All about Structure: Adapting Structural Information across Domains for Boosting Semantic Segmentation
Supplementary Materials

Wei-Lun Chang* Hui-Po Wang* Wen-Hsiao Peng Wei-Chen Chiu
National Chiao Tung University, Taiwan
{luckchang.ee06g, a88575847.cs06g, wpeng, walon}@nctu.edu.tw

Table 1. Ablation study in terms of mIoU for the case of adapting SYNTHIA [2] to Cityscapes [1]. We present results for no adaptation (Source Only), adaptation at the output space only (Seg-map Adaptation), adaptation at the output space together with structure and texture disentanglement (our DISE w/o Label Transfer), and adaptation with all losses considered (our DISE).

Method	A	B	C	D	mIoU
Source Only	✓				31.7
Seg-map Adaptation	✓	✓			35.1
DISE w/o Label Transfer	✓	✓	✓		38.8
DISE	✓	✓	✓	✓	41.5

A: $\mathcal{L}_{seg}^{s}$
B: $\mathcal{L}_{seg\_adv}$
C: $\mathcal{L}_{rec} + \mathcal{L}_{trans\_str} + \mathcal{L}_{trans\_tex} + \mathcal{L}_{trans\_adv}$
D: $\mathcal{L}_{seg}^{s2t}$

1. Ablation Study on SYNTHIA

As an extension to Section 4.3 in our main paper, this session provides additional results for an ablation study of our DISE on SYNTHIA dataset. We follow the same settings as in the main paper to test the four variants: Source Only, Seg-map Adaptation, DISE w/o Label Transfer, and DISE. From Table 1, it is seen that the performance of these four variants exhibit a similar upward trend in mIoU to what we have seen on Cityscapes dataset with the incremental use of output-space domain adaptation (Seg-map Adaptation), structure and texture disentanglement (DISE w/o Label Transfer), and label transfer (DISE). As expected, our full DISE model reaches the highest performance.

2. Visualization for SYNTHIA to Cityscapes

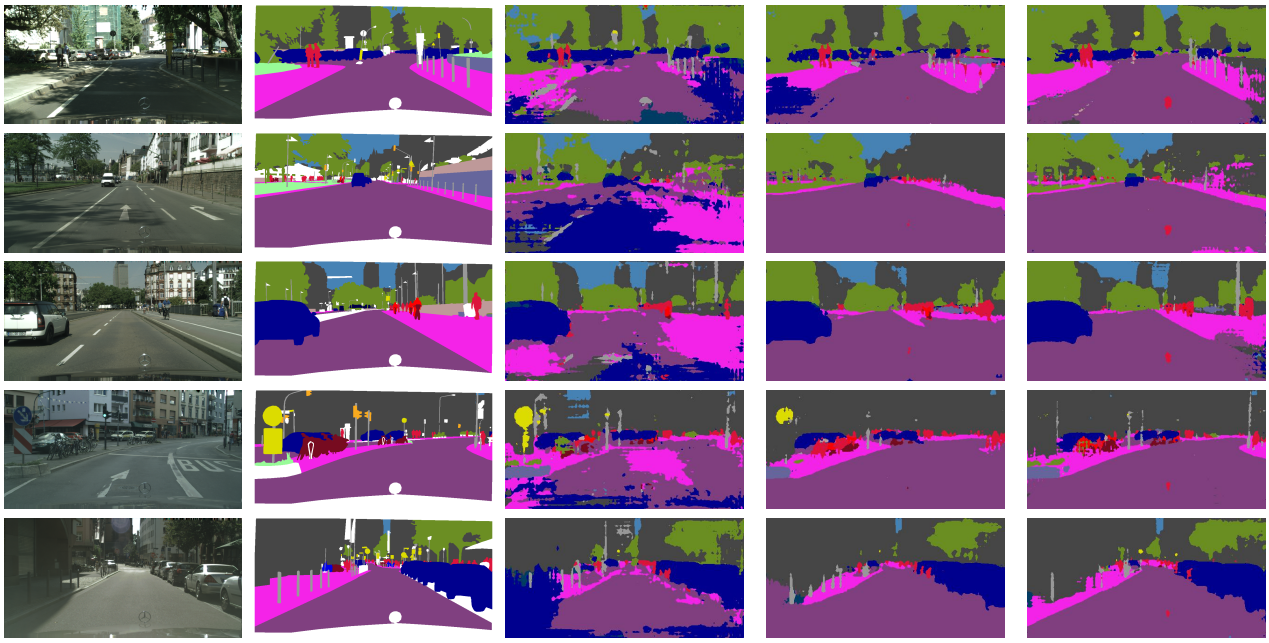
Similar to Figure 3 in our main paper, Figure 1 provides qualitative results for adaption from SYNTHIA to Cityscapes, comparing DISE against Source Only (i.e. no adaptation) and Conventional Adaptation (i.e. without disentanglement of structure and texture). We make the same observation as in Figure 3 in our main paper that the segmentations predicted by our model resemble more closely the ground truths than the other baselines. In particular, our DISE is able to distinguish clearly between "Sidewalk" and "Road", which is often difficult if not impossible with the conventional approach.

3. Image-to-Image Translation on SYNTHIA

Qualitative results showing image-to-image translation on SYNTHIA and Cityscapes datasets are likewise presented in Figure 2 for the two settings, S2T and T2S. Recall that the S2T (respectively, T2S) is to combine the structure content of a source image (respectively, a target image) with the texture appearance of another target image (respectively, source image). In the present case, the source domain is SYNTHIA and the target domain is Cityscapes. We observe that although the diversity of generated images tends to decrease due to the larger discrepancy between Cityscapes and SYNTHIA, the DISE is still effective in translating images from one domain to another with high quality while preserving well the desired structure content.

4. Structure and Texture Disentanglement

This session provides an additional set of simulation results not appearing in any part of the main paper to validate the DISEs effectiveness in disentangling the structure and texture components of an image. Specifically, we show that an image can preserve well its structure content and texture appearance even in a multi-cycle translation process. The experiment proceeds as follows: (1) given an image $I^a = \{z_c^a, z_p^a\}$ in one domain and another image $I^b = \{z_c^b, z_p^b\}$ in the other domain, we first produce in the first translation cycle a translated image $I^{a2b} = \{z_c^a, z_p^b\}$ by combining the structure content z_c^a of the image I^a and the texture appearance z_p^b of the image I^b, and then (2) at the end of the second cycle, in which the structure component, ideally z_c^a, of the translated image $I^{a2b} = \{z_c^a, z_p^b\}$ is extracted and combined with the texture component z_p^a of I^a



(a) Target Image (b) Ground Truth (c) Source Only (d) Conventional Adapt. (e) DISE (ours)

Figure 1. Segmentation results on Cityscapes when adapted from SYNTHIA. From left to right, (a) Target Image, (b) Ground Truth, (c) Source Only, (d) Conventional Adaptation, (e) and DISE.

obtained in the first cycle to reconstruct an image $\hat{I}^a$, we show that the reconstructed image $\hat{I}^a$ can approximate well the input image I^a in both structure content and texture appearance. For this to hold true, not only must the disentanglement of structure and texture components be successful, both components need to be preserved well when subjected to another cycle of decomposition and reconstruction.

Figure 3 presents qualitative results for all four possible combinations of datasets, namely, $\{I^a \in \text{GTA5}, I^b \in \text{Cityscapes}\}$, $\{I^a \in \text{Cityscapes}, I^b \in \text{GTA5}\}$, $\{I^a \in \text{SYNTHIA}, I^b \in \text{Cityscapes}\}$, and $\{I^a \in \text{Cityscapes}, I^b \in \text{SYNTHIA}\}$. It is observed that the translated images in columns (b) and (d) preserve well the structure content of the source images in column (a) while showing a variety of texture appearances depending on the target datasets. Furthermore, the reconstructions in column (c) and (e) do provide good approximations to their counterparts in column (a). These observations validate the disentanglement ability of DISE.

References

[1] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1

[2] G. Ros, L. Sellart, J. Materzynska, D. Vazquez, and A. 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 (CVPR)*, 2016. 1

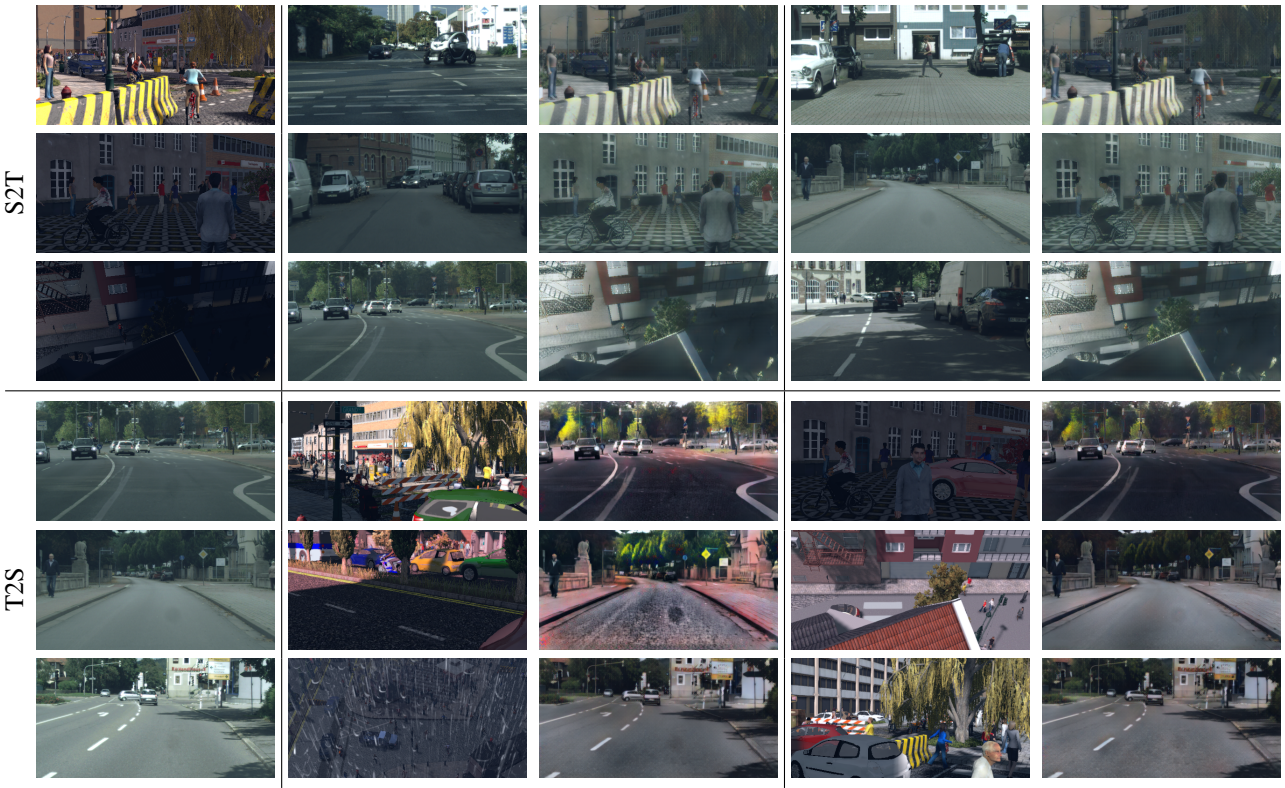


Figure 2. Sample results of translated images for SYNTHIA adapted to Cityscapes. S2T: the structure content of SYNTHIA images in (a) is combined with the texture appearance of Cityscapes images in (b) and (d) to output translated images in (c) and (e), respectively. T2S: the structure content of Cityscapes images in (a) is combined with the texture appearance of SYNTHIA images in (b) and (d) to output translated images in (c) and (e), respectively.

Figure 3. Sample results of multi-cycle image-to-image translation for GTA5 (source) to Cityscapes (target), Cityscapes (source) to GTA5 (target), SYNTHIA (source) to Cityscapes (target), and Cityscapes (source) to SYNTHIA (target). In the first cycle, the translated images in columns (b) and (d) are produced by combining the the structure content of the source images in column (a) and the texture appearance of some images in the target domain. In the second cycle, the reconstructed images in columns (c) and (e) are obtained by extracting the structure content of the translated images in columns (b) and (d), respectively, and combining it with the texture appearance of the source images in column (a).