APPENDIX

Xin Sun, Haojie Li

Shandong University of Science and Technology
Qingdao, Shandong, China

Xinchen Ye, Rui Xu*

Dalian University of Technology
Dalian, Liaoning, China

1 The Pretrained Encoder of STD

Inspired by some recent works [1], [2], [3] that use disentangled representations to decompose an image into domain-invariant structure and domain-specific style components for effective image translation, style transfer, and generalization improvement, we apply a similar strategy in our DSR scenario. Since we lack ground truth labels for structure and texture to supervise the network in learning disentangled representations, we adopt an unsupervised approach with adversarial losses to separate structure and texture components.

As illustrated in Figure 1, given two visual domains X and Y without paired training data, the framework consists of structure encoders E_X^c, E_Y^c, style encoders E_X^a, E_Y^a, and generators G_X, G_Y for both domains. Taking domain X as an example, the structure encoder E_X^c maps images onto a shared, domain-invariant structure space while the style encoder E_X^a maps images onto a domain-specific style space. The generator G_X then generates images conditioned on both structure and style features.

We add an adversarial loss $L_{adv}^{content}$ containing a structure discriminator D^s to distinguish the extracted structure representations between two domains, which encourages the structure features not to carry domain-specific cues:

$$\begin{aligned}&L_{adv}^{content} = \\ &\mathbb{E}_x[\frac{1}{2}\log D^s(E_X^c(x)) + \frac{1}{2}\log(1 - D^s(E_X^c(x)))] + \\ &\mathbb{E}_y[\frac{1}{2}\log D^s(E_Y^c(y)) + \frac{1}{2}\log(1 - D^s(E_Y^c(y)))].\end{aligned} \tag{1}$$

Besides, we impose adversarial loss L_{adv}^{domain} accompanied with discriminator D^d to discriminate between real images and translated images in each domain while G_X and G_Y attempt to generate realistic images.

Then, a cross-cycle consistency loss L_1^{cc} is introduced to exploit the disentangled structure and style representations for cyclic reconstruction:

$$\begin{aligned}L_1^{cc} = &\|G_X(E_Y^c(v), E_X^a(u)) - x\| + \\ &\|G_Y(E_Y^c(u), E_X^a(v)) - y\|,\end{aligned} \tag{2}$$

where $u = G_X(E_Y^c(y), E_X^a(x))$ and $v = G_Y(E_X^c(x), E_Y^a(y))$.

Finally, the full objective function of our network is:

$$\min_{G, E^c, E^a} \max_{D^s, D^d} \lambda_s L_{adv}^{content} + \lambda_c L_1^{cc} + \lambda_d L_{adv}^{domain}, \tag{3}$$

where the hyper-parameters $\lambda_s, \lambda_c, \lambda_d$ control the importance of each term.

Once the network is well trained, the structure encoder E_X^c (or E_Y^c) can be used for extracting the scene structure features for subsequent DSR training.

2 Training Details

For the structure encoder and style encoder, we use an architecture consisting of three convolution layers followed by four residual blocks. For the generator G, we use an architecture containing four residual blocks followed by three fractionally strided convolution layers.

We construct a dedicated dataset based on the aforementioned data sources. The key aspect of this dataset is the division into two different visual domains. Since each original dataset contains different categories with varying styles (e.g., different "Bathroom" categories in NYUv2 or "Living Room" category in both NYUv2 and RGB-D-D), we adopt two strategies: either selecting the same category from different datasets as separate visual domains, or splitting data of similar category in one dataset based on different capturing conditions (e.g., styles, lighting). Ultimately, we assemble approximately 1,500 images per visual domain for training the network. We empirically set $\lambda_s = 1$, $\lambda_c = 10$, and $\lambda_d = 1$ in the loss function.

3 References

[1] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang, "Diverse image-to-image translation via disentangled representations," in *ECCV*, 2018, pp. 35–51.

[2] Xiaotian Chen, Yuwang Wang, Xuejin Chen, and Wenjun Zeng, "S2r-depthnet: Learning a generalizable depth-specific structural representation," in *IEEE CVPR*, 2021, pp. 3034–3043.

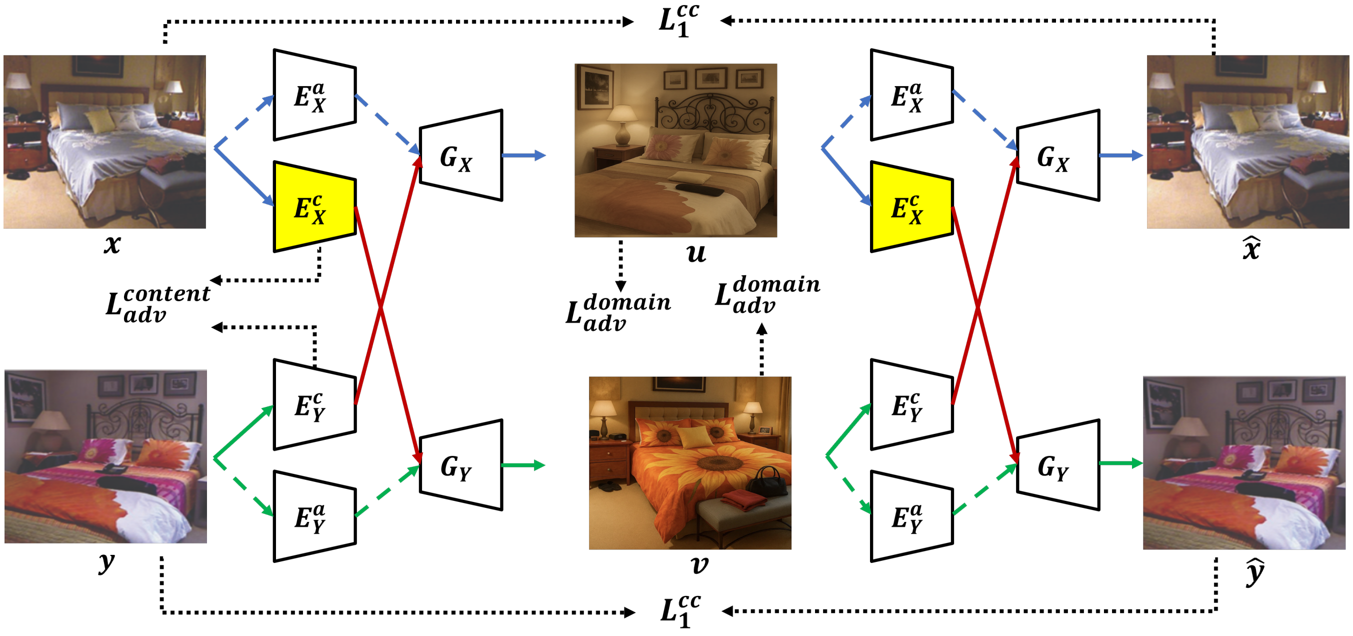

Fig. 1. The overall framework to train a domain-invariant structure encoder.

[3] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz, "Multimodal unsupervised image-to-image translation," in *ECCV*, 2018, pp. 172–189.