Learning Semantic Segmentation from Synthetic Data: A Geometrically Guided Input-Output Adaptation Approach

Yuhua Chen[1] Wen Li[1*] Xiaoran Chen[1] Luc Van Gool[1,2]

[1]Computer Vision Laboratory, ETH Zurich [2]VISICS, ESAT/PSI, KU Leuven

{yuhua.chen,liwen,chenx,vangool}@vision.ee.ethz.ch

Abstract

As an alternative to manual pixel-wise annotation, synthetic data has been increasingly used for training semantic segmentation models. Such synthetic images and semantic labels can be easily generated from virtual 3D environments. In this work, we propose an approach to cross-domain semantic segmentation with the auxiliary geometric information, which can also be easily obtained from virtual environments. The geometric information is utilized on two levels for reducing domain shift: on the input level, we augment the standard image translation network with the geometric information to translate synthetic images into realistic style; on the output level, we build a task network which simultaneously performs semantic segmentation and depth estimation. Meanwhile, adversarial training is applied on the joint output space to preserve the correlation between semantics and depth. The proposed approach is validated on two pairs of synthetic to real dataset: Virtual KITTI→KITTI, and SYNTHIA→Cityscapes, where we achieve a clear performance gain compared to the baselines and various competing methods, demonstrating the effectiveness of the geometric information for cross-domain semantic segmentation. Our implementation is available at http://github.com/yuhuayc/gio-ada.

1. Introduction

Semantic segmentation refers to the task of classifying each pixel in a given image to its semantic category, *e.g. sky*, *road*, *car*. The task provides pixel-wise semantic understanding of scenes, and leads to many attractive applications such as robotics, autonomous driving *etc*. Like many other visual perception tasks, deep neural networks [28] excel at semantic segmentation when trained on large-scale labeled datasets. However, building such labeled datasets for semantic segmentation is not an easy task, in terms of both collecting and annotating: it is non-trivial to collect images with large diversity of scenes and conditions; annotating them can be even more costly due to the process of acquiring pixel-wise labels.

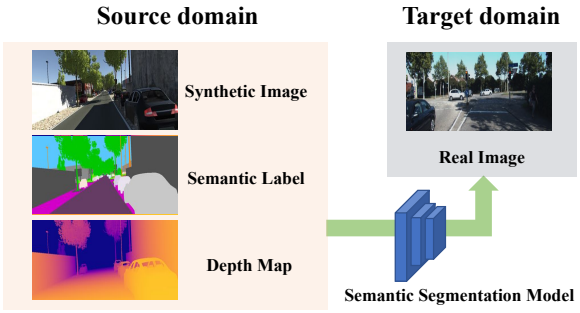

Figure 1. We aim to adapt a semantic segmentation model learned from synthetic data to real data. The semantic label is only available in the source domain(synthetic data). Only unlabeled images are given in the target domain(real data). The domain adaptation process is strengthened by auxiliary geometric information in synthetic data, which can be obtained easily from virtual environment.

To address these bottlenecks, synthetic data becomes a charming alternative to supervise semantic segmentation models. Recent advances in computer graphics make it possible to automatically generate synthetic images, with the corresponding per-pixel labels from virtual 3D environments [46, 45]. Training on synthetic data seems to be a tempting way to reduce annotation cost, however, the mismatch in appearance often leads to a significant performance drop when the learned models are applied to real data.

Many works have been proposed to tackle this issue from the domain distribution shift perspective using various domain adaptation techniques [21, 57, 5]. On the other hand, image translation techniques [60] have also been widely used to transform synthetic images into realistic style. This can be seen as aligning the domain distribution at pixel level [20]. Nevertheless, these works typically utilize only synthetic images and the corresponding semantic labels. However, a significant advantage of synthetic data has been

*corresponding author.

unfortunately overlooked: one can actually obtain rich information from virtual environments, such as depth, surface norm, optical flow, *etc*., at a much lower cost than obtaining the information of the same kind in the real world.

As illustrated in Figure 1, the aim of this work is to exploit the supplementary geometric information from the synthetic domain for improving cross-domain semantic segmentation in real data. We are motivated by the fact that geometry and semantics are naturally coupled. Geometric cue can usually imply semantics and vice verse. As shown in previous works [54], joint reasoning the semantics and depth improves the performance of both tasks. Moreover, unlike the large gap between synthetic and real images, the correlation between depth and semantics is more domain-invariant and suffers less from domain shift. For example, a road is usually flat, the sky is far away, poles are vertical. These facts hold regardless in synthetic data and real data. Thus, the correlation between semantics and geometry is highly favoured for reducing the domain gap. Besides, depth information is relatively easy to acquire from synthetic data, as one can simply generate the depth from virtual 3D environments, and no special equipment (*e.g*. Lidar, calibrated stereo cameras) is needed.

We present a new approach called *Geometrically Guided Input-Output Adaptation* (GIO-Ada), in which we leverage depth information for the domain adaptation task on two levels: 1) on the input level, an augmented image transform network takes synthetic image and its corresponding semantic and depth map as input, and is trained to produce images with realistic style by exploiting the intrinsic connection between raw images, semantic and geometric information; and 2) on the output level, a task network jointly performs semantic segmentation and depth estimation using supervision from synthetic domain. Further, adversarial training is applied on the joint output space of semantic segmentation and depth estimation, thus preserving domain-invariant correlation between semantics and depth. With the aforementioned modules, geometric information not only improves the semantic segmentation, but also helps to alleviate the domain gap between synthetic and real data.

The proposed approach is validated through extensive experiments on Virtual KITTI [11], KITTI [13], SYNTHIA [47], and Cityscapes [7] datasets, where we achieve significant performance improvement over the non-adaptive baseline and competing methods that don't leverage geometric information. The experiments demonstrate that our approach can improve cross-domain semantic segmentation by incorporating geometric information from synthetic data.

2. Related Works

Semantic Segmentation is a highly active research field. Recent approaches are mostly based on fully convolutional network [35], with modifications for pixel-wise prediction, such as DilatedNet [56], DeepLab [3], PSPNet [59] *etc*. Such models are generally trained on datasets with pixel-wise annotation, *e.g*., PASCAL [9], COCO [34], and Cityscapes [7]. However, building such labeled datasets is expensive and laborious. With the development of computer graphics techniques, synthetic data enables an alternative approach to training semantic segmentation models at a lower cost. To this end, several synthetic datasets have been built, for example, GTAV [46], SYNTHIA [47], Virtual KITTI [11], *etc*. These datasets are typically generated from virtual 3D environments, meaning that modalities other than the semantic label can be generated easily as well. Such modalities include optical flow, depth, surface normal *etc*. Our work is motivated to leverage such free supervision in synthetic data in order to effectively perform cross-domain semantic segmentation.

Domain Adaptation is a classic problem in machine learning and computer vision. It aims to mitigate the performance drop caused by the distribution mismatch between training and test data. It is mostly studied in image recognition problems by both conventional approaches [29, 18, 15, 10, 33] and CNN-based approaches [36, 12, 14, 50, 42, 39, 31, 19, 37, 38]. We refer to [43, 8] for comprehensive surveys. Besides image classification, domain adaptation has been studied in other vision tasks such as object detection [4], depth estimation [1] *etc*.

Our work is mostly related to cross-domain semantic segmentation [21, 57, 20, 53, 48, 5, 62, 61]. Hoffman *et al.* [21] propose to improve the cross-domain semantic segmentation, by aligning the features from two domains with adversarial training. Following this line, many works have been proposed to address the domain shift problem in semantic segmentation using different techniques, such as curriculum style learning [57], distillation loss [5], output space alignment [53], class-balanced self-training [62], conservative loss [61], *etc*. Moreover, inspired by the success of generative adversarial network [44, 17] and image translation techniques [60, 22], a few works have also suggested to transform synthetic images to realistic style, thus reducing domain gap on raw-pixel level [51, 20, 41, 20, 48, 16] and to boost the semantic segmentation performance in real world.

The aforementioned works typically only leverage labeled source images and unlabeled target images while neglecting other information in the dataset, such as geometric information. In this work we take advantage of privilege depth information in the target domain. Similar idea has been exploited for image recognition [32, 2], and by a concurrent work [30] for semantic segmentation.

Depth Aided Semantic Segmentation Depth estimation and semantic segmentation are two fundamental tasks of scene understanding. Many works have been proposed

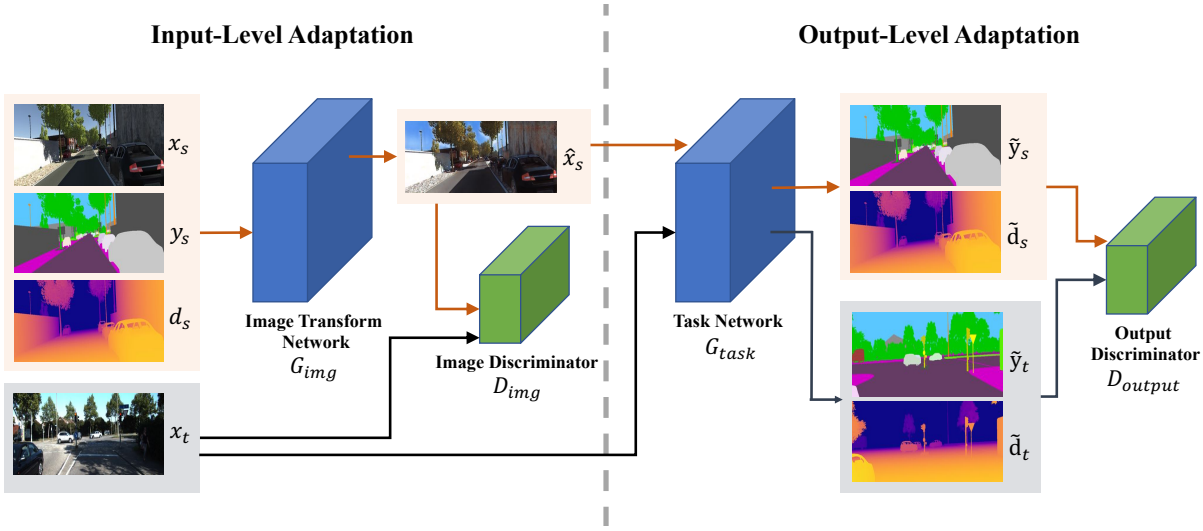

Figure 2. **Overview of the proposed architecture.** The flow of source data is shown in orange line, while the flow of target data is shown in black line. The image transform network transforms synthetic images to realistic style, and the task network is used to perform semantic segmentation and depth estimation simultaneously.

to jointly learn the two tasks in a mutually beneficial manner. Wang *et al.* [54] build a hierarchical CRF with CNN to leverage the geometric cue, and Kendall *et al.* [26] propose a cross-task uncertainty to weight losses for the two tasks. Besides, various techniques have been used for the task, including fine-tuning [40], cross-modality influence [23], task distillation module with intermediate auxiliary tasks [55], recursive estimation [58], task attention loss [24]. More broadly speaking, it can be related to multi-tasking [27]. In this work, the correlation between semantics and depth is also leveraged, but for the purpose of domain adaptation.

3. Methodology

In this section, we present our approach to learning semantic segmentation models from synthetic data, with the aid of depth information. Following unsupervised domain adaptation protocol, synthetic data is utilized as the source domain S, and real data as the target domain T. In the source domain, we have access to synthetic images $x_s \in S$ along with their corresponding ground-truth labels, including semantic segmentation labels y_s and depth labels d_s. In the target domain, only unlabeled images $x_t \in T$ are available.

3.1. Overview of the Proposed Approach

The overview of our proposed Geometrically Guided Input-Output Adaptation (GIO-Ada) approach is illustrated in Figure 2. To address the domain gap between synthetic and real domains, domain adaptation is performed jointly on two levels, namely input level and output level. Depth information (*i.e.*, geometric information) is exploited for improving adaptation on both levels.

Input-level adaptation aims to reduce visual differences at raw pixel level. The output from input-level adaptation are later used as input to the following task network. For this purpose, we deploy an image transform network G_{img} which takes a synthetic image x_s, along with its corresponding depth d_s and semantic labels y_s as input. The transform network G_{img} is supposed to produce transformed images $\hat{x}_s$ with visually similar appearances to the images in the target domain, and at the same time preserves useful information for semantic segmentation and depth estimation.

Most of the existing pixel-level adaptation methods do not consider the depth information of the source domain. This is apparently not optimal for several reasons: geometric information becomes more difficult to recover once discarded in the rendering process. On the other hand, geometric information is highly correlated with semantic information. Due to these reasons, we use the depth information as an auxiliary input of the image transform network to better preserve information during image translation.

Output-level adaptation aims to align the outputs of the task network for two domains, and also retain the coherent correlation between tasks. The output-level adaptation includes a task network G_{task} and an output-level discriminator D_{output}. G_{task} takes real images x_t or transformed synthetic images $\hat{x}_s$ as input, then simultaneously predicts semantic segmentation $\widetilde{y}$ and depth prediction $\widetilde{d}$. D_{output} tries to determine if the outputs (semantics and depth) are predicted from a transformed synthetic image or a real one.

Utilizing geometric information in output-level adaptation brings several benefits. First, by learning depth estimation as an auxiliary task, we can learn representation which

is more robust against domain shift. Second, the correlation between semantics and depth can be used as a powerful cue for domain alignment. Since no ground-truth label is given in the target domain, aligning the output space between the two domains can be a highly useful supervision signal to guide the training. Unlike the previous work [53] which only aligns the output space of a single task, here we consider the joint output space of semantic segmentation and depth estimation. In this way, we align not only the output distributions of each individual task, but also the underlying interconnection between different tasks. This is proven to be effective for boosting the performance of the two tasks. It is also consistent with our motivation that such connections suffer less from domain shift, for instance, the sky is always far away, cars are usually on the street *etc*. Hereby, we respectively elaborate the adaptation on the two levels in the following sections.

3.2. Input-Level Adaptation

To transform synthetic images into the real-style images, we build an image transform network G_{img} with synthetic image x_s, semantic segmentation label y_s and depth map d_s as input. In particular, the depth map is normalized into a range of $[0, 1]$ among all images in the dataset, and the semantic label is represented as a one-hot map of C channels where C is the total number of categories. The network produces the transformed image $\hat{x}_s = G_{img}(x_s, y_s, d_s)$, which is expected to be realistic-looking and still contains vital information for the task networks (*e.g.*, semantic segmentation, depth estimation.).

Inspired by recent works on generative adversarial networks (GANs) [17], we apply a discriminator D_{img} to guarantee the realism of generated images. The discriminator D_{img} is trained to distinguish between transformed synthetic images and real images. At the same time, D_{img} is also used to guide the training of the image transform network in a similar way to the adversarial training strategy in GANs. Similar with previous works [20, 60], we use PatchGAN [22] to operate on patches, from which we obtain the discriminator output in the form of a two-dimensional map. The loss for training D_{img} can be written as follows:

$$\begin{aligned}\mathcal{L}_{input} = &\mathbb{E}_{x_t \sim X_T}\left[\log D_{img}(x_t)\right] \\ &+\mathbb{E}_{x_s \sim X_S}\left[\log(1 - D_{img}(\hat{x}_s))\right],\end{aligned} \tag{1}$$

in which we omit the image width and height dimension for simplicity.

As mentioned above, the transformed images are expected to be useful for the tasks at hand. This is achieved by joint training the image transform network with the task network (details are provided in the next section). Since the image transform network is differentiable, the gradients from the task network can guide the transform network to ensure the preservation of useful information from synthetic data.

3.3. Output-Level Adaptation

Our task network G_{task} concurrently performs semantic segmentation and depth estimation for a given input image. The network is shared between two domains and takes either a transformed synthetic image $\hat{x}_s$ or a real image x_t as input. Specifically, a feature extractor is shared between the two tasks, with two decoders on top of it respectively for each task, namely one decoder for semantic segmentation output and the other one for depth estimation output.

The semantic segmentation task is learned by minimizing a standard cross-entropy loss:

$$\mathcal{L}_{seg} = \mathbb{E}_{x_s \sim X_S}[CE(y_s, \widetilde{y}_s)], \tag{2}$$

where y_s stands for ground-truth semantic labels, and $\widetilde{y}_s$ stands for predicted labels. With regard to the semantic segmentation task, depth estimation can be seen as an auxiliary task. As a common practice, we deploy the ℓ_1 loss for the depth estimation task as follows:

$$\mathcal{L}_{depth} = \mathbb{E}_{x_s \sim X_S}[||d_s - \widetilde{d}_s||_1], \tag{3}$$

where d_s stands for ground-truth depth, and $\widetilde{d}_s$ stands for predicted depth. Note that both losses only apply to the source domain, where supervision is available.

To ensure that the task network performs well in the target domain, we further apply a discriminator D_{task} on the outputs as inspired by [53]. However, instead of using only the semantic segmentation output, our work jointly considers both semantics and depth, as the inherent correlation between semantics and depth information could be a helpful cue to effectively reduce domain difference. In particular, we concatenate the output of semantic segmentation prediction $\widetilde{y}_s$ (*resp.* $\widetilde{y}_t$) and the output of depth estimation map $\widetilde{d}_s$ (*resp.* $\widetilde{d}_t$), which leads to a total of $C + 1$ channels in the concatenated output. We use the concatenated output to train the discriminator D_{task} which distinguishes outputs of the source domain from those of the target domain. Similar to D_{img}, D_{task} is also formulated as a PatchGAN in favour of its awareness of spatial contextual relations. The loss for D_{task} can be written as follows:

$$\begin{aligned}\mathcal{L}_{output} = &\mathbb{E}_{x_t \sim X_T}\left[\log D_{output}(\widetilde{d}_t, \widetilde{y}_t)\right] \\ &+\mathbb{E}_{x_s \sim X_S}\left[\log(1 - D_{output}(\widetilde{d}_s, \widetilde{y}_s))\right].\end{aligned} \tag{4}$$

3.4. Overall Training Objective

Bring together the input-level and the output-level modules, we jointly train all networks G_{task}, G_{img}, D_{img} and D_{output}. The overall objective is written as follows:

	road	building	pole	traffic light	traffic sign	vegetation	terrain	sky	car	truck	mIoU
non-adapt	79.3	60.5	0.0	0.3	9.5	66.8	8.3	85.9	59.2	4.8	37.5
input-level adapt	83.2	67.4	10.8	21.9	24.5	68.8	6.5	88.3	77.8	9.3	45.9
output-level adapt	81.1	69.1	7.1	8.6	28.3	79.5	43.3	86.0	79.3	17.8	50.0
GIO-Ada	81.4	71.2	11.3	26.6	23.6	82.8	56.5	88.4	80.1	12.7	**53.5**

Table 1. **Quantitative results on Virtual KITTI→KITTI.** The results are reported using mIoU over 10 categories. The best result is denoted in bold.

$$\min_{\substack{G_{img}\\G_{task}}} \max_{\substack{D_{img}\\D_{output}}} \{\mathcal{L}_{seg} + \lambda_{depth}\mathcal{L}_{depth} + \lambda_{input}\mathcal{L}_{input} + \lambda_{output}\mathcal{L}_{output}\}, \tag{5}$$

where λs act as the trade-off weights to balance different loss terms. The min-max problem is optimized with the adversarial training strategy. Note that domain adaptation procedure is only performed in the training phase. During test time, only G_{task} is applied on real images, and other components such as G_{img}, D_{img} and D_{output} are removed during inference.

3.5. Implementation Details

In our GIO-Ada approach, the image transform network G_{img} resembles the generator in CycleGAN [60], which is based on the network in [25] with several convolutional layers and residual blocks. For the task network, we deploy similar architecture with DeepLab-v2 model [3] with VGG backbone [52]. In more details, on the top of shared VGG encoder, we build two separate decoders: one for depth estimation, and the other for semantic segmentation. ASPP module from DeepLab v2 are used in both decoders, where the only difference is the number of output channel. The task network is initialized with the ImageNet pre-trained weights. Moreover, the discriminators are based on PatchGAN [22], for which the weights are randomly initialized from a Gaussian distribution.

In the training, the trade-off parameters are set as $\lambda_{depth} = 0.1, \lambda_{input} = 0.1, \lambda_{output} = 0.001$. Each mini-batch contains two images, one from the source domain and the other sampled from the target domain. Random horizontal flip is used for data augmentation. We use Adam optimizer with an initial learning rate of 2×10^{-4}. The network is trained for 10 epochs.

4. Experiments

In this section, we verify the effectiveness of our proposed GIO-Ada approach for semantic segmentation from synthetic data to real scenarios.

4.1. Experiment Settings

Following the common unsupervised domain adaptation protocol, we use a synthetic dataset as the source domain, and a real dataset as the target domain. For the synthetic datasets, we use Virtual KITTI [11] and SYNTHIA [47], as depth information is available for these two datasets. Accordingly, KITTI [13] and Cityscapes [7] are used as the real datasets, which results in two adaptation pairs: Virtual KITTI→KITTI, and SYNTHIA→Cityscapes. We briefly introduce the datasets used in our experiments as below.

KITTI [13] is a dataset on autonomous driving, which consists of images depicting several driving urban scenarios. It is collected by moving vehicles in multiple cities. The official split for semantic segmentation is used in our experiment, which contains 200 training images, and 200 test images. The images have a spatial resolution around 1242×375. As the ground-truth label is only available in the training set, thus we use the official unlabeled test images to adapt our model, and we report the results on the official training set.

Virtual KITTI [11] is a photo-realistic synthetic dataset which contains 21,260 images. Each image is densely annotated at pixel level with category and depth information. It is designed to mimic the conditions of KITTI dataset and has similar scene layout, camera viewpoint, and image resolution as KITTI dataset, thus making it ideal to study the domain adaptation problems between synthetic and real data.

Cityscapes [7] consists of $2,975$ images in the training set, and 500 images in the validation set. The images have a fixed spatial resolution of 2048×1024 pixels. Due to the large size of image, as a common practice we down-size the images to half resolution (at 1024×512 pixels). The training set is used to adapt the model, and we report our results on the validation set.

SYNTHIA [47] is a dataset with synthetic images of urban scenes and pixel-wise annotations. The render-

na	cg	gd	+d	+s	+sd
37.5	39.8	43.5	44.2	44.7	**45.9**

Table 2. **Ablation study on input-level adaptation.** mIoU over 10 categories is reported. **na**: non-adaptive baseline; **cg**: image translation with CycleGAN [60]; **gd**: image transform network is guided by the task network; **+d**: with additional depth input to the image transform network; **+s**: with additional semantic label input; **+sd**: with both semantic and depth labels as additional input, which is also our final model for input-level adaptation.

na	ss	depth	sep	joint
37.5	45.9	43.8	46.3	**50.0**

Table 3. **Ablation study on output-level adaptation.** mIoU over 10 categories is reported. **na**: the non-adaptive baseline; **ss**: aligning the semantic segmentation output; **depth**: aligning the depth estimation output; **sep**: individually aligning outputs of semantic segmentation and depth estimation; **joint**: aligning the joint output space of semantic segmentation and depth estimation, which is also our final model for output-level adaptation.

ing covers a variety of environments and weather conditions. In our experiment, we adopt the SYNTHIA-RAND-CITYSCAPES subset, which contains 9,400 images compatible with the Cityscapes categories.

4.2. Results on Virtual KITTI→KITTI

We first evaluate the proposed method for learning semantic segmentation from the Virtual KITTI dataset to the KITTI dataset. The 10 common categories between two datasets are used for performance evaluation. We report the results using mean of Intersection over Union (mIoU), which are summarized in Table 1. Overall, our GIO-Ada improves the mIoU over the non-adaptive baseline by $+16\%$, which confirms the effectiveness of our method for cross-domain semantic segmentation. To further study the benefits of the adaptation modules on different levels, we break down the performance by testing the ablated versions of our approach: the input-level adaptation achieves $+8.4\%$ performance gain, while the output-level adaptation achieves $+12.5\%$ improvements. This demonstrates the effectiveness of both modules for adapting segmentation models form the synthetic domain to the real domain. Moreover, the two adaptation modules are also shown to be complementary, as combining them can further reduce the domain gap.

We also provide a few qualitative examples in Figure 3. From those results, we observe that the segmentation results generally get improved with our GIO-Ada approach. Especially, by leveraging the geometric cues, our model produce improved segmentation quality on objects with geometric structure, such as poles, traffic signs, *etc*., which are usually challenging for existing methods.

To further investigate the different design variants, especially with a focus on the importance geometric cue in the two components. We conduct further ablation studies on the two adaptation modules individually in below.

Ablation study on input-level adaptation: In our final input-level adaptation model, we use an image transform network, which takes an image and its corresponding semantic and depth label as input. To investigate the benefits of using additional inputs, we test three variants of input-level adaptation module with only depth, with only semantic label, or with none as additional input. We also include [60], an image translation model commonly adapted for domain adaptation for comparison.

The results are summarized in Table 2. We observe that all other methods outperform the non-adaptive baseline, demonstrating the importance of input-level adaptation. However, CycleGAN only improves the baseline result by $+2.3\%$, which is less effective compared to the improvement of $+6\%$ achieved by the task network. This indicates that the gradient from the task network is a useful guidance for the image transform network to preserve useful information. Nevertheless, the performance can be further boosted when additional information is further taken as input. Adding individually depth and semantic segmentation as the additional input gives an improvement of $+6.7\%$ and $+7.2\%$, respectively, and integrating them together produces $+8.4\%$ performance gain. The results suggest that the geometric information can be very useful in the image transformation process in the sense that it helps to preserve rich information in the raw 3D environment.

We further demonstrate this by providing a few examples of translated images with CycleGAN and our approach in Figure 4, in which we clearly observe that our model is able to preserves more of the geometric and semantic consistency during the translation process. More specifically, CycleGAN is observed to hallucinate buildings and trees in the sky (row 1,2,4), the poles turn into trees (row 5), and cars turn to road (row 3). In comparison, our approach is able to preserve the semantic and geometric consistency.

Ablation study on output-level adaptation: We also study different variants of the output-level adaptation. There are several possible alternatives to our joint output space adaptation. For example, performing the output space alignment proposed by [53] in semantic segmentation space and depth estimation space separately. Additionally, we try to build two discriminators to individually align the two output spaces, without considering the correlation between the two tasks. We compare these variants to our final model which aligns the joint output space of semantic segmenta-

	road	sidewalk	building	wall*	fence*	pole*	traffic light	traffic sign	vegetation	sky	person	rider	car	bus	motorbike	bicycle	mIoU	mIoU excl.*
FCNs Wld [21]	11.5	19.6	30.8	4.4	0.0	20.3	0.1	11.7	42.3	68.7	51.2	3.8	54.0	3.2	0.2	0.6	20.1	22.9
Curriculum [57]	65.2	26.1	74.9	0.1	0.5	10.7	3.7	3.0	76.1	70.6	47.1	8.2	43.2	20.7	0.7	13.1	29.0	34.8
Cross-City [6]	62.7	25.6	78.3	-	-	-	1.2	5.4	81.3	81.0	37.4	6.4	63.5	16.1	1.2	4.6	-	35.7
ROAD-Net [5]	77.7	30.0	77.5	9.6	0.3	25.8	10.3	15.6	77.6	79.8	44.5	16.6	67.8	14.5	7.0	23.8	36.1	41.7
Tsai *et al.* [53]	78.9	29.2	75.5	-	-	-	0.1	4.8	72.6	76.7	43.4	8.8	71.1	16.0	3.6	8.4	-	37.6
Sankaranarayanan *et al.* [49]	80.1	29.1	77.5	2.8	0.4	26.8	11.1	18.0	78.1	76.7	48.2	15.2	70.5	17.4	8.7	16.7	36.1	42.1
CBST [62]	69.6	28.7	69.5	12.1	0.1	25.4	11.9	13.6	82.0	81.9	49.1	14.5	66.0	6.6	3.7	32.4	35.4	40.7
non-adapt	9.7	14.1	58.5	4.7	0.3	22.7	1.9	12.9	70.7	60.9	50.2	7.2	32.2	17.4	1.3	8.0	23.3	26.5
input-level adapt	77.0	29.3	67.9	0.1	0.1	24.7	10.7	17.4	79.4	78.8	49.2	13.7	70.3	4.3	5.8	12.8	33.8	39.7
output-level adapt	79.6	29.7	75.7	11.4	0.3	25.3	11.1	14.8	76.7	76.9	45.3	15.9	67.7	15.8	4.8	13.5	35.3	40.6
GIO-Ada	78.3	29.2	76.9	11.4	0.3	26.5	10.8	17.2	81.7	81.9	45.8	15.4	68.0	15.9	7.5	30.4	**37.3**	**43.0**

Table 4. **Comparison with state-of-the-arts methods for cross-domain semantic segmentation from SYNTHIA to Cityscapes.** All results are based on VGG as the backbone architecture. Some works only report on 13 classes, we hereby mark these excluded categories with *. We also report the average performance over 13 classes as *mIoU excl. **. The best results are denoted in bold.

tion and depth estimation.

The results are shown in Table 3. First, we observe that all variants achieve significant gain over the baseline, showing the effectiveness of domain adaptation techniques in general. Particularly, output space alignment on semantic segmentation prediction [53] achieves performance gain of $+8.4\%$, while the improvement of the same output space adaptation module on depth prediction is $+6.3\%$. This is not surprising, considering our final objective is semantic segmentation. Aligning the semantic segmentation output has a more direct influence on the segmentation results. We then combine depth alignment and semantic segmentation alignment, which gives an improvement of $+8.8\%$ over the baseline, marginally better than using only semantic segmentation alignment. This suggests that trivially optimizing each task can not bring in performance gain without modeling the correlation between the tasks. Finally, by aligning the joint output space of semantic segmentation and depth estimation, we achieve a notable improvement of $+12.5\%$, showing that joint correlation is highly effective for reducing domain shift, which also verifies our motivations.

4.3. Results on SYNTHIA→Cityscapes

To facilitate the comparison with other state-of-the-art works, we further evaluate the proposed method on SYNTHIA to Cityscapes setting following [21, 57, 6, 5, 53, 49, 62]. The results of all methods are summarized in Table 4. For a fair comparison, all methods are based on VGG-16 backbones.

Similarly to the setting of Virtual KITTI → KITTI, the adaptation at both input and output levels is helpful for performance improvement: the input-level adaptation improves the baseline by $+10.5\%$, while the output-level adaptation improves it by $+12.0\%$. Integrating the two modules gives a larger performance gain of $+14.0\%$ over the non-adaptive baseline. This again verifies the effectiveness of our adaptation modules in both the input and output levels.

Our GIO-Ada outperforms all other competing methods by a notable margin. We attribute this to the supplement of geometric cues to the semantic segmentation task during domain adaptation. Nevertheless, our method takes the complementary information of the geometric cue, which is often overlooked by other methods. Our method has the potential to be integrated with other techniques for potential improvement.

5. Conclusion

In this paper, we have introduced *Geometrically Guided Input-Output Adaptation* (GIO-Ada) approach, which effectively leverages the geometric information in synthetic data to tackle the cross-domain semantic segmentation problem. Geometrically guided adaptation is performed on two different levels: 1) on the input level, depth information together with the semantic annotation is used as additional input for guiding the image transform network to reduce the domain shift on raw pixels, and 2) on the output level, depth prediction and semantic prediction are used to form a joint output space, on which an adversarial training strategy is applied to reduce the domain shift. We have experimentally validated our method on two pairs of datasets. The results demonstrate effectiveness of our GIO-Ada for cross-domain semantic segmentation with leveraged geometric information from virtual data.

Acknowledgments The authors gratefully acknowledge the support by armasuisse.

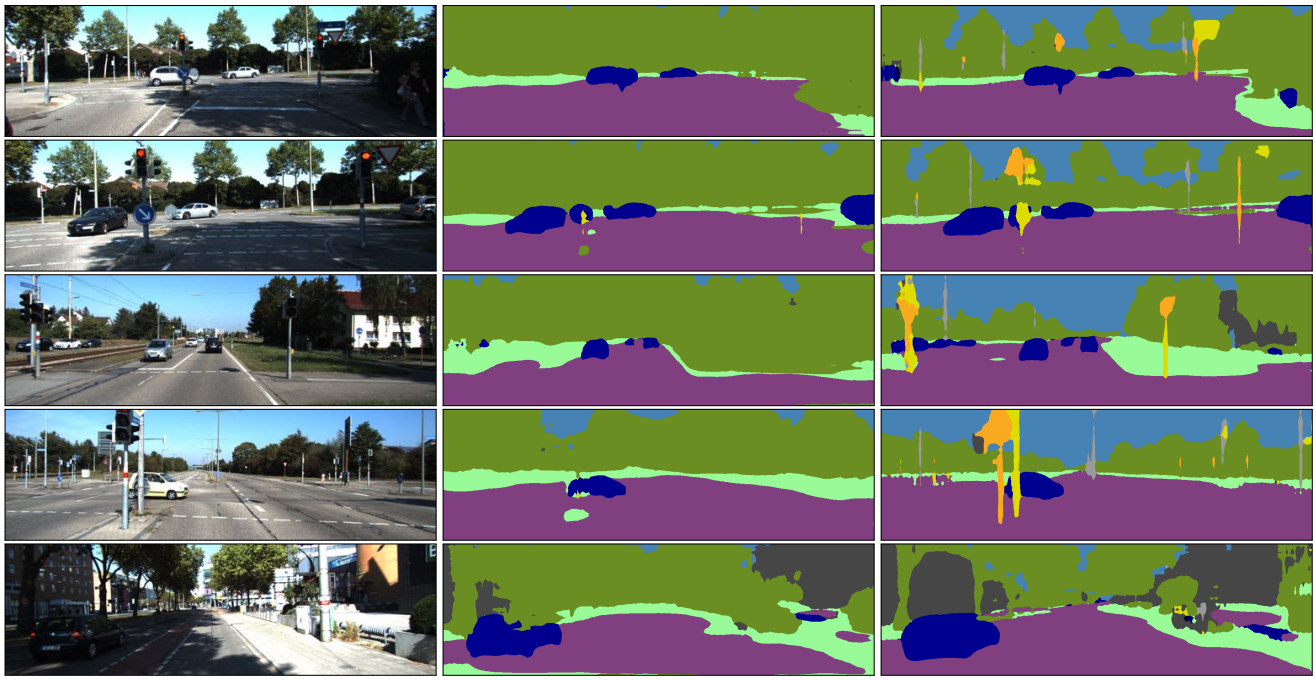

Figure 3. **Semantic segmentation qualitative results on KITTI dataset.** We follow the color encoding scheme of Cityscapes to colorize the label map. From left to right: **left:** input image, **middle:** non-adaptive results, and **right:** GIO-Ada results. Note that our approach yields noticeable improvements for objects with geometric structure, such as poles, traffic signs, *etc*.

Figure 4. **Qualitative results on input-level adaptation.** From left to right: **left:** input synthetic image, we compare the image translation results of **middle:** CycleGAN, with **right:** GIO-Ada result. Note that CycleGAN hallucinates objects in the transformation process, while GIO-Ada is able to preserve the semantic and geometric information.

References

[1] Amir Atapour-Abarghouei and Toby P Breckon. Real-time monocular depth estimation using synthetic data with domain adaptation via image style transfer. *CVPR*, 2018. 2

[2] Lin Chen, Wen Li, and Dong Xu. Recognizing rgb images by learning from rgb-d data. In *CVPR*, 2014. 2

[3] 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. In *T-PAMI*, volume 40, pages 834–848. IEEE, 2017. 2, 5

[4] Yuhua Chen, Wen Li, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Domain adaptive Faster R-CNN for object detection in the wild. *CVPR*, 2018. 2

[5] Yuhua Chen, Wen Li, and Luc Van Gool. ROAD: Reality oriented adaptation for semantic segmentation of urban scenes. *CVPR*, 2018. 1, 2, 7

[6] Yi-Hsin Chen, Wei-Yu Chen, Yu-Ting Chen, Bo-Cheng Tsai, Yu-Chiang Frank Wang, and Min Sun. No more discrimination: Cross city adaptation of road scene segmenters. *ICCV*, 2017. 7

[7] 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. *CVPR*, 2016. 2, 5

[8] Gabriela Csurka. Domain adaptation for visual applications: A comprehensive survey. *arXiv:1702.05374*, 2017. 2

[9] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. In *IJCV*, volume 88, pages 303–338. Springer, 2010. 2

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

[11] Adrien Gaidon, Qiao Wang, Yohann Cabon, and Eleonora Vig. Virtual worlds as proxy for multi-object tracking analysis. *CVPR*, 2016. 2, 5

[12] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. *ICML*, 2015. 2

[13] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The KITTI dataset. In *IJRR*, volume 32, pages 1231–1237. Sage Publications Sage UK: London, England, 2013. 2, 5

[14] Muhammad Ghifary, W Bastiaan Kleijn, Mengjie Zhang, David Balduzzi, and Wen Li. Deep reconstruction-classification networks for unsupervised domain adaptation. *ECCV*, 2016. 2

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

[16] Rui Gong, Wen Li, Yuhua Chen, and Luc Van Gool. DLOW: Domain flow for adaptation and generalization. In *CVPR*, 2019. 2

[17] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *NIPS*, 2014. 2, 4

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

[19] Philip Haeusser, Thomas Frerix, Alexander Mordvintsev, and Daniel Cremers. Associative domain adaptation. *ICCV*, 2017. 2

[20] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei A Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. *ICML*, 2018. 1, 2, 4

[21] Judy Hoffman, Dequan Wang, Fisher Yu, and Trevor Darrell. FCNs in the wild: Pixel-level adversarial and constraint-based adaptation. *arXiv:1612.02649*, 2016. 1, 2, 7

[22] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. *CVPR*, 2017. 2, 4, 5

[23] Omid Hosseini Jafari, Oliver Groth, Alexander Kirillov, Michael Ying Yang, and Carsten Rother. Analyzing modular CNN architectures for joint depth prediction and semantic segmentation. *ICRA*, 2017. 3

[24] Jianbo Jiao, Ying Cao, Yibing Song, and Rynson Lau. Look deeper into depth: Monocular depth estimation with semantic booster and attention-driven loss. *ECCV*, 2018. 3

[25] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. *ECCV*, 2016. 5

[26] Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. *CVPR*, 2018. 3

[27] Iasonas Kokkinos. UberNet: Training a universal convolutional neural network for low-, mid-, and high-level vision using diverse datasets and limited memory. *CVPR*, 2017. 3

[28] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *NIPS*, 2012. 1

[29] Brian Kulis, Kate Saenko, and Trevor Darrell. What you saw is not what you get: Domain adaptation using asymmetric kernel transforms. *CVPR*, 2011. 2

[30] Kuan-Hui Lee, German Ros, Jie Li, and Adrien Gaidon. SPIGAN: Privileged adversarial learning from simulation. *arXiv:1810.03756*, 2018. 2

[31] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. *ICCV*, 2017. 2

[32] Wen Li, Lin Chen, Dong Xu, and Luc Van Gool. Visual recognition in rgb images and videos by learning from rgb-d data. *T-PAMI*, 40(8):2030–2036, 2018. 2

[33] Wen Li, Zheng Xu, Dong Xu, Dengxin Dai, and Luc Van Gool. Domain generalization and adaptation using low rank exemplar svms. *T-PAMI*, 40(5):1114–1127, 2018. 2

[34] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. *ECCV*, 2014. 2

[35] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. *CVPR*, 2015. 2

[36] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I Jordan. Learning transferable features with deep adaptation networks. *ICML*, 2015. 2

[37] Hao Lu, Lei Zhang, Zhiguo Cao, Wei Wei, Ke Xian, Chunhua Shen, and Anton van den Hengel. When unsupervised domain adaptation meets tensor representations. *ICCV*, 2017. 2

[38] Fabio Maria Carlucci, Lorenzo Porzi, Barbara Caputo, Elisa Ricci, and Samuel Rota Bulo. AutoDIAL: Automatic domain alignment layers. *ICCV*, 2017. 2

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

[40] Arsalan Mousavian, Hamed Pirsiavash, and Jana Košecká. Joint semantic segmentation and depth estimation with deep convolutional networks. *3DV*, 2016. 3

[41] Zak Murez, Soheil Kolouri, David Kriegman, Ravi Ramamoorthi, and Kyungnam Kim. Image to image translation for domain adaptation. *CVPR*, 2018. 2

[42] Pau Panareda Busto and Juergen Gall. Open set domain adaptation. *ICCV*, 2017. 2

[43] V. M. Patel, R. Gopalan, R. Li, and R. Chellappa. Visual domain adaptation: A survey of recent advances. *IEEE Signal Processing Magazine*, 30(3):53–69, 2015. 2

[44] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *ICLR*, 2016. 2

[45] Stephan R Richter, Zeeshan Hayder, and Vladlen Koltun. Playing for benchmarks. *ICCV*, 2017. 1

[46] Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. *ECCV*, 2016. 1, 2

[47] 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. *CVPR*, 2016. 2, 5

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

[49] Swami Sankaranarayanan, Yogesh Balaji, Arpit Jain, Ser Nam Lim, and Rama Chellappa. Learning from synthetic data: Addressing domain shift for semantic segmentation. *CVPR*, 2018. 7

[50] Ozan Sener, Hyun Oh Song, Ashutosh Saxena, and Silvio Savarese. Learning transferrable representations for unsupervised domain adaptation. *NIPS*, 2016. 2

[51] Ashish Shrivastava, Tomas Pfister, Oncel Tuzel, Joshua Susskind, Wenda Wang, and Russell Webb. Learning from simulated and unsupervised images through adversarial training. *CVPR*, 2017. 2

[52] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv:1409.1556*, 2014. 5

[53] Yi-Hsuan Tsai, Wei-Chih Hung, Samuel Schulter, Kihyuk Sohn, Ming-Hsuan Yang, and Manmohan Chandraker. Learning to adapt structured output space for semantic segmentation. *CVPR*, 2018. 2, 4, 6, 7

[54] Peng Wang, Xiaohui Shen, Zhe Lin, Scott Cohen, Brian Price, and Alan L Yuille. Towards unified depth and semantic prediction from a single image. *CVPR*, 2015. 2, 3

[55] Dan Xu, Wanli Ouyang, Xiaogang Wang, and Nicu Sebe. PAD-Net: Multi-tasks guided prediction-and-distillation network for simultaneous depth estimation and scene parsing. *CVPR*, 2018. 3

[56] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *ICLR*, 2016. 2

[57] Yang Zhang, Philip David, and Boqing Gong. Curriculum domain adaptation for semantic segmentation of urban scenes. *ICCV*, 2017. 1, 2, 7

[58] Zhenyu Zhang, Zhen Cui, Chunyan Xu, Zequn Jie, Xiang Li, and Jian Yang. Joint task-recursive learning for semantic segmentation and depth estimation. *ECCV*, 2018. 3

[59] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. *CVPR*, 2017. 2

[60] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *ICCV*, 2017. 1, 2, 4, 5, 6

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

[62] Yang Zou, Zhiding Yu, BVK Vijaya Kumar, and Jinsong Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. *ECCV*, 2018. 2, 7