SEDiff: Structure Extraction for Domain Adaptive Depth Estimation via Denoising Diffusion Models

Dongseok Shim[1] and H. Jin Kim[1,2]

[1] Interdisciplinary Program in AI, Seoul National University
[2] Artificial Intelligence Institute, Seoul National University (AIIS)

Abstract. In monocular depth estimation, it is challenging to acquire a large amount of depth-annotated training data, which leads to a reliance on synthetic datasets. However, the inherent discrepancies between the synthetic environment and the real-world result in a domain shift and sub-optimal performance. In this paper, we introduce SEDiff which firstly leverages a diffusion-based generative model to extract essential structural information for accurate depth estimation. SEDiff wipes out the domain-specific components in the synthetic data and enables structural-consistent style transfer to mitigate the performance degradation due to the domain gap. Extensive experiments demonstrate the superiority of SEDiff over state-of-the-art methods in various scenarios for domain-adaptive depth estimation.

Keywords: unsupervised domain adaptation · monocular depth estimation · denoising diffusion models

1 Introduction

Monocular depth estimation, which aims to estimate a dense depth map from a single image, plays a critical role in various computer vision and robotics applications such as VR/AR (Virtual/Augmented Reality), autonomous driving, and navigation. The advent of deep learning has revolutionized this field which enables generating high-quality depth predictions through supervised training on large datasets paired with densely annotated ground truth (GT) depth labels.

However, the challenge often lies in the arduous and costly acquisition of datasets containing the annotated depth labels that require depth sensors like LiDAR or RGB-D cameras. In response to this challenge, unsupervised learning algorithms have emerged as promising alternatives, utilizing geometric cues such as stereo images or monocular sequences, thereby mitigating the dependence on ground-truth depth labels. Unfortunately, the availability of stereo images and monocular image sequences is also not always guaranteed.

To overcome data limitations, some studies shift their focus to training neural depth estimators using data from graphics-based synthetic environments, where

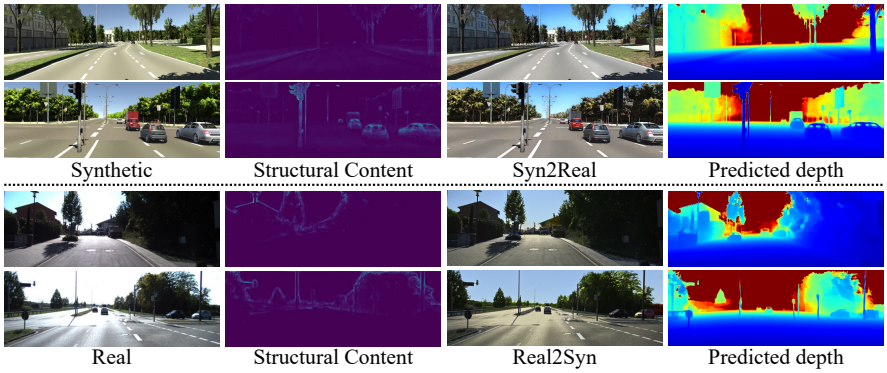

Fig. 1: Visualization of SEDiff outputs. We set the synthetic data as images from Virtual-KITTI [6] and real data from KITTI [7]. Syn2Real and Real2Syn denote synthetic-to-real depth-consistent style trasnfer and vice versa respectively.

obtaining densely annotated depth labels is more feasible. Despite this advantage, a critical issue arises due to inherent disparities between synthetic environments and the real-world, i.e., domain gap, resulting in sub-optimal performance of neural depth estimators when they are deployed in real-world applications.

In recent studies, a compelling solution has emerged to address the challenge of domain shift in depth estimation. Researchers propose a novel domain adaptation strategy that involves the style transfer of synthetic images to closely resemble their real-world counterparts [42, 43]. On the other hand, some approaches adopt a dual-domain projection technique, wherein images from synthetic and real domains are mapped onto a unified, shared domain [2, 27]. The underlying objective for these methods is to narrow the domain gap between training and inference environments, facilitating more accurate depth estimation. The effectiveness of these methods hinges on their ability to adeptly disentangle domain-specific style components irrelevant for depth estimation from synthetic images, while preserving structural content for accurate depth estimation.

In this paper, we introduce a novel domain adaptive depth estimation framework, SEDiff, which employs a diffusion-based generative model to extract the domain-invariant structural content for depth estimation. By incorporating the domain-specific learnable style parameters and the alternate adaptive instance normalization (AdaIN) layers into the U-Net-based denoising network, SEDiff enables disentangling domain-specific style components which only aggravates domain shift from the input image.

Furthermore, SEDiff achieves high-quality depth-consistent style transfer, presenting a distinct advantage in estimating attention maps for domain-invariant structural contents with style-transferred images from synthetic to the real domain. The resulting attention map, based on this depth-consistent style transfer, contributes to the performance of domain adaptive depth estimation alongside the domain-invariant structure extraction.

It is important to note that our proposed framework, SEDiff, harnesses the power of structure-style disentanglement in diffusion-based models restrictively

during the training phase. This strategic approach ensures that the computational complexity and prolonged inference times associated with the iterative denoising process of diffusion-based models do not affect the inference speed of our domain adaptive depth estimation framework.

We demonstrate the effectiveness of SEDiff in domain adaptive depth estimation through extensive experimentation across both indoor and outdoor scenarios. In addition, we extend the applicability of our method to scenarios where real-world geometric cues, such as stereo images, are available. We also validate better domain generalization performance of SEDiff on unseen dataset compared to competitive domain adaptation methods.

In short, our contribution can be summarized as follows:

- We propose SEDiff which firstly integrates diffusion models into the domain-adaptive depth estimation framework, facilitating the extraction of domain-invariant structural content from synthetic images.
- SEDiff leverages depth-consistent style transfer to estimate attention maps by minimizing the domain gap between synthetic environments and the real-world.
- We demonstrate the performance of SEDiff over existing domain-adaptive depth estimation methods in both indoor and outdoor environments, exhibiting better domain generalization performance on unseen datasets.

2 Related Work

2.1 Domain Adaptation for Depth Estimation

As it is difficult to acquire the data for training the depth estimation network in the real-world, there emerge some methods which try to collect the training data in the synthetic environment where it is much easier to collect data for training even with densely paired depth labels compared to the real-world. The problem is that there is a domain gap between the synthetic environment and the real-world, which leads to sub-optimal performance when the depth estimation network trained with synthetic data is deployed in the real-world applications. Therefore, several studies [2, 22, 27, 40, 42, 43] focus on reducing the domain gap between the synthetic environment and real-world.

T^2Net [43] introduces a domain adaptation strategy by jointly training an unsupervised image translation network to transform synthetic images to a realistic domain and a depth estimation network that utilize the style-transferred images from synthetic to real as inputs. Building upon this, GASDA [42] incorporates epipolar geometric constraints for both image translation and depth estimation using stereo-pair images from the real-world, augmenting the synthetic data with additional geometric cues. Similarly, SharinGAN [27] also utilizes stereo pairs as additional geometric cues, but instead of translating images from synthetic to real environments, it projects both synthetic and real images into a shared domain. Meanwhile, 3D-PL [40], which is a current state-of-the-art method, adopts an AdaIN [15]-based style transfer approach and generates

pseudo-labels for real-world images using a pre-trained depth estimation network. Some recent approaches have extended beyond synthetic and real-world data improve the domain adaptation performance. S2R-DepthNet [2] proposes a structure extraction module inspired by multi-modal image translation (MUNIT [38]) trained on synthetic data and the WikiArt dataset [17], which contains images from various styles of paintings. DESC [22] utilizes semantic segmentation labels to detect cars in images and utilizes the car height as depth priors.

In this paper, we leverage synthetic images with depth labels and independently sampled real-world images in order to extract domain-invariant structural information for domain-adaptive depth estimation via diffusion-based generative models.

2.2 Diffusion Probabilistic Models

Recently, diffusion-based generative models [33], represented by DDPM [13], have achieved significant success across various computer vision tasks, including image generation [4, 14, 34], editing [1, 24], inpainting [23, 37], text-to-image synthesis [11, 26, 30, 31], and 3D rendering [25, 28, 32]. These approaches have demonstrated remarkable performance on producing high-quality image outputs compared to earlier generative models such as VAEs [18], GANs [10], and Normalizing Flows [29].

In this paper, we firstly leverage the capabilities of diffusion-based denoising models to extract domain-invariant structural information and enable depth-consistent domain translation from synthetic environments to the real-world for domain adaptive depth estimation. Our work shares the similar motivation to StyleDiffusion [36], which harnesses diffusion models to extract image content for style transfer. However, unlike StyleDiffusion which involves an iterative denoising process inherent to diffusion-based models, our proposed method, SEDiff, performs the iterative denoising process only during the training phase. It means that SEDiff does not entail any iterative denoising process during its deployment, having advantages in time and computational complexity.

3 Preliminaries

3.1 Diffusion Probabilistic Model

Diffusion Probabilistic Models (DPMs) belong to the class of latent variable models which are designed to estimate the target distribution from Gaussian distribution through iterative denoising processes. It consists of two essential stages: the forward diffusion process and the reverse denoising process.

The forward diffusion process gradually adds infinitesimal Gaussian noise ϵ into the data as follows:

$$\begin{aligned} q(x_1, \cdots, x_T|x_0) &:= \prod_{t=1}^{T} q(x_t|x_{t-1}) \\ q(x_t|x_{t-1}) &:= \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t\mathbf{I}) \end{aligned} \tag{1}$$

Here, T represents the total diffusion steps, and $\mathbf{I}$ denotes the identity matrix. β_t is a fixed noise scheduling strategy for diffusion process. This process can be alternatively represented to directly sample x_t from x_0 as:

$$\begin{aligned} q(x_t|x_0) &\coloneqq \mathcal{N}(\sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)\mathbf{I}) \\ x_t &\coloneqq \sqrt{\bar{\alpha}_t}x_0 + (1-\bar{\alpha}_t)\epsilon \end{aligned} \tag{2}$$

where $\alpha \coloneqq 1-\beta_t$ and $\bar{\alpha}_t \coloneqq \prod_{s=1}^{t} \alpha_s$.

The reverse denoising process estimates the target distribution from the Gaussian distribution by estimating the reverse process $q(x_{t-1}|x_t)$. By Bayes' Theorem, the posterior $q(x_{t-1}|x_t, x_0)$ is formulated as:

$$\begin{aligned} \tilde{\beta}_t &\coloneqq \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t \\ \tilde{\mu}_t(x_t, x_0) &\coloneqq \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}x_0 + \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}x_t \\ q(x_{t-1}|x_t, x_0) &\coloneqq \mathcal{N}(\tilde{\mu}_t(x_t, t), \tilde{\beta}_t\mathbf{I}). \end{aligned} \tag{3}$$

Unfortunately, a direct calculation of the posterior requires knowledge of x_0, which is usually unavailable. Thus, DPM employs a neural network p_θ to estimate the posterior as:

$$p_\theta(x_{t-1}|x_t) \simeq q(x_{t-1}|x_t, x_0). \tag{4}$$

To train the DPM, the model minimizes a loss function formulated as a re-weighted variational lower bound:

$$\mathcal{L} = \mathbb{E}_{t,x_0,\epsilon}[||\epsilon - \epsilon_\theta(x_t, t)||_2^2]. \tag{5}$$

Here, $\epsilon_\theta(x_t, t)$ represents the estimated infinitesimal noise between consecutive timesteps.

3.2 Latent Diffusion Model

As DPMs require the iterative denosing process to generate the target data from the noise sampled from the Gaussian distribution, it requires a huge amount of computation and time, i.e., DDPM requires 1000 denoising steps.

To address these challenges, Latent Diffusion Model (LDM) [30] has emerged as a promising alternative. LDM enhances computational and memory efficiency by conducting denoising in a lower-dimensional latent space, rather than directly in pixel space. First, LDMs train a compression model, typically a regularized autoencoder, which maps the input image to a spatially reduced latent representation. This compression is achieved through encoding the input x into a latent vector z, followed by decoding z back into x as follows:

$$\hat{x} = \mathcal{D}(\mathcal{E}(x)) \approx x, \tag{6}$$

where $\mathcal{E}$ and $\mathcal{D}$ indicate the encoder and the decoder of regularized autoencoder respectively. To ensure photo-realistic reconstructions, LDM utilizes a GAN-based framework, incorporating a patch-based discriminator [16] and preceptual loss [41]. By replacing x in DPMs with the latent vector $z = \mathcal{E}(x)$, LDM obtains efficiency gains in both time and computational complexity.

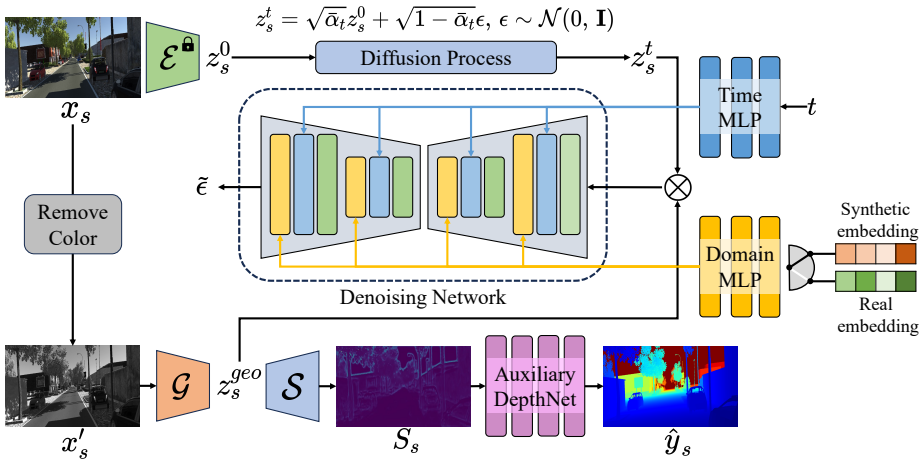

Fig. 2: SEDiff Overview. We extract the domain-invariant structural content S_s from the grayscale synthetic image x'_s. We decouple the domain-specific style component and domain-invariant structural content via domain AdaIN layer in the denoising network. In addition, to preserve the depth information in S_s, we adopt auxiliary depth supervision. $\otimes$ denotes a channel-wise concatenation.

4 Method

In this section, we provide a comprehensive overview of SEDiff for extracting a domain-invariant structural content for depth estimation through latent diffusion models (LDMs). Additionally, we delve into the process of a depth-consistent style transfer and employ its output to accurately estimate the attention map for the structural content.

4.1 Structure Extraction

To facilitate the training of diffusion probabilistic models in the latent space, we first train an autoencoder-based encoder $\mathcal{E}$ and decoder $\mathcal{D}$ following [30] with KL-regularization. The input images acquired from both synthetic environments x_s and real-world scenarios x_r are compressed using the pre-trained $\mathcal{E}$, which effectively reduces the spatial dimensions of the input image:

$$x_d \in \mathbb{R}^{H\times W\times 3}, z_d = \mathcal{E}(x_d) \in \mathbb{R}^{\frac{H}{4}\times\frac{W}{4}\times 3}, \tag{7}$$

where d indicates a domain label, i.e., synthetic domain s and real domain r.

Subsequently, we employ a geometric feature network $\mathcal{G}$ to extract domain-invariant geometric features, $z_s^{geo}, z_r^{geo} \in \mathbb{R}^{\frac{H}{4}\times\frac{W}{4}\times 3}$, from x_s and x_r in the latent space. To ensure compatibility across domains, we first preprocess the inputs, x_s and x_r, by removing color information and converting them to grayscale images, x'_s and x'_r, using the ITU-R 601-2 luma transform [9]. This approach is grounded in the understanding that color distribution often embodies integral aspects of domain-specific image styles [20,36], and this grayscale transformation

can roughly wipe out domain-specific components within the image. These features are then concatenated with the noisy diffusion outputs, z_s^t and z_r^t, and fed to the denosing network ϵ_θ such that it can provide a domain-invariant structure information of input images from different domains to the LDM. By doing so, LDM can synthesize images which share the same structural (or depth) representations as the input image.

Even though we extract geometric features irrelevant to domain-specific components from grayscale input, there could be remaining domain-specific components in the output of geometric feature network $\mathcal{G}$. Therefore, in order to disentangle the domain-invariant geometric features, z_s^{geo} and z_r^{geo}, from domain-specific style components, denoted as w_s and w_r, we introduce two learnable parameters where each parameter encapsulates components specific to the synthetic and real domains, respectively. These parameters are then processed through a shared domain MLP to estimate w_s and w_r. Following, these domain-specific style components are utilized to estimate the scale and the shift parameters, γ_w and β_w, for the domain Adaptive Instance Normalization (domain AdaIN) layer of the denoising network ϵ_θ.

Specifically, we implement an alternating arrangement of time AdaIN and domain AdaIN layers to strongly condition the diffusion time information and domain-specific components to the denoising network ϵ_θ. This arrangement ensures that the time AdaIN layer effectively injects denoising timestep information using the temporal scale and shift parameters, γ_t and β_t, while the domain AdaIN layer proficiently removes any remaining domain-specific components present in the geometric features, z_s^{geo} and z_r^{geo}. Additionally, the inputs of the domain AdaIN layer are re-stylized based on domain embeddings, w_s and w_r, facilitating domain transfer to either the synthetic or real domain.

The denoising function ϵ_θ estimates the noise between two successive diffusion timesteps following LDM. It achieves this by utilizing noisy latent variables z_d^t from the previous denoising step, the current diffusion timesteps t, geometric features z_d^{geo} of the input image x_d, and the target domain-specific component w_d, where d denotes the domain label, synthetic domain s or real domain r. This comprehensive integration allows LDM to generate an image that aligns with the geometric features of the input image, irrespective of its domain. Furthermore, by incorporating domain-specific styling through the parameter w_d, SEDiff ensures that the generated images include the stylistic characteristics inherent to the specified domain, thereby enhancing the fidelity of the generated outputs.

The loss function to train this geometry-domain-conditioned LDM, i.e., SEDiff, can be formulated as below:

$$\mathcal{L}_{ldm} = ||\epsilon - \epsilon_\theta(z_d^t, z_d^{geo}, t, w_d)||_2^2 \tag{8}$$

From the domain-invariant geometric features z_d^{geo}, we construct the structural content $S_d \in \mathbb{R}^{H\times W\times 1}$ with structural decoder $\mathcal{S}$, specifically tailored to include only the essential elements of the input image x_d for accurate depth estimation.

In order to ensure that the derived structural content S indeed encapsulates essential information for precise depth estimation, we jointly train auxiliary

Algorithm 1 Depth-consistent style transfer

Input: A synthetic image x_s, real domain embedding w_r, total diffusion step T
Load: geometric feature network $\mathcal{G}$, noise predictor ϵ_θ, and LDM decoder $\mathcal{D}$
Extract $z_s^{geo} = \mathcal{G}(x_s)$
Sample $z^T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
for $t = T, T-1, \cdots, 1$ **do**
  Sample $z_{noise} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ if $t > 1$, else $z_{noise} = 0$
  Compute x_{t-1} using Eq. (3):
  $z^{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(z^t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} \epsilon_\theta(z^t, z_s^{geo}, t, w_r) \right) + \sigma_t z_{noise}$
end for
Decode $x_{s \to t} = \mathcal{D}(z^0)$
return $x_{s \to t}$ as depth-consistent style transfer

depth estimation network which estimates the depth value from S. We leverage the availability of dense depth labels y_s paired with synthetic images x_s and train the auxiliary depth estimation network in a supervised manner as

$$\begin{aligned} S_s = \mathcal{S}(z_s^{geo}),\ \hat{y}_{\text{aux}} = f_{\text{aux}}(S_s) \\ \mathcal{L}_{\text{aux}} = ||\hat{y}_{\text{aux}} - y_s||_1, \end{aligned} \tag{9}$$

where f_{aux} indicates the auxiliary depth estimation network.

We opt for an end-to-end training approach for all networks within the structure extraction module of SEDiff, except for $\mathcal{E}$ and $\mathcal{D}$. This methodology allows us to train the framework, facilitating the extraction of domain-invariant geometric features z^{geo} essential for domain-adaptive depth estimation, while ensuring structural consistency between the conditioned input x_d and the output image of LDM. The total loss function is formulated as the linear combination of two loss functions,

$$\mathcal{L}_{se} = \mathcal{L}_{ldm} + \lambda_{\text{aux}} \mathcal{L}_{\text{aux}}, \tag{10}$$

where λ_{aux} is set to 0.1.

4.2 Attention Map via Depth-Consistent Style Transfer

After training the feature extraction network $\mathcal{G}$ and the structural decoder $\mathcal{S}$, we can extract the domain-invariant structural content, S_s and S_r, from both synthetic images x_s and real-world images x_r. Subsequently, we proceed to train the depth estimation network in a supervised manner, which utilizes the domain-invariant structure S_s as input and its paired dense depth labels y_s as ground truth.

To mitigate the domain gap between synthetic and real-world images and thus minimize potential performance degradation due to domain shift, we remove color information from input images through grayscale transformation before they are fed into the feature extraction network $\mathcal{G}$. However, it is worth noting that the rich semantic information for depth estimation is often embedded in the RGB color space. For instance, the sky is typically blue, while paved roads in urban or highway settings tend to appear gray.

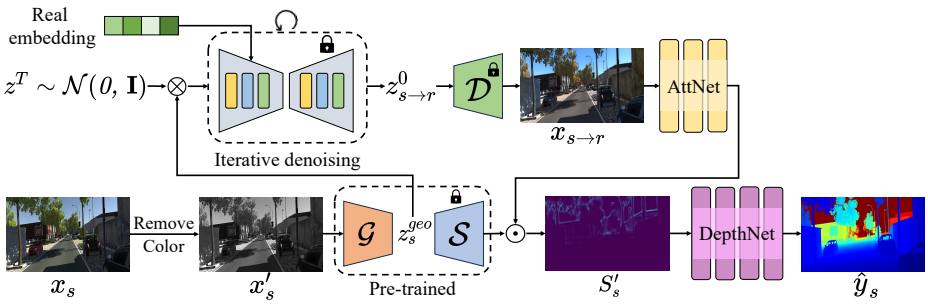

Fig. 3: A pipeline for training the domain-adaptive depth estimation network with the domain-invariant structural content and depth-consistent style-transfer-based attention map. $\otimes$ and $\odot$ denote a channel-wise concatenation and a Hadamard product respectively.

Hence, we also integrate RGB images to provide additional cues for accurate depth estimation. However, to prevent potential adverse effects of domain-specific content inherent in RGB images on domain adaptation performance, we restrict their use solely to deriving the attention map [2], $A \in \mathbb{R}^{H\times W\times 1}$, of the domain-invariant structural content S. This approach ensures that RGB images contribute exclusively to refining the focus for accurate depth prediction without directly influencing the estimation of the value of S itself.

Furthermore, instead of directly leveraging synthetic images x_s to estimate the attention map A, we employ the outputs of depth-consistent style transfer $x_{s\to r}$ utilizing the geometry-domain-conditioned LDM trained for the structure extraction module. During the iterative denoising process of LDM for image synthesis, we incorporate the geometric features z_s of the synthetic image x_s and the domain embedding w_r that represents the real-world stylistic characteristics as described in Algorithm 1. This approach helps minimize the domain shift by ensuring that the output of the style transfer $x_{s\to r}$ shares the same depth information as the input synthetic image x_s, while also including the visual style characteristic of the real-world.

The loss function to train the structure-based depth estimation network and the style-transfer-based attention map is formulated as:

$$\begin{aligned} S_s = \mathcal{S}(\mathcal{G}(x_s)),\ A = f_{\text{attn}}(x_{s\to r}), \\ \hat{y}_s = f_{\text{depth}}(A \odot S_s), \\ \mathcal{L}_{\text{dep}} = ||\hat{y}_s - y_s||_1, \end{aligned} \tag{11}$$

where f_{attn} and f_{depth} indicate the attention network and the depth estimation network respectively, and $\odot$ denotes the Hadamard product.

4.3 Inference

Integrating diffusion-based generative models into depth estimation tasks demands careful consideration due to the iterative denoising process inherent in such models, which often entails significant time and computational complexity.

SEDiff also incorporates the iterative denoising process for depth-consistent style transfer from the synthetic environment to the real-world during training. It is important to note, however, that during the inference phase, SEDiff streamlines the process by excluding any iterative denoising steps, thereby avoiding the associated slowdown in inference time. Instead, we directly estimate the structure S_r and the attention map A from the real-world image x_r, ensuring efficient and rapid depth estimation,

$$\begin{aligned} S_r = \mathcal{S}(\mathcal{G}(x_r)),\ A = f_{\text{attn}}(x_r), \\ \hat{y}_r = f_{\text{depth}}(A \odot S_r). \end{aligned} \tag{12}$$

5 Experiment

In this section, we present the effectiveness of our proposed framework, SEDiff, in both single-view depth estimation and stereo-pair settings. We perform an extensive experiment on synthetic-to-real domain adaptive depth estimation which shows the superiority of SEDiff over existing methods across indoor and outdoor scenarios. Additionally, we validate the contributions of each component within SEDiff to the overall depth estimation performance and better domain generalization capabilities in the unseen environments.

5.1 Dataset

For outdoor scenarios, we utilize Virtual KITTI (vKITTI) [6] as our synthetic dataset and KITTI [7] as the real-world dataset. vKITTI offers 21,260 synthetic images paired with dense depth labels, which we leverage for training alongside KITTI's 22,600 training images. For indoor scenarios, Replica [35] serves as our synthetic dataset, while NYU Depth v2 [3] serves as the real-world dataset. We evaluate the domain generalization performance on unseen real-world datasets, we employ DrivingStereo [39] and DDAD [12], specifically collected from outdoor scenes, for evaluation purposes only.

5.2 Comparison to state-of-the-art

In our evaluation on KITTI with 697 test images from Eigen's train/test split [5], we conduct a thorough comparison between SEDiff and existing domain-adaptive depth estimation methods. We observe that our proposed SEDiff achieves compelling improvements over competitive methods which employs domain adaptation strategies across a range of evaluation metrics. Especially compared to current state-of-the-art, i.e., 3D-PL [40], we validate that our method outperforms the current state-of-the-art in majority of the metrics.

Furthermore, we demonstrate that SEDiff produces comparable performance to methods such as S2R-DepthNet [2] and DESC [22], which leverage additional data beyond synthetic images paired with depth labels from vKITTI and

Table 1: Performance on KITTI. All results are computed using the test split from [5]. The Dataset column specifies the data sources: V for synthetic supervision with vKITTI, K(M) for monocular images from KITTI, K(sem) for semantic labels from KITTI, and W for the WikiArt dataset [17]. The best results are highlighted in **bold**, while the second-best are underlined. Methods incorporating domain adaptation without requiring additional data are shaded in gray.

Method	Dataset	cap	Higher is better			Lower is better			
			$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$	Abs Rel	Sq Rel	RMSE	RMSE_{log}
All synthetic	V	80m	0.635	0.856	0.937	0.253	2.303	6.953	0.328
AdaDepth [19]	V + K(M)	80m	0.665	0.882	0.950	0.214	1.932	7.157	0.295
T^2Net [43]	V + K(M)	80m	0.757	0.918	0.969	0.171	1.351	5.944	0.247
3D-PL [40]	V + K(M)	80m	0.753	0.918	0.968	0.169	**1.262**	6.034	0.249
SEDiff (ours)	V + K(M)	80m	**0.773**	**0.932**	**0.973**	**0.165**	1.301	**5.686**	**0.237**
S2R-DepthNet [2]	V + W	80m	0.781	0.931	0.972	0.165	1.351	5.695	0.236
DESC [22]	V + K(Sem)	80m	0.787	0.924	0.970	0.156	1.067	5.628	0.237
All synthetic	V	50m	0.647	0.866	0.943	0.244	1.771	5.354	0.313
AdaDepth [19]	V + K(M)	50m	0.687	0.899	0.958	0.203	1.734	6.251	0.284
T^2Net [43]	V + K(M)	50m	0.773	0.928	0.974	0.164	1.019	4.469	0.231
3D-PL [40]	V + K(M)	50m	0.770	0.932	0.975	0.161	**0.936**	**4.398**	0.230
SEDiff (ours)	V + K(M)	50m	**0.786**	**0.939**	**0.977**	**0.159**	1.013	4.417	**0.225**

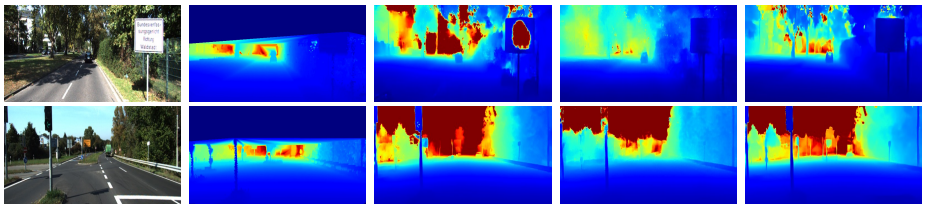

Fig. 4: Qualitative results on KITTI test split. Compared to T^2Net [43] and 3D-PL [40], SEDiff produces better depth estimation results with distinct object boundaries.

randomly sampled monocular images from KITTI . Notably, these methods incorporate data sources like additional images featuring various paintings with diverse styles from various artists, i.e., WikiArt [17], or employ semantic labels from KITTI.

In addition to quantitative results, we validate the superiority of SEDiff in a qualitative way by visualizing the predicted depth maps in Fig. 4. Our proposed method not only exhibits clear boundaries between objects and backgrounds but also accurately captures the depth values of small objects critical for autonomous driving, such as traffic signs or traffic lights.

5.3 Training with Stereo-Pairs

Recent domain-adaptive depth estimation algorithms, such as GASDA [42], SharinGAN [27], and DESC [22], have integrated stereo images from the real-world as additional geometric cues. These algorithms employ self-supervised stereo supervision [8] alongside synthetic supervision to ensure geometric consistency between virtual and real domains. We also leverage additional stereo images from real-world datasets in SEDiff to validate its effectiveness in diverse domain adaptation scenarios. As demonstrated in Table 2, SEDiff consistently

Table 2: Performance on KITTI with additional stereo pairs during training. All results are computed on KITTI Eigen [5] test split. K(S) and V denote stereo self-supervision with KITTI and synthetic supervision with vKITTI respectively. The best results are marked in **bold**.

Method	Dataset	cap	Higher is better			Lower is better			
			$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$	Abs Rel	Sq Rel	RMSE	RMSE_{log}
All synthetic	V	80m	0.635	0.856	0.937	0.253	2.303	6.953	0.328
All real	K(S)	80m	0.811	0.934	0.970	0.158	1.151	5.285	0.238
GASDA [42]	V + K(S)	80m	0.824	0.941	0.973	0.149	1.003	4.995	0.227
SharinGAN [27]	V + K(S)	80m	0.850	0.948	0.978	0.116	0.939	5.068	0.203
3D-PL [40]	V + K(S)	80m	0.859	0.952	0.979	0.113	0.903	4.902	0.201
SEDiff (ours)	V + K(S)	80m	**0.871**	**0.956**	**0.980**	**0.110**	**0.899**	**4.816**	**0.195**
All synthetic	V	50m	0.647	0.866	0.943	0.244	1.771	5.354	0.313
All real	K(S)	50m	0.824	0.940	0.973	0.151	0.856	4.043	0.227
GASDA [42]	V + K(S)	50m	0.836	0.946	0.976	0.143	0.756	3.846	0.217
SharinGAN [27]	V + K(S)	50m	0.864	0.954	0.981	0.109	0.673	3.770	0.190
3D-PL [40]	V + K(S)	50m	0.872	0.958	**0.982**	0.106	**0.641**	3.643	0.189
SEDiff (ours)	V + K(S)	50m	**0.884**	**0.962**	**0.982**	**0.104**	0.649	**3.617**	**0.184**

outperforms not only the methods trained solely on synthetic data (All synthetic) and real stereo pairs (All real) but also the existing state-of-the-art methods that utilize both synthetic data and real stereo pairs during training [27,40,42] across a majority of evaluation metrics.

5.4 Ablation Study

For better understanding of our proposed SEDiff, we perform an ablation study in Table 3. This study includes three key variations: (a) the input of the depth estimation network, (b) the training method for the geometric feature extractor $\mathcal{G}$, and (c) the domain of the input for the attention network f_{attn}.

To demonstrate the effectiveness of domain-specific style removal process in the input image of the depth estimation network, we conduct a comparative analysis including three different representations: the raw RGB image x_s, the grayscale image x'_s obtained through the ITU-R 601-2 luma transform [9], and the structural content S_s extracted using SEDiff, as shown in Table 3(a). Our findings reveal that the proposed structural content S_s yields the most promising results, surpassing both the RGB and grayscale images due to its superior domain-invariant properties. Furthermore, the result that grayscale input outperforms RGB input validate the adoption of grayscale transformation as a pre-processing step before feeding the RGB image into the geometric feature extractor $\mathcal{G}$. By converting the RGB image to grayscale, we can roughly wipe out the influence of domain-specific information containing in the RGB space.

We evaluate the impact of the geometric feature extractor $\mathcal{G}$ on the depth estimation results, as presented in Table 3(b). Three methods are compared: Firstly, we directly employ the encoder $\mathcal{E}$ of LDM for image compression, assuming that the domain-invariant features may stem from the robust data compression capability of $\mathcal{E}$ rather than the intrinsic properties of diffusion probabilistic models. Secondly, we train $\mathcal{G}$ using pixel-level diffusion-based probabilistic models such

Table 3: Ablation. Results for different variants of SEDiff on Eigen [5] test split. Best results are marked in **bold**.

	Method	cap	Higher is better			Lower is better			
			$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$	Abs Rel	Sq Rel	RMSE	RMSE_{log}
(a)	RGB	80m	0.630	0.859	0.940	0.268	3.302	7.839	0.327
	Gray	80m	0.682	0.888	0.953	0.231	2.451	6.776	0.297
	SC (S)	80m	**0.773**	**0.932**	**0.973**	**0.165**	**1.301**	**5.686**	**0.237**
(b)	Encoder ($\mathcal{E}$)	80m	0.735	0.921	0.971	0.183	1.416	5.864	0.249
	Pixel	80m	0.767	0.930	**0.974**	0.166	**1.230**	**5.629**	**0.236**
	Latent	80m	**0.773**	**0.932**	0.973	**0.165**	1.301	5.686	0.237
(c)	Synthetic	80m	0.761	0.924	0.970	0.173	1.481	5.952	0.247
	Syn2Real	80m	0.770	0.928	0.971	0.168	1.366	5.899	0.243
	Ensemble	80m	**0.773**	**0.932**	**0.973**	**0.165**	**1.301**	**5.686**	**0.237**

Method	Abs Rel	RMSE	log 10	$\delta < 1.25$
Li *et al.* [21]	0.232	0.821	0.094	0.621
Eigen *et al.* [5]	0.215	0.907	-	0.611
All synthetic	1.717	0.983	0.133	0.499
T^2Net [43]	0.293	0.843	0.115	0.585
SharinGAN [27]	0.339	0.795	0.101	0.613
SEDiff (ours)	**0.221**	**0.762**	**0.096**	**0.638**

Table 4: Performance on NYU Depth

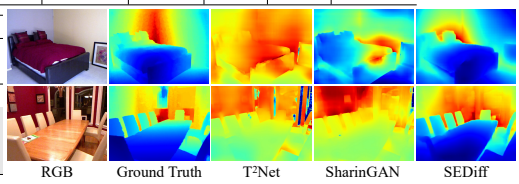

Fig. 5: Qualitative results on NYU Depth v2

as DDPM, instead of in the feature space. Our observations indicate that utilizing $\mathcal{E}$ as the geometric feature extractor $\mathcal{G}$ produces sub-optimal performance compared to diffusion-based approaches, including both pixel-level and latent-level diffusion. This suggests that while the encoder $\mathcal{E}$ excels at compressing data, it does not directly contribute to domain generalization performance. Furthermore, we find no significant performance difference between pixel-level and latent-level diffusion models. This implies the influence of the denoising process on the domain-invariant properties of geometric features z^{geo}, irrespective of whether the diffusion process operates at the pixel or latent level. Given the comparable performance, we opt for latent-level diffusion for its faster sampling capabilities during the training phases.

In Table 3(c), regarding the attention map, we observe that estimating it using the style-transferred output $x_{s\to r}$ shows better performance compared to using the raw synthetic image x_s, primarily because $x_{s\to r}$ exhibits a reduced domain gap with real-world images. Also, we find that leveraging images from both domains, x_s and $x_{s\to r}$, further enhances performance. To this end, we employ two distinct attention networks f_{attn} and depth estimation networks f_{depth}, with each network processing x_s and $x_{s\to r}$, respectively. By ensembling the depth estimation results from both inputs, we achieve optimal performance compared to methods that utilize a single domain input.

5.5 Indoor Environment

The indoor environment usually exhibits unique challenges distinct from outdoor settings, and therefore, we compare SEDiff with existing supervised [5, 21] and domain-adaptive depth estimation algorithms [27, 43] on the NYU Depth v2 [3] test split. As shown in Table 4, SEDiff consistently outperforms existing domain adaptation-based depth estimation algorithms across all metrics in a quantitative manner. Furthermore, qualitative evaluations in Fig. 5 show that SEDiff also

Method	Dataset	Abs Rel	Sq Rel	RMSE	$RMSE_{log}$
T^2Net [43]	DS	0.419	4.802	9.870	0.660
3D-PL [40]	DS	0.243	3.491	9.551	0.307
SEDiff (ours)	DS	**0.193**	**2.295**	**8.363**	**0.250**
T^2Net [43]	DDAD	0.475	5.591	10.474	0.731
3D-PL [40]	DDAD	0.220	2.564	10.348	0.315
SEDiff (ours)	DDAD	**0.216**	**2.449**	**10.048**	**0.302**

Table 5: Quantitative results on unseen dataset, i.e., DrivingStereo (DS) [39] , and DDAD [12]

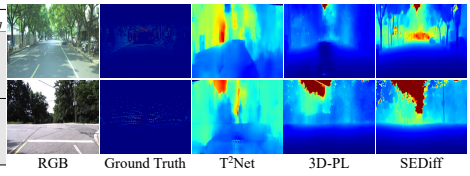

Fig. 6: Depth prediction results on DrivingStereo [39] (upper row) and DDAD [12] (lower row).

produces visually superior depth predictions, closely resembling ground truth depth labels.

5.6 Domain Generalization

We also evaluate the domain generalization performance of SEDiff in comparison to competitive domain-adaptive depth estimation approaches [40, 43] on unseen datasets, i.e., DrivingStereo dataset [39] and DDAD dataset [12]. Notably, we refrain from training or fine-tuning SEDiff and its competitive methods on these datasets, opting instead to directly evaluate models trained solely on vKITTI and KITTI. As illustrated in Table 5 and Fig. 6, our proposed SEDiff demonstrates superior performance over state-of-the-art domain adaptive depth estimation methods on unseen datasets, both qualitatively and quantitatively, despite encountering a larger domain gap.

6 Conclusion

In this paper, we introduce SEDiff, which firstly integrates diffusion probabilistic models into the domain-adaptive depth estimation framework to extract domain-invariant structural content, mitigating the domain shift between synthetic environments and the real-world. Additionally, SEDiff enables depth-consistent style transfer to estimate the attention map while minimizing the domain gap. Our experiments demonstrate SEDiff's effectiveness over state-of-the-art methods in both indoor and outdoor scenarios, highlighting its potential across various computer vision applications.

Acknowledgements. This work was partly supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) [NO.RS-2021-II211343, Artificial Intelligence Graduate School Program (Seoul National University)] and by Unmanned Vehicles Core Technology Research and Development Program through the National Research Foundation of Korea(NRF) and Unmanned Vehicle Advanced Research Center(UVARC) funded by the Ministry of Science and ICT, the Republic of Korea(NRF-2020M3C1C1A010864).

References

1. Avrahami, O., Lischinski, D., Fried, O.: Blended diffusion for text-driven editing of natural images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18208–18218 (2022) 4
2. Chen, X., Wang, Y., Chen, X., Zeng, W.: 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) 2, 3, 4, 9, 10, 11
3. Couprie, C., Farabet, C., Najman, L., LeCun, Y.: Indoor semantic segmentation using depth information. arXiv preprint arXiv:1301.3572 (2013) 10, 13
4. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Advances in neural information processing systems **34**, 8780–8794 (2021) 4
5. Eigen, D., Puhrsch, C., Fergus, R.: Depth map prediction from a single image using a multi-scale deep network. Advances in neural information processing systems **27** (2014) 10, 11, 12, 13
6. Gaidon, A., Wang, Q., Cabon, Y., Vig, E.: 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) 2, 10
7. Geiger, A., Lenz, P., Urtasun, R.: 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) 2, 10
8. Godard, C., Mac Aodha, O., Brostow, G.J.: Unsupervised monocular depth estimation with left-right consistency. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 270–279 (2017) 11
9. Gonzalez, R.C.: Digital image processing. Pearson education india (2009) 6, 12
10. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. Advances in neural information processing systems **27** (2014) 4
11. Gu, S., Chen, D., Bao, J., Wen, F., Zhang, B., Chen, D., Yuan, L., Guo, B.: Vector quantized diffusion model for text-to-image synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10696–10706 (2022) 4
12. Guizilini, V., Ambrus, R., Pillai, S., Raventos, A., Gaidon, A.: 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) 10, 14
13. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020) 4
14. Ho, J., Saharia, C., Chan, W., Fleet, D.J., Norouzi, M., Salimans, T.: Cascaded diffusion models for high fidelity image generation. The Journal of Machine Learning Research **23**(1), 2249–2281 (2022) 4
15. Huang, X., Belongie, S.: Arbitrary style transfer in real-time with adaptive instance normalization. In: Proceedings of the IEEE international conference on computer vision. pp. 1501–1510 (2017) 3
16. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1125–1134 (2017) 5
17. K Nichol., W.K.: Painter by numbers (2016), https://kaggle.com/competitions/painter-by-numbers 4, 11

18. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013) 4
19. Kundu, J.N., Uppala, P.K., Pahuja, A., Babu, R.V.: Adadepth: Unsupervised content congruent adaptation for depth estimation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2656–2665 (2018) 11
20. Lang, B.: The concept of style. Cornell University Press (1987) 6
21. Li, B., Shen, C., Dai, Y., Van Den Hengel, A., He, M.: Depth and surface normal estimation from monocular images using regression on deep features and hierarchical crfs. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1119–1127 (2015) 13
22. Lopez-Rodriguez, A., Mikolajczyk, K.: Desc: Domain adaptation for depth estimation via semantic consistency. International Journal of Computer Vision **131**(3), 752–771 (2023) 3, 4, 10, 11
23. Lugmayr, A., Danelljan, M., Romero, A., Yu, F., Timofte, R., Van Gool, L.: Repaint: Inpainting using denoising diffusion probabilistic models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11461–11471 (2022) 4
24. Meng, C., He, Y., Song, Y., Song, J., Wu, J., Zhu, J.Y., Ermon, S.: Sdedit: Guided image synthesis and editing with stochastic differential equations. In: International Conference on Learning Representations (2021) 4
25. Müller, N., Siddiqui, Y., Porzi, L., Bulo, S.R., Kontschieder, P., Nießner, M.: Diffrf: Rendering-guided 3d radiance field diffusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4328–4338 (2023) 4
26. Nichol, A.Q., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., Mcgrew, B., Sutskever, I., Chen, M.: Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In: International Conference on Machine Learning. pp. 16784–16804. PMLR (2022) 4
27. PNVR, K., Zhou, H., Jacobs, D.: 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) 2, 3, 11, 12, 13
28. Poole, B., Jain, A., Barron, J.T., Mildenhall, B.: Dreamfusion: Text-to-3d using 2d diffusion. In: The Eleventh International Conference on Learning Representations (2022) 4
29. Rezende, D., Mohamed, S.: Variational inference with normalizing flows. In: International conference on machine learning. pp. 1530–1538. PMLR (2015) 4
30. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022) 4, 5, 6
31. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. Advances in Neural Information Processing Systems **35**, 36479–36494 (2022) 4
32. Seo, J., Jang, W., Kwak, M.S., Ko, J., Kim, H., Kim, J., Kim, J.H., Lee, J., Kim, S.: Let 2d diffusion model know 3d-consistency for robust text-to-3d generation. In: International Conference on Learning Representations (2024) 4
33. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: International conference on machine learning. pp. 2256–2265. PMLR (2015) 4

34. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: International Conference on Learning Representations (2020) 4
35. Straub, J., Whelan, T., Ma, L., Chen, Y., Wijmans, E., Green, S., Engel, J.J., Mur-Artal, R., Ren, C., Verma, S., et al.: The replica dataset: A digital replica of indoor spaces. arXiv preprint arXiv:1906.05797 (2019) 10
36. Wang, Z., Zhao, L., Xing, W.: Stylediffusion: Controllable disentangled style transfer via diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7677–7689 (2023) 4, 6
37. Xie, S., Zhang, Z., Lin, Z., Hinz, T., Zhang, K.: Smartbrush: Text and shape guided object inpainting with diffusion model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22428–22437 (2023) 4
38. Xu, D., Ricci, E., Ouyang, W., Wang, X., Sebe, N.: Multi-scale continuous crfs as sequential deep networks for monocular depth estimation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5354–5362 (2017) 4
39. Yang, G., Song, X., Huang, C., Deng, Z., Shi, J., Zhou, B.: Drivingstereo: A large-scale dataset for stereo matching in autonomous driving scenarios. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 899–908 (2019) 10, 14
40. Yen, Y.T., Lu, C.N., Chiu, W.C., Tsai, Y.H.: 3d-pl: Domain adaptive depth estimation with 3d-aware pseudo-labeling. In: European Conference on Computer Vision. pp. 710–728. Springer (2022) 3, 10, 11, 12, 14
41. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018) 5
42. Zhao, S., Fu, H., Gong, M., Tao, D.: 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) 2, 3, 11, 12
43. Zheng, C., Cham, T.J., Cai, J.: 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) 2, 3, 11, 13, 14