Supplementary Material for DLOW: Domain Flow for Adaptation and Generalization

Rui Gong[1] Wen Li[1] Yuhua Chen[1] Luc Van Gool[1,2]
[1]Computer Vision Laboratory, ETH Zurich [2]VISICS, ESAT/PSI, KU Leuven
gongr@student.ethz.ch {liwen, yuhua.chen, vangool}@vision.ee.ethz.ch

In this Supplementary, we provide additional information for,

- enlarged version of DLOW translated images for GTA5 to Cityscapes,
- the adaptation and generalization performance of our DLOW model on SYNTHIA to Cityscapes,
- the detailed network structure of our DLOW model for style generalization with four target domains,
- more examples for style generalization,
- the qualitative comparison of different methods on style transfer and style generalization.

1. Comparison of DLOW Translated Images with Brightness Adjusted Images

In Section 4.1 of the main paper, we show the examples of intermediate domain images between the source domain GTA5 and the target domain Cityscapes. The main change in those images at the first glance might be the image brightness. Here we provide an enlarged version of intermediate images to show that not only the brightness but also the subtle texture are adjusted to mimic the Cityscapes style. For comparison, we adjust the brightness of the translated image with $z = 0$ to match it with the brightness of the corresponding translated image with $z = 1$. The enlarged translated image with $z = 1$ and the corresponding the brightness adjusted image($z = 0$) are shown in Fig 2, from which we observe that the brightness adjusted image still exhibits obvious features of the game style such as the high contrast textures of the road and the red curb, while our DLOW translated image well mimics the texture of Cityscapes style.

2. Additional Results for Domain Adaptation and Generalization

In Section 4.1 of the main paper, we show the adaptation and the generalization performance of the DLOW model on the GTA5 to Cityscapes dataset. In this Supplementary, we further present the experimental results of our DLOW model on the SYNTHIA to Cityscapes dataset. The SYNTHIA dataset [6] is used as the source domain while the Cityscapes dataset [1] is used as the target domain. Similar to the experiment on GTA5, we also evaluate the generalization ability of learnt segmentation models to unseen domains on the KITTI [2], WildDash [9] and BDD100K [8] datasets.

SYNTHIA-RAND-CITYSCAPES is a dataset comprising 9400 photo-realistic images rendered from a virtual city and the semantic labels of the images are precise and compatible with Cityscapes test set.

The same training parameters and scheme as GTA5 are applied to SYNTHIA dataset, while the only difference lies in that we resize the training images to 1280×760 for the segmentation network.

Similar to GTA5, our DLOW model based on SYNTHIA dataset also exhibits excellent performance for the domain adaptation and the domain generalization. Following [7], the segmentation performance based on SYNTHIA dataset is tested on the Cityscapes validation dataset with 13 classes. As shown in Table 2, all pixel-level adaptation methods outperform the “NonAdapt” baseline, which verifies the effectiveness of the image translation for cross-domain segmentation. In particular, our “DLOW($z = 1$)” model achieves 41.6%, gaining 3% improvment compared to the ‘NonAdapt” baseline. After using the intermediate domain images, the adaptation performance can be further improved from 41.6% to 42.8%. The Table 1 also reports the result of our DLOW model adaptation performance combining with the AdaptSegNet method and the domain generalization performance for the unseen domains. The Original* in Table 1 denotes our retrained multi-level AdaptSegNet model in [7]. Compared with the retraining AdaptSegNet model, our DLOW model could improve the adaptation performance from 45.7% to 47.1%. The domain generalization results show that the intermediate domain images could improve the generalization ability of the adapted model.

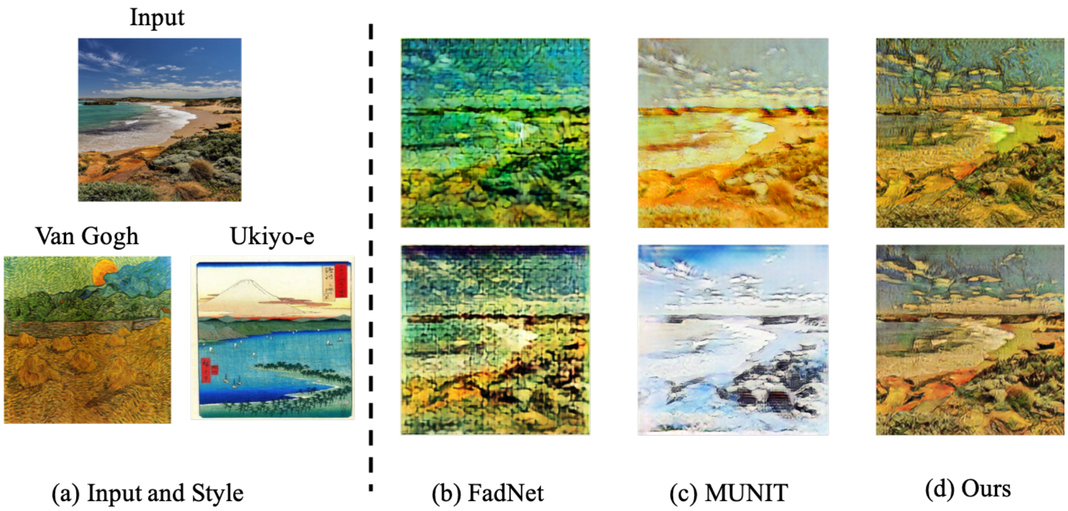

Figure 1: Comparison of our model with existing methods on style transfer and style generalization. The left part (a) shows the given input photo image and the example images of the target style. The translated results with different methods FadNet [5], MUNIT [4] and our DLOW are shown in right part (b), (c) and (d). The first row of the right part is the Van Gogh style transfer result while the second row is the style generalization result aiming at mixing the Van Gogh and Ukiyo-e style.

Table 1: Comparison of the performance of AdaptSegNet [7] when using original source images and intermediate domain images translated with our DLOW model for semantic segmention under domain adaptation (1st column) and domain generalization (2nd to 4th columns) scenarios. The Original* denotes our retrained multi-level AdaptSegNet model. The Original model is provided by the author of AdaptSegNet. The results are reported on mIoU over 13 categories. The best result is denoted in bold.

	Cityscapes	KITTI	WildDash	BDD100K
Original [7]	46.7	33.3	20.6	30.8
Original* [7]	45.7	**34.4**	20.0	30.8
DLOW	**47.1**	**34.4**	**24.4**	**35.3**

3. Network Structure for Style Generalization

In Section 3.6 of the main paper, we introduce that our DLOW model can be adapted for style generalization when there are multiple target domains available. We present the details in this section. The network structure of our DLOW model for style generalization is shown in Fig 3, where we have four target domains, each of which represents an image style. For the direction of $\mathcal{S} \rightarrow \mathcal{T}$, shown in Fig 3a, the style generalization model consists of two modules, the adversarial module and the image reconstruction module. For each target domain $\mathcal{T}_i$, there is one corresponding discriminator $D_{\mathcal{T}_i}$ measuring the distribution distance between the intermediate domain $\mathcal{M}^{(z)}$ and the target domain $\mathcal{T}_i$. Accordingly, the domainness variable z is expanded as a 4-dim vector $\mathbf{z} = [z_1, \dots, z_4]'$. For the other direction $\mathcal{T} \rightarrow \mathcal{S}$, shown in Fig 3b, the adversarial module is similar to that of the direction $\mathcal{S} \rightarrow \mathcal{T}$. However, the image reconstruction module is slightly different, since the image reconstruction loss should be weighted by the domainness vector $\mathbf{z}$.

4. Additional Results for Style Generalization

We provide an example for style generalization in Fig 6 of the main paper. Here we provide more experimental results in Fig 4 and Fig 5. The images with red bounding boxes are translated images in four target domains, *i.e.*, Monet, Van Gogh, Cezanne, and Ukiyo-e. Those can be considered as the “seen” styles. Our model gives similar translation results to CycleGAN model for each target domain. But the difference is that we only need one unified model for the four target domains while the CycleGAN should train four models. Moreover, the images with green bounding boxes are the mixed style images of their neighboring target styles and the image in the center is the mixed style image of all the four target styles, which are new styles that are never seen in the training data. We can observe that our DLOW model could generalize well across different styles, which proves the good domain generalization ability of our model.

5. Qualitative Comparison for Style Transfer and Style Generalization

In Section 4.2 of the main paper, we show the quantitative comparison results of our DLOW model with the Fad-

Table 2: Results of semantic segmentation on the CityScapes dataset based on DeepLab-v2 model with ResNet-101 backbone using the images translated with different models. The results are reported on mIoU over 13 categories. The best result is denoted in bold.

						SYNTHIA → Cityscapes							
Method	road	sidewalk	building	traffic light	traffic sign	vegetation	sky	person	rider	car	bus	motorbike	bicycle	mIoU
NonAdapt[7]	55.6	23.8	74.6	6.1	12.1	74.8	79.0	**55.3**	19.1	39.6	23.3	13.7	25.0	38.6
CycleGAN[3]	69.4	**28.3**	73.8	12.7	15.2	74.0	78.9	46.2	18.0	62.2	**27.6**	14.2	27.2	42.1
DLOW($z = 1$)	**71.0**	26.8	74.0	**13.9**	**17.5**	75.6	79.9	43.5	17.0	63.5	16.7	**14.5**	27.4	41.6
DLOW	65.3	22.4	**75.5**	9.1	13.2	**76.1**	**80.4**	52.0	**21.1**	**70.5**	26.3	10.7	**33.5**	**42.8**

Net [5] and the MUNIT [4] on the style transfer and the style generalization task. In this Supplementary, we further provide the qualitative result comparison in Fig 1. It can be observed that the FadNet fails to translate the photo to painting while the MUNIT and our DLOW model both could get reasonable results. For the Van Gogh style transfer result shown in Fig 1, our DLOW model could not only learn the typical color of the painting but also the details such as the brushwork and lines while the MUNIT only learns the main colors. For the Van Gogh and Ukiyo-e style generalization results shown in Fig 1, our DLOW model could combine the color and the stroke of the two styles while the MUNIT just fully changes the main colors from one style to another. The qualitative comparison result also demonstrates that our DLOW model performs better on both of the style transfer and generalization task compared with the FadNet and the MUNIT.

References

[1] 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 *CVPR*, 2016. 1

[2] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, 2012. 1

[3] 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 *ICML*, 2018. 3

[4] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *ECCV*, 2018. 2, 3

[5] Guillaume Lample, Neil Zeghidour, Nicolas Usunier, Antoine Bordes, Ludovic DENOYER, et al. Fader networks: Manipulating images by sliding attributes. In *NIPS*, 2017. 2, 3

[6] 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 *CVPR*, 2016. 1

[7] Y.-H. Tsai, W.-C. Hung, S. Schulter, K. Sohn, M.-H. Yang, and M. Chandraker. Learning to adapt structured output space for semantic segmentation. In *CVPR*, 2018. 1, 2, 3

[8] Fisher Yu, Wenqi Xian, Yingying Chen, Fangchen Liu, Mike Liao, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving video database with scalable annotation tooling. *arXiv:1805.04687*, 2018. 1

[9] Oliver Zendel, Katrin Honauer, Markus Murschitz, Daniel Steininger, and Gustavo Fernandez Dominguez. Wilddash - creating hazard-aware benchmarks. In *ECCV*, 2018. 1

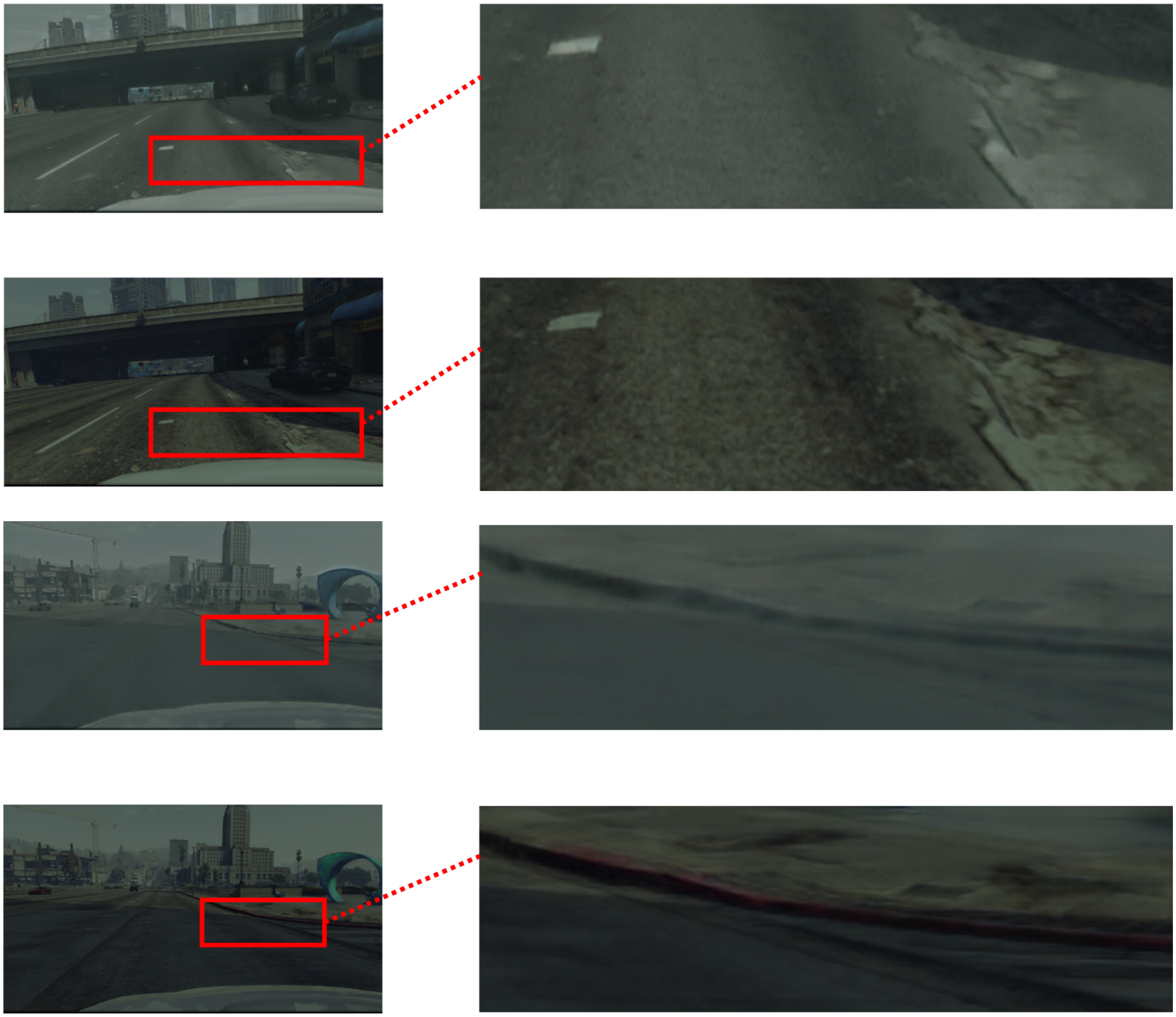

Figure 2: Examples of comparison between the DLOW translated image and the brightness adjusted image. We adjust the brightness of the DLOW translated source image($z = 0$) to make its brightness match the corresponding DLOW translated target image($z = 1$). The lower one in each group is the brightness adjusted image while the upper one is the DLOW translated target image($z = 1$). Part of the image is enlarged and shown in the right to prove that our DLOW translation not only change the brightness but also change the details such as the texture of the road and the style of the curb to mimic the feature of the Cityscapes image.

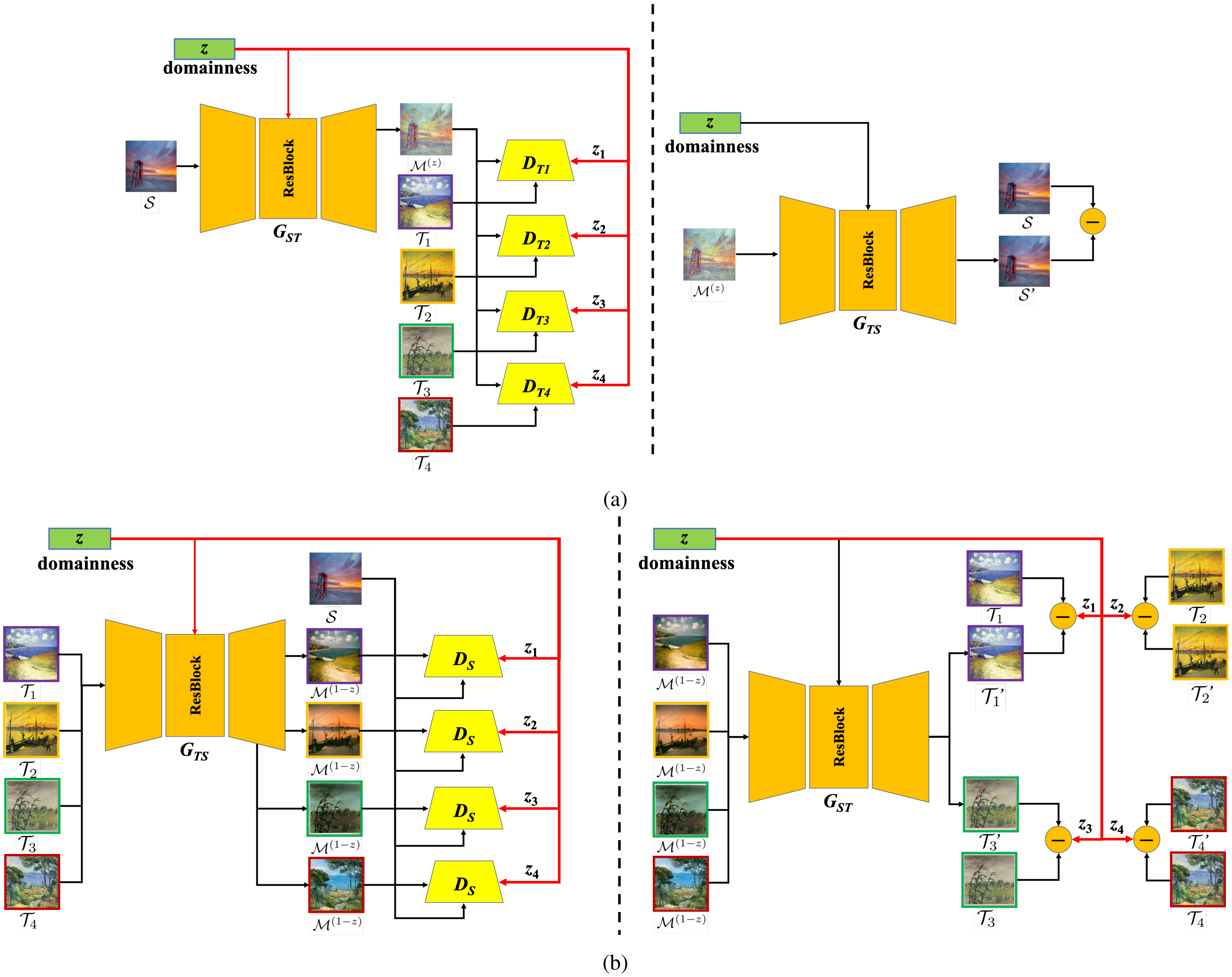

Figure 3: Network structure of DLOW model for style generalization with four target domains: (a) direction from $\mathcal{S} \rightarrow \mathcal{T}$; (b) direction from $\mathcal{T} \rightarrow \mathcal{S}$.

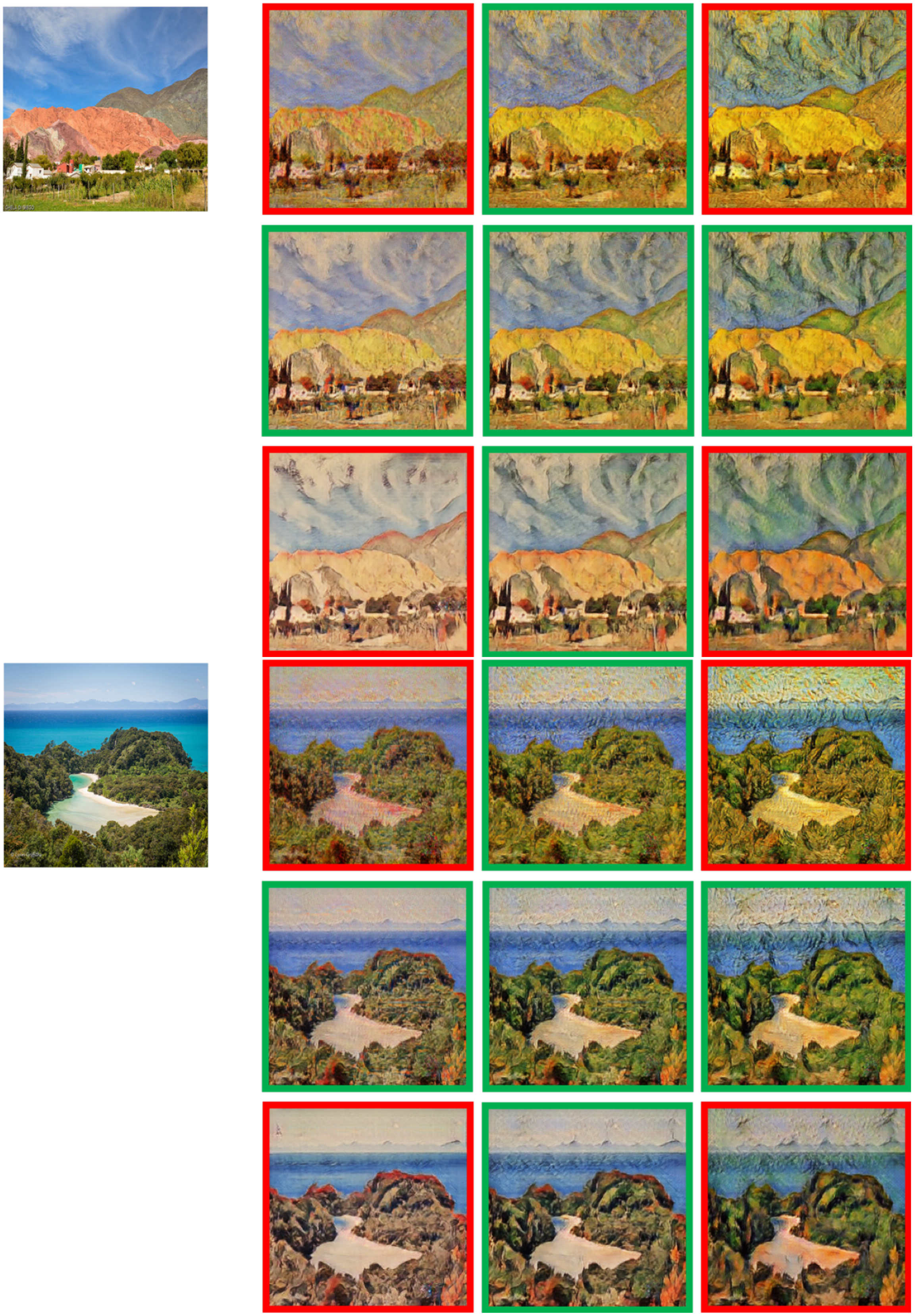

Figure 4: Examples of style generalization I. Results with red rectangles at four corners are images translated into the four target domains, and those with green rectangles in between are images translated into intermediate domains. The results show that our DLOW model generalizes well across styles, and produces new images styles smoothly.

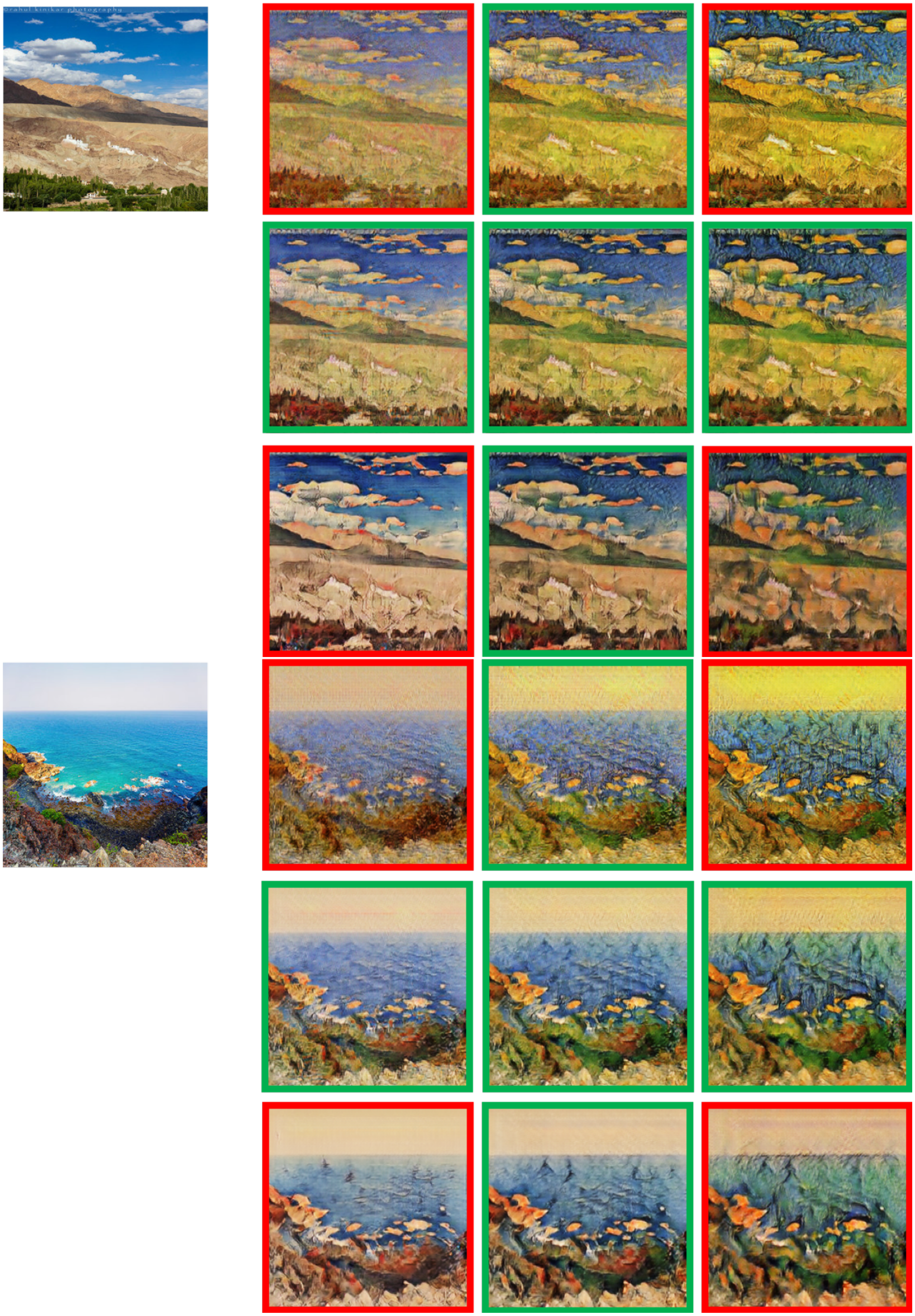

Figure 5: Examples of style generalization II. Results with red rectangles at four corners are images translated into the four target domains, and those with green rectangles in between are images translated into intermediate domains. The results show that our DLOW model generalizes well across styles, and produces new images styles smoothly.