Supplementary Materials to "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

This supplementary file provides a detailed explanation of the full objective, implementation details, subjective test, and additional visual comparison results of SemST.

1. Full Objective

The objective of image-to-image (I2I) translation is to learn a mapping from a source domain to a target domain, enabling the input images X in the source domain to be transferred to the target domain. This can be achieved by aligning the distribution between the domain of output translated images and the target domain. The standard adversarial loss implements this alignment:

$$L_{GAN}(G, D, X, Y) = \mathbb{E}_{y\sim Y} log D(y) + \mathbb{E}_{x\sim X} log(1 - D(\hat{y})), \quad (1)$$

where y denotes images from the target domain, $\hat{y} = G(x)$ refers to the translated images, while G and D symbolize the generator and discriminator, respectively.

However, aligning the distribution may lead to semantic distortions. To address this, our proposed method, SemST, incorporates two loss functions: the proposed texture-structure consistency constraint (TS loss) and the semantics-aided decoupled InfoNCE (hDCE) loss. Utilizing these loss functions ensures the semantic consistency between the input and output images, implying that the content of the images remains unaltered.

SemST employs a multi-scale framework with two branches to learn global large-scale embeddings and local small-scale embeddings, respectively. Since the purpose of both the hDCE loss L_{hDCE} and the TS loss L_{TS} is to improve the correlation between the input and output, and a mixture of embeddings at different scales results in inaccuracies in the computation of their correlation, these loss functions are applied separately in local and global embeddings:

$$L_{TS} = L^g_{Ts} + L^l_{Ts}, \quad (2)$$

$$L_{hDCE} = L^g_{hDCE} + L^l_{hDCE}, \quad (3)$$

where superscripts g and l indicate whether the loss function is computed in global or local embeddings, respectively. L_{hDCE} and L_{TS} are the final hDCE loss and TS loss, respectively.

In summary, by integrating the loss functions that account for both semantics and distribution alignment, the full objective function for our SemST approach takes the form:

$$L_{full} = \lambda_{TS} L_{TS} + \lambda_{hDCE} L_{hDCE} + \lambda_{GAN} L_{GAN}, \quad (4)$$

where λ_{TS}, λ_{hDCE} and λ_{GAN} represent the weights of different loss functions, respectively.

2. Implementation Details

Our codes are based on the source code of SRC [4] and SCC [2]. Detailed implementations are explained in this section.

2.1. Network Architecture

2.1.1 Generator Archtecture

The generator, $G_{enc-dec}$, contains one block with a 7×7 Convolution-InstanceNorm-ReLU structure and stride 1, two downsampling blocks with a 3×3 Convolution-InstanceNorm-ReLU structure and stride 2, nine residual blocks with a residual connected 3×3 Convolution-InstanceNorm-ReLU-Convolution-Normalization structure, two upsampling blocks with a Deconvolution-InstanceNorm-ReLU structure and stride 2, and finally, one block with a 7×7 Convolution-InstanceNorm-ReLU structure and stride 1 [6]. The first half of the generator is the encoder G_{enc} and the remainder is the decoder G_{dec}. The global and local crop prediction branches share the same generator structure, but their weights differ and are represented by $G^g_{enc-dec}$ and $G^l_{enc-dec}$, respectively.

2.1.2 Fully Connected Layers Archtecture

We employ G_{enc} to both input and output images to extract features from different layers. Specifically, we extract features from the 0th, 4th, 8th, 12th, and 16th layers, which correspond to receptive fields of sizes 1×1, 9×9, 15×15, 35×35, and 99×99. Following CUT [5], we randomly sample 256 locations and apply a 2-layer MLP F to these features to generate a shared 256-dimensional embedding space between input and output images. The TS and hDCE losses are applied to this shared embedding space to constrain the semantics between input and output. Similar to the generator, the global and local crop prediction branches share the same structure but have different weights, represented by F_g and F_l, respectively.

2.1.3 Scale Attention Archtecture

The scale attention is applied to the embedding space obtained by G^g_{enc}. We utilize Atrous Spatial Pyramid Pooling (ASPP) [1] to learn the scale map, M_s, composed of one block of a 1×1 AdaptivePool-Convolution structure with stride 1, four 4×4 convolutional layers with dilations of 1, 6, 12, and 18, and a convolutional layer with stride 3 applied on the concatenated features from all convolutional layers resized to the same resolution. Finally, the features are input into a convolutional layer to predict a scale map with one channel and size equal to the resized global crops h_g. By utilizing ASPP to predict the scale map on global predictions, scale maps can learn from different scales and assist in deciding which region should rely more on local or global predictions.

2.1.4 Discriminator Archtecture

We utilize the 70×70 PatchGAN [3], which classifies whether each 70×70 patch is real or fake and averages all results as the output of the discriminator. PatchGAN comprises one block with a 4×4 Convolution-LeakyReLU structure and stride 2, three blocks with a 4×4 Convolution-InstanceNorm-LeakyReLU and stride 2, and finally, one convolutional layer.

2.1.5 Training Paramters

In our experiments, we set the global crop size at $h_g = 512$ and the local crop size at $h_l = 256$. Both the global and local crops are extracted from resized input images to cover the full scale of input images. The batch size is fixed at 1. We employ the Adam optimizer with exponential decay rates set for the first-moment estimates $\beta_1 = 0.5$ and the second-moment estimates $\beta_2 = 0.999$. The learning rate is initiated from 0.0002.

3. Subjective Test

To further demonstrate the superior performance of our method in comparison to 6 benchmarking methods. We conducted a subjective assessment involving 30 participants. In this study, we presented participants with 12 sets of blind A/B tests, including the results generated by our method and prior works. Each of the benchmarking methods appears twice within our testing. The result of this subjective test is shown in Figure 1, unveiling a preference range of 83% to 100% among users favoring our method over the other benchmarking methods.

4. Additional Results

We present further visual results comparing various methods: Cityscapes Parsings $\rightarrow$ Images (Figure 2), Summer $\rightarrow$ Winter (Figure 3), Horse $\rightarrow$ Zebra (Figure 4), and Domain Adaptation of GTA $\rightarrow$ Cityscapes (Figure 5). These comparisons clearly demonstrate the prominent improvements of SemST over other methods.

References

[1] 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. 2

[2] 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. 1

[3] 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. 2

[4] 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. 1

[5] 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. 2

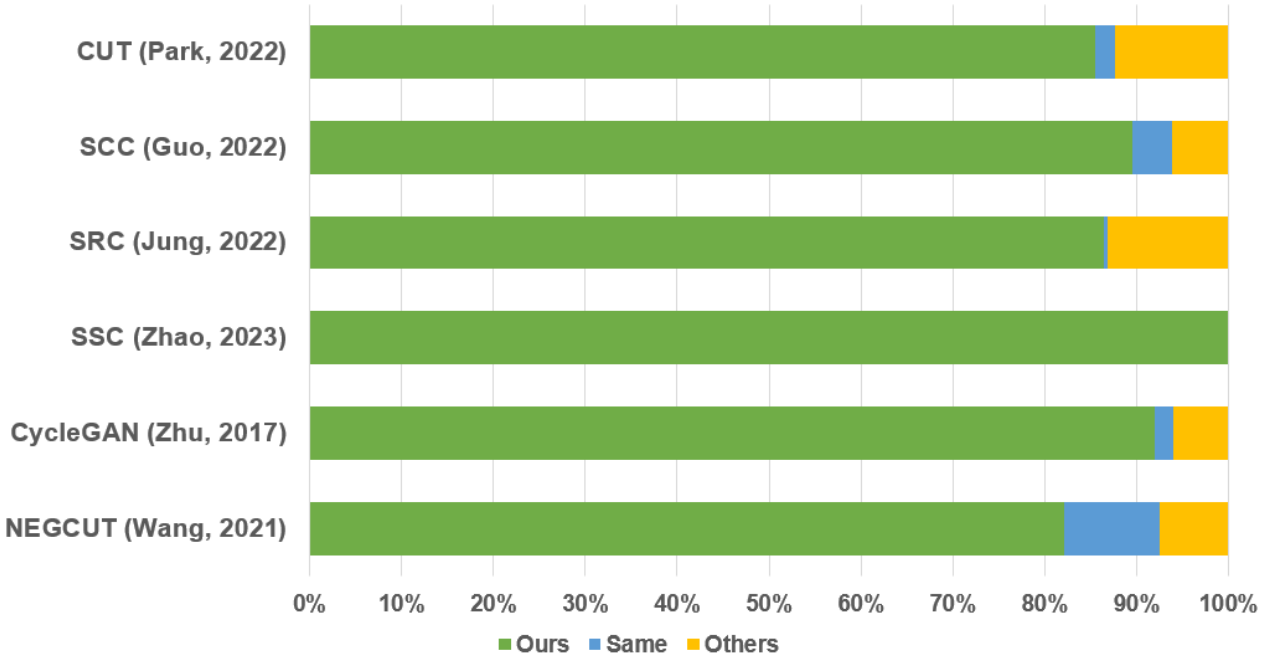

Figure 1. User study results: we show the user's preference in pair-wise comparisons between our method and six benchmarking methods.

Figure 2. A visual comparison of images refined by our SemST method and other benchmarking methods on Cityscapes Parsings → Images. As highlighted by bounding boxes, other methods exhibit artifacts and semantic distortions, while our results effectively mitigate these issues, resulting in higher-quality images

[6] 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

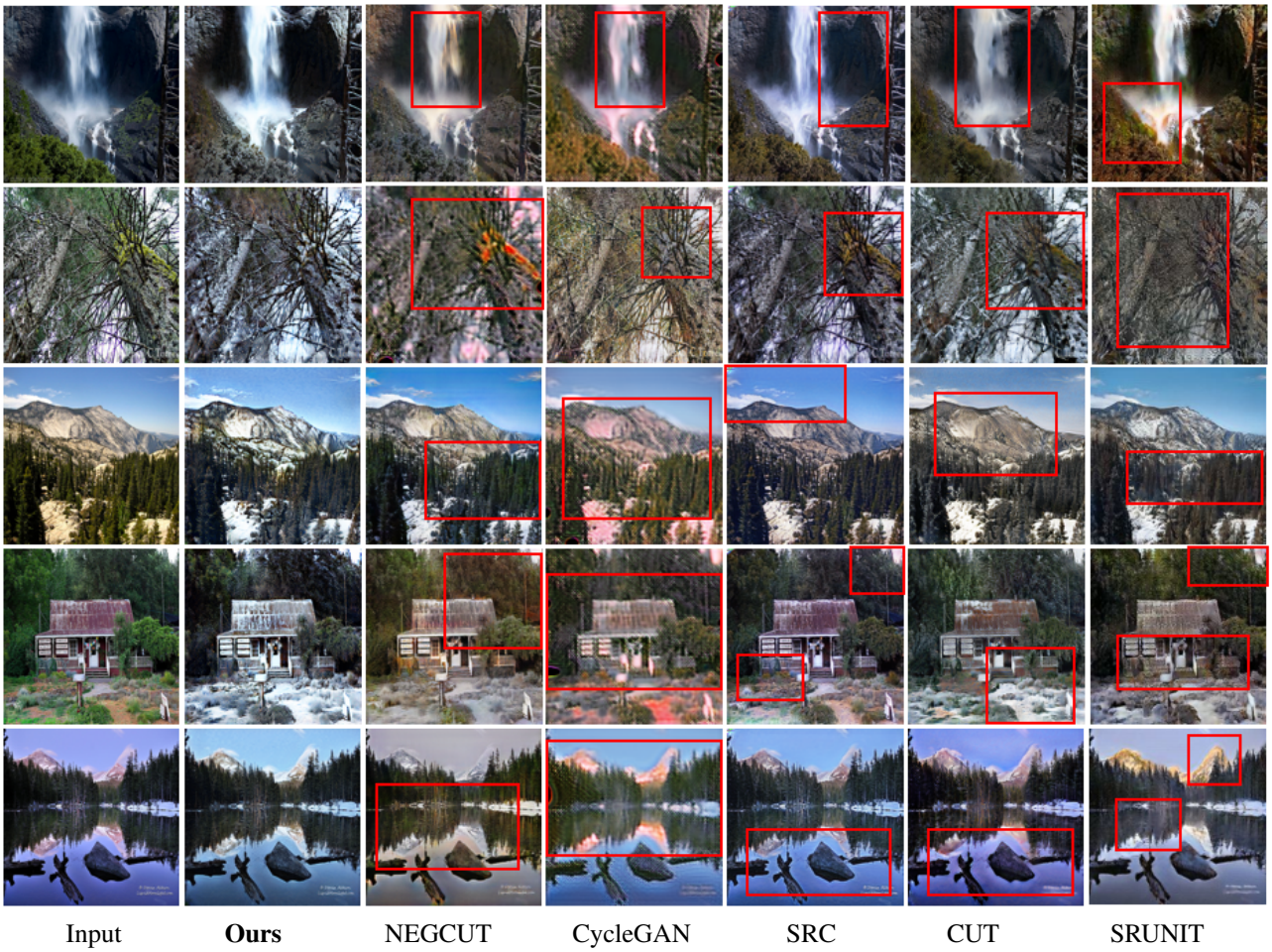

Figure 3. A visual comparison of images refined by our SemST method versus other benchmarking methods on summer → winter. Our results realistically render buildings, leaves, and mountains with snow, and provide more natural colors. The artifacts, highlighted by bounding boxes, are effectively reduced by our SemST approach.

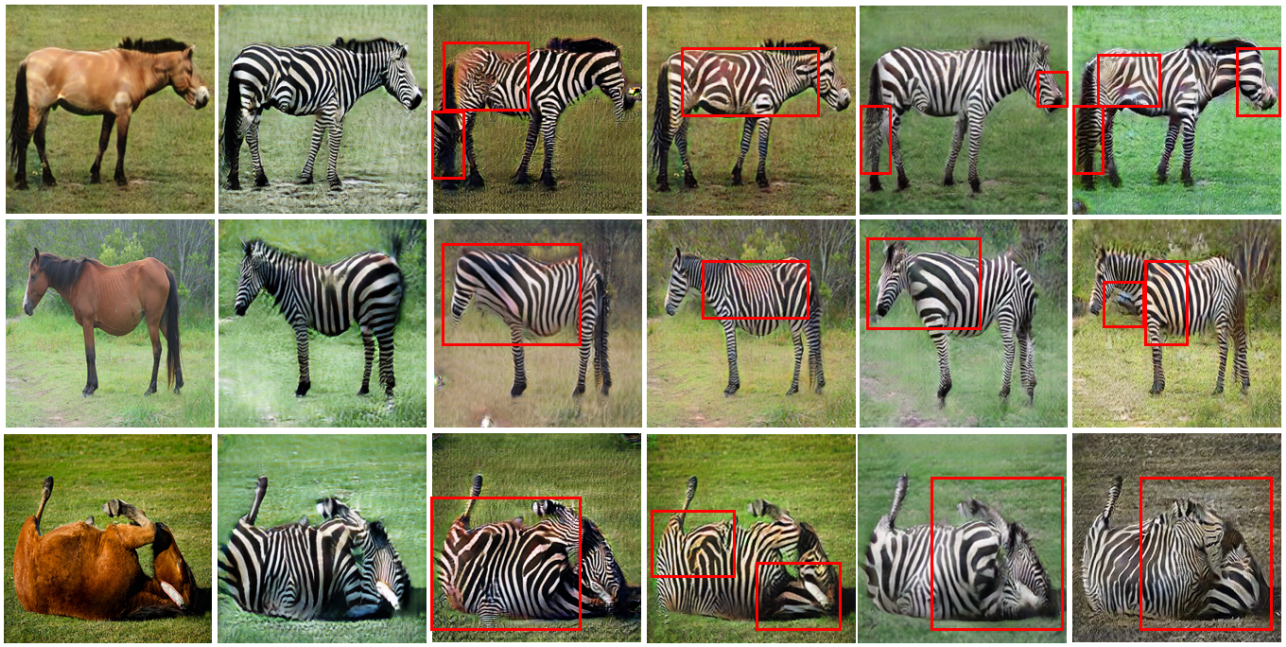

Figure 4. A visual comparison of images refined by our SemST method versus other benchmarking methods on horse → zebra. We generate better or comparable results. certain artifacts, such as the brown color on zebras and object shape distortion, are highlighted by bounding boxes in results generated by other methods. However, these issues are effectively mitigated in our approach.

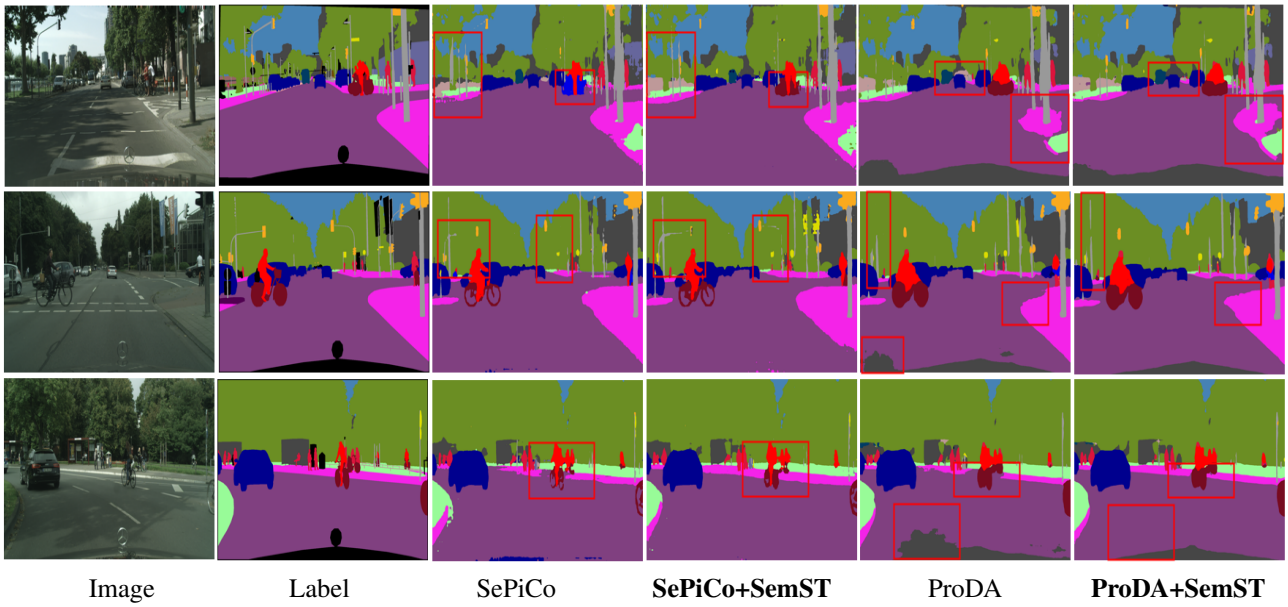

Figure 5. A visual comparison of the results of domain adaptation on GTA5 → Cityscapes using benchmarking methods and those methods trained in combination with SemST-refined images. As shown in the bounding boxes, our SemST corrects wrongly classified objects, accurately predicts finer details (e.g. street lamps), and rectifies region labels.