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

Abstract

In this work, we present a domain flow generation(DLOW) model to bridge two different domains by generating a continuous sequence of intermediate domains flowing from one domain to the other. The benefits of our DLOW model are two-fold. First, it is able to transfer source images into different styles in the intermediate domains. The transferred images smoothly bridge the gap between source and target domains, thus easing the domain adaptation task. Second, when multiple target domains are provided for training, our DLOW model is also able to generate new styles of images that are unseen in the training data. We implement our DLOW model based on CycleGAN. A domainness variable is introduced to guide the model to generate the desired intermediate domain images. In the inference phase, a flow of various styles of images can be obtained by varying the domainness variable. We demonstrate the effectiveness of our model for both cross-domain semantic segmentation and the style generalization tasks on benchmark datasets. Our implementation is available at https://github.com/ETHRuiGong/DLOW.

1. Introduction

The domain shift problem is drawing increasing attention in recent years [21, 64, 54, 52, 15, 8]. In particular, there are two tasks that are of interest in computer vision community. One is the *domain adaptation* problem, where the goal is to learn a model for a given task from a label-rich data domain (*i.e*., source domain) to perform well in a label-scarce data domain (*i.e*., target domain). The other one is the *image translation* problem, where the goal is to transfer images in the source domain to mimic the image style in the target domain.

Generally, most existing works focus on the target domain only. They aim to learn models that well fit the target data distribution, *e.g*., achieving good classification accuracy in the target domain, or transferring source images into the target style. In this work, we instead are interested in the intermediate domains between source and target domains.

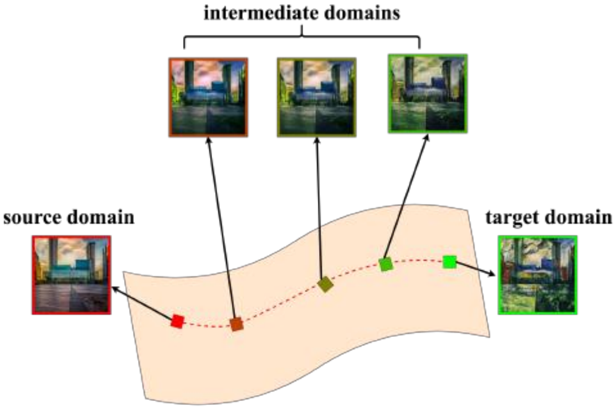


Figure 1: Illustration of data flow generation. Traditional image translation methods directly map the image from the source domain to the target domain, while our DLOW model is able to produce a sequence of intermediate domains shifting from the source domain to the target domain.

We present a new *domain flow generation* (DLOW) model, which is able to translate images from the source domain into an arbitrary intermediate domain between source and target domains. As shown in Fig 1, by translating a source image along the domain flow from the source domain to the target domain, we obtain a sequence of images that naturally characterize the distribution shift from the source domain to the target domain.

The benefits of our DLOW model are two-fold. First, those intermediate domains are helpful to bridge the distribution gap between two domains. By translating images into intermediate domains, those translated images can be used to ease the domain adaptation task. We show that the traditional domain adaptation methods can be boosted to achieve better performance in target domain with intermediate domain images. Moreover, the obtained models also exhibit good generalization ability on new datasets that are not seen in the training phase, benefiting from the diverse intermediate domain images.

Second, our DLOW model can be used for style generalization. Traditional image-to-image translation works [64, 28, 30, 38] mainly focus on learning a deterministic one-to-one mapping that transfers a source image into the target style. In contrast, our DLOW model allows to translate a source image into an intermediate domain that is related to multiple target domains. For example, when performing the photo to painting transfer, instead of obtaining a Monet or Van Gogh style, our DLOW model could produce a mixed style of Van Gogh, Monet, etc. Such mixture can be customized in the inference phase by simply adjusting an input vector that encodes the relatedness to different domains.

We implement our DLOW model based on CycleGAN [64], which is one of the state-of-the-art unpaired image-to-image translation methods. We augment the CycleGAN to include an additional input of domainness variable. On one hand, the domainness variable is injected into the translation network using the conditional instance normalization layer to affect the style of output images. On the other hand, it is also used as weights on discriminators to balance the relatedness of the output images to different domains. For multiple target domains, the domainness variable is extended as a vector containing the relatedness to all target domains. Extensive results on benchmark datasets demonstrate the effectiveness of our proposed model for domain adaptation and style generalization.

2. Related Work

Image to Image Translation: Our work is related to the image-to-image translation works. The image-to-image translation task aims at translating the image from one domain into another domain. Inspired by the success of Generative Adversarial Networks(GANs) [17], many works have been proposed to address the image-to-image translation based on GANs [28, 56, 64, 38, 39, 20, 65, 27, 1, 8, 33, 58, 37]. The early works [28, 56] assume that paired images between two domains are available, while the recent works such as CycleGAN [64], DiscoGAN [30] and UNIT [38] are able to train networks without using paired images. However, those works focus on learning deterministic image-to-image mappings. Once the model is learnt, a source image can only be transferred to a fixed target style.

A few recent works [39, 20, 65, 27, 1, 8, 33, 58, 37, 32] concentrate on learning a unified model to translate images into multiple styles. These works can be divided into two categories according to the controllability of the target styles. The first category, such as [27, 1], realizes the multimodal translation by sampling different style codes which are encoded from the target style images. However, those works focus on modelling intra-domain diversity, while our DLOW model aims at characterizing the inter-domain diversity. Moreover, they cannot explicitly control the translated target style using the input codes.

The second category, such as [8, 32], assigns the domain labels to different target domains and the domain labels are proven to be effective in controlling the translation direction. Among those, [32] shows that they could make interpolation between target domains by continuously shifting the different domain labels to change the extent of the contribution of different target domains. However, these methods only use the discrete binary domain labels in the training. Unlike the above work, the domainness variable proposed in this work is derived from the data distribution distance, and is used explicitly to regularize the style of output images during training.

Domain Adaptation and Generalization: Our work is also related to the domain adaptation and generalization works. Domain adaptation aims to utilize a labeled source domain to learn a model that performs well on an unlabeled target domain [13, 18, 12, 55, 29, 3, 31, 16, 6, 61, 57]. Domain generalization is a similar problem, which aims to learn a model that could be generalized to an unseen target domain by using multiple labeled source domains [42, 15, 45, 41, 44, 34, 36, 35].

Our work is partially inspired by [18, 16, 10], which have shown that the intermediate domains between source and target domains are useful for addressing the domain adaptation problem. They represent each domain as a subspace or covariance matrix, and then connect them on the corresponding manifold to model intermediate domains. Different from those works, we model the intermediate domains by directly translating images on pixel level. This allows us to easily improve the existing deep domain adaptation models by using the translated images as training data. Moreover, our model can also be applied to image-level domain generalization by generating mixed-style images.

Recently, there is an increasing interest to apply domain adaptation techniques for semantic segmentation from synthetic data to the real scenario [22, 21, 7, 67, 40, 25, 11, 46, 51, 53, 23, 47, 62, 54, 43, 50, 52, 66, 5]. Most of those works conduct the domain adaptation by adversarial training on the feature level with different priors. The recent Cycada [21] also shows that it is beneficial to perform pixel-level domain adaptation firstly by transferring source image into the target style based on the image-to-image translation methods like CycleGAN [64]. However, those methods address domain shift by adapting to only the target domain. In contrast, we aim to perform pixel-level adaptation by transferring source images to a flow of intermediate domains. Moreover, our model can also be used to further improve the existing feature-level adaptation methods.

3. Domain Flow Generation

3.1. Problem Statement

In the domain shift problem, we are given a source domain $\mathcal{S}$ and a target domain $\mathcal{T}$ containing samples from two different distributions P_S and P_T, respectively. Denoting a source sample as $\mathbf{x}^s \in \mathcal{S}$ and a target sample as $\mathbf{x}^t \in \mathcal{T}$, we have $\mathbf{x}^s \sim P_S$, $\mathbf{x}^t \sim P_T$, and $P_S \neq P_T$.

Such distribution mismatch usually leads to a significant performance drop when applying the model trained on $\mathcal{S}$ to $\mathcal{T}$. Many works have been proposed to address the domain shift for different vision applications. A group of recent works aim to reduce the distribution difference on the feature level by learning domain-invariant features [13, 18, 31, 16], while others work on the image level to transfer source images to mimic the target domain style [64, 38, 65, 27, 1, 8].

In this work, we also propose to address the domain shift problem on image level. However, different from existing works that focus on transferring source images into only the target domain, we instead transfer them into all intermediate domains that connect source and target domains. This is partially motivated by the previous works [18, 16, 10], which have shown that the intermediate domains between source and target domains are useful for addressing the domain adaptation problem.

In the follows, we first briefly review the conventional image-to-image translation model CycleGAN. Then, we formulate the intermediate domain adaptation problem based on the data distribution distance. Next, we present our DLOW model based on the CycleGAN model. We then show the benefits of our DLOW model with two applications: 1) improve existing domain adaptation models with the images generated from DLOW model, and 2) transfer images into arbitrarily mixed styles when there are multiple target domains.

3.2. The CycleGAN Model

We build our model based on the state-of-the-art CycleGAN model [64] which is proposed for unpaired image-to-image translation. Formally, the CycleGAN model learns two mappings between $\mathcal{S}$ and $\mathcal{T}$, *i.e.*, $G_{ST} : \mathcal{S} \rightarrow \mathcal{T}$ which transfers the images in $\mathcal{S}$ into the style of $\mathcal{T}$, and $G_{TS} : \mathcal{T} \rightarrow \mathcal{S}$ which acts in the inverse direction. We take the $\mathcal{S} \rightarrow \mathcal{T}$ direction as an example to explain CycleGAN.

To transfer source images into the target style and also preserve the semantics, the CycleGAN employs an adversarial training module and a reconstruction module, respectively. In particular, the adversarial training module is used to align the image distributions for two domains, such that the style of mapped images matches the target domain. Let us denote the discriminator as D_T, which attempts to distinguish the translated images and the target images. Then the objective function of the adversarial training module can be written as,

$$\begin{aligned} \min_{G_{ST}} \max_{D_T} \quad & \mathbb{E}_{\mathbf{x}^t \sim P_T} \left[\log(D_T(\mathbf{x}^t)) \right] \\ & + \mathbb{E}_{\mathbf{x}^s \sim P_S} \left[\log(1 - D_T(G_{ST}(\mathbf{x}^s))) \right]. \end{aligned} \tag{1}$$

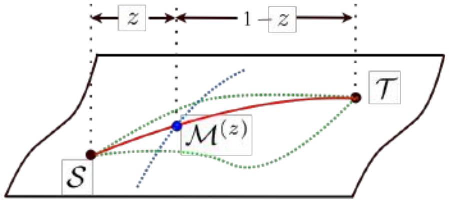


Figure 2: Illustration of domain flow. Many possible paths (the green dash lines) connect source and target domains, while the domain flow is the shortest one (the red line). There are multiple domains (the blue dash line) keeping the expected relative distances to source and target domains. An intermediate domain (the blue dot) is the point at the domain flow that keeps the right distances to two domains.

Moreover, the reconstruction module is to ensure the mapped image $G_{ST}(\mathbf{x}^s)$ to preserve the semantic content of the original image $\mathbf{x}^s$. This is achieved by enforcing a cycle consistency loss such that $G_{ST}(\mathbf{x}^s)$ is able to recover $\mathbf{x}^s$ when being mapped back to the source style, *i.e.*,

$$\min_{G_{ST}} \quad \mathbb{E}_{\mathbf{x}^s \sim P_S} \left[\| G_{TS}(G_{ST}(\mathbf{x}^s)) - \mathbf{x}^s \|_1 \right]. \tag{2}$$

Similar modules are applied to the $\mathcal{T} \rightarrow \mathcal{S}$ direction. By jointly optimizing all modules, CycleGAN model is able to transfer source images into the target style and v.v.

3.3. Modeling Intermediate Domains

Intermediate domains have been shown to be helpful for domain adaptation [18, 16, 10], where they model intermediate domains as a geodesic path on Grassmannian or Riemannian manifold. Inspired by those works, we also characterize the domain shift using intermediate domains that connect the source and target domains. Diffrent from those works, we directly operate at the image level, *i.e.*, translating source images into different styles corresponding to intermediate domains. In this way, our method can be easily integrated with deep learning techniques for enhancing the cross-domain generalization ability of models.

In particular, let us denote an intermediate domain as $\mathcal{M}^{(z)}$, where $z \in [0, 1]$ is a continous variable which models the relatedness to source and target domains. We refer to z as the domainness of intermediate domain. When $z = 0$, the intermediate domain $\mathcal{M}^{(z)}$ is identical to the source domain $\mathcal{S}$; and when $z = 1$, it is identical to the target domain $\mathcal{T}$. By varying z in the range of $[0, 1]$, we thus obtain a sequence of intermediate domains that flow from $\mathcal{S}$ to $\mathcal{T}$.

There are many possible paths to connect the source and target domains. As shown in Fig 2, assuming there is a manifold of domains, where a domain with given data distribution can be seen as a point residing at the manifold. We expect the domain flow $\mathcal{M}^{(z)}$ to be the shortest geodesic path connecting $\mathcal{S}$ and $\mathcal{T}$. Moreover, given any z, the distance from $\mathcal{S}$ to $\mathcal{M}^{(z)}$ should also be proportional to the

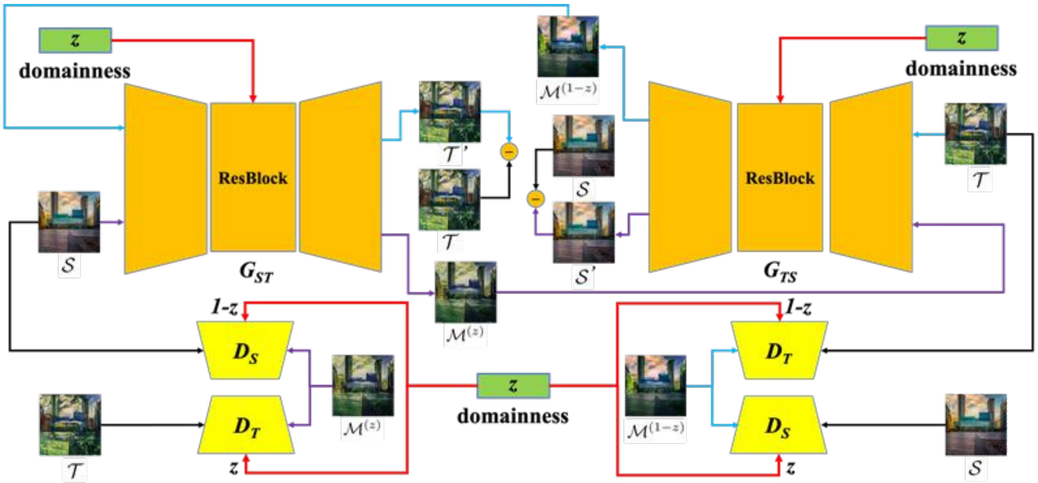


Figure 3: The overview of our DLOW model: the generator takes domainness z as additional input to control the image translation and to reconstruct the source image; The domainness z is also used to weight the two discriminators.

distance between $\mathcal{S}$ to $\mathcal{T}$ by the value of z. Denoting the data distribution of $\mathcal{M}^{(z)}$ as $P_M^{(z)}$, we expect that,

$$\frac{dist\left(P_S, P_M^{(z)}\right)}{dist\left(P_T, P_M^{(z)}\right)} = \frac{z}{1-z}, \tag{3}$$

where $dist(\cdot,\cdot)$ is a valid distance measurement over two distributions. Thus, generating an intermediate domain $\mathcal{M}^{(z)}$ for a given z becomes finding the point satisfying Eq. (3) that is closet to $\mathcal{S}$ and $\mathcal{T}$, which leads to minimize the following loss,

$$\mathcal{L} = (1-z)\cdot dist\left(P_S, P_M^{(z)}\right) + z \cdot dist\left(P_T, P_M^{(z)}\right). \tag{4}$$

As shown in [2], many types of distance have been exploited for image generation and image translation. The adversarial loss in Eq. (1) can be seen as a lower bound of the Jessen-Shannon divergence. We also use it to measure distribution distance in this work.

3.4. The DLOW Model

We now present our DLOW model to generate intermediate domains. Given a source image $\mathbf{x}^s \sim P_s$, and a domainness variable $z \in [0,1]$, the task is to transfer $\mathbf{x}^s$ into the intermediate domain $\mathcal{M}^{(z)}$ with the distribution $P_M^{(z)}$ that minimizes the objective in Eq. (4). We take the $\mathcal{S} \to \mathcal{T}$ direction as an example, and the other direction can be similarly applied.

In our DLOW model, the generator G_{ST} no longer aims to directly transfer $\mathbf{x}^s$ to the target domain $\mathcal{T}$, but to move $\mathbf{x}^s$ towards it. The interval of such moving is controlled by the domainness variable z. Let us denote $\mathcal{Z} = [0,1]$ as the domain of z, then the generator in our DLOW model can be represented as $G_{ST}(\mathbf{x}^s, z) : \mathcal{S} \times \mathcal{Z} \to \mathcal{M}^{(z)}$ where the input is a joint space of $\mathcal{S}$ and $\mathcal{Z}$.

Adversarial Loss: As discussed in Section 3.3, We deploy the adversarial loss as the distribution distance measurement to control the relatedness of an intermediate domain to the source and target domains. Specifically, we introduce two discriminators, $D_S(\mathbf{x})$ to distinguish $\mathcal{M}^{(z)}$ and $\mathcal{S}$, and $D_T(\mathbf{x})$ to distinguish $\mathcal{M}^{(z)}$ and $\mathcal{T}$, respectively. Then, the adversarial losses between $\mathcal{M}^{(z)}$ and $\mathcal{S}$ and $\mathcal{T}$ can be written respectively as,

$$\begin{aligned}\mathcal{L}_{adv}(\ G_{ST}\ , D_S) = {} & \mathbb{E}_{\mathbf{x}^s \sim P_S}\left[\log(D_S(\mathbf{x}^s))\right] \\ + {} & \mathbb{E}_{\mathbf{x}^s \sim P_S}\left[\log(1 - D_S(G_{ST}(\mathbf{x}^s, z)))\right]\end{aligned} \tag{5}$$

$$\begin{aligned}\mathcal{L}_{adv}(\ G_{ST}\ , D_T) = {} & \mathbb{E}_{\mathbf{x}^t \sim P_T}\left[\log(D_T(\mathbf{x}^t))\right] \\ + {} & \mathbb{E}_{\mathbf{x}^s \sim P_S}\left[\log(1 - D_T(G_{ST}(\mathbf{x}^s, z)))\right].\end{aligned} \tag{6}$$

By using the above losses to model $dist(P_S, P_M^{(z)})$ and $dist(P_T, P_M^{(z)})$ in Eq. (4), we derive the following loss,

$$\mathcal{L}_{adv} = (1-z)\mathcal{L}_{adv}(G_{ST}, D_S) + z\mathcal{L}_{adv}(G_{ST}, D_T). \tag{7}$$

Image Cycle Consistency Loss: Similarly as in CyleGAN, we also apply a cycle consistency loss to ensure the semantic content is well-preserved in the translated image. Let us denote the generator on the other direction as $G_{TS}(\mathbf{x}^t, z) : \mathcal{T} \times \mathcal{Z} \to \mathcal{M}^{(1-z)}$, which transfers a sample $\mathbf{x}^t$ from the target domain towards the source domain by a interval of z. Since G_{TS} acts in an inverse direction to G_{ST}, we can use it to recover $\mathbf{x}^s$ from the translated version $G_{ST}(\mathbf{x}^s, z)$, which gives the following loss,

$$L_{cyc} = \mathbb{E}_{\mathbf{x}^s \sim P_s}\left[\| G_{TS}(G_{ST}(\mathbf{x}^s, z), z) - \mathbf{x}^s \|_1\right]. \tag{8}$$

Full Objective: We integrate the losses defined above, then the full objective can be defined as,

$$\mathcal{L} = \mathcal{L}_{adv} + \lambda_1 \mathcal{L}_{cyc}, \tag{9}$$

where λ_1 is a hyper-parameter used to balance the two losses in the training process.

Similar loss can be defined for the other direction $\mathcal{T} \rightarrow \mathcal{S}$. Due to the usage of adversarial loss $\mathcal{L}_{adv}$, the training is performed in an alternating manner. We first minimize the full objective with regard to the generators, and then maximize it with regard to the discriminators.

Implementation: We illustrate the network structure of of our DLOW model in Fig 3. First, the domainness variable z is taken as the input of the generator G_{ST}. This is implemented with the Conditional Instance Normalization (CN) layer [1, 26]. We first use one deconvolution layer to map the domainness variable z to the vector with dimension $(1, 16, 1, 1)$, and then use this vector as the input for the CN layer. Moreover, the domainness variable also plays the role of weighting discriminators to balance the relatedness of the generated images to different domains. It is also used as input in the image cycle consistency module. During the training phase, we randomly generate the domainess parameter z for each input image. As inspired by [24], we force the domainness variable z to obey the beta distribution, i.e. $f(z, \alpha, \beta) = \frac{1}{B(\alpha,\beta)} z^{\alpha-1}(1-z)^{\beta-1}$, where β is fixed as 1, and α is a function of the training step $\alpha = e^{\frac{t-0.5T}{0.25T}}$ with t being the current iteration and T being the total number of iterations. In this way, z tends to be sampled more likely as small values at the beginning, and gradually shift to larger values at the end, which gives slightly more stable training than uniform sampling.

3.5. Boosting Domain Adaptation Models

With the DLOW model, we are able to translate each source image $\mathbf{x}^s$ into an arbitrary intermediate domain $\mathcal{M}^{(z)}$. Let us denote the source dataset as $\mathcal{S} = \{(\mathbf{x}_i^s, y_i)|_{i=1}^n\}$ where y_i is the label of $\mathbf{x}_i^s$. By feeding each of the image $\mathbf{x}_i^s$ combined with z_i randomly sampled from the uniform distribution $\mathcal{U}(0, 1)$, we then obtain a translated dataset $\tilde{\mathcal{S}} = \{(\tilde{\mathbf{x}}_i^s, y_i)|_{i=1}^n\}$ where $\tilde{\mathbf{x}}_i^s = G_{ST}(\mathbf{x}_i^s, z_i)$ is the translated version of $\mathbf{x}_i^s$. The images in $\tilde{\mathcal{S}}$ spread along the domain flow from source to target domain, and therefore become much more diverse. Using $\tilde{\mathcal{S}}$ as the training data is helpful to learn domain-invariant models for computer vision tasks. In Section 4.1, we demonstrate that model trained on $\tilde{\mathcal{S}}$ achieves good performance for the cross-domain semantic segmentation problem.

Moreover, the translated dataset $\tilde{\mathcal{S}}$ can also be used to boost the existing adversarial training based domain adaptation approaches. Images in $\tilde{\mathcal{S}}$ fill the gap between the source and target domains, and thus ease the domain adaptation task. Taking semantic segmentation as an example, a typical way is to append a discriminator to the segmentation model, which is used to distinguish the source and target samples. Using the adversarial training strategy to optimize the discriminator and the segmentation model, the segmentation model is trained to be more domain-invariant.

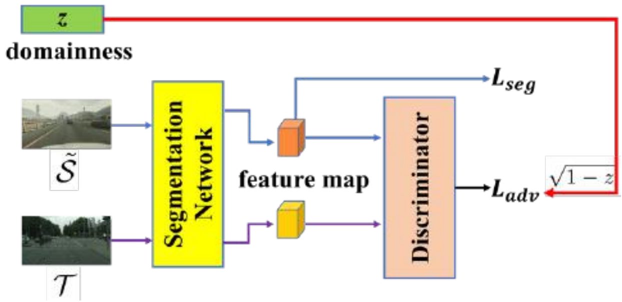


Figure 4: Illustration of boosting domain adaptation model for corss-domain semantic segmentation with DLOW model. Intermediate domain images are used as source dataset, and the adversarial loss is weighted by domainness.

As shown in Fig 4, we replace the source dataset $\mathcal{S}$ with the translated version $\tilde{\mathcal{S}}$, and apply a weight $\sqrt{1-z_i}$ to the adversarial loss. The motivation is as follows, for each sample $\tilde{\mathbf{x}}_i^s$, if the domainness z_i is higher, it is closer to the target domain, then the weight of adversarial loss can be reduced. Otherwise, we should enhance the loss weight.

3.6. Style Generalization

Most existing image-to-image translation works learn a deterministic mapping between two domains. After the model is learnt, source images can only be translated to a fixed style. In contrast, our DLOW model takes an random z to translate images into various styles. When multiple target domains are provided, it is also able to transfer the source image into a mixture of different target styles. In other words, we are able to generalize to an unseen intermediate domain that is related to existing domains.

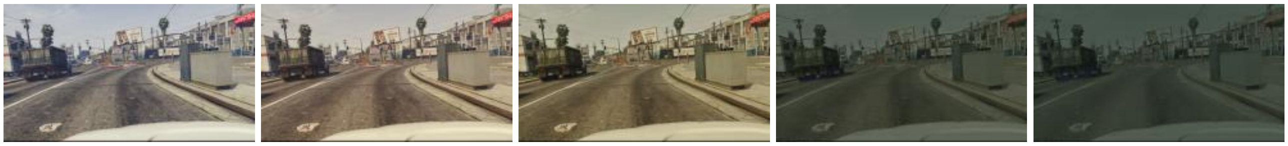
In particular, suppose we have K target domains, denoted as $\mathcal{T}_1, \ldots, \mathcal{T}_K$. Accordingly, the domainness variable z is expanded as a K-dim vector $\mathbf{z} = [z_1, \ldots, z_K]'$ with $\sum_{k=1}^{K} z_k = 1$. Each elelment z_k represents the relatedness to the k-th target domain. To map an image from the source domain to the intermediate domain defined by $\mathbf{z}$, we need to optimize the following objective,

$$\mathcal{L} = \sum_{k=1}^{K} z_k \cdot dist(P_M, P_{\mathcal{T}_k}), \quad \text{s.t.} \quad \sum_{1}^{K} z_k = 1 \tag{10}$$

where P_M is the distribution of the intermediate domain, P_{T_K} is the distribution of T_k. The network structure can be easily adjusted from our DLOW model to optimize the above objective. We leave the details in the Supplementary due to the space limitation.

4. Experiments

In this section, we demonstrate the benefits of our DLOW model with two tasks. In the first task, we ad-



(a) $z = 0$ (b) $z = 0.3$ (c) $z = 0.6$ (d) $z = 0.8$ (e) $z = 1$

Figure 5: Examples of intermediate domain images from GTA5 to Cityscapes. As the domainness variable increases from 0 to 1, the styles of the translated images shift from the synthetic GTA5 style to the realistic Cityscapes style gradually.

GTA5 → Cityscapes																				
Method	road	sidewalk	building	wall	fence	pole	traffic light	traffic sign	vegetation	terrian	sky	person	rider	car	truck	bus	train	motorbike	bicycle	mIoU
NonAdapt[54]	75.8	16.8	77.2	12.5	**21.0**	25.5	**30.1**	20.1	81.3	24.6	70.3	53.8	**26.4**	49.9	17.2	25.9	**6.5**	**25.3**	**36.0**	36.6
CycleGAN[21]	81.7	27.0	**81.7**	**30.3**	12.2	28.2	25.5	27.4	82.2	**27.0**	77.0	**55.9**	20.5	**82.8**	30.8	38.4	0.0	18.8	32.3	41.0
DLOW($z = 1$)	**88.5**	**33.7**	80.7	26.9	15.7	27.3	27.7	**28.3**	80.9	26.6	74.1	52.6	25.1	76.8	30.5	27.2	0.0	15.7	**36.0**	40.7
DLOW	87.1	33.5	80.5	24.5	13.2	**29.8**	29.5	26.6	**82.6**	26.7	**81.8**	**55.9**	25.3	78.0	**33.5**	**38.7**	0.0	22.9	34.5	**42.3**

Table 1: 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 19 categories. The best result is denoted in bold.

	Cityscapes	KITTI	WildDash	BDD100K
Original [54]	42.4	30.7	18.9	37.0
DLOW	**44.8**	**36.6**	**24.9**	**39.1**

Table 2: Comparison of the performance of AdaptSegNet [54] 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 results are reported on mIoU over 19 categories. The best result is denoted in bold.

dress the domain adaptation problem, and train our DLOW model to generate the intermediate domain samples to boost the domain adaptation performance. In the second task, we consider the style generalization problem, and train our DLOW model to transfer images into new styles that are unseen in the training data.

4.1. Domain Adaptation and Generalization

4.1.1 Experiments Setup

For the domain adaptation problem, we follow [22, 21, 7, 67] to conduct experiments on the urban scene semantic segmentation by learning from synthetic data to real scenario. The GTA5 dataset [48] is used as the source domain while the Cityscapes dataset [9] as the target domain. Moreover, we also evaluate the generalization ability of learnt segmentation models to unseen domains, for which we take the KITTI [14], WildDash [60] and BDD100K [59] datasets as additional unseen datasets for evaluation. We also conduct experiments using the SYNTHIA dataset [49] as the source domain, and provide the results in Supplementary.

Cityscapes is a dataset consisting of urban scene images taken from some European cities. We use the $2,993$ training images without annotation as unlabeled target samples in training phase, and 500 validation images with annotation for evaluation, which are densely labelled with 19 classes.

GTA5 is a dataset consisting of $24,966$ densely labelled synthetic frames generated from the computer game whose scenes are based on the city of Los Angeles. The annotations of the images are compatible with the Cityscaps.

KITTI is a dataset consisting of images taken from mid-size city of Karlsruhe. We use 200 validation images densely labeled and compatible with Cityscapes.

WildDash is a dataset covers images from different sources, different environments(place, weather, time and so on) and different camera characteristics. We use 70 labeled and Cityscapes annotation compatible validation images.

BDD100K is a driving dataset covering diverse images taken from US whose label maps are with training indices specified in Cityscapes. We use $1,000$ densely labeled images for validation in our experiment.

In this task, we first train our proposed DLOW model using the GTA5 dataset as the source domain, and Cityscapes as the target domain. Then, we generate a translated GTA5 dataset with the learnt DLOW model. Each source image is fed into DLOW with a random domainness variable z. The new translated GTA5 dataset contains exactly the same number of images as the original one, but the styles of images randomly drift from the synthetic style to the real style. We then use the translated GTA dataset as the new source domain to train segmentation models.

We implement our model based on Augmented CycleGAN [1] and CyCADA [21]. Following their setup, all images are resized to have width 1024 while keeping the aspect ratio and the crop size is set as 400×400. When training the DLOW model, the image cycle consistency loss weight is set as 10. The learning rate is fixed as 0.0002. For the segmentation network, we use the AdaptSegNet [54] model, which is based on DeepLab-v2 [4] with ResNnet-101 [19] as the backbone network. The training images are resized to 1280×720. We follow the exact the same training policy as in the AdaptSegNet.

4.1.2 Experimental Results

Intermediate Domain Images: To verify the ability of our DLOW model to generate intermediate domain images, in the inference phase, we fix the input source image, and vary the domainness variable from 0 to 1. A few examples are shown in Fig 5. It can be observed that the styles of translated images gradually shift from the synthetic style of GTA5 to the real style of Cityscapes, which demonstrates the DLOW model is capable of modeling the domain flow to bridge the source and target domains as expected. Enlarged images and more discussion are provided in Supplementary.

Cross-Domain Semantic Segmentation: We further evaluate the usefulness of intermediate domain images in two settings. In the first setting, we compare with the CycleGAN model [64], which is used in the CycADA approach [21] for performing pixel-level domain adaptation. The difference between CycleGAN and our DLOW model is that CycleGAN transfers source images to mimic only the target style, while our DLOW model transfers source images into random styles flowing from the source domain to the target domain. We first obtain a translated version of the GTA5 dataset with each model. Then, we respectively use the two transalated GTA5 datasets to train DeepLab-v2 models, which are evaluated on the Cityscapes dataset for semantic segmentation. We also include the "NonAdapt" baseline which uses the original GTA5 images as training data, as well as a special case of our approach, "DLOW($z = 1$)", where we set $z = 1$ for all source images when making image translation using the learnt DLOW model.

The results are shown in Table 1. We observe that all pixel-level adaptation methods outperform the "NonAdapt" baseline, which verifies that image translation is helpful for training models for cross-domain semantic segmentation. Moreover, "DLOW($z = 1$)" is a special case of our model that directly translates source images into the target domain, which non-surprisingly gives comparable result as the CycADA-pixel method (40.7% v.s. 41.0%). By further using intermediate domain images, our DLOW model is able to improve the result from 40.7% to 42.3%, which demonstrates that intermediate domain images are helpful for learning a more robust domain-invariant model.

In the second setting, we further use intermediate domain images to improve the feature-level domain adpatation model. We conduct experiments based on the AdaptSegNet method [54], which is open source and has reported the state-of-the-art result for GTA5→CityScapes. It consists of multiple levels of adversarial training, and we augment each level with the loss weight discussed in Section 3.5. The results are reported in Table 2. The "Original" method denotes the AdaptSegNet model that is trained using GTA5 as the source domain, for which the results are obtained using their released pretrained model. The "DLOW" method is AdaptSegNet trained using translated dataset with our DLOW model. From the first column, we observe that the intermediate domain images are able to improve the AdaptSegNet model by 2.5% from 42.3% to 44.8%. More interestingly, we show that the AdaptSegNet model with DLOW translated images also exhibits excellent domain generalization ability when being applied to unseen domains, which achieves significantly better results than the original AdaptSegNet model on the KITTI, WildDash and BDD100K datasets as reported in the second to the fourth columns, respectively. This shows that intermediate domain images are useful to improve the model's cross-domain generalization ability.

4.2. Style Generalization

We conduct the style generalization experiment on the Photo to Artworks dataset[64], which consists of real photographs ($6,853$ images) and artworks from Monet($1,074$ images), Cezanne(584 images), Van Gogh(401 images) and Ukiyo-e($1,433$ images). We use the real photographs as the source domain, and the remaining as four target domains. As discussed in Section 3.6, The domainness variable in this experiment is expanded as a 4-dim vector $[z_1, z_2, z_3, z_4]'$ meeting the condition $\sum_{i=1}^{4} z_i = 1$. Also, z_1, z_2, z_3 and z_4 corresponds to Monet, Van Gogh, Ukiyo-e and Cezanne, respectively. Each element z_i can be seen as how much each style contributes to the final mixture style. In every 5 steps of the training, we set the domainness variable z as $[1, 0, 0, 0]$, $[0, 1, 0, 0]$, $[0, 0, 1, 0]$, $[0, 0, 0, 1]$ and uniformly distributed random variable. The qualitative results of the style generalization are shown in Fig 6. From the qualitative results, it is shown that our DLOW model can translate the photo image to corresponding artworks with different styles. When varying the values of domainness vector, we can also successfully produce new styles related to different painting styles, which demonstrates the good generalization ability of our model to unseen domains. Note, different from [63, 26], we do not need any reference image

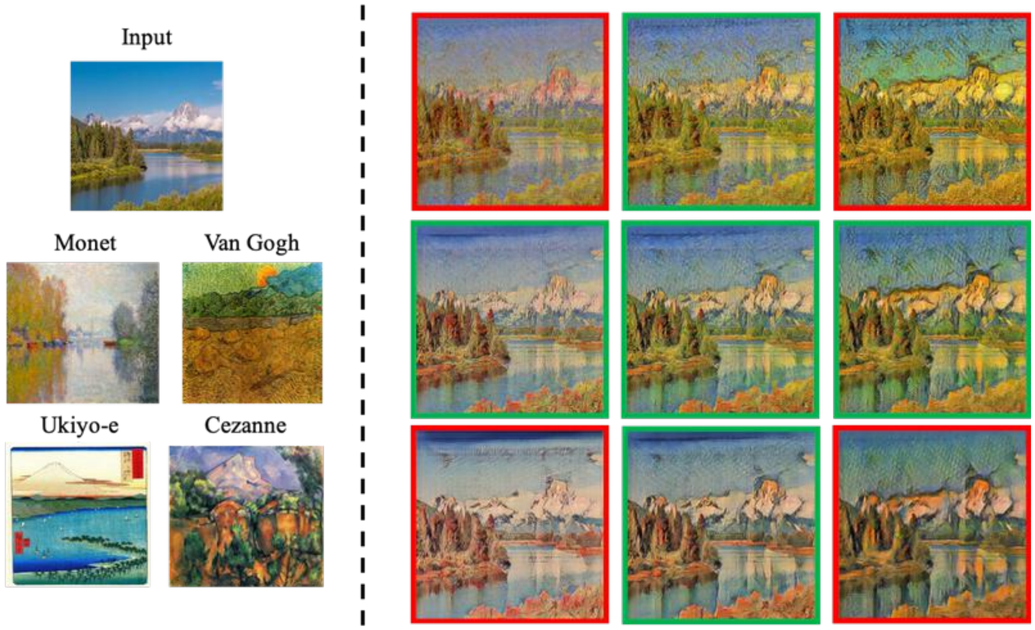


Figure 6: Examples of style generalization. Results with red rectangles at four corners are images translated into the four target domains, and those with green rectangles in between are images translated into intermediate domains. The results show that our DLOW model generalizes well across styles, and produces new images styles smoothly.

in the test phase, and the domainness vector can be changed instantly to generate different new styles of images. We provide more examples in Supplementary.

Quantitative Results: To verify the effectiveness of our model for style generalization, we conduct an user study on Amazon Mechanical Turk (AMT) to compare with the existing methods FadNet [32] and MUNIT [27]. Two cases are considered, style transfer to Van Gogh, and style generalization to mixed Van Gogh and Ukiyo-e. For FadNet, domain labels are treated as attributes. For MUNIT, we mix Van Gogh and Ukiyo-e as the target domain. The data for each trial is gathered from 10 participants and there are 100 trials in total for each case. For the first case, participants are shown the example Van Gogh style painting and are required to choose the image whose style is more similar to the example. For the second case, participants are shown the example Van Gogh and Ukiyo-e style painting and are required to choose the image with a style that is more like the mixed style of the two example paintings. The user preference is summarized in Table 3, which shows that DLOW outperforms FadNet and MUNIT on both tasks. Qualitative comparison between different methods is provided in Supplementary due to the space limitation.

	FadNet[32] / DLOW	MUNIT[27] / DLOW
Van Gogh	1.4% / 98.6%	21.4% / 78.6%
Van Gogh + Ukiyo-e	1.6% / 98.4%	15.3% / 84.7%

Table 3: User preference for style transfer and generalization. It is shown that more users prefer our translated results on both of the style transfer and generalization tasks compared with the existing methods FadNet and MUNIT.

5. Conclusion

In this paper, we have presented the DLOW model to generate intermediate domains for bridging different domains. The model takes a domainness variable z (or domainness vector $\mathbf{z}$) as the conditional input, and transfers images into the intermediate domain controlled by z or $\mathbf{z}$. We demonstrate the benefits of our DLOW model in two scenarios. Firstly, for the cross-domain semantic segmentation task, our DLOW model can improve the performance of the pixel-level domain adaptation by taking the translated images in intermediate domains as training data. Secondly, our DLOW model also exhibits excellent style generalization ability for image translation and we are able to transfer images into a new style that is unseen in the training data. Extensive experiments on benchmark datasets have verified the effectiveness of our proposed model.

Acknowledgments The authors gratefully acknowledge the support by armasuisse.

References

[1] Amjad Almahairi, Sai Rajeswar, Alessandro Sordoni, Philip Bachman, and Aaron Courville. Augmented CycleGAN: Learning many-to-many mappings from unpaired data. In *ICML*, 2018. 2, 3, 5, 7

[2] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan. *arXiv:1701.07875*, 2017. 4

[3] Mahsa Baktashmotlagh, Mehrtash T Harandi, Brian C Lovell, and Mathieu Salzmann. Unsupervised domain adaptation by domain invariant projection. In *ICCV*, 2013. 2

[4] 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, 2018. 7

[5] Yuhua Chen, Wen Li, Xiaoran Chen, and Luc Van Gool. Learning semantic segmentation from synthetic data: A geometrically guided input-output adaptation approach. *arXiv:1812.05040*, 2018. 2

[6] Yuhua Chen, Wen Li, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Domain adaptive faster r-cnn for object detection in the wild. In *CVPR*, 2018. 2

[7] Yuhua Chen, Wen Li, and Luc Van Gool. Road: Reality oriented adaptation for semantic segmentation of urban scenes. In *CVPR*, 2018. 2, 6

[8] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation. In *CVPR*, 2018. 1, 2, 3

[9] 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. 6, 11

[10] Zhen Cui, Wen Li, Dong Xu, Shiguang Shan, Xilin Chen, and Xuelong Li. Flowing on riemannian manifold: Domain adaptation by shifting covariance. *IEEE transactions on cybernetics*, 44(12):2264–2273, 2014. 2, 3

[11] Aysegul Dundar, Ming-Yu Liu, Ting-Chun Wang, John Zedlewski, and Jan Kautz. Domain stylization: A strong, simple baseline for synthetic to real image domain adaptation. *arXiv:1807.09384*, 2018. 2

[12] Basura Fernando, Amaury Habrard, Marc Sebban, and Tinne Tuytelaars. Unsupervised visual domain adaptation using subspace alignment. In *ICCV*, 2013. 2

[13] Yaroslav Ganin and Victor S. Lempitsky. Unsupervised domain adaptation by backpropagation. In *ICML*, 2015. 2, 3

[14] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, 2012. 6, 11

[15] Muhammad Ghifary, W Bastiaan Kleijn, Mengjie Zhang, and David Balduzzi. Domain generalization for object recognition with multi-task autoencoders. In *ICCV*, 2015. 1, 2

[16] Boqing Gong, Yuan Shi, Fei Sha, and Kristen Grauman. Geodesic flow kernel for unsupervised domain adaptation. In *CVPR*, 2012. 2, 3

[17] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*, 2014. 2

[18] Raghuraman Gopalan, Ruonan Li, and Rama Chellappa. Domain adaptation for object recognition: An unsupervised approach. In *ICCV*, 2011. 2, 3

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 7

[20] Zhenliang He, Wangmeng Zuo, Meina Kan, Shiguang Shan, and Xilin Chen. Arbitrary facial attribute editing: Only change what you want. *arXiv:1711.10678*, 2017. 2

[21] 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. 1, 2, 6, 7, 12

[22] Judy Hoffman, Dequan Wang, Fisher Yu, and Trevor Darrell. Fcns in the wild: Pixel-level adversarial and constraint-based adaptation. *arXiv:1612.02649*, 2016. 2, 6

[23] Weixiang Hong, Zhenzhen Wang, Ming Yang, and Junsong Yuan. Conditional generative adversarial network for structured domain adaptation. In *CVPR*, 2018. 2

[24] Yann N. Dauphin David Lopez-Paz Hongyi Zhang, Moustapha Cisse. mixup: Beyond empirical risk minimization. In *ICLR*, 2018. 5

[25] Haoshuo Huang, Qixing Huang, and Philipp Krähenbühl. Domain transfer through deep activation matching. In *ECCV*, 2018. 2

[26] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*, 2017. 5, 7

[27] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *ECCV*, 2018. 2, 3, 8, 12, 13

[28] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017. 2

[29] I-Hong Jhuo, Dong Liu, DT Lee, and Shih-Fu Chang. Robust visual domain adaptation with low-rank reconstruction. In *CVPR*, 2012. 2

[30] Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jung Kwon Lee, and Jiwon Kim. Learning to discover cross-domain relations with generative adversarial networks. In *ICML*, 2017. 2

[31] Elyor Kodirov, Tao Xiang, Zhenyong Fu, and Shaogang Gong. Unsupervised domain adaptation for zero-shot learning. In *ICCV*, 2015. 2, 3

[32] Guillaume Lample, Neil Zeghidour, Nicolas Usunier, Antoine Bordes, Ludovic DENOYER, et al. Fader networks: Manipulating images by sliding attributes. In *NIPS*, 2017. 2, 8, 12, 13

[33] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Kumar Singh, and Ming-Hsuan Yang. Diverse image-to-image translation via disentangled representations. In *ECCV*, 2018. 2

[34] Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C Kot. Domain generalization with adversarial feature learning. In *CVPR*, 2018. 2

[35] Wen Li, Zheng Xu, Dong Xu, Dengxin Dai, and Luc Van Gool. Domain generalization and adaptation using low rank exemplar svms. *IEEE transactions on pattern analysis and machine intelligence*, 40(5):1114–1127, 2018. 2

[36] Ya Li, Xinmei Tian, Mingming Gong, Yajing Liu, Tongliang Liu, Kun Zhang, and Dacheng Tao. Deep domain generalization via conditional invariant adversarial networks. In *ECCV*, 2018. 2

[37] Jianxin Lin, Yingce Xia, Tao Qin, Zhibo Chen, and Tie-Yan Liu. Conditional image-to-image translation. In *CVPR*, 2018. 2

[38] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. In *NIPS*, 2017. 2, 3

[39] Yongyi Lu, Yu-Wing Tai, and Chi-Keung Tang. Conditional cyclegan for attribute guided face image generation. *arXiv:1705.09966*, 2017. 2

[40] Yawei Luo, Liang Zheng, Tao Guan, Junqing Yu, and Yi Yang. Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation. *arXiv:1809.09478*, 2018. 2

[41] Saeid Motiian, Marco Piccirilli, Donald A Adjeroh, and Gianfranco Doretto. Unified deep supervised domain adaptation and generalization. In *ICCV*, 2017. 2

[42] Krikamol Muandet, David Balduzzi, and Bernhard Schlkopf. Domain generalization via invariant feature representation. In *ICML*, 2013. 2

[43] Zak Murez, Soheil Kolouri, David Kriegman, Ravi Ramamoorthi, and Kyungnam Kim. Image to image translation for domain adaptation. *arXiv:1712.00479*, 2017. 2

[44] Li Niu, Wen Li, and Dong Xu. Multi-view domain generalization for visual recognition. In *ICCV*, 2015. 2

[45] Li Niu, Wen Li, and Dong Xu. Visual recognition by learning from web data: A weakly supervised domain generalization approach. In *CVPR*, 2015. 2

[46] Xingang Pan, Ping Luo, Jianping Shi, and Xiaoou Tang. Two at once: Enhancing learning and generalization capacities via ibn-net. In *ECCV*, 2018. 2

[47] Xingchao Peng, Ben Usman, Neela Kaushik, Dequan Wang, Judy Hoffman, Kate Saenko, Xavier Roynard, Jean-Emmanuel Deschaud, Francois Goulette, Tyler L Hayes, et al. Visda: A synthetic-to-real benchmark for visual domain adaptation. In *CVPR Workshops*, 2018. 2

[48] Stephan R. Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *ECCV*, 2016. 6

[49] 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. 6, 11

[50] Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. Maximum classifier discrepancy for unsupervised domain adaptation. *arXiv:1712.02560*, 2017. 2

[51] Fatemeh Sadat Saleh, Mohammad Sadegh Aliakbarian, Mathieu Salzmann, Lars Petersson, and Jose M Alvarez. Effective use of synthetic data for urban scene semantic segmentation. In *ECCV*, 2018. 2

[52] Swami Sankaranarayanan, Yogesh Balaji, Arpit Jain, Ser Nam Lim, and Rama Chellappa. Unsupervised domain adaptation for semantic segmentation with gans. *arXiv preprint arXiv:1711.06969*, 2017. 1, 2

[53] Swami Sankaranarayanan, Yogesh Balaji, Arpit Jain, Ser Nam Lim, and Rama Chellappa. Learning from synthetic data: Addressing domain shift for semantic segmentation. In *CVPR*, 2018. 2

[54] 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, 6, 7, 11, 12

[55] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *CVPR*, 2017. 2

[56] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *CVPR*, 2018. 2

[57] Markus Wulfmeier, Alex Bewley, and Ingmar Posner. Incremental adversarial domain adaptation for continually changing environments. In *ICRA*, 2018. 2

[58] Zili Yi, Hao (Richard) Zhang, Ping Tan, and Minglun Gong. Dualgan: Unsupervised dual learning for image-to-image translation. In *ICCV*, 2017. 2

[59] 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. 6, 11

[60] Oliver Zendel, Katrin Honauer, Markus Murschitz, Daniel Steininger, and Gustavo Fernandez Dominguez. Wilddash - creating hazard-aware benchmarks. In *ECCV*, 2018. 6, 11

[61] Weichen Zhang, Wanli Ouyang, Wen Li, and Dong Xu. Collaborative and adversarial network for unsupervised domain adaptation. In *CVPR*, 2018. 2

[62] Yiheng Zhang, Zhaofan Qiu, Ting Yao, Dong Liu, and Tao Mei. Fully convolutional adaptation networks for semantic segmentation. In *CVPR*, 2018. 2

[63] Yexun Zhang, Ya Zhang, and Wenbin Cai. A unified framework for generalizable style transfer: Style and content separation. *arXiv:1806.05173*, 2018. 7

[64] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, 2017. 1, 2, 3, 7

[65] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. Toward multimodal image-to-image translation. In *NIPS*, 2017. 2, 3

[66] Xinge Zhu, Hui Zhou, Ceyuan Yang, Jianping Shi, and Dahua Lin. Penalizing top performers: Conservative loss for semantic segmentation adaptation. *arXiv:1809.00903*, 2018. 2

[67] Yang Zou, Zhiding Yu, BVK Vijaya Kumar, and Jinsong Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *ECCV*, 2018. 2, 6

6. Supplementary

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.

6.1. 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 8, 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.

6.2. 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 [49] is used as the source domain while the Cityscapes dataset [9] 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 [14], WildDash [60] and BDD100K [59] 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.

Table 4: Comparison of the performance of AdaptSegNet [54] 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 [54]	46.7	33.3	20.6	30.8
Original* [54]	45.7	**34.4**	20.0	30.8
DLOW	**47.1**	**34.4**	**24.4**	**35.3**

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.

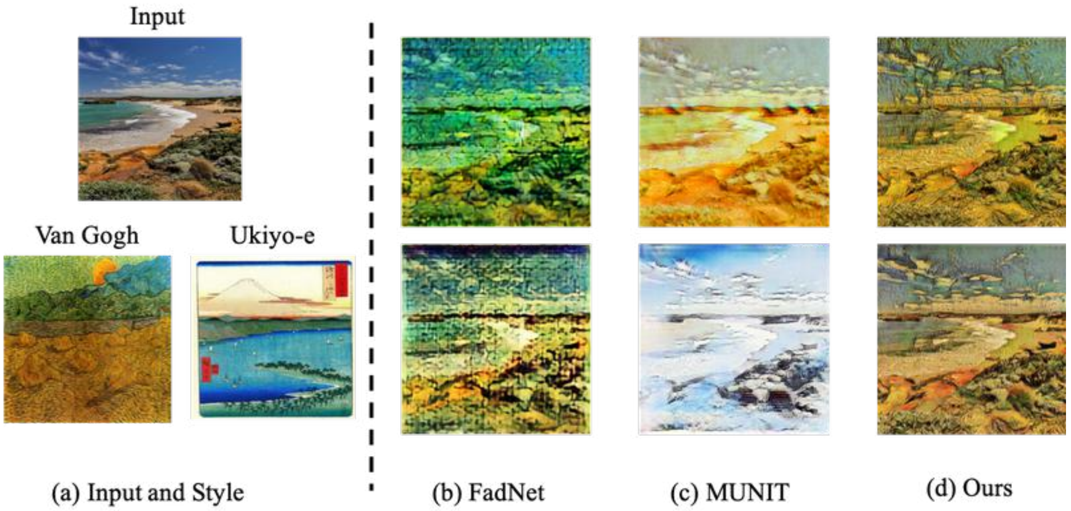
Similar to GTA5, our DLOW model based on SYNTHIA dataset also exhibits excellent performance for the domain adaptation and the domain generalization. Following [54], the segmentation performance based on SYNTHIA dataset is tested on the Cityscapes validation dataset with 13 classes. As shown in Table 5, 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 4 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 4 denotes our retrained multi-level AdaptSegNet model in [54]. 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.

6.3. 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 9, 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 9a, the style generalization model consists of two modules, the adversarial module and the image reconstruction module. For

Table 5: 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[54]	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[21]	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**



(a) Input and Style (b) FadNet (c) MUNIT (d) Ours

Figure 7: 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 [32], MUNIT [27] 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.

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 $\mathbf{z} = [z_1, \ldots, z_4]'$. For the other direction $\mathcal{T} \rightarrow \mathcal{S}$, shown in Fig 9b, 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 $\mathbf{z}$.

6.4. 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 10 and Fig 11. 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.

6.5. 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 FadNet [32] and the MUNIT [27] on the style transfer and the style generalization task. In this Supplementary, we further provide the qualitative result comparison in Fig 7. 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 7, 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 7, 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.

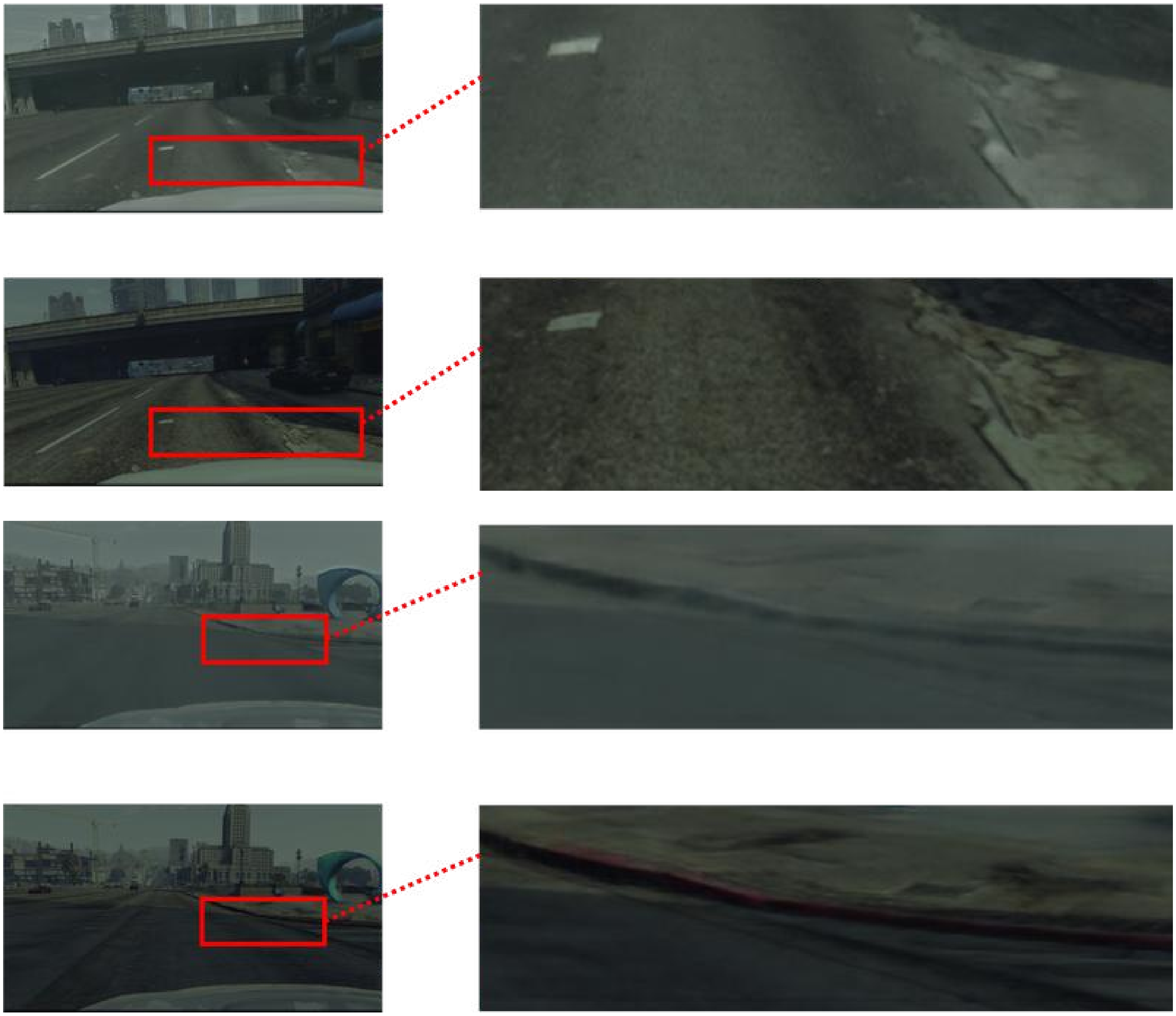
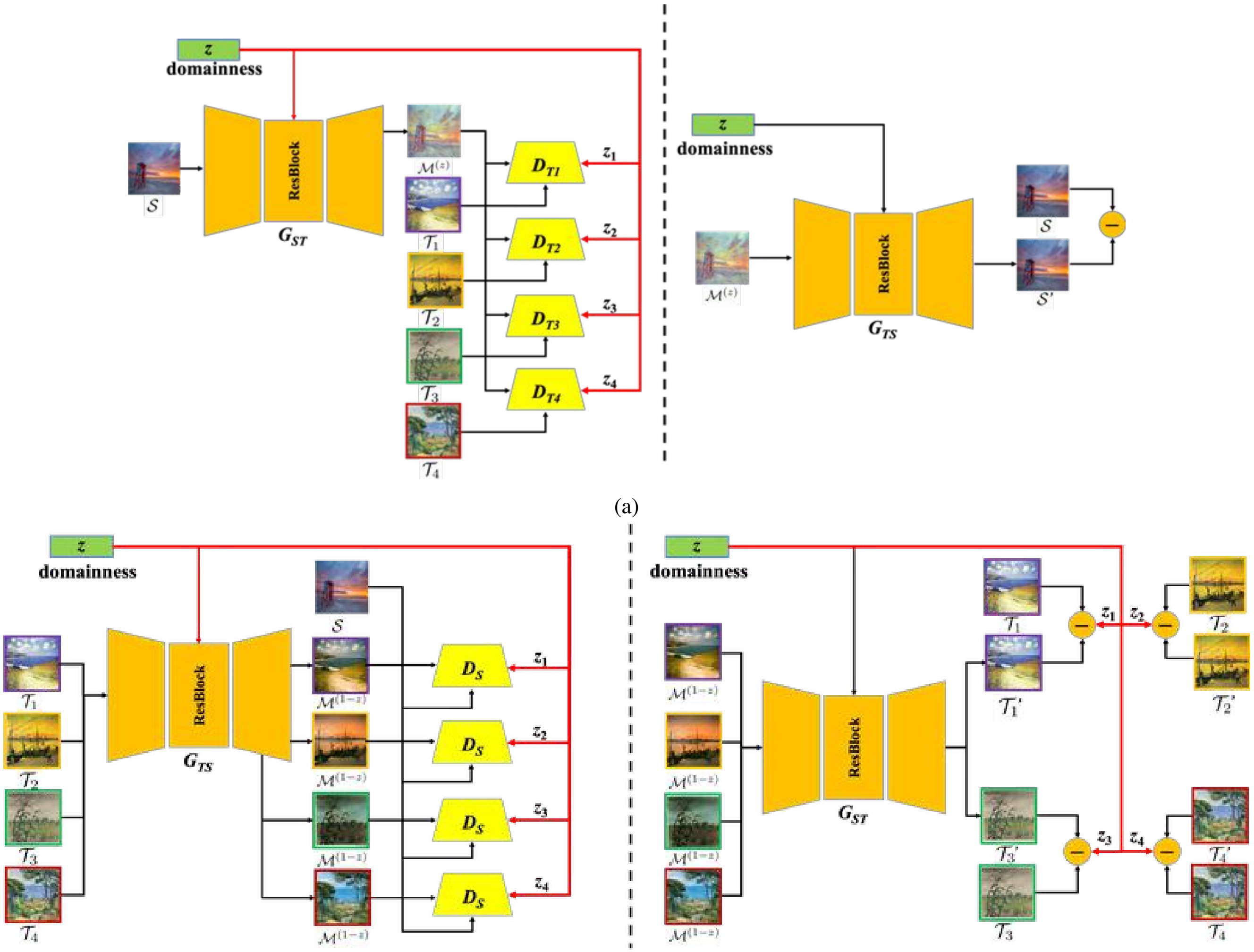


Figure 8: Examples of comparison between the DLOW translated image and the brightness adjusted image. We adjust the brightness of the DLOW translated source image($z = 0$) to make its brightness match the corresponding DLOW translated target image($z = 1$). The lower one in each group is the brightness 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.



(a)

(b)

Figure 9: Network structure of DLOW model for style generalization with four target domains: (a) direction from $\mathcal{S} \rightarrow \mathcal{T}$; (b) direction from $\mathcal{T} \rightarrow \mathcal{S}$.

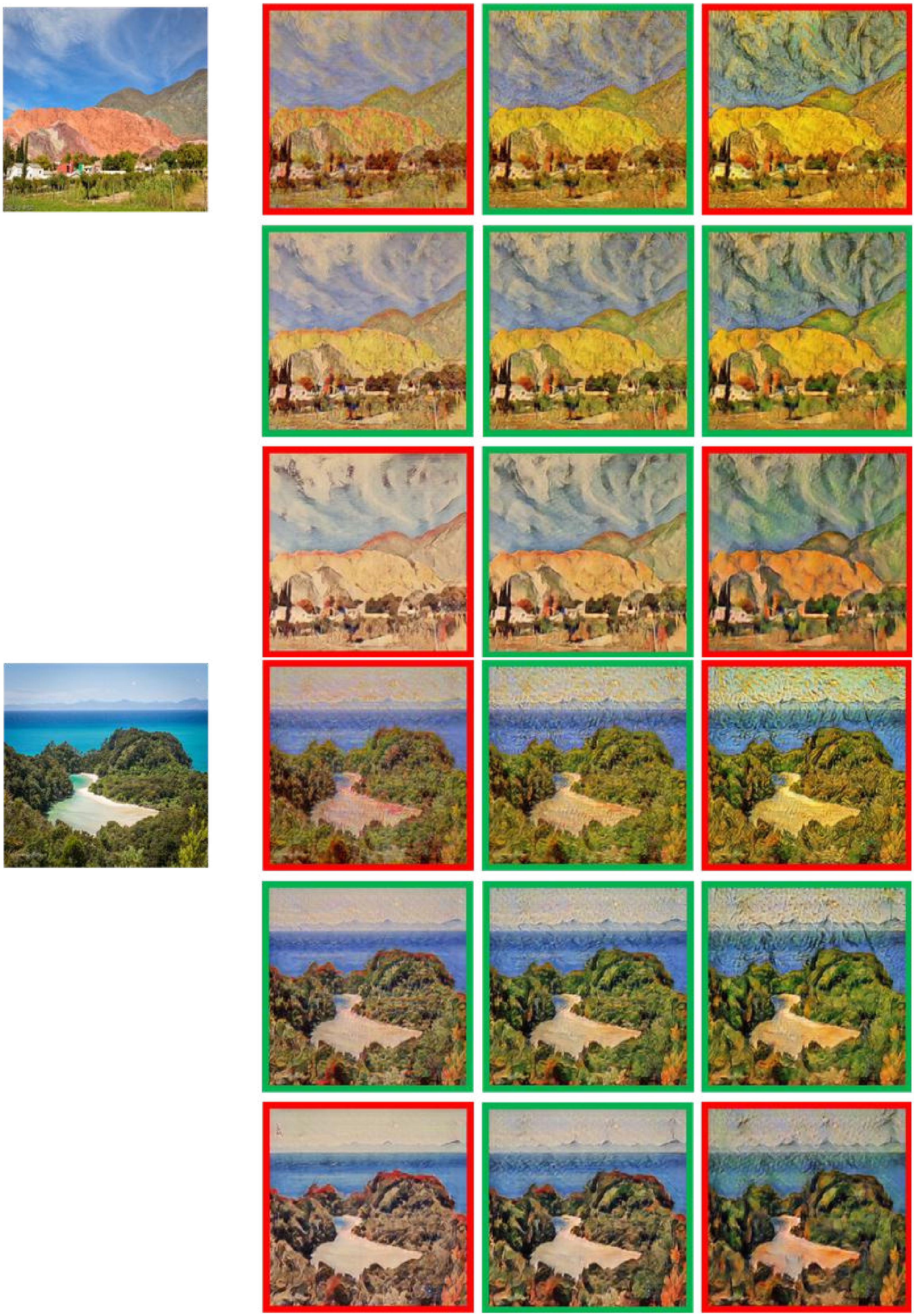


Figure 10: Examples of style generalization I. Results with red rectangles at four corners are images translated into the four target domains, and those with green rectangles in between are images translated into intermediate domains. The results show that our DLOW model generalizes well across styles, and produces new images styles smoothly.

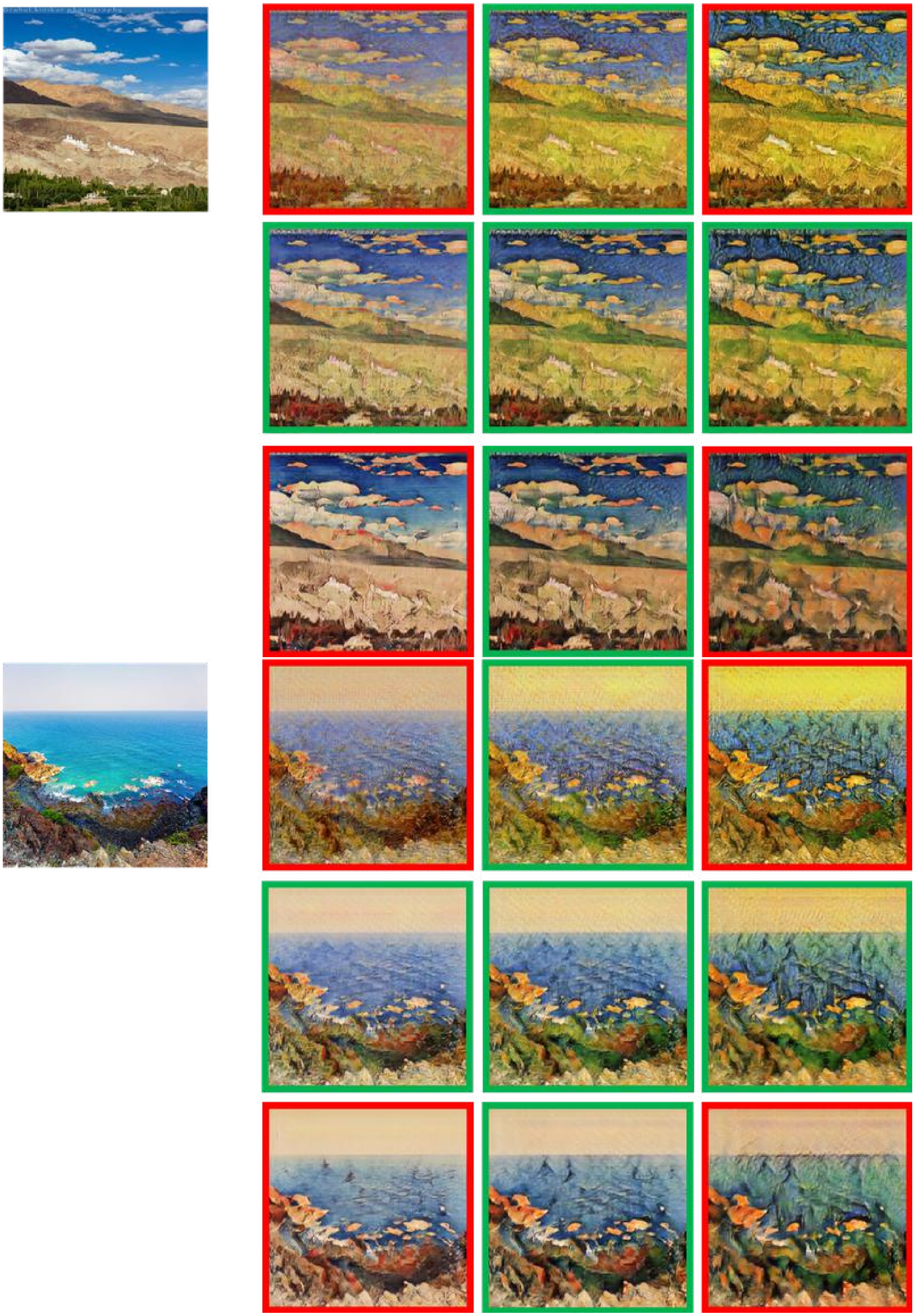


Figure 11: Examples of style generalization II. Results with red rectangles at four corners are images translated into the four target domains, and those with green rectangles in between are images translated into intermediate domains. The results show that our DLOW model generalizes well across styles, and produces new images styles smoothly.