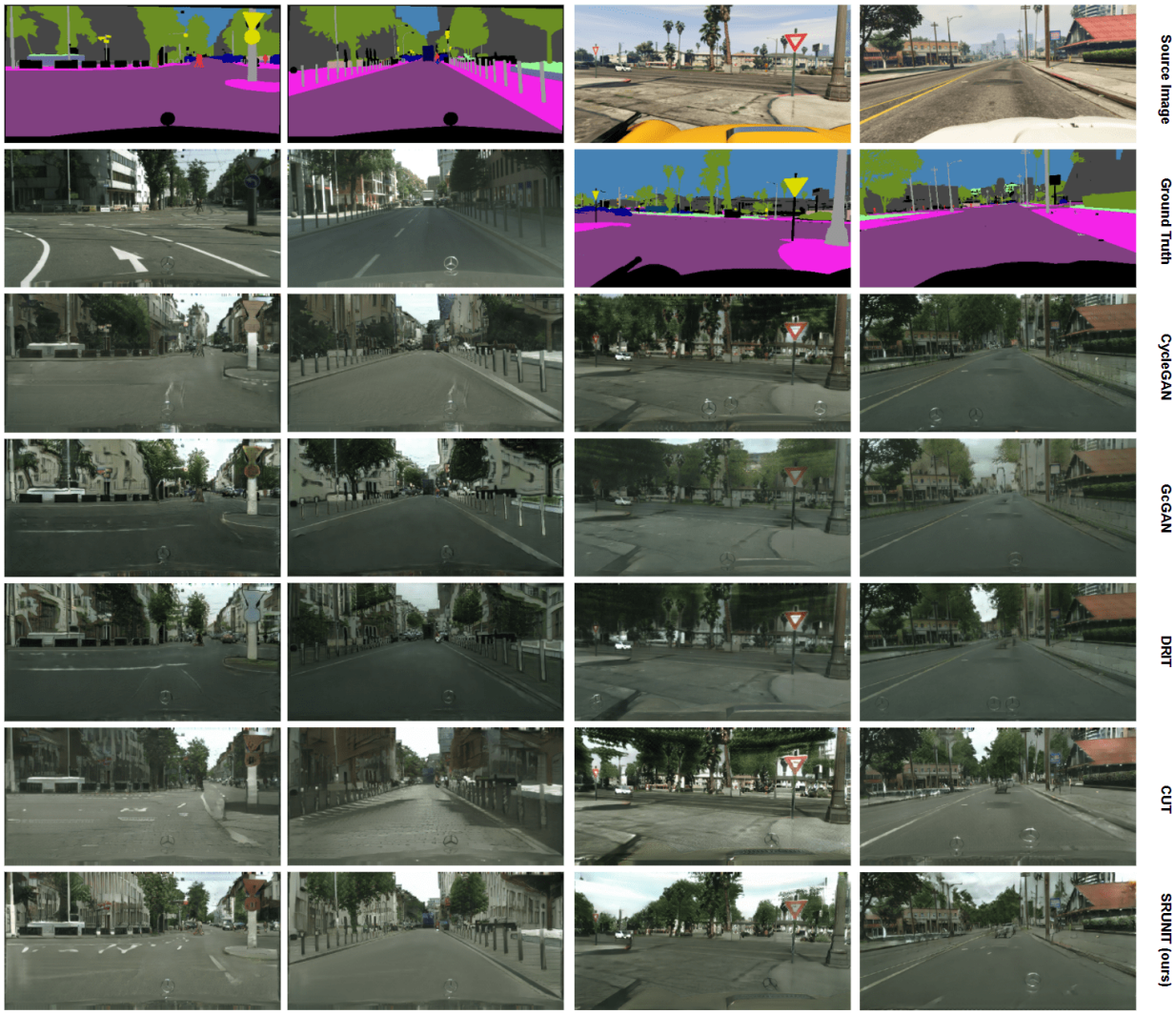

Figure 4. Additional visual results for the Label to Image and the GTA to Cityscapes tasks.

E1 We aim to compare SRUNIT with the constraint proposed in DistanceGAN [1]. E1 is trained by adding the self-distance loss given as:

$$\mathcal{L}_{\text{dist}}(G) = \mathbb{E}_x\Big[\frac{1}{\sigma_X}\left(\|L(x) - R(x)\|_1 - \mu_X\right) - \frac{1}{\sigma_Y}\left(\|L\left(G(x)\right) - R\left(G(x)\right)\|_1 - \mu_Y\right)\Big]$$

where $L, R : \mathbb{R}^{H\times W\times 3} \rightarrow \mathbb{R}^{H\times \frac{W}{2}\times 3}$ are the operators that given an input image x return the left or right part of it and σ_*, μ_* are the pre-computed image statistics from the two domain X and Y (see [1]).

E2 Similarly we compare SRUNIT with (a modified version of) the constraint proposed in HarmonicGAN [10]. We train E2 by adding the smoothness loss:

$$\mathcal{L}_{smooth} = \mathbb{E}_{x_1,x_2}\left[||d(x_1, x_2) - d(G(x_1), G(x_2))||_1\right]$$

where x_1, x_2 are image patches from the same input image, $d(\cdot,\cdot)$ is the distance function measured by using the histograms of two input patches, and $G(x_1)$ refers to image patch of the translation $G(x)$ corresponding to the patch x_1 from the input image x (see details in [10]). In each iteration, we randomly sample 256 image patches to form 128 pairs from the input image to compute $\mathcal{L}_{smooth}$ (assuming the batch size is 1).

E3 We aim to verify the necessity of using feature extractor F_k in $\mathcal{L}_{robust}$. We train E3 by removing F_k in $\mathcal{L}_{robust}$,

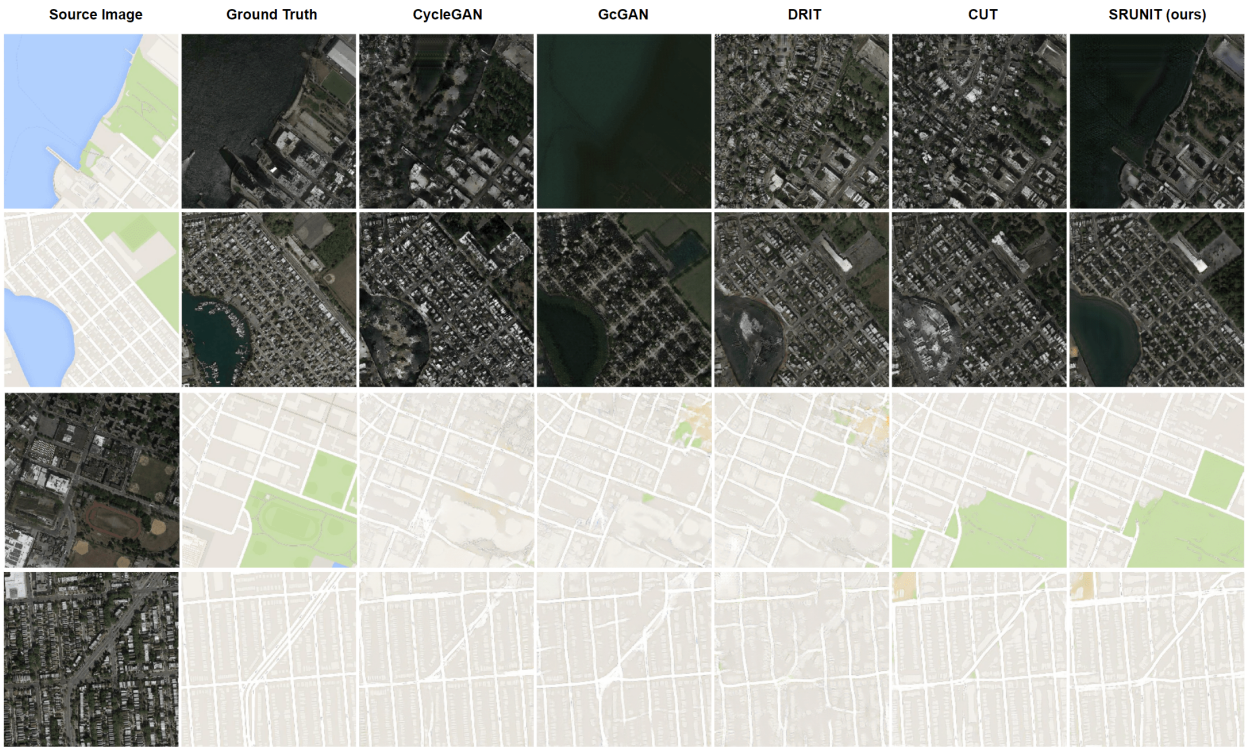

Figure 5. Additional visual results for the Map to Photo and the Photo to Map tasks.

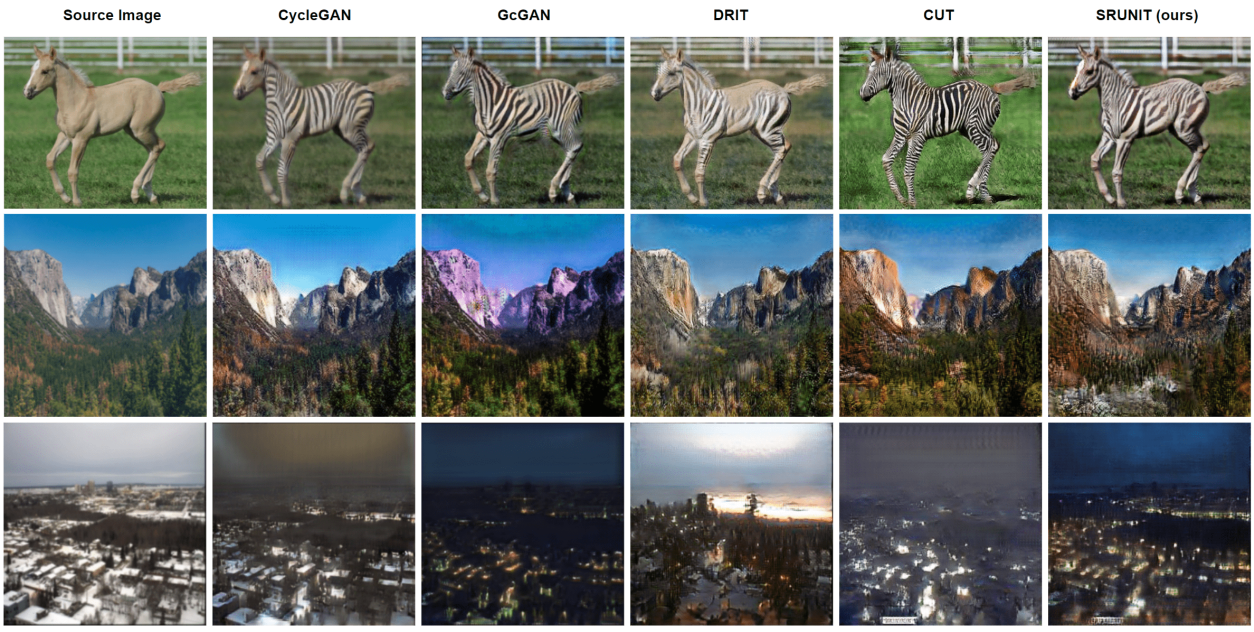

Figure 6. Additional visual results for the Horse to Zebra, the Summer to Winter and the Day to Night tasks.

i.e., defining $\mathcal{L}_{robust} = \frac{1}{K}\sum_{k=1}^{K}\widetilde{\mathcal{L}}_k$, where

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

E4 We aim to empirically show the advantage of Eqn. a over Eqn. b (see Sec. 4.2 in the main paper and Appendix A for a discussion). We train E4 by setting $\mathcal{L}_{robust} = \frac{1}{K}\sum_{k=1}^{K}\mathcal{L}'_k$ instead of $\frac{1}{K}\sum_{k=1}^{K}\mathcal{L}_k$.

E5 We aim to show that directly minimizing the distance between semantics extracted by F_k of the input image and that of the corresponding translated image is not an effective way to reduce semantics flipping. We train E5 by adding the semantics consistency term $\mathcal{L}^{sc} = \frac{1}{K}\sum_{k=1}^{K}\mathcal{L}_k^{sc}$, where

$$\mathcal{L}_k^{sc} = \mathbb{E}_x \| F_k(\mathcal{G}_1^k(x)) - F_k(\mathcal{G}_1^k(G(x)))\|_2$$

Since the semantics extractors $\{F_k\}$ are learned in an unsupervised manner, they are not accurate enough for direct enforcement of the preservation of semantics during the translations.

E6 Another direction of efforts to reduce semantics flipping is to pose constraints on the discriminator instead of on the generator. We train E6 by applying the Lipschitz penalty [3] to the discriminator in CUT. Namely, we add the Lipschitz loss:

$$\mathcal{L}_{lip} = \mathbb{E}_x[||\nabla_x D_Y(\overline{x})||_2 + ||\nabla_{\overline{G(x)}} D_Y(G(x))||_2]$$

where $\overline{x}$ is the mean value of x and $\overline{G(x)}$ the mean value of $G(x)$. We do so since CUT utilizes a patch-based discriminator and the gradient computation ∇ is much cheaper than the Jacobian computation.

Remark: In E3, E4, and E5, x refers to the image patches instead of the entire images in the similar way as to how we adopt the patch-based approach in CUT to optimize our proposed semantic robustness loss $\mathcal{L}_{robust}$.

References

[1] Sagie Benaim and Lior Wolf. One-sided unsupervised domain mapping. In *Advances in neural information processing systems*, pages 752–762, 2017. 4

[2] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 1

[3] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. In *Advances in neural information processing systems*, pages 5767–5777, 2017. 6

[4] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017. 3

[5] 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. 3

[6] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 3

[7] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2794–2802, 2017. 3

[8] Taesung Park, Alexei A Efros, Richard Zhang, and Jun-Yan Zhu. Contrastive learning for unpaired image-to-image translation. In *European Conference on Computer Vision*, pages 319–345. Springer, 2020. 1, 3

[9] 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. 3

[10] Rui Zhang, Tomas Pfister, and Jia Li. Harmonic unpaired image-to-image translation. *arXiv preprint arXiv:1902.09727*, 2019. 4