CD-DEPTH: UNSUPERVISED DOMAIN ADAPTATION FOR DEPTH ESTIMATION VIA CROSS DOMAIN INTEGRATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Despite the efficiency of data collecting for depth estimation in the synthetic environment, we cannot take full advantage of such benefit due to the distribution gap between the synthetic and the real world. In this paper, we introduce a new unsupervised domain adaptation framework, CD-Depth, for depth estimation to alleviate domain shift by extracting structure-consistent and domain-agnostic latents using following methods. (1) We propose domain-agnostic latent mapping which projects images from different domains to the shared latent space by removing redundant domain features for estimating monocular depth. (2) We also fuse visual signals from both RGB and latent domains to fully exploit multi domain information with adaptive-window-based cross-attention. Our proposed framework achieves state-of-the-art results in unsupervised domain adaptation for depth estimation both on indoor and outdoor datasets and produces better generalization performance on an unseen dataset.

1 INTRODUCTION

Training a depth estimation network in a supervised manner produces outstanding results with high accuracy, but it requires a large number of images paired with densely annotated depth labels. So, in recent depth estimation, self-supervised approaches have become mainstream, either utilizing stereo images or monocular sequences. However, self-supervised methods usually rely on photometric reprojection loss, which is only applicable to the static scene and susceptible to illumination change, dynamic objects, and so on.

Fortunately, we can easily obtain a large number of accurate depth labels in a synthetic world to train a neural network in a supervised setting. The problem is that it is difficult to render a photorealistic scene in the synthetic environment, resulting in the domain gap between the synthetic and the real world. Previous Unsupervised Domain Adaptation (UDA) methods for depth estimation (Atapour-Abarghouei & Breckon (2018); Zheng et al. (2018); Zhao et al. (2019; 2020b); Akada et al. (2022)) leverage unpaired style transfer networks with cyclic consistent loss, e.g. CycleGAN (Zhu et al., 2017), to generate photorealistic images from synthetic data and overcome domain shift while keeping their geometric structure. However, they cannot perfectly fetch images from one domain to another, and some undesired distortion occurs during style transfer. It is due to the unsupervised setting during training and the trade-off between geometric consistency and domain fidelity when utilizing cyclic consistency loss. Also, most of the UDA approaches for depth estimation exploit geometric features from a *single* domain, e.g. transferred target domain, which leads to imperfect depth results due to the inaccurate structural information during domain transfer or domain-specific irrelevant features for depth estimation.

In this paper, we introduce a Cross-Domain Depth estimation network (CD-Depth), which is the UDA framework with two main contributions, **Domain-Agnostic Latent Mapping (DALM)** and **Cross-Domain Disparity Network (CD-DispNet)**. (1) DALM alleviates the domain shift by detouring the direct domain transfer from one to the other but instead projecting images from different domains to a single shared latent domain $\mathcal{Z}$ as shown in Figure 1. There exist some prior arts (PNVR et al. (2020); Chen et al. (2021)) which also attempt to learn domain-generalized representation. However, the generalized representation from PNVR et al. (2020) remains within RGB

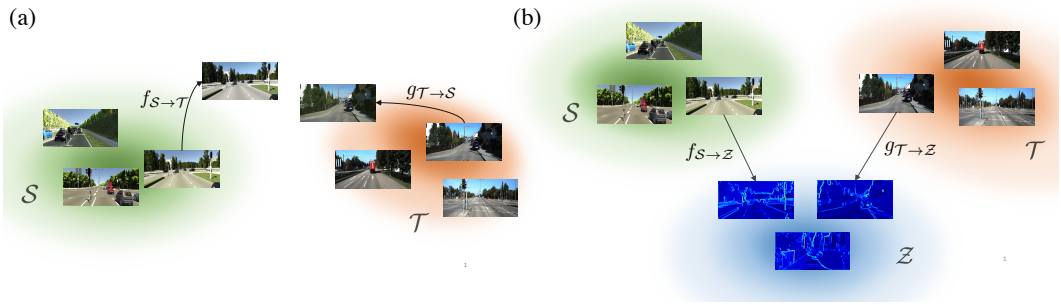


Figure 1: (a) Style transfer methods cannot perfectly transfer image from one to the other domain with its geometric structure consistent. (b) DALM maps the image from other domain to the shared latent space by removing domain relevant features.

domain which leads to little change between input and output images. Chen et al. (2021) learns domain-generalized representation for depth estimation, but it requires the encoder pre-trained with data that contains various image styles and densely annotated depth labels. On the contrary, our proposed DALM only requires unpaired images from the source (synthetic, $\mathcal{S}$) and the target (real, $\mathcal{T}$) domain to learn domain-agnostic projection via self-reconstruction and feature-level adversarial loss.

(2) CD-DispNet fuses visual signals from two different domains, the latent domain $\mathcal{Z}$ and the real domain $\mathcal{T}$, to reinforce cross-domain interaction and effectively exploit structural representation from both domains using cross-attention mechanism. We also utilize non-overlapping window-based attention for computation and memory efficiency due to the high resolution input. Previous models with window-based attention, e.g. Swin Transformer (Liu et al., 2021) and HRFormer (Yuan et al., 2021), fix the size of the attention window. However, the attention window with the fixed size has difficulty in capturing global context which induces performance degradation. Therefore, we propose an adaptive attention window whose size changes according to the feature map resolution. By doing so, the attention mechanism enables learning both local and global correlation between two different domains.

To sum up, our contribution can be summarized as below:

- We introduce an unsupervised domain adaptation framework for depth estimation, CD-Depth, which achieves high performance both in domain generalization and monocular depth estimation.
- We propose Domain-Agnostic Latent Mapping to alleviate domain shift between synthetic and real data by projecting images to the shared latent space without any additional data.
- We propose Cross-Domain Disparity Network with adaptive-window-based cross-attention mechanism so that the network can effectively fuse signals from different domains for depth estimation.
- Our CD-Depth outperforms the state-of-the-art unsupervised domain adaptation for depth estimation both in indoor and outdoor datasets and also generalize well in an unseen dataset compared to prior arts.

2 Related Work

2.1 Monocular Depth Estimation

Monocular Depth Estimation (MDE) plays a critical role in various computer vision applications such as AR (Augmented Reality), VR (Virtual Reality), autonomous driving, and robotics. In recent years, MDE has achieved satisfactory performance thanks to the exceptional development in deep learning. Especially, depth estimators trained in a supervised manner (Fu et al., 2018; Bhat et al., 2021) produce high-quality results with a huge amount of densely annotated depth labels. However, it is both expensive and inefficient to collect a large amount of data with paired depth labels for supervised learning. Fortunately, due to the large-scale virtual environment, it becomes popular to

train the depth estimation network with high-quality synthetic data in s supervised setting (Zheng et al., 2018; Zhao et al., 2019; PNVR et al., 2020; Chen et al., 2021). T2Net (Zheng et al., 2018) adopts cyclic consistency loss from CycleGAN (Zhu et al., 2017) and transfer the synthetic data to the realistic target domain. Following, GASDA (Zhao et al., 2019) exploits stereo images in the real-world for additional cues to improve the geometric consistency. Rather than transferring the synthetic image to the target domain, SharinGAN (PNVR et al., 2020) and S2R-DepthNet (Chen et al., 2021) transfer both synthetic and real-world images to the generalized space with pre-trained encoders.

The difference between aforementioned approaches and our method can be summarized into the following: whether they train the depth estimator in the target domain or the generalized domain, they exploit information only from a single domain. On the other hand, our proposed CD-Depth leverages geometric features of the data from *both* domains, i.e. target and generalized domain, with Cross-Domain Disparity Network (CD-DispNet) to generate structural-consistent latent vectors without domain-specific redundant features.

2.2 Unsupervised Domain Adaptation

Unsupervised Domain Adaptation (UDA) has been developed along with Generative Adversarial Networks (GAN) and cyclic consistency loss (Zhu et al., 2017; Kim et al., 2017; Hoffman et al., 2018). CycleGAN (Zhu et al., 2017) firstly proposes cyclic consistency loss with min-max optimization in the adversarial loss for unpaired image-to-image translation. Cycle consistency loss makes it possible to transfer image domain without paired data, but it is too restrictive that there exists only a difference in the color tone between input and output images with little change in the entire image. Following, CUT (Park et al. (2020)) points out such limitation in cyclic consistency and only employs contrastive learning based PatchNCE loss. StyleGAN (Karras et al., 2019; 2020; 2021) firstly proposes style-based image generation in the disentangled latent space and manipulates (or conditions) images on semantic levels. Recently, Denoising Diffusion Probabilistic Models (DDPM, Ho et al. (2020)) synthesize images iteratively with superior quality both in unconditional and conditional settings. ILVR (Choi et al., 2021) and SDEdit (Meng et al., 2022) introduce unpaired image-to-image translation leveraging only unconditional DDPM pre-trained on the source domain and the single reference image from the target domain. ILVR utilizes High-pass Filter (HPF) to remove low-frequency signals which contain semantic information in the DDPM and add low-frequency information from the reference image. SDEdit leverages generative Stochastic Differential Equation (SDE) to inject signals from the reference image via stroke-based guiding.

Unfortunately, it is not possible to perfectly transform the image from one domain to another while keeping its structural information due to the trade-off between the scene consistency and the domain fidelity. Therefore, rather than directly transferring the image from one to the other domain, we alternatively project the image to the shared latent space in which the distribution distance between the source domain is much closer than the distance between the source and the target RGB domains.

3 CD-Depth

In this section, we introduce our proposed unsupervised domain adaptation framework for depth estimation called, CD-Depth. CD-Depth consists of three different modules, Domain-Agnostic Latent Mapping, Single-Domain Disparity Network, and Cross-Domain Disparity Network. Each component takes part in generalizing domain representation, estimating depth, and fusing signals from individual domains, i.e., source $\mathcal{S}$, target $\mathcal{T}$, and latent domain $\mathcal{Z}$. The diagram of the overall framework is illustrated in Figure 2.

3.1 Domain-Agnostic Latent Mapping

Previously, a priority of unsupervised domain adaptation approaches for depth estimation (Zheng et al., 2018; Zhao et al., 2019; 2020b; Akada et al., 2022) was synthesizing the source domain image $I_{\mathcal{S}}$ indistinguishable from the target domain images $I_{\mathcal{T}}$. The aforementioned methods mostly adopt unpaired image-to-image translation algorithms with cyclic consistency loss proposed in CycleGAN (Zhu et al., 2017). Unfortunately, the constraint from cyclic consistency is too strong that it is insufficient to perfectly overcome discrepancy between two data distribution. Rather than trans-

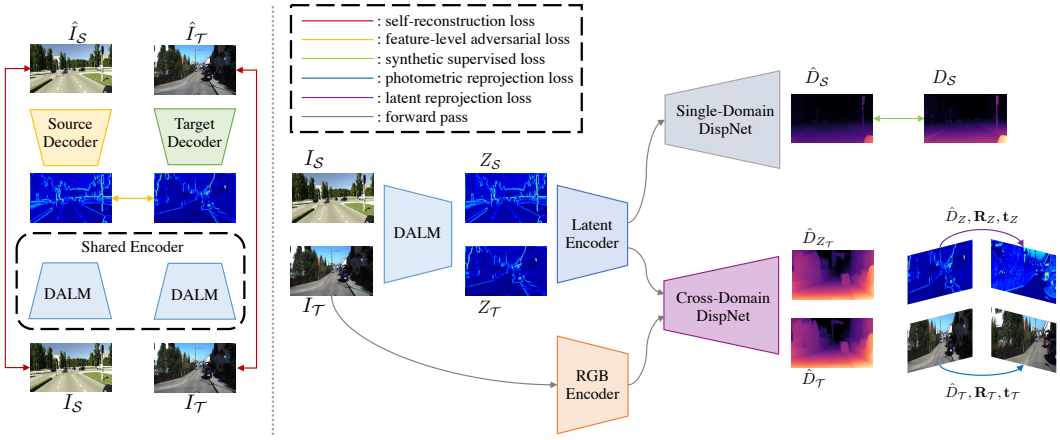


Figure 2: Overall CD-Depth Framework. It consists of three main parts: (1) domain mapping from RGB to latent, (2) single-domain depth estimation, and (3) cross-domain signal fusion and depth estimation. More details are demonstrated in Section 3.

ferring the domain from source to target domain, we propose Domain-Agnostic Latent Mapping (DALM), which projects images from both domains to a single shared latent space $\mathcal{Z}$. Unlike previous approaches which also generalize domain representation, (PNVR et al., 2020; Chen et al., 2021), DALM does not require additional data or depth labels, but it only employs images from the source and the target domain. For structural (scene) consistency, we leverage self-reconstruction loss so that the latent z from DALM represents intact structural information of the image to produce the same depth outputs. We design encoder-decoder architecture for self-reconstruction without skip-connection, and only the encoder $\mathcal{E}$ is leveraged during depth estimation. DALM adopts the shared encoder to remove domain-relevant features while maintaining the geometric context, and domain-specific decoders to recover the input image. The self-reconstruction loss $\mathcal{L}_{recon}$ is formulated as:

$$\begin{aligned} \hat{I}_{\mathcal{S}} = \mathcal{D}_{\mathcal{S}}(\mathcal{E}(I_{\mathcal{S}})),\ \hat{I}_{\mathcal{T}} = \mathcal{D}_{\mathcal{T}}(\mathcal{E}(I_{\mathcal{T}})), \\ \mathcal{L}_{recon} = \frac{1}{2}(|\hat{I}_{\mathcal{S}} - I_{\mathcal{S}}|_1 + |\hat{I}_{\mathcal{T}} - I_{\mathcal{T}}|_1), \end{aligned} \tag{1}$$

where $\mathcal{D}_{\mathcal{S}}$ and $\mathcal{D}_{\mathcal{T}}$ are domain-specific decoders for image reconstruction.

Along with self-reconstruction loss, we also utilize feature-level adversarial loss where the discriminator D distinguishes whether the latent z comes from the source or the target domain. With the feature-level adversarial loss, the encoder $\mathcal{E}$ can project the images from different domains to the same distribution and remove the domain-relevant image features which are redundant for estimating depth.

$$\begin{aligned} z_{\mathcal{S}} = \mathcal{E}(I_{\mathcal{S}}),\ z_{\mathcal{T}} = \mathcal{E}(I_{\mathcal{T}}) \\ \mathcal{L}_{gan} = \mathbb{E}[\log D(z_{\mathcal{T}})] + \mathbb{E}[\log(1 - D(z_{\mathcal{S}}))] \end{aligned} \tag{2}$$

Taken together, the objective loss for traing DALM is:

$$\mathcal{L}_z = \mathcal{L}_{recon} + \alpha_z \mathcal{L}_{gan} \tag{3}$$

where α_z is a hyperparameter to balance the projected latent between geometric consistency and domain indistinguishability.

3.2 Single-Domain Disparity Network

For domain-agnostic depth estimation, we first project the input image to the latent space $\mathcal{Z}$ with DALM. Then, the projected image propagates to the latent encoder $\psi_{\mathcal{Z}}$ to extract geometric features for depth estimation. With the output of $\psi_{\mathcal{Z}}$, Single-Domain Disparity Network (SD-DispNet) produces multi-scale depth outputs. As we have acquired synthetic images $I_{\mathcal{S}}$ paired with densely annotated depth label $D_{\mathcal{S}}$, we train SD-DispNet in a supervised manner. A synthetic supervised loss $\mathcal{L}_{syn}$ for single-domain disparity network f_θ is as below:

$$\begin{aligned} z_{\mathcal{S}} = \mathcal{E}(I_{\mathcal{S}}),\ \hat{D}_{\mathcal{S}} = f_\theta(\psi_{\mathcal{Z}}(z_{\mathcal{S}})), \\ \mathcal{L}_{syn} = |\hat{D}_{\mathcal{S}} - D_{\mathcal{S}}|_1. \end{aligned} \tag{4}$$

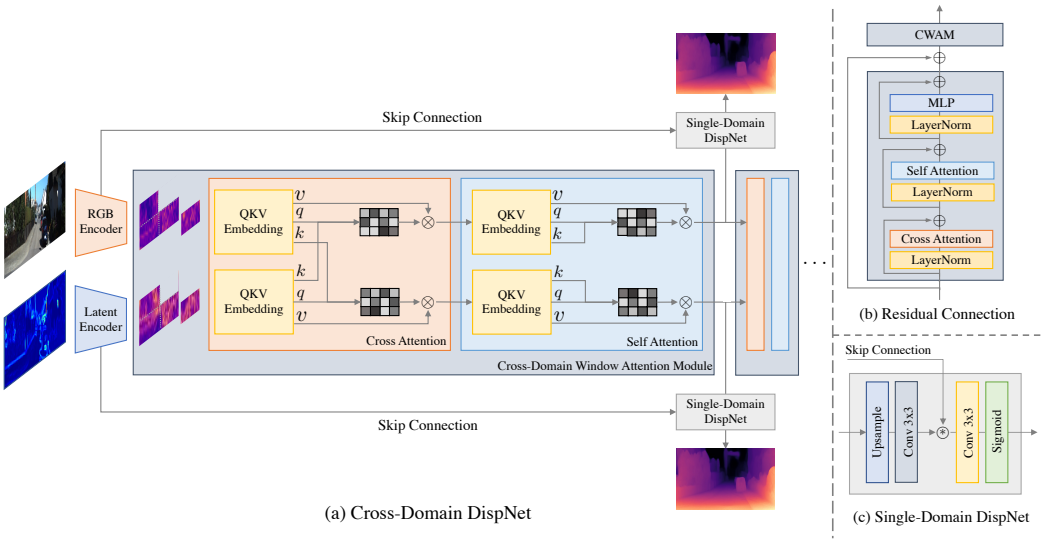


Figure 3: (a) Illustration of Cross-Domain Disparity Network with sequential cross-attention and self-attention functions. (b) Residual connection inside the Cross-Domain Window Attention Module. (c) The model architecture of Single-Domain Disparity Network.

3.3 Cross-Domain Disparity Network

Aside from synthetic depth labels, we also employ monocular sequences from $\mathcal{T}$ and $\mathcal{Z}$ as additional geometric cues. To leverage monocular sequences from different domains, we propose Cross-Domain Disparity Network (CD-DispNet), which produces domain-specific depths, $\hat{D}_{\mathcal{T}}$ and $\hat{D}_{\mathcal{Z}}$. We also utilize two pose estimation networks with the same architecture, $p_{\mathcal{T}}$ and $p_{\mathcal{Z}}$, for learning Structure-from-Motion (SfM) in different domains. Photometric reproejction loss $\mathcal{L}_{photo}$ (and latent reprojection loss $\mathcal{L}_{latent}$) calculates per-pixel intensity error between the warpped images (latents) and the adjacent frames (latents from adjacent frames). Each loss function is formulated as:

$$\begin{aligned} I_{t'\to t} &= I_{t'} \left\langle proj(\hat{D}_{\mathcal{T}}),\, \mathbf{R}_{\mathcal{T}},\, \mathbf{t}_{\mathcal{T}},\, \mathbf{K} \right\rangle \\ z_{t'\to t} &= z_{t'} \left\langle proj(\hat{D}_{\mathcal{Z}}),\, \mathbf{R}_{\mathcal{Z}},\, \mathbf{t}_{\mathcal{Z}},\, \mathbf{K} \right\rangle \end{aligned} \tag{5}$$

$$\begin{aligned} \mathcal{L}_{photo} &= \alpha_{photo} \left(\frac{1 - SSIM(I_t, I_{t'\to t})}{2} \right) + (1 - \alpha_{photo})(|I_t - I_{t'\to t}|_1) \\ \mathcal{L}_{latent} &= |z_t - z_{t'\to t}|_1 \\ \mathcal{L}_{proj} &= \mathcal{L}_{photo} + \alpha_{proj}\mathcal{L}_{latent} \end{aligned} \tag{6}$$

where $\mathbf{R}$, $\mathbf{t}$ denote rotation and translation from each domain-specific pose network, and $\mathbf{K}$ indicates intrinsic parameter of the camera. $proj(\cdot)$ and $\langle\cdot\rangle$ indicate 2D coordinates of input and sampling operator respectively. α_{photo} and α_{proj} are set to 0.85 and 0.001 respectively.

CD-DispNet consists of two modules, Cross-domain Window Attention Module (CWAM) and SD-DispNet, each for aggregating information from two different domains and estimating depth using features from CWAM as shown in Figure 3.

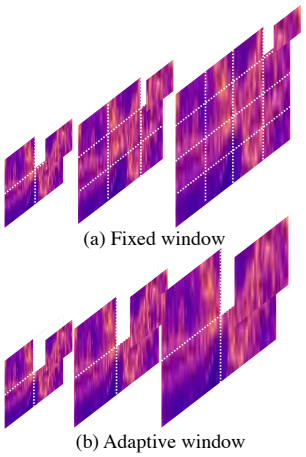
Images $I_{\mathcal{T}}$ from $\mathcal{T}$ and latents $z_{\mathcal{T}}$ from $\mathcal{Z}$ separately pass through the RGB encoder $\psi_{\mathcal{T}}$ and the latent encoder $\psi_{\mathcal{Z}}$. These two encoders are followed by CWAM, which encourages cross-domain learning for depth estimation by utilizing a sequential attention mechanism. First, the cross-attention effectively fuses the visual signals from $\psi_{\mathcal{T}}$ and $\psi_{\mathcal{Z}}$ to enforce the interaction between two different domains. Then, the output of the cross-attention function sequentially propagates to the self-attention for exploring the inner-domain representation. The cross-attention and the self-attention are implemented with $\text{Attn}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d}}\right) \cdot V$ as:

$$\begin{aligned} &\text{CrossAttn}_{RGB} = \text{Attn}(Q_{RGB}, K_Z, V_{RGB}), \text{CrossAttn}_Z = \text{Attn}(Q_Z, K_{RGB}, V_Z) \\ &\text{SelfAttn}_{RGB} = \text{Attn}(Q_{RGB}, K_{RGB}, V_{RGB}),\ \text{SelfAttn}_Z = \text{Attn}(Q_Z, K_Z, V_Z) \end{aligned} \tag{7}$$

where

$$Q_{RGB} = W_Q \cdot \psi_{\mathcal{T}}(I_{\mathcal{T}}), K_{RGB} = W_K \cdot \psi_{\mathcal{T}}(I_{\mathcal{T}}), V_{RGB} = W_V \cdot \psi_{\mathcal{T}}(I_{\mathcal{T}}),$$
$$Q_Z = W_Q \cdot \psi_{\mathcal{T}}(z_{\mathcal{T}}), K_Z = W_K \cdot \psi_{\mathcal{Z}}(z_{\mathcal{T}}), V_Z = W_V \cdot \psi_{\mathcal{T}}(z_{\mathcal{T}}), \tag{8}$$

and d indicates the number of attention head for normalization. We note that the linear transformations for query, key, and value are not distinguished for notation simplicity, but they are specific for the input domain in the actual implementation.



(a) Fixed window

(b) Adaptive window

Figure 4: Fixed window vs Adaptive window

Additionally, we deploy non-overlapping window-based attention (Huang et al., 2019) rather than full attention to harness computational and memory efficiency for the high-resolution input image. Unfortunately, naïve window-based attention suffers from performance degradation due to the insufficiency of capturing global context. So, we utilize two different mechanisms, i.e. Shifted window and Adaptive window, to alleviate such undesired property. (1) We leverage the shifted window partitioning strategy (Liu et al., 2021) so that the attention window can interact with adjacent windows and broadcast information from one window to another. (2) We propose an **adaptive attention window** in which the size of the attention window changes according to its feature map resolution. The choice for the size of the fixed attention window (Liu et al., 2021; Yuan et al., 2021; Li et al., 2022a;b) is quite limited in that it should be the common divisor of the width and the height of the smallest feature map. On the other hand, in the proposed method, the size of the adaptive attention window grows as the feature map gets larger. Specifically, the adaptive attention window grows by keeping the ratio of the area of the attention window to the area of the feature map the same as shown in Figure 4. Adaptive attention window democratizes the window size from the smallest feature map and leads to larger attention field, which enables learning the representation of both local and global correlation within the window.

3.4 Inference

During the inference phase, we aim to estimate depth from both a single monocular image $I_{\mathcal{T}}$ in the real world and its corresponding projected latent $z_{\mathcal{T}}$ with resultant models, SD-DispNet f_θ and CD-DispNet g_θ. The final prediction is the weighted sum of outputs from each function as below:

$$\hat{D}_{S_{\mathcal{Z}}} = f_\theta(\psi_{\mathcal{Z}}(z_{\mathcal{T}})),\ \hat{D}_{C_{\mathcal{T}}}, \hat{D}_{C_{\mathcal{Z}}} = g_\theta(\psi_{\mathcal{T}}(I_{\mathcal{T}}), \psi_{\mathcal{Z}}(z_{\mathcal{T}})),$$
$$\hat{D} = \alpha_1 \hat{D}_{S_{\mathcal{Z}}} + \alpha_2 \hat{D}_{C_{\mathcal{T}}} + \alpha_3 \hat{D}_{C_{\mathcal{Z}}}, \tag{9}$$

where $\alpha_1 + \alpha_2 + \alpha_3 = 1$.

4 Experiment

In this section, we present the effectiveness of our proposed framework, CD-Depth, on the challenging datasets for a single-view depth estimation. We perform an extensive experiment on KITTI (Geiger et al., 2012) dataset and NYU Depth v2 (Nathan Silberman & Fergus, 2012) dataset as outdoor and indoor environments, respectively. In terms of domain generalization, we experiment CD-Depth on the unseen real-world dataset, Make3D (Saxena et al., 2008). We also validate that (1) **DALM** outperforms previous unpaired domain adaptation (UDA) and unpaired domain generalization (UDG) methods, (2) **cross-attention** generates better performance in multi-domain settings than self-attention, and (3) **adaptive attention window** achieves high-quality depth outputs compared to fixed attention window.

4.1 KITTI Dataset

In the outdoor scenario, we adopt KITTI dataset as a realistic target domain and Virtual KITTI (vKITTI) (Gaidon et al., 2016) as a synthetic source domain for evaluating unpaired domain adaptation performance for depth estimation. For fair evaluation, all the images in KITTI are resized to 640×192, and the regions with the ground truth depth over the max value (80 m) are masked out.

Table 1: Quantitative Result on KITTI. Best results are in **bold**, and the second best are underlined. Sup column indicates the supervision level during training and Data column denote D: Depth Supervision, M: Monocular self-supervision, S: Stereo self-supervision, vK: synthetic supervision with virtual KITTI respectively. Methods with unsupervised domain adaptation are shaded in gray.

Method	Sup	Data	Lower is Better				Higher is Better		
			Abs Rel	Sq Rel	RMSE	RMSE_{log}	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Fu et al. (2018)	Yes	D	0.072	0.307	2.727	0.120	0.932	0.984	0.994
Lee et al. (2019)	Yes	D	0.096	-	2.756	0.059	0.956	0.993	0.998
Zhou et al. (2017)	No	M	0.215	1.515	7.156	0.270	0.678	0.885	0.957
Godard et al. (2019)	No	M	0.115	0.882	4.701	0.190	0.879	0.961	0.982
Guizilini et al. (2020)	No	M	0.111	0.785	4.601	0.189	0.878	0.960	0.982
Shu et al. (2020)	No	M	0.109	0.923	4.819	-	0.886	-	0.981
Wang et al. (2021)	No	M	0.109	0.779	4.641	0.186	0.883	0.962	0.982
Hui (2022)	No	M	0.108	0.710	4.513	0.183	0.884	0.964	0.983
Kundu et al. (2018)	Semi	vK+D	0.167	1.275	5.578	0.237	0.771	0.922	0.971
Zheng et al. (2018)	No	vK	0.174	1.410	6.046	0.253	0.754	0.916	0.966
Zhao et al. (2019)	No	vK+S	0.149	1.003	4.995	0.227	0.824	0.941	0.973
Zhao et al. (2020b)	No	vK	0.145	1.003	5.333	0.229	0.811	0.934	0.972
PNVR et al. (2020)	No	vK+S	0.116	0.939	5.068	0.203	0.850	0.948	0.978
Chen et al. (2021)	No	vK	0.165	1.351	5.695	0.236	0.781	0.931	0.972
Guizilini et al. (2021)	No	vK+M	0.114	0.875	4.808	-	0.871	-	-
Akada et al. (2022)	No	vK	0.168	1.228	5.498	0.235	0.771	0.921	0.973
Ours	No	vK+M	**0.106**	**0.771**	**4.520**	**0.182**	**0.890**	**0.964**	**0.983**

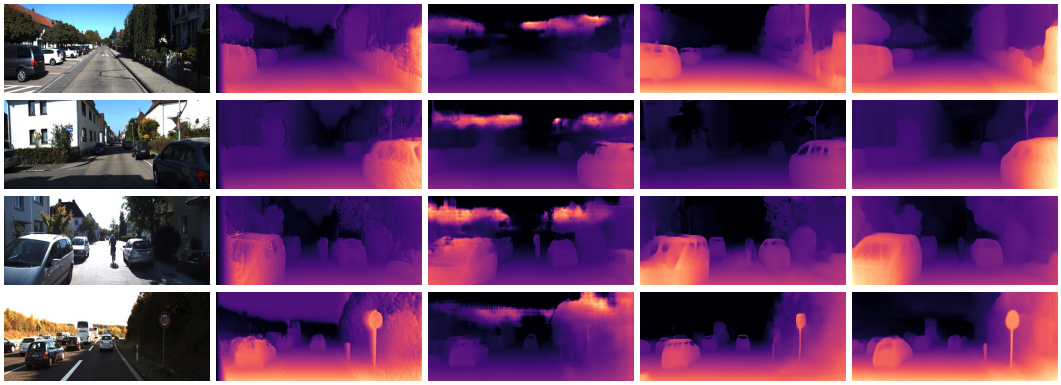


Figure 5: Qualitative Results. Our CD-Depth produces better results with distinct boundaries compared to prior state-of-the-art domain adaptation methods (GASDA (Zhao et al., 2019), ARC (Zhao et al., 2020b), and S2R-Depth (Chen et al., 2021)).

In Table 1, we report results of our prposed method compared to prior state-of-the-art algorithms in self-supervised depth estimation and unpaired domain adaptation on 697 test images from Eigen et al. (2014) train/test split. We observe that CD-Depth produces convincing improvements over both self-supervised and unpaired domain adaptation algorithms in most of the metrics. Specifically, our model achieves 13% better Sq Rel error and 7% better RMSE error compared to Guizilini et al. (2021) which also utilizes both synthetic depth labels from vKITTI and monocular sequences from KITTI for domain adaptation in depth estimation. Apart from quantitative results, we present qualitative results on KITTI compared to the recent domain adaptation algorithms in Figure 5. Our method exhibits relatively edge-consistent, smooth invariant depth outputs with fewer holes in reflective surface.

4.2 Ablation Study

For better understanding, we ablate the components of CD-Depth one by one to figure out how each component contributes to the model performance.

Table 2: Comparison on FID score. DALM produces the smallest image distribution distance between two domains compared to state-of-the-art UDA (CycleGAN (Zhu et al., 2017), CUT (Park et al., 2020), ILVR (Choi et al., 2021), and SDEdit (Meng et al., 2022)) and UDG (SharinGAN (PNVR et al., 2020) and S2R-Depth (Chen et al., 2021)) algorithms. In the sampling hyperparameters for diffusion-based models, we set the downsampling factor N for ILVR as 4, and t_0 for SDEdit as 0.4.

Type	UDA				UDG		
Method	CycleGAN	CUT	ILVR	SDEdit	SharinGAN	S2R-Depth	Ours
FID (↓)	98.39	62.27	82.91	81.78	101.39	47.00	**43.96**

Table 3: Ablation. The comparison between different modifications of CD-Depth in (a) domain adaptation strategies and (b) attention mechanisms.

	Method	Lower is Better				Higher is Better		
		Abs Rel	Sq Rel	RMSE	RMSE_{log}	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
(a)	RGB Only	0.115	0.873	4.747	0.191	0.881	0.962	0.982
	+CycleGAN	0.109	0.783	4.537	0.182	0.885	**0.964**	**0.983**
	+CUT	0.109	0.806	4.603	0.185	0.886	0.963	0.982
	+DALM	**0.106**	**0.771**	**4.520**	**0.182**	**0.890**	**0.964**	**0.983**
(b)	Baseline	0.109	0.803	4.581	0.185	0.884	0.963	**0.983**
	+WSA (fixed)	0.108	0.796	4.538	0.183	0.886	0.963	**0.983**
	+CDWA (fixed)	0.108	0.796	4.556	0.183	0.888	**0.964**	**0.983**
	+CDWA (adaptive)	**0.106**	**0.771**	**4.520**	**0.182**	**0.890**	**0.964**	**0.983**

DALM. First, we demonstrate how DALM generalizes the domain well compared to both the previous unpaired domain adaptation (UDA) and unpaired domain generalization (UDG) methods. We quantitatively evaluate the domain adaptation/generalization performance by leveraging FID (Fréchet Inception Distance) score (Heusel et al., 2017) which estimates the visual quality and calculates the distance between two image distributions. Table 2 reports FID score of UDA and UDG algorithms compared to our DALM. We measure the FID score in the target domain $\mathcal{T}$ for UDA methods and in the generalized domain $\mathcal{Z}$ for UDG methods. Our method achieves the best result in FID score compared to state-of-the-art domain adaptation/generalization approaches. It indicates that generalizing the domain with DALM is much effective than directly transferring the source domain $\mathcal{S}$ to the target domain $\mathcal{T}$. In Table 3(a), we experiment how domain adaptation/generalization performance is relevant to the depth estimation performance. We observe that our DALM, which achieves the highest score in domain adaptation/generalization, also shows the best result in depth estimation. Interestingly, CUT achieves better results in FID score compared to CycleGAN, but it cannot outperform CycleGAN-based depth estimation because of the distortion in structural information during domain transfer. It indicates that not only the domain similarity but also the geometric consistency plays an important role in depth estimation with different domain data.

Cross Domain Window Attention. In Table 3(b), we manipulate the attention module in Cross-Domain Window Attention (CDWA) by setting Baseline as DALM domain generalization and no attention module in the depth decoder. To verify how effective the cross-attention is compared to self-attention, we replace the cross-attention mechanism in CDWA with Window-based Self-Attention (WSA). We report the quantitative performance drop when we utilize self-attention rather than the cross-attention module. It is due to the restricted feature propagation as the signals from two different domains, $\mathcal{T}$ and $\mathcal{Z}$, cannot interact with each other when we solely leverage self-attention module. We also demonstrate that the performance of the adaptive attention window achieves better results compared to the fixed attention window. It is attributed to the property of adaptive window which represents global context information from the larger receptive field within the attention window. To sum up, the model which combines the proposed concepts, i.e. DALM, cross domain attention and adaptive attention window, altogether achieves the best results.

4.3 NYU V2 DATASET

For evaluating CD-Depth on the indoor environment, we choose Scene-RGBD (McCormac et al., 2017) as synthetic dataset, and NYU Depth v2 (Nathan Silberman & Fergus, 2012) as real-world

Table 4: Quantitative results on NYU Depth v2 dataset.

Method	Lower is Better		Higher is Better		
	Abs Rel	RMSE	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Zhou et al. (2017)	0.208	0.712	0.674	0.900	0.968
Godard et al. (2019)	0.160	0.601	0.767	0.949	0.988
Bian et al. (2019)	0.147	0.536	0.804	0.950	0.986
Zhao et al. (2020a)	0.189	0.686	0.701	0.912	0.978
Ji et al. (2021)	<u>0.134</u>	<u>0.526</u>	<u>0.823</u>	<u>0.958</u>	**0.989**
Ours	**0.133**	**0.498**	**0.831**	**0.959**	**0.989**

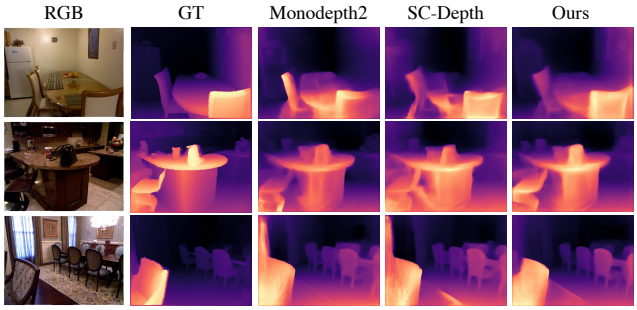


Figure 6: Qualitative results on NYU Depth v2 compared to state-of-the-art indoor self-supervised depth estimation algorithms, i.e. Monodepth2 (Godard et al., 2019) and SC-Depth (Bian et al., 2019).

dataset. The results of the Figure 6 and Table 4 are reported on the NYU Depth v2 official test split, and the input images are resized to 320×256 during training and inference for a fair evaluation. Our CD-Depth outperforms existing indoor self-supervised depth estimation algorithms in all the metrics in quantitative results. Also, in qualitative results, our method produces the best results which are the most similar to the ground truth depth labels with distinct boundaries between the object and the background.

4.4 MAKE3D DATASET

In Table 5, we report the domain generalization performance for depth estimation on the unseen outdoor dataset, Make3D. Train column indicates whether the model is trained on Make3D dataset with the depth labels in a supervised manner. We observe that our CD-Depth achieves the best result in every error metrics compared to state-of-the-art UDA algorithms and shows on par performance with supervised approaches.

Table 5: Quantitative results on Make3D.

Method	Train	Lower is Better		
		Abs Rel	Sq Rel	RMSE
Karsch et al. (2014)	Yes	0.398	4.723	7.801
Laina et al. (2016)	Yes	0.198	1.655	5.461
Kundu et al. (2018)	Yes	0.452	5.71	9.559
Zhao et al. (2019)	No	0.403	6.709	10.424
PNVR et al. (2020)	No	0.377	4.900	8.388
Zhao et al. (2020b)	No	0.516	8.009	10.031
Chen et al. (2021)	No	0.656	11.664	12.917
Ours	No	**0.306**	**3.098**	**6.884**

5 CONCLUSION

In this paper, we introduce a new unsupervised domain adaptation framework for depth estimation, CD-Depth. It enables producing structurally consistent and domain-agnostic features from different domains and effectively integrate them with the following ideas. We present DALM which projects images from multiple domains to a single shared latent space for alleviating the domain shift by removing domain-specific redundant components for depth estimation. We also propose an adaptive-window-based cross-attention module to effectively fuse signals from different domains and reinforce their interaction. Our CD-Depth outperforms prior state-of-the-art methods for depth estimation on both outdoor and indoor environments and achieves the best result on the unseen dataset.

REFERENCES

Hiroyasu Akada, Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. Self-supervised learning of domain invariant features for depth estimation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3377–3387, 2022.

Amir Atapour-Abarghouei and Toby P Breckon. Real-time monocular depth estimation using synthetic data with domain adaptation via image style transfer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2800–2810, 2018.

Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. Adabins: Depth estimation using adaptive bins. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4009–4018, 2021.

Jiawang Bian, Zhichao Li, Naiyan Wang, Huangying Zhan, Chunhua Shen, Ming-Ming Cheng, and Ian Reid. Unsupervised scale-consistent depth and ego-motion learning from monocular video. *Advances in neural information processing systems*, 32, 2019.

Xiaotian Chen, Yuwang Wang, Xuejin Chen, and Wenjun Zeng. S2r-depthnet: Learning a generalizable depth-specific structural representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3034–3043, 2021.

Jooyoung Choi, Sungwon Kim, Yonghyun Jeong, Youngjune Gwon, and Sungroh Yoon. Ilvr: Conditioning method for denoising diffusion probabilistic models. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14347–14356. IEEE Computer Society, 2021.

David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. *Advances in neural information processing systems*, 27, 2014.

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*, pp. 2002–2011, 2018.

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*, pp. 4340–4349, 2016.

Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, pp. 3354–3361. IEEE, 2012.

Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow. Digging into self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3828–3838, 2019.

Vitor Guizilini, Rares Ambrus, Sudeep Pillai, Allan Raventos, and Adrien Gaidon. 3d packing for self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2485–2494, 2020.

Vitor Guizilini, Jie Li, Rareș Ambruș, and Adrien Gaidon. Geometric unsupervised domain adaptation for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8537–8547, 2021.

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*, pp. 770–778, 2016.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *International conference on machine learning*, pp. 1989–1998. Pmlr, 2018.

Lang Huang, Yuhui Yuan, Jianyuan Guo, Chao Zhang, Xilin Chen, and Jingdong Wang. Interlaced sparse self-attention for semantic segmentation. *arXiv preprint arXiv:1907.12273*, 2019.

Tak-Wai Hui. Rm-depth: Unsupervised learning of recurrent monocular depth in dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1675–1684, 2022.

Pan Ji, Runze Li, Bir Bhanu, and Yi Xu. Monoindoor: Towards good practice of self-supervised monocular depth estimation for indoor environments. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12787–12796, 2021.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4401–4410, 2019.

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8110–8119, 2020.

Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. *Advances in Neural Information Processing Systems*, 34:852–863, 2021.

Kevin Karsch, Ce Liu, and Sing Bing Kang. Depth transfer: Depth extraction from video using non-parametric sampling. *IEEE transactions on pattern analysis and machine intelligence*, 36 (11):2144–2158, 2014.

Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jung Kwon Lee, and Jiwon Kim. Learning to discover cross-domain relations with generative adversarial networks. In *International conference on machine learning*, pp. 1857–1865. PMLR, 2017.

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*, pp. 2656–2665, 2018.

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)*, pp. 239–248. IEEE, 2016.

Jin Han Lee, Myung-Kyu Han, Dong Wook Ko, and Il Hong Suh. From big to small: Multi-scale local planar guidance for monocular depth estimation. *arXiv preprint arXiv:1907.10326*, 2019.

Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. Exploring plain vision transformer backbones for object detection. *arXiv preprint arXiv:2203.16527*, 2022a.

Yanghao Li, Chao-Yuan Wu, Haoqi Fan, Karttikeya Mangalam, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Mvitv2: Improved multiscale vision transformers for classification and detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4804–4814, 2022b.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10012–10022, 2021.

Reza Mahjourian, Martin Wicke, and Anelia Angelova. Unsupervised learning of depth and ego-motion from monocular video using 3d geometric constraints. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5667–5675, 2018.

John McCormac, Ankur Handa, Stefan Leutenegger, and Andrew J.Davison. Scenenet rgb-d: Can 5m synthetic images beat generic imagenet pre-training on indoor segmentation? 2017.

Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*, 2022.

Pushmeet Kohli Nathan Silberman, Derek Hoiem and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012.

Taesung Park, Alexei A Efros, Richard Zhang, and Jun-Yan Zhu. Contrastive learning for unpaired image-to-image translation. In *European conference on computer vision*, pp. 319–345. Springer, 2020.

Koutilya PNVR, Hao Zhou, and David Jacobs. Sharingan: Combining synthetic and real data for unsupervised geometry estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13974–13983, 2020.

Ashutosh Saxena, Min Sun, and Andrew Y Ng. Make3d: Learning 3d scene structure from a single still image. *IEEE transactions on pattern analysis and machine intelligence*, 31(5):824–840, 2008.

Chang Shu, Kun Yu, Zhixiang Duan, and Kuiyuan Yang. Feature-metric loss for self-supervised learning of depth and egomotion. In *European Conference on Computer Vision*, pp. 572–588. Springer, 2020.

Lijun Wang, Yifan Wang, Linzhao Wang, Yunlong Zhan, Ying Wang, and Huchuan Lu. Can scale-consistent monocular depth be learned in a self-supervised scale-invariant manner? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12727–12736, 2021.

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*, pp. 1983–1992, 2018.

Yuhui Yuan, Rao Fu, Lang Huang, Weihong Lin, Chao Zhang, Xilin Chen, and Jingdong Wang. Hrformer: High-resolution vision transformer for dense predict. *Advances in Neural Information Processing Systems*, 34:7281–7293, 2021.

Shanshan Zhao, Huan Fu, Mingming Gong, and Dacheng Tao. Geometry-aware symmetric domain adaptation for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9788–9798, 2019.

Wang Zhao, Shaohui Liu, Yezhi Shu, and Yong-Jin Liu. Towards better generalization: Joint depth-pose learning without posenet. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9151–9161, 2020a.

Yunhan Zhao, Shu Kong, Daeyun Shin, and Charless Fowlkes. Domain decluttering: Simplifying images to mitigate synthetic-real domain shift and improve depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3330–3340, 2020b.

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)*, pp. 767–783, 2018.

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*, pp. 1851–1858, 2017.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pp. 2223–2232, 2017.

A Evaluation Protocol

A.1 Median Scaling

During evaluation, we recover the scale of the predicted depth maps with the ratio of median values of the prediction $\hat{D}$ and its corresponding ground truth labels D_{gt} because the monocular self-supervised methods do not contain the absolute scale information.

$$\hat{D} = \hat{D} \times median(D_{gt})/median(\hat{D}) \tag{10}$$

A.2 Evaluation Metrics

We use the standard error and accuracy metrics defined below in order to quantitatively evaluate the depth performance of the network. "*Lower is better*" for the four error metrics (Abs Rel, RMSE, Sq Rel, RMSE log), and "*higher is better*" for the three accuracy metrics ($\delta < 1.25$, $\delta < 1.25^2$, $\delta < 1.25^3$).

- Abs Rel: $\frac{1}{|\mathcal{D}|}\Sigma_{d\in\mathcal{D}}|d^* - d|/d^*$;
- RMSE: $\sqrt{\frac{1}{|\mathcal{D}|}\Sigma_{d\in\mathcal{D}}||d^* - d||^2}$;
- Sq Rel: $\frac{1}{|\mathcal{D}|}\Sigma_{d\in\mathcal{D}}||d^* - d||^2/d^*$;
- RMSE log: $\sqrt{\frac{1}{|\mathcal{D}|}\Sigma_{d\in\mathcal{D}}||\log d^* - \log d||^2}$;
- $\delta < 1.25^i$: $\frac{1}{|\mathcal{D}|}|\max_{d\in\mathcal{D}}(\frac{d}{d^*}, \frac{d^*}{d}) < 1.25^i|$ $i \in \{1, 2, 3\}$;

where $\mathcal{D}$ indicates a set of all the predicted depth maps of an image. d and d^* denote the predicted depth maps and ground truth depth labels respectively, and $|\cdot|$ represents the number of the elements.

B Model Architecture

layer	input res	win size	feat size
1	6×20	2	256
2	12×40	4	128
3	24×80	8	64
4	48×160	16	32
5	96×320	32	16

Table 6: Model Specification of CD-DispNet

In this section, we demonstrate the details of the network architecture of CD-Depth. We adopt ResNet 50 (He et al., 2016) as feature extractor for Domain Agnostic Latent Mapping (DALM) encoder, RGB and latent encoder to extract domain-agnostic and structural consistent features from the shared latent space. Features from the image and the latnet encoder propagates to the Cross-Domain Disparity Network (CD-DispNet) which is composed of Adaptive Window-based Cross Attention (A-WCA), Shifted Adaptive Window-based Self Attention (SA-WSA), Layer Normalization (LN) and MLP layer. Each layer in the CD-DispNet is residually connected as below,

$$\begin{aligned}
\hat{z}_l &= \text{A-WCA}(\text{LN}(z_{l-1}), \text{LN}(z'_{l-1})) + z_{l-1}, \\
z_l &= \text{MLP}(\text{LN}(z_l)) + \hat{z}_l, \\
\hat{z}_{l+1} &= \text{SA-WSA}(\text{LN}(z_l)) + z_l, \\
z_{l+1} &= \text{MLP}(\text{LN}(\hat{z}_{l+1})) + \hat{z}_{l+1},
\end{aligned} \tag{11}$$

where $\hat{z}$, z and z' indicate the outputs of attention module, the MLP layer and the feature from the other domain respectively. The input resolution, the window size and the attention feature size following the layer are shown in the Table 6. These features are passed to the Single Domain Disparity Network (SD-DispNet) which consists 2 layers of ConvNet with the kernel size of 3 and sigmoid activation function to produce scaled depth outputs.

C ODOMETRY EVALUATION

Method	Seq. 09	Seq. 10	Mean	# frames
ORB-SLAM (full)	0.014	0.012	0.013	-
ORB-SLAM (short)	0.064	0.064	0.063	5
Mean Odom.	0.032	0.028	0.030	-
SfMLearner (Zhou et al., 2017)	0.021	0.020	0.021	5
Mahjourian et al. (2018)	0.013	0.012	0.012	3
GeoNet (Yin & Shi, 2018)	0.012	0.012	0.012	5
SfMLearner (Zhou et al., 2017)	0.050	0.034	0.042	2
Monodepth2 (Godard et al., 2019)	0.017	0.015	0.016	2
Ours (CD-Depth)	0.015	0.015	**0.015**	2

Table 7: Absolute Trajectory Error (ATE) on KITTI Odometry dataset in meters.

We additionally evaluate the performance of the pose estimation network which is the byproduct of estimating depth with Structure-from-Motion (SfM) and consecutive monocular images. The depth and pose esitmation networks are simultaneously trained with the 00-08 image sequences from KITII Odometry dataset and tested with 09, 10 sequences. In Table 7, we report the quantitative results of visual odometry performance with Absolute Trajectory Error (ATE) in meters. Following evaluation protocol in Zhou et al. (2017), ATE is calculated over average value of all overlapping 5 frame snippets in test sequences for a fair evaluation. Similar to Godard et al. (2019), we only leverage two image frames as input and produce a single transformation matrix $\mathbf{T}$ between the two images pair. As shown in the Table 7, our method outperforms other SfM methods which leverages 2 input frames in average. In indicates that the domain adaptation with CD-Depth also improves and preserves the pose estimation performance in the real-world while significantly increases the depth estimation performance in the real-world.

D QUALITATIVE RESULTS ON MAKE3D

We prove that our CD-Depth generalizes well to the unseen dataset Make3D (Saxena et al., 2008) compared to other state-of-the-art unsupervised domain adaptation for depth estimation algorithms in the Table 5 from the main paper. To strengthen our claim, we also provide qualitative results in Figure 7. CD-Depth produces the most similar depth outputs to the ground truth from unseen environment. Also, our method distinguishes the objects from the background better than competitive models.

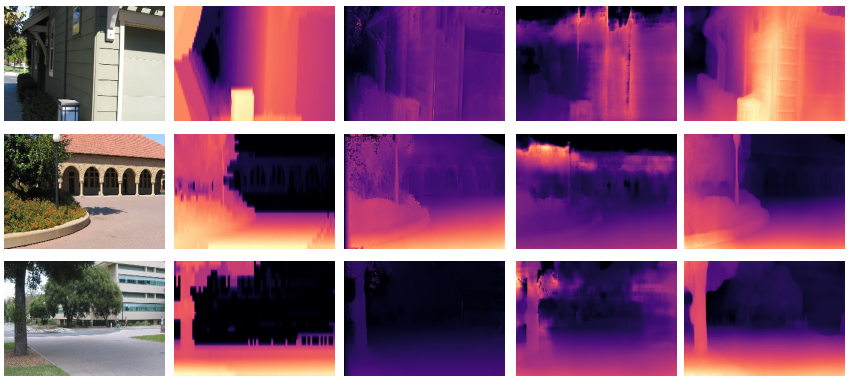


Figure 7: Qualitative results on Make3D

E EFFECT OF ADAPTIVE WINDOW IN QUALITATIVE RESULT

To clarify the effectiveness of adaptive attention window compared to prior fixed window, we not only evaluate the effectiveness of the shape of the attention window quantitatively in the Table 3 from the main paper, but also visualize the depth outputs from each ablated models in Figure 8. We compare three different ablated models which are the same as Table 3, i.e. self-attention mechanism with fixed-size window (F-WSA), cross-attention with fixed-size window (F-WCA), and cross-attention with adaptive window (A-WCA). By leveraging adaptive attention window, the network is able to capture small, but important objects during driving such as traffic light and the traffic signs as shown in Figure 8.

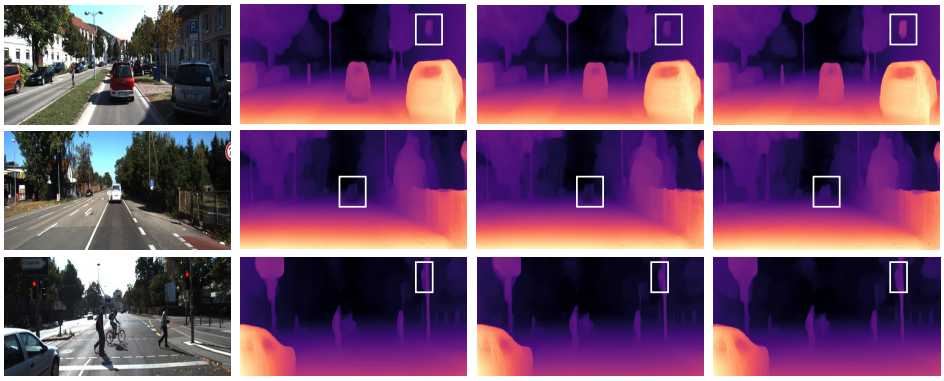


Figure 8: Qualitative results on KITTI test split. We demonstrate the effectiveness of cross-attention and adaptive window by comparing self-attention with fixed size window (F-WSA), cross-attention with fixed size window (F-WCA), and cross-attention with adaptive window (A-WCA)

F TRAINING DETAILS

We implement our model in official deep learning framework PyTorch, trained with 20 epochs by Adam optimizer. We use the batch size of 12 and the input image resolution is fixed to 640×192. We adopt learning rate of the optimizer as 10^{-4} for initial 15 epochs and becomes 10 times lower for the rest of the iterations. An image augmentation strategy of random horizontal flip is applied with the probability of 0.5, and color jittering, random brightness, contrast, saturation, and hue jittering, is also adopted with 50% of chance. All the experiments have been done on a single Nvidia RTX A6000 GPU, AMD Ryzen Threadripper 3960X, 6*32GB DDR4 RAM, and 1TB M.2 NVMe SSD. These specifications are sufficient to run all the experiments with the same configuration we have elaborated in the text.