DESC: Domain Adaptation for Depth Estimation via Semantic Consistency

Adrian Lopez-Rodriguez
al4415@imperial.ac.uk
Krystian Mikolajczyk
k.mikolajczyk@imperial.ac.uk

MatchLab
Imperial College London
London, UK

Abstract

Accurate real depth annotations are difficult to acquire, needing the use of special devices such as a LiDAR sensor. Self-supervised methods try to overcome this problem by processing video or stereo sequences, which may not always be available. Instead, in this paper, we propose a domain adaptation approach to train a monocular depth estimation model using a fully-annotated source dataset and a non-annotated target dataset. We bridge the domain gap by leveraging semantic predictions and low-level edge features to provide guidance for the target domain. We enforce consistency between the main model and a second model trained with semantic segmentation and edge maps, and introduce priors in the form of instance heights. Our approach is evaluated on standard domain adaptation benchmarks for monocular depth estimation and show consistent improvement upon the state-of-the-art.

1 Introduction

State-of-the-art depth estimation methods are capable of inferring an accurate depth map from a monocular image by relying on deep learning methods that require a large amount of data with annotations [16, 36]. Annotations in the form of precise depth measurements are typically provided by special tools such as a LiDAR sensor [20] or structured light devices [55]. Thus, obtaining depth annotations is costly and time-consuming. Much research has focused on developing methods not relying on directly acquired depth annotations by leveraging stereo [19, 22] or video sequences [5, 23, 60] for self-supervision. These research directions have shown promise, but a stereo pair or video sequence may not always be available in existing datasets. The use of synthetic data provides a way to obtain a large amount of accurate ground truth depth in a fast manner, however, synthetic data and real data have usually a domain gap due to the difficulty of generating photorealistic synthetic images. To that end, domain adaptation techniques [47, 63] can help to transfer the models trained on an annotated source dataset $\mathcal{S}$ to a target dataset $\mathcal{T}$, reducing the burden of training a model for a new environment or camera.

Research results have shown that the domain gap for semantic segmentation and instance detection can be reduced by introducing depth information during training [9, 41]. A different direction, which leverages semantic information to reduce the domain gap in depth estimation, has been less studied and mainly in multi-task scenarios [2, 35]. Existing

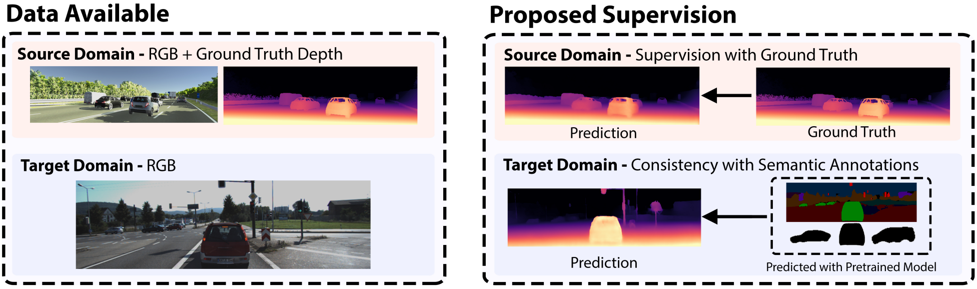
Data Available
Source Domain - RGB + Ground Truth Depth
Target Domain - RGB
Proposed Supervision
Source Domain - Supervision with Ground Truth
Prediction
Ground Truth
Target Domain - Consistency with Semantic Annotations
Prediction
Predicted with Pretrained Model

Figure 1: Overview of the data available and proposed supervision. The source domain $\mathcal{S}$ contains both RGB and ground truth depth data, and the target domain $\mathcal{T}$ contains RGB data only. We train a depth estimation model to achieve high performance in $\mathcal{T}$ by leveraging semantic annotations to introduce semantic consistency in $\mathcal{T}$. The semantic annotations are obtained using a panoptic segmentation model trained with external data.

datasets with semantic annotations are large and diverse in scenes as well as cameras used, hence models trained on these diverse semantic datasets are capable of generalizing to different settings [37]. Several works [6, 39] have shown that using pretrained models to obtain semantic annotations can also bring improvements in the depth estimation task. Motivated by these findings, we exploit readily-available panoptic segmentation models as guidance to bridge the gap between two different domains and to improve monocular depth estimation.

Domain adaptation approaches benefit from pseudo-labelling [6, 53] and consistency of predictions in the source and target domains [10, 62]. Therefore, we propose an approach that leverages semantic annotations to enforce consistency for depth estimation between the two domains, and to provide depth pseudo-labels to the target domain by using the size of the detected objects. Figure 1 shows an overview of the task. Our main contributions are: (1) the proposal of an approach to form depth pseudo-labels in the target domain by using object size priors, which are learnt in an instance-based manner in the annotated source domain; (2) the introduction of a consistency constraint with predictions from a second model trained on high-level semantics and low-level edge maps; (3) state-of-the-art results in the task of monocular depth estimation with domain adaptation from VirtualKITTI [17] to KITTI [20].

2 Related Work

2.1 Monocular Depth Estimation

Self-Supervision. Early depth estimation methods rely on supervised training, using annotations from LiDAR [20] or structured light scanners [55]. Due to the difficulty of obtaining depth annotations, several works have focused on using either stereo pairs or video self-supervision. Xie *et al.* [59] regressed a discretized disparity map and used a pixel-wise consistency loss with a second camera view, and Garg *et al.* [19] extended it to predict continuous depth values. The accuracy was further improved in Monodepth [22] by forcing the network to predict from a single image both left and right disparities and adding a consistency term. A stereo pair was used in Luo *et al.* [42] to supervise a model that synthesized the right view from the left image, and then processing both views by a stereo-matching network. Other notable approaches include the use of adversarial techniques and cycle-

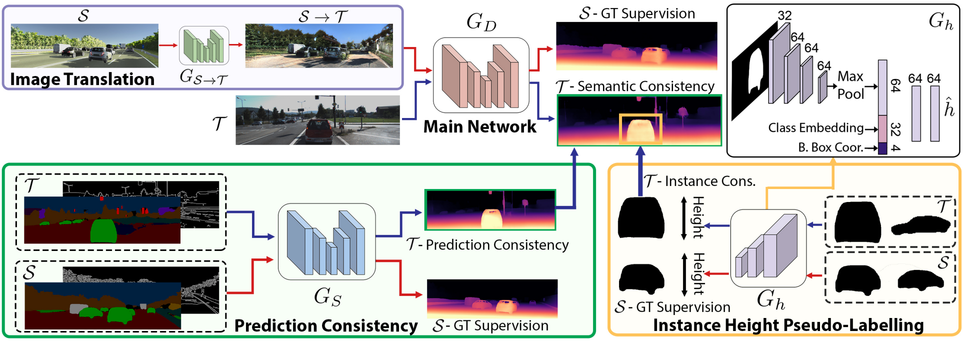
$\mathcal{S}$
$\mathcal{S} \to \mathcal{T}$
Image Translation
$G_{\mathcal{S}\to\mathcal{T}}$
$\mathcal{T}$
G_D
Main Network
$\mathcal{S}$ - GT Supervision
$\mathcal{T}$ - Semantic Consistency
G_h
32
64
64
64
Max Pool
64
64 64
$\hat{h}$
Class Embedding
32
B. Box Coor.
4
$\mathcal{T}$ - Instance Cons.
Height
Height
$\mathcal{T}$
$\mathcal{S}$
$\mathcal{S}$ - GT Supervision
G_h
Instance Height Pseudo-Labelling
$\mathcal{T}$
$\mathcal{S}$
$\mathcal{T}$ - Prediction Consistency
G_S
Prediction Consistency
$\mathcal{S}$ - GT Supervision

Figure 2: Overview of the approach. We train a depth estimation network G_D with both target $\mathcal{T}$ and source $\mathcal{S}$ images. Source images are adapted to the style of the target images. For $\mathcal{S}$, we use ground truth supervision, while we enforce consistency with semantic information in $\mathcal{T}$. The consistency is enforced with (1) predictions from a second network G_S trained with edges and semantic maps as input, and (2) depth pseudo-labels formed using an instance height $\hat{h}$ predicted by G_h. Both G_S and G_h are trained using ground truth data from $\mathcal{S}$. The architecture of G_h is given in the top right. We use ReLU between the layers of G_h.

consistency [48, 49]. Stereo images are not always available, hence video self-supervision has also been researched. Simultaneous learning of depth and pose was addressed in Zhou *et al.* [64], which given three video frames, projected the t+1 and t-1 views to the reference view t. Joint pose, depth and optical flow learning was proposed in GeoNet [60], and Monodepth2 [23] focused on improving the pixel reprojection loss and the multi-scale loss.

Depth and Semantic Information. Mousavian *et al.* [46] trained a single network for both semantic and depth prediction in a multi-task manner by using a shared backbone and task-specific layers. In that direction, Chen *et al.* [7] trained a network capable of selecting between depth or semantic segmentation output by only changing an intermediate task layer. In Zhang *et al.* [61] the two tasks, semantic segmentation and depth estimation, were refined alternately in a progressive manner by using a task attention module to propagate information from one task to the other. Jiao *et al.* [31] proposed a novel unit to share information between the two tasks. Another method, Guizilini *et al.* [24] used a pretrained semantic segmentation network to guide the feature maps of the depth network using pixel-adaptive convolutions. In MegaDepth [39], a new diverse depth dataset was collected from the internet using Multi-View Stereo and Structure-from-Motion to retrieve depth information, where semantic information was used to filter spurious depth values and to define ordinal labels. Atapour-Abarghouei and Breckon [2] assumed the availability of temporal information both in training and test time, where the different video frames were fused together to predict depth and semantic segmentation in a multi-task approach. Struct2Depth [5], which is more related to our work, used precomputed masks of object instances to tackle the problem of dynamic objects in video self-supervision by imposing object size constraints.

2.2 Domain Adaptation

Domain adaptation is attracting more and more attention due to the lack of sufficient volume of annotated data for supervised training. It showed some success in areas such as classification [53, 57] and semantic segmentation [10, 56]. Popular approaches include style adaptation of the source images to match the target images [28], enforcing consistency of predictions [15, 52, 54], adversarial approaches to match either the features [18, 57] or the outputs [56] of the two domains, and using pseudo-labels [6, 53].

Depth Estimation. Image translation techniques have also been used for domain adaptation for depth estimation tasks [1, 47, 62, 63]. Atapour-Abarghouei and Breckon [1] generated synthetic data using the video-game GTA V and used a cycle-consistency approach. Additionally, during inference [1] translated the target domain images to the style of the source domain before estimating the depth, adding computational burden. Our approach builds upon T^{2}Net [63], which also uses an image translation network but without a cycle-consistency loss, reducing the complexity due to the lower number of networks needed. In T^{2}Net, the target domain images are not translated during inference contrary to [1]. GASDA [62] focused on the scenario where stereo supervision is available in the target domain, and added stereo photometric guidance and depth prediction consistency between original and style-transferred target domain images. GASDA [62] averages during test time the depth predicted for a given target image and its corresponding style-transferred image, increasing the inference-time complexity. AdaDepth [47] used an adversarial approach to align both output and feature distributions between the source and target domain, along with feature consistency to avoid mode collapse. In a multi-task setup, Kundu *et al.* [35] developed a cross-task distillation module and contour-based content regularization to extract feature representations with greater transferability. Several synthetics datasets have been generated that can be used for depth estimation. Virtual KITTI [17] provides a synthetic version of KITTI. SYNTHIA [51] provides multi-camera images and depth annotations, whereas CARLA [13] offers a simulated environment where virtual cameras can be placed arbitrarily. In non-driving settings, some synthetic datasets that provide depth annotations are also available [38, 44].

3 Method

In this section we introduce our domain adaptation for Depth Estimation via Semantic Consistency (DESC) approach. An overview is presented in Figure 2. During inference we only apply our depth estimation network G_D to our target images. Semantic annotations are predicted for our source and target datasets using a panoptic segmentation model [34] trained with external data, providing per image detected instances and a semantic segmentation map.

3.1 Pseudo-Labelling using Instance Height

The height of the detected object instances can provide a strong cue for distance estimation. Struct2Depth [5] used the instance height to deal with moving objects in video self-supervision. Thus, Struct2Depth retrieved an approximate distance to the objects by solving

$$\hat{D} \approx \frac{f \cdot h}{H} \tag{1}$$

where $\hat{D}$ is an approximate distance to the object, f is the focal length in pixels, H is the predicted instance size in pixels and h is the physical height of the object. It is assumed that the entire object instance is placed at a distance $\hat{D}$, that f is known, and that the real object size h is unknown. In Struct2Depth [5], the object size was set as a shared learnable parameter $\hat{h}$ for the class *car*, *i.e.*, all of the detected instances of class *car* were assumed to have the same height. We argue that predicting a $\hat{h}$ per object instance rather than class can provide a better height estimate, as it can take into account both intra-class variations and occlusions in the detected instances. Furthermore, instead of learning $\hat{h}$ in an unsupervised manner as in Struct2Depth [5], we can improve the estimation using source domain data. Therefore, we use a network G_h, with a simple architecture presented in Figure 2, to predict a $\hat{h}_i$ for an instance i from the dimensions of its bounding box, the detected binary instance mask and the predicted class label. We train G_h using labels in the source data by retrieving $h_{GT,i}$, which is the ground truth physical object size for instance i. To retrieve $h_{GT,i}$ we use $h_{GT,i} = \frac{H_i \cdot \hat{D}_{\mathcal{S},i}}{f_{\mathcal{S}}}$, where the instance depth $\hat{D}_{\mathcal{S},i}$ is obtained directly from the depth ground truth. To obtain $\hat{D}_{\mathcal{S},i}$ we use $\hat{D}_i = median(M_{\mathcal{S},i} \odot y_{\mathcal{S}})$, where $M_{\mathcal{S},i}$ is the binary segmentation instance mask for a source domain detected instance i, $\odot$ refers to the Hadamard product, $y_{\mathcal{S}}$ is the ground truth depth, and the median operation is performed only for non-zero values. Thus, G_h is trained in the source domain with $\mathcal{L}_{I,\mathcal{S}} = \frac{1}{n_I} \sum_i |\hat{h}_{\mathcal{S},i} - h_{GT,i}|$, where n_I is the number of detected instances. In the target domain, G_h is used to predict a height $\hat{h}_{\mathcal{T},i}$ for a detected instance i, and then $\hat{h}_{\mathcal{T},i}$ is used to retrieve a depth pseudo-label $\hat{D}_{\mathcal{T},i}$ computed using Equation 1. We use the depth pseudo-labels $\hat{D}_{\mathcal{T},i}$ to provide supervision for G_D in the target domain using a sum of pixel-wise L_1 losses over all detected instances i,

$$\mathcal{L}_{I,\mathcal{T}} = \frac{\phi}{p_I} \sum_i \|(\frac{\hat{D}_{\mathcal{T},i}}{\phi} - G_D(x_{\mathcal{T}})) \odot M_{\mathcal{T},i}\|_1 \tag{2}$$

where p_I is the sum of non-zero pixels for all the binary segmentation masks $M_{\mathcal{T},i}$, $x_{\mathcal{T}}$ is an image from $\mathcal{T}$ and ϕ is a learnable scalar. The scalar ϕ is used to correct any scale mismatch in the predictions of $G_D(x_{\mathcal{T}})$ due to camera differences between $\mathcal{S}$ and $\mathcal{T}$ [26]. When computing $\hat{D}_{\mathcal{T},i}$ we use the focal length $f_{\mathcal{T}}$ of the target domain camera, although as we will show in Section 4, ϕ automatically scales the values to the correct range even for unknown $f_{\mathcal{T}}$. As we use a panoptic segmentation model trained with external data to extract semantic annotations, some of the classes detected may be present in $\mathcal{T}$ but not in $\mathcal{S}$, *e.g.*, *person* in Virtual KITTI→KITTI. For those classes, G_h can also learn an instance-based height prior in an unsupervised manner via consistency with G_D in $\mathcal{L}_{I,\mathcal{T}}$.

3.2 Consistency of Predictions using Semantic Information

Many works [15, 52, 54] have shown that constraining the learning process by requiring consistency in a domain adaptation setting reduces the performance gap. Similar observations have been made in semi-supervised learning [8], where a contrastive loss is used between different views of the same scene obtained via data augmentation. Following these findings, we enforce consistency between the predictions generated by our main depth estimation network, G_D, and a secondary network, G_S, whose input data x_{Sem} is formed by two channels that have a low domain gap: a semantic segmentation map and an edge map.

Semantic Structure. A semantic segmentation map provides information on the high-level structure of the scene, and this high-level structure helps to predict the depth structure. The information is introduced in the form of an integer corresponding to the semantic class label, as we experimentally found it to yield better performance than one-hot encoding.

Edge Map. Deep learning networks tend to use texture cues [21] for predictions. We use an edge map to reduce the impact of the texture differences between domains, and to provide a different modality of the data to the network. Edges include information about the shapes of objects, and this shape information is valuable in depth related tasks [29, 30]. Edges also present less variation and need less adaptation in domains with semantically similar scenes.

	Lower is better				Higher is better		
Method	Abs Rel	Sq Rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Cap 80m							
AdaDepth [47]	0.214	1.932	7.157	0.295	0.665	0.882	0.950
T^{2}Net [63]	0.173	1.396	6.041	0.251	0.757	0.916	0.966
DESC	**0.156**	**1.067**	**5.628**	**0.237**	**0.787**	**0.924**	**0.970**
AdaDepthS [47]	0.167	1.257	5.578	0.237	0.771	0.922	0.971
Cap 50m							
AdaDepth [47]	0.203	1.734	6.251	0.284	0.687	0.899	0.958
T^{2}Net [63]	0.165	1.034	4.501	0.235	0.772	0.927	0.972
DESC	**0.149**	**0.819**	**4.172**	**0.221**	**0.805**	**0.934**	**0.975**
AdaDepthS [47]	0.162	1.041	4.344	0.225	0.784	0.930	0.974

Table 1: Results for Virtual KITTI→KITTI in KITTI [20] Eigen [14] split. Results from *T^{2}Net* are recomputed using median scaling and the official pretrained model. *AdaDepthS* is a semi-supervised method using additionally 1000 annotated KITTI images for training.

Consistency. As both networks G_D and G_S receive different input modalities, forcing consistency between them for the predictions of the target domain can significantly increase the target-domain performance of both models. We propose to supervise G_S with source domain depth ground truth $y_\mathcal{S}$ by using a pixel-wise L_1 loss, $\mathcal{L}_{Con,\mathcal{S}}$, and then force consistency of predictions in the target domain via $\mathcal{L}_{Con,\mathcal{T}}$. Then, assuming N is the total number of pixels,

$$\mathcal{L}_{Con,\mathcal{S}} = \frac{1}{N}\|G_S(x_{Sem,\mathcal{S}}) - y_\mathcal{S}\|_1, \qquad \mathcal{L}_{Con,\mathcal{T}} = \frac{1}{N}\|G_D(x_\mathcal{T}) - G_S(x_{Sem,\mathcal{T}})\|_1 \tag{3}$$

3.3 Training Loss

We now present the modules used in DESC in addition to our semantic consistency losses.

Depth Estimation Loss. Our model G_D outputs a multiscale prediction that is supervised using source domain ground truth with $\mathcal{L}_D$, which is a pixel-wise multiscale L_1 loss [62, 63].

Image Translation. Image translation has been demonstrated to effectively reduce the domain gap [62, 63]. We adopt the approach from T^{2}Net [63], where a network $G_{\mathcal{S}\to\mathcal{T}}$ translates the source image to the target domain without cycle consistency. T^{2}Net [63] uses a least-squares adversarial term $\mathcal{L}_{GAN}$ [43] to produce examples $x_{\mathcal{S}\to\mathcal{T}}$ having a similar distribution to $x_\mathcal{T}$, and leverages the constraint imposed by $\mathcal{L}_D$ to ensure $x_{\mathcal{S}\to\mathcal{T}}$ is geometrically consistent with $x_\mathcal{S}$. The method also uses a L_1 identity loss $\mathcal{L}_{IDT} = \frac{1}{N}\|G_{\mathcal{S}\to\mathcal{T}}(x_\mathcal{T}) - x_\mathcal{T}\|_1$ to force $G_{\mathcal{S}\to\mathcal{T}}(x_\mathcal{T}) \approx x_\mathcal{T}$, *i.e.*, $\mathcal{L}_{IDT}$ forces $G_{\mathcal{S}\to\mathcal{T}}$ to behave as an identity mapping for $x_\mathcal{T}$.

Smoothing. We use for the target data the smoothing term $\mathcal{L}_{Sm}$ introduced in Monodepth [22], and successfully used in domain adaptation [62, 63] methods for depth estimation.

Overall Loss. Our final model is trained using the following loss

$$\mathcal{L} = \lambda_\mathcal{S}(\mathcal{L}_\mathcal{D} + \mathcal{L}_{Con,\mathcal{S}} + \mathcal{L}_{I,\mathcal{S}}) + \lambda_\mathcal{T}(\mathcal{L}_{Con,\mathcal{T}} + \mathcal{L}_{I,\mathcal{T}}) + \lambda_{Sm}\mathcal{L}_{Sm} + \lambda_{IDT}\mathcal{L}_{IDT} + \lambda_{GAN}\mathcal{L}_{GAN} \tag{4}$$

where $\lambda_\mathcal{S}, \lambda_\mathcal{T}, \lambda_{Sm}, \lambda_{IDT}, \lambda_{GAN}$ are hyperparameters to balance the different terms.

4 Experiments

We discuss the experimental setup before presenting our evaluation results.

Setup. We use Pytorch 1.4 and an NVIDIA 1080TI GPU. We obtain the semantic annotations in both $\mathcal{S}$ and $\mathcal{T}$, by using a ResNet-101 [25] panoptic segmentation model [33] trained

	Lower is better				Higher is better		
Method	Abs Rel	Sq Rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Only Source [62]	0.223	2.205	7.055	0.305	0.672	0.872	0.945
+ Img.	0.199	2.436	7.137	0.280	0.753	0.890	0.950
+ Img. + Con. (only edges)	0.187	1.330	6.094	0.258	0.708	0.905	0.966
+ Img. + Con.	0.173	1.235	5.776	0.244	0.748	0.919	0.969
+ Img. + Ins.	0.171	1.332	5.818	0.250	0.771	0.918	0.966
+ Img. + Ins. ($\lambda_{Sm} = 0.1$)	0.165	1.157	5.670	0.245	0.774	0.921	0.968
DESC - Full (1 h per class [8])	0.160	1.107	5.746	0.243	0.780	0.920	0.968
DESC - Full (unknown $f_{\mathcal{T}}$)	**0.156**	1.084	5.654	**0.237**	0.783	**0.926**	**0.971**
DESC - Full	**0.156**	**1.067**	**5.628**	**0.237**	**0.787**	0.924	0.970
G_S	0.186	2.164	7.011	0.282	0.763	0.894	0.949

Table 2: Ablation study of DESC for Virtual KITTI→KITTI in Eigen split [14] capped at 80m. *Img.* refers to using image translation, *Ins.* to using instance-height pseudo-labels (Section 3.1) and *Con.* to the consistency of predictions constraint (Section 3.2).

in COCO-Stuff [3, 40] from the *Detectron 2* library [58]. We employ a U-Net [50] for G_D and G_S, and a ResNet-based model for $G_{S\to\mathcal{T}}$. Both image translation and depth estimation architectures are the same as the architectures used in [62, 63]. Following [62], we set $\lambda_{\mathcal{S}} = 50$, $\lambda_{GAN} = 1$, $\lambda_{Sm} = 0.01$, and following [63] we set $\lambda_{IDT} = 100$. Similarly to the original implementation of [62], we first pretrain the networks to reach good performance in $\mathcal{S}$ before introducing the consistency terms, *i.e.*, with $\lambda_{\mathcal{T}} = 0$. Afterwards, we freeze $G_{\mathcal{S}\to\mathcal{T}}$ to reduce the memory footprint, and we introduce the semantic consistency terms by setting $\lambda_{\mathcal{T}} = 1$ unless stated otherwise. The batch size is set to 4, with a 50/50 target and source data ratio, we use Adam [32] with learning rate 10^{-4} and we train for 20,000 iterations after pretraining. To obtain the edge map for G_S we use a Canny Edge detector [4]. We randomly change the brightness, saturation and contrast of the images for data augmentation.

Virtual KITTI→KITTI. We follow the same experimental settings as in [62, 63]. Both Virtual KITTI [17] and KITTI [20] images are downscaled to 640x192, and following [63] we cap the Virtual KITTI [17] ground truth depth at 80m.

Cityscapes→KITTI. Cityscapes [11] provides disparity maps computed using Semi-Global Matching [27]. We use the official training set, consisting of 2975 images of size 2048x1024. We set the horizon line approximately in the center by cropping the upper part, resulting in images of 2048x964. We then take the 2048x614 center crop to have the same aspect ratio as in KITTI and rescale the images to 640x192. We use $\lambda_{\mathcal{T}} = 5$ for this experiment.

Evaluation in KITTI. We follow the same evaluation protocol, metrics and splits as in Eigen *et al.* [14] for KITTI, using the evaluation code from Monodepth2 [23]. The predictions are upscaled to match the ground truth size. The results are reported using median scaling as in past methods [5, 47, 64], except when using stereo supervision in KITTI. We provide results for both ground truth depth capped at 80m and between 1-50m as done in [62, 63].

4.1 Quantitative Results

Comparison with State-of-the-Art. Table 1 compares the performance of DESC with the Virtual KITTI→KITTI state-of-the-art methods not using stereo nor video self-supervision in KITTI. DESC performs better than AdaDepth [47] and T^{2}Net [63], with a Sq. Rel. error almost 24% lower than T^{2}Net. We also include *AdaDepthS*, which is a version of AdaDepth that uses 1000 annotated KITTI images for training in addition to Virtual KITTI ground truth. We improve upon *AdaDepthS* in most metrics without using any KITTI annotations.

	Lower is better				Higher is better		
Method	Abs Rel	Sq Rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Virtual KITTI→ KITTI							
Source + Stereo	0.131	1.154	5.518	0.227	0.837	0.937	0.971
T^{2}Net [63] + Stereo	0.126	1.114	5.429	0.223	0.839	0.938	0.971
GASDA [62]	0.124	1.018	5.202	**0.217**	**0.846**	**0.944**	0.973
DESC + Stereo	**0.119**	**0.935**	**5.050**	**0.217**	0.843	0.942	**0.974**
Only KITTI							
Monodepth2 (w/o pre.) [23]	0.130	1.144	5.485	0.232	0.831	0.932	0.968
Monodepth2 (ImageNet pre.) [23]	0.109	0.873	4.960	0.209	0.864	0.948	0.975

Table 3: Results in KITTI Eigen split (80m cap) for methods using stereo data in KITTI. Due to an evaluation error in [62], results from GASDA are recomputed using the official pretrained models. We include the state-of-the-art stereo-trained method *Monodepth2* [23].

Ablation Study. Table 2 shows an ablation study of DESC. The result marked with *+Img* correspond to T^{2}Net [63] without the adversarial feature module, and with a lower smoothing weight λ_{Sm} as we use $\lambda_{Sm} = 0.01$ instead of the $\lambda_{Sm} = 0.1$ used for the T^{2}Net implementation shown in Table 1. The lower λ_{Sm} we use accounts for the better results of T^{2}Net in Table 1. We chose a smaller λ_{Sm} for our experiments because a larger λ_{Sm} blurs the predictions, leading to a worse result after enforcing consistency with G_S due to the loss of detail. However, a larger λ_{Sm} is beneficial when consistency with G_S is not applied as shown by the improved results of *+Img.+Ins. ($\lambda_{Sm} = 0.1$)* compared to *+Img.+Ins.*. Both the instance-based pseudo-labelling and consistency with G_S modules bring an improvement as shown in *+Img.+Ins.* and *+Img.+Con.* compared to *+Img*. Using the consistency term in the case where only edge maps are input into G_S improves most metrics as shown in *+ Img.+Con. (only edges)*, although it also shows that inputting the semantic map into G_S is largely beneficial. We argue that the better results of *+Img.+Con.* compared to *Img* are not due to a distillation process, *i.e.*, due to G_S having a higher accuracy than G_D after source data pretraining. Table 2 shows in the line G_S the accuracy when evaluating G_S after source data pretraining (*i.e.*, before G_D consistency), and its lower performance compared to *+Img.+Con.* suggests that consistency is the reason for the accuracy increase. *DESC - Full* shows an improvement in all metrics, also compared to learning a single h per class as in Struct2Depth [5]. For *DESC - Full (unknown $f_{\mathcal{T}}$)* we set $f_{\mathcal{T}}$ to half the actual value, obtaining comparable results to when using the correct value of $f_{\mathcal{T}}$, *i.e.*, in *DESC - Full*. This result shows that ϕ in Equation 2 automatically scales the instance size pseudo-labels to the correct range for unknown $f_{\mathcal{T}}$.

Stereo Supervision. Although DESC focuses on the setting where no self-supervision is used in $\mathcal{T}$, our approach can also bring an improvement in such a scenario. We train DESC adding stereo supervision in KITTI by adding the same multiple-scale pixel-wise reconstruction method as in GASDA [62] with the same loss weight of $\lambda_{St} = 50$. To account for the introduced supervision in $\mathcal{T}$, we increase $\lambda_{\mathcal{T}} = 5$ and the number of training iterations to 100,000. Table 3 shows that, compared to *T^{2}Net+Stereo*, our method with stereo supervision, *DESC + Stereo*, achieves better results in all metrics and also outperforms GASDA [62] in most metrics. GASDA is a domain adaptation method tailored for stereo supervision that uses two depth estimation networks and an image-translation network during inference. We report better performance than the state-of-the-art for stereo supervision, Monodepth2 [23] without ImageNet [12] pretraining in *Monodepth2 (w/o pre.)*. However, ImageNet pretraining has a positive effect on the accuracy, shown in *Monodepth2 (ImageNet pre.)*.

Evaluation on KITTI Stereo. KITTI Stereo 2015 [45] provides images annotated in a process combining (1) static background retrieval via egomotion compensation and (2) fitting

	Lower is better				Higher is better		
Method	Abs Rel	Sq Rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Virtual KITTI→KITTI							
T^{2}Net [63]	0.151	1.535	6.177	0.224	0.817	0.935	0.975
DESC	0.120	0.968	5.597	0.206	0.839	0.937	0.977
GASDA [62]	0.095	1.068	5.015	0.168	0.906	0.966	**0.986**
DESC + Stereo	**0.085**	**0.781**	**4.490**	**0.158**	**0.909**	**0.967**	**0.986**
Only KITTI							
Monodepth2 (w/o pre.) [23]	0.096	1.163	5.161	0.179	0.898	0.959	0.981
Monodepth2 (ImageNet pre.) [23]	0.082	0.908	4.698	0.158	0.919	0.970	0.986

Table 4: Results on the KITTI 2015 stereo 200 training set disparity images [20, 45]. We include *Monodepth2* [23], the state-of-the-art stereo method trained only in KITTI.

	Lower is better				Higher is better		
Method	Abs Rel	Sq Rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Only Cityscapes							
Source Baseline	0.189	1.717	6.478	0.257	0.740	0.919	0.968
Struct2Depth (M) [5]	0.188	1.354	6.317	0.264	0.714	0.905	0.967
Struct2Depth (M+R) [5]	0.153	1.109	5.557	0.227	0.796	0.934	0.975
Cityscapes→ KITTI							
T^{2}Net [63]	0.173	1.335	5.640	0.242	0.773	0.930	0.970
DESC (Img.+Ins.)	0.174	1.480	5.920	0.240	0.782	0.931	0.971
DESC (Img.+Con.)	0.150	0.981	5.359	**0.222**	0.805	0.938	**0.976**
DESC (Full, $\phi = 1$)	0.169	1.142	5.936	0.261	0.741	0.919	0.967
DESC (Full, ϕ learnt)	**0.149**	**0.967**	**5.236**	0.223	**0.810**	**0.940**	**0.976**

Table 5: Cityscapes→KITTI results, evaluated in KITTI [20] Eigen split (80m cap). *Struct2Depth (M+R)* [5] uses three consecutive frames for refinement.

of CAD models to account for dynamic objects. The result is a denser ground truth compared to the LiDAR depth annotations provided in KITTI, especially in the cars. DESC, which uses instances pseudo-labels, benefits from evaluating in images with denser annotation in the vehicles, as shown in Table 4 in the larger accuracy gap between *DESC* and *T^{2}Net*, and also between *DESC + Stereo* and *GASDA* compared to Table 1 and Table 3. *DESC + Stereo* achieves either better (Sq Rel, RMSE) or equal (RMSE log) squared metrics results than the state-of-the-art *Monodepth2 (ImageNet pre.)* without pretraining G_D in ImageNet.

Cityscapes→ KITTI. Table 5 shows the results for this benchmark. We improve upon T^{2}Net for all metrics, with a 13.9% lower absolute relative error. Most of the accuracy improvement comes from the consistency term as shown in *DESC (Img.+Con.)* and *DESC (Full, ϕ learnt)*. Due to the camera difference between the datasets, the learnable scalar ϕ is necessary for good performance, as shown for fixed $\phi = 1$ in *DESC (Full, ϕ=1)*. Struct2Depth [5] also uses precomputed semantic annotations to improve its self-supervised video learning, although Struct2Depth is not a domain adaptation method as it only trains with Cityscapes [11] data, *i.e.*, it does not use KITTI for training. Struct2Depth also uses a different crop for Cityscapes. Table 5 shows that we achieve better accuracy than *Struct2Depth (M+R)*, which uses three frames at test time for refinement, whereas we only need a single image for inference.

4.2 Qualitative Results

Figure 3 shows predictions using DESC without stereo supervision. Compared to T^{2}Net [63], we find that our method contains less high-error regions due to the guidance provided by G_S, as shown in the upper-right wall of the predictions in the first row. The geometry of the

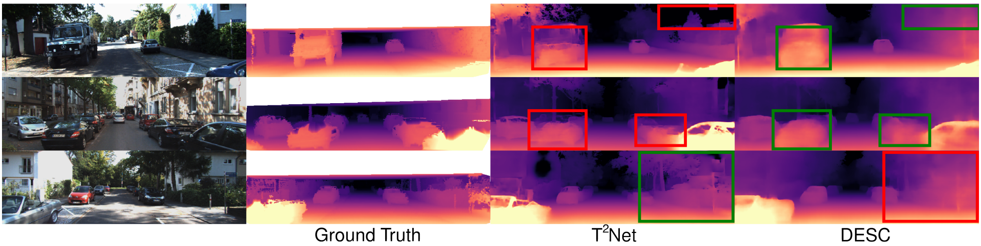

Figure 3: Qualitative results in KITTI for models trained in Virtual KITTI→KITTI. Ground truth depth is linearly interpolated for visualization. Green bounding boxes refer to areas of the prediction more accurate compared to the corresponding red bounding boxes.

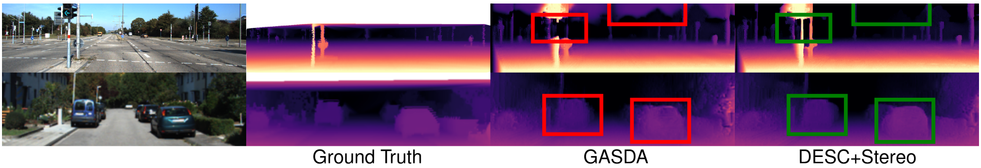

Figure 4: Qualitative results in KITTI for models trained in Virtual KITTI→KITTI with stereo supervision in KITTI. Bottom row corresponds to a center crop of the original image.

instances in our method tends to be complete, *e.g.*, the cars of the second row and the larger car in the first row, which has large missing parts in the T^{2}Net prediction. Figure 4 shows predictions for domain adaptation methods using stereo supervision in KITTI. Compared to GASDA, we observe a better recovery of fine structures, shown in the pole of the first row of Figure 4, and better predictions of further object instances, shown in the bottom row. DESC also predicts a better depth for the sky, as shown in the first row of Figure 4.

Limitations. Due to the consistency term with G_S, our method shows some loss of detail in fine structures compared to T^{2}Net [63], as shown in the last row of Figure 3. Additionally, DESC is more computationally demanding than T^{2}Net due to the added G_S. Furthermore, our method relies on the quality of the computed semantic data, hence in settings where the extracted annotations are of low quality, the performance of the method may degrade.

5 Conclusion

We proposed a method that leverages semantic annotations to improve the performance of a depth estimation model in a domain adaptation setting. We used the relationship between instance size and depth to provide pseudo-labels in the target domain. A segmentation map and an edge map were input to a second network, whose prediction was forced to be consistent with the prediction of the main network. These additions led to higher accuracy in the settings studied. In the Virtual KITTI to KITTI benchmark, we showed a 9.8% lower absolute relative error and a 23.6% lower squared relative error compared to the state-of-the-art. As we use automatically extracted semantic annotations, our method can be easily added to current approaches to improve their accuracy in a domain adaptation setting, as we showed in the improvement achieved with stereo self-supervision. Approaches aiming to reduce the detail loss due to the enforced consistency of predictions could improve the method.

Acknowledgement. This work was supported by UK EPSRC funding EP/S032398/1.

References

[1] Amir Atapour-Abarghouei and Toby P. Breckon. Real-time monocular depth estimation using synthetic data with domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, 2018.

[2] Amir Atapour-Abarghouei and Toby P Breckon. Veritatem dies aperit-temporally consistent depth prediction enabled by a multi-task geometric and semantic scene understanding approach. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3373–3384, 2019.

[3] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1209–1218, 2018.

[4] John Canny. A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, (6):679–698, 1986.

[5] Vincent Casser, Soeren Pirk, Reza Mahjourian, and Anelia Angelova. Depth prediction without the sensors: Leveraging structure for unsupervised learning from monocular videos. In *Thirty-Third AAAI Conference on Artificial Intelligence (AAAI-19)*, 2019.

[6] Chaoqi Chen, Weiping Xie, Wenbing Huang, Yu Rong, Xinghao Ding, Yue Huang, Tingyang Xu, and Junzhou Huang. Progressive feature alignment for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 627–636, 2019.

[7] Po-Yi Chen, Alexander H Liu, Yen-Cheng Liu, and Yu-Chiang Frank Wang. Towards scene understanding: Unsupervised monocular depth estimation with semantic-aware representation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2624–2632, 2019.

[8] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020.

[9] Yuhua Chen, Wen Li, Xiaoran Chen, and Luc Van Gool. Learning semantic segmentation from synthetic data: A geometrically guided input-output adaptation approach. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1841–1850, 2019.

[10] Yun-Chun Chen, Yen-Yu Lin, Ming-Hsuan Yang, and Jia-Bin Huang. Crdoco: Pixel-level domain transfer with cross-domain consistency. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1791–1800, 2019.

[11] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3213–3223, 2016.

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[13] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. Carla: An open urban driving simulator. In *Proceedings of the Conference on Robot Learning (CoRL)*, pages 1–16, 2017.

[14] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 2366–2374, 2014.

[15] Geoffrey French, Michal Mackiewicz, and Mark Fisher. Self-ensembling for visual domain adaptation. In *Proceedings of the International Conference on Learning Representations (ICLR)*, number 6, 2018.

[16] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. Deep Ordinal Regression Network for Monocular Depth Estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[17] Adrien Gaidon, Qiao Wang, Yohann Cabon, and Eleonora Vig. Virtual worlds as proxy for multi-object tracking analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4340–4349, 2016.

[18] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 1180–1189, 2015.

[19] Ravi Garg, Vijay Kumar BG, Gustavo Carneiro, and Ian Reid. Unsupervised cnn for single view depth estimation: Geometry to the rescue. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 740–756, 2016.

[20] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3354–3361. IEEE, 2012.

[21] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2019.

[22] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 270–279, 2017.

[23] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow. Digging into self-supervised monocular depth estimation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 3828–3838, 2019.

[24] Vitor Guizilini, Rui Hou, Jie Li, Rares Ambrus, and Adrien Gaidon. Semantically-guided representation learning for self-supervised monocular depth. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2020.

[25] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[26] Lei He, Guanghui Wang, and Zhanyi Hu. Learning depth from single images with deep neural network embedding focal length. *IEEE Transactions on Image Processing*, 27 (9):4676–4689, 2018.

[27] Heiko Hirschmuller. Stereo processing by semiglobal matching and mutual information. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 30(2): 328–341, 2007.

[28] 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. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2018.

[29] Junjie Hu, Yan Zhang, and Takayuki Okatani. Visualization of convolutional neural networks for monocular depth estimation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 3869–3878, 2019.

[30] Yu-Kai Huang, Tsung-Han Wu, Yueh-Cheng Liu, and Winston H Hsu. Indoor depth completion with boundary consistency and self-attention. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV) Workshops*, pages 0–0, 2019.

[31] Jianbo Jiao, Ying Cao, Yibing Song, and Rynson Lau. Look deeper into depth: Monocular depth estimation with semantic booster and attention-driven loss. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 53–69, 2018.

[32] Diederick P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015.

[33] Alexander Kirillov, Ross Girshick, Kaiming He, and Piotr Dollar. Panoptic feature pyramid networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[34] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9404–9413, 2019.

[35] Jogendra Nath Kundu, Nishank Lakkakula, and R Venkatesh Babu. Um-adapt: Unsupervised multi-task adaptation using adversarial cross-task distillation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1436–1445, 2019.

[36] Iro Laina, Christian Rupprecht, Vasileios Belagiannis, Federico Tombari, and Nassir Navab. Deeper depth prediction with fully convolutional residual networks. In *2016 Fourth international conference on 3D vision (3DV)*, pages 239–248. IEEE, 2016.

[37] John Lambert, Liu Zhuang, Ozan Sener, James Hays, and Vladlen Koltun. MSeg: A composite dataset for multi-domain semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[38] Wenbin Li, Sajad Saeedi, John McCormac, Ronald Clark, Dimos Tzoumanikas, Qing Ye, Yuzhong Huang, Rui Tang, and Stefan Leutenegger. Interiornet: Mega-scale multi-sensor photo-realistic indoor scenes dataset. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2018.

[39] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2041–2050, 2018.

[40] 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. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 740–755, 2014.

[41] Keng-Chi Liu, Yi-Ting Shen, Jan P Klopp, and Liang-Gee Chen. What synthesis is missing: Depth adaptation integrated with weak supervision for indoor scene parsing. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 7345–7354, 2019.

[42] Yue Luo, Jimmy Ren, Mude Lin, Jiahao Pang, Wenxiu Sun, Hongsheng Li, and Liang Lin. Single view stereo matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 155–163, 2018.

[43] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2794–2802, 2017.

[44] Nikolaus Mayer, Eddy Ilg, Philip Hausser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[45] Moritz Menze and Andreas Geiger. Object scene flow for autonomous vehicles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3061–3070, 2015.

[46] Arsalan Mousavian, Hamed Pirsiavash, and Jana Košecká. Joint semantic segmentation and depth estimation with deep convolutional networks. In *2016 Fourth International Conference on 3D Vision (3DV)*, pages 611–619. IEEE, 2016.

[47] Jogendra Nath Kundu, Phani Krishna Uppala, Anuj Pahuja, and R Venkatesh Babu. Adadepth: Unsupervised content congruent adaptation for depth estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2656–2665, 2018.

[48] Andrea Pilzer, Stephane Lathuiliere, Nicu Sebe, and Elisa Ricci. Refine and distill: Exploiting cycle-inconsistency and knowledge distillation for unsupervised monocular depth estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[49] Andrea Pilzer, Stéphane Lathuilière, Dan Xu, Mihai Marian Puscas, Elisa Ricci, and Nicu Sebe. Progressive fusion for unsupervised binocular depth estimation using cycled networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2019.

[50] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 234–241, 2015.

[51] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[52] Subhankar Roy, Aliaksandr Siarohin, Enver Sangineto, Samuel Rota Bulo, Nicu Sebe, and Elisa Ricci. Unsupervised domain adaptation using feature-whitening and consensus loss. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9471–9480, 2019.

[53] Kuniaki Saito, Yoshitaka Ushiku, and Tatsuya Harada. Asymmetric tri-training for unsupervised domain adaptation. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 2988–2997, 2017.

[54] Mehdi Sajjadi, Mehran Javanmardi, and Tolga Tasdizen. Regularization with stochastic transformations and perturbations for deep semi-supervised learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1163–1171, 2016.

[55] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 746–760, 2012.

[56] 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. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7472–7481, 2018.

[57] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7167–7176, 2017.

[58] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019.

[59] Junyuan Xie, Ross Girshick, and Ali Farhadi. Deep3d: Fully automatic 2d-to-3d video conversion with deep convolutional neural networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 842–857, 2016.

[60] Zhichao Yin and Jianping Shi. Geonet: Unsupervised learning of dense depth, optical flow and camera pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1983–1992, 2018.

[61] Zhenyu Zhang, Zhen Cui, Chunyan Xu, Zequn Jie, Xiang Li, and Jian Yang. Joint task-recursive learning for semantic segmentation and depth estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 235–251, 2018.

[62] Shanshan Zhao, Huan Fu, Mingming Gong, and Dacheng Tao. Geometry-aware symmetric domain adaptation for monocular depth estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9788–9798, 2019.

[63] Chuanxia Zheng, Tat-Jen Cham, and Jianfei Cai. T2net: Synthetic-to-realistic translation for solving single-image depth estimation tasks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 767–783, 2018.

[64] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G Lowe. Unsupervised learning of depth and ego-motion from video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1851–1858, 2017.