V2Depth: Monocular Depth Estimation via Feature-Level Virtual-View Simulation and Refinement

Zizhang Wu*
ZongmuTech
wuzizhang87@gmail.com

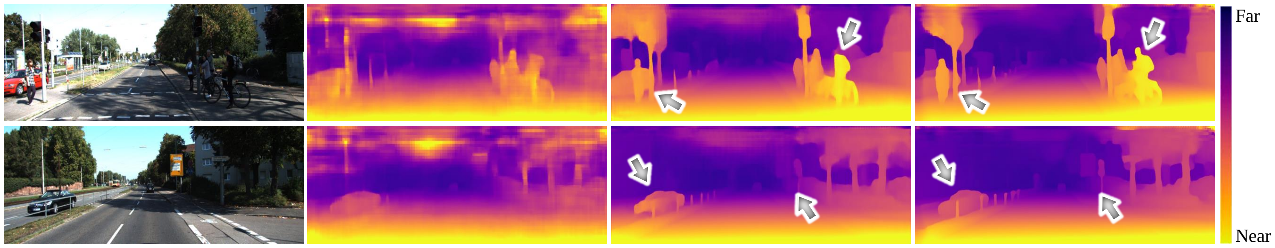
Zhuozheng Li*
ZongmuTech
zhuozheng.li@zongmutech.com

Zhi-Gang Fan*
ZongmuTech
zhigang.fan@zongmutech.com

Yunzhe Wu
ZongmuTech
nelson.wu@zongmutech.com

Jian Pu
Fudan University
jianpu@fudan.edu.cn

Xianzhi Li
Huazhong University of Science and Technology
xzli@hust.edu.cn



(a) Input images (b) Results w/o simulator and refiner (c) Results of DRO (d) Results of our V2Depth

Figure 1: Given input images (a), the proposed V2Depth predicts more accurate depth (d) compared to the results (b) without our virtual-view feature simulator and refiner, as well as the results of baseline method DRO [23] (c) without our simulator and 3DVA, especially for objects with occlusion relationships such as cars and pedestrians (marked by gray arrows).

ABSTRACT

Due to the lack of spatial cues giving merely a single image, many monocular depth estimation methods have been developed to leverage stereo or multi-view images to learn the spatial information of a scene in a self-supervised manner. However, these methods have limited performance gain since they are not able to exploit sufficient 3D geometry cues during inference, where only monocular images are available. In this work, we present **V2Depth**, a novel coarse-to-fine framework with **V**irtual **V**iew feature simulation for supervised monocular **Depth** estimation. Specifically, we first design a virtual-view feature simulator by leveraging the technique of novel view synthesis and contrastive learning to generate virtual view feature maps. In this way, we explicitly provide representative spatial geometry for subsequent depth estimation in both the training and inference stages. Then we introduce a 3DVA-Refiner to iteratively optimize the predicted depth map. During the optimization process, 3D-aware virtual attention is developed to capture the global spatial-context correlations to maintain the feature consistency of different views and estimation integrity of the 3D scene such as objects with occlusion relationships. Decisive improvements over state-of-the-art approaches on three benchmark datasets across all metrics demonstrate the superiority of our method.

CCS CONCEPTS

- **Computing methodologies → Computer vision tasks**.

KEYWORDS

Monocular Depth Estimation, Virtual-View Feature Simulation, Contrastive Learning, Refinement Network

ACM Reference Format:
Zizhang Wu*, Zhuozheng Li*, Zhi-Gang Fan*, Yunzhe Wu, Jian Pu, and Xianzhi Li. 2023. V2Depth: Monocular Depth Estimation via Feature-Level Virtual-View Simulation and Refinement. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23), October 29-November 3, 2023, Ottawa, ON, Canada.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3581783.3611751

* These authors contributed equally to this work.

1 INTRODUCTION

Monocular depth estimation is a fundamental task in computer vision with the goal of predicting per-pixel depth values from a single RGB image. Benefits from its nature of low cost and easy implementation, monocular depth estimation contributes to various applications including robot navigation [54], 3D scene reconstruction [37, 82], augmented reality [13, 15], and autonomous driving [43, 84], etc. Significant progress has been achieved by exploiting neural networks to learn the mapping from image to depth map [6, 41, 46, 66, 89]. However, inferring an accurate depth map

MM '23, October 29-November 3, 2023, Ottawa, ON, Canada

ACM ISBN 979-8-4007-0108-5/23/10…$15.00
https://doi.org/10.1145/3581783.3611751

from merely a single image is quite challenging due to its inherent ambiguity and ill-posed nature [26, 65], where many possible 3D scenes can actually be projected into the same 2D scene [58, 89]. Monocular depth estimation task turns into a difficult fitting problem due to the lack of *3D geometric information in monocular images*, i.e. without available spatial cues.

A crucial intuition for humans' perception of depth is to draw on prior knowledge of left-right views. Hence, a number of approaches develop the self-supervised training schemes[19, 22, 79], which take stereo pairs as input for network training. It adopts left-right depth reprojection with photometric cost minimization to learn depth from binocular data. More recently, several distillation-based methods [8, 83] have been used to train a stereo teacher network on image pairs and distill the learned knowledge or structure priors to the monocular student model. These methods, however, achieve limited performance gain since they are not able to exploit sufficient 3D geometry cues during the inference stage, where only monocular images are available as test input.

In this paper, we address supervised monocular **Depth** estimation by introducing the **V**irtual-**V**iew feature simulation and refinement to enable the learning of 3D spatial cues, which is the first of this kind. Our approach (named **V2Depth**) first generates a high-quality feature map of a synthetic image as taken from a virtual camera placed in a different viewpoint. More specifically, given a reference input image (e.g., left-view image), we design our network to generate not only the reference feature map but also the feature map from a new perspective (e.g., right-view). In specific, we develop a *virtual-view feature simulator* (abbr. VVF-Simulator), based on contrastive representation learning [57]. During network training, given two identical images of the same view, we optimize the output feature embedding space to *maximize* the cosine similarity of the two positive features, since we expect our network to convert one of the two identical images into a new view. Conversely, given two images of different views, we *minimize* the cosine similarity of the negative feature pairs, since we expect our network to convert one of the two different images to be identical to the other. During inference, VVF-Simulator consumes two identical monocular images and generates a pair of features, i.e., a (real) reference feature map and a (virtual) novel-view feature map. Therefore, compared to the teacher-student knowledge distillation approach [83], our VVF-Simulator can *explicitly* make full use of 3D geometric information during both training and testing.

Afterwards, based on the above 3D-aware virtual features, we develop a 3DVA-Refiner to iteratively optimize our initial depth predictions, as inspired by the deep recurrent optimizer (DRO) [23]. Specifically, we develop our 3D-aware virtual attention (3DVA) module with the cross-attention mechanism that assigns the 3D-aware virtual features as values, and the depth context features as queries and keys. In this way, it effectively interacts with the global 3D-aware virtual features and depth context features, thus achieves feature consistency and depth estimation integrity. Extensive experiments demonstrate that by rectifying the mismatching problem, our V2Depth predicts the accurate depth and contours of objects, especially for objects with occlusion relationships and in messy environment.

Qualitatively, as shown in Fig. 1 (b), we directly predict depth by learning features from monocular input without our VVF-Simulator and 3DVA-Refiner. Clearly, it is difficult to recover the complete geometry of objects. On the other hand, compared to the baseline method DRO (c), the predictions from our V2Depth (d) demonstrate its robustness and precision at occluded objects; see particularly the cars and pedestrians marked by gray arrows. Quantitatively, we conduct extensive experiments on three challenging benchmarks to validate the effectiveness of our pipeline against state-of-the-art models and sufficient ablation studies to verify the contribution of each major component of our V2Depth. Please refer to the experiments section for more detailed comparisons.

In a nutshell, we summarize our main contributions in threefold:

- To the best of our knowledge, we are the first to introduce monocular depth estimation with the feature-level virtual-view simulation and refinement. Our proposed V2Depth predicts accurate depth for mutually occluding objects and achieves state-of-the-art performance on challenging KITTI, Virtual KITTI 2, and DrivingStereo datasets.
- We design a novel virtual-view feature simulator by leveraging contrastive learning to produce a pair of features, including a real reference feature map and a virtual novel-view feature map. It stimulates the applicability of monocular depth estimation to real 3D understanding with representative spatial geometric information.
- We develop an 3D-aware virtual attention based refiner to learn the 3D geometric constraints from the virtual-view features and the reference features. It learns the feature consistency along with the estimation integrity, and iteratively optimizes the depth through attention-based integration.

2 RELATED WORK

2.1 Monocular Depth Estimation

Monocular depth estimation methods aim at predicting the pixel-wise depth values from a single image. Depending on whether the input images are monocular or stereo images during the network training, we divide those methods into *monocular training* and *stereo training* methods.

Monocular training refers to training and inference with monocular images only for the task of depth estimation. Eigen et al. [14] first proposed a coarse-to-fine framework that uses the deep convolutional neural network (CNN) to regress the depth map from the extracted image features. Afterwards, Lee et al. [38] introduced the multi-stage local planar guidance layer for stimulating the learning of depth features. BANet [1] proposed a bidirectional attention module to utilize the feed-forward feature maps and incorporate the global context to filter depth ambiguity. Moreover, the dense prediction transformer (DPT) was proposed [63] that leverages the vision transformer (ViT) [12] as the backbone for accomplishing global-aware feature extraction. NewCRFs [89] adopted the swin-transformer [48] for image encoding and developed the fully-connected conditional random fields for depth decoding. Particularly, DRO [23] introduced the deep recurrent optimizer with the gated recurrent unit (GRU) [76] to alternately update depth and camera poses. Besides, some of those approaches attempt to estimate the depth values from the results of classification or ordinal regression[6, 17]. Fu et al. [17] introduced a spacing-increasing

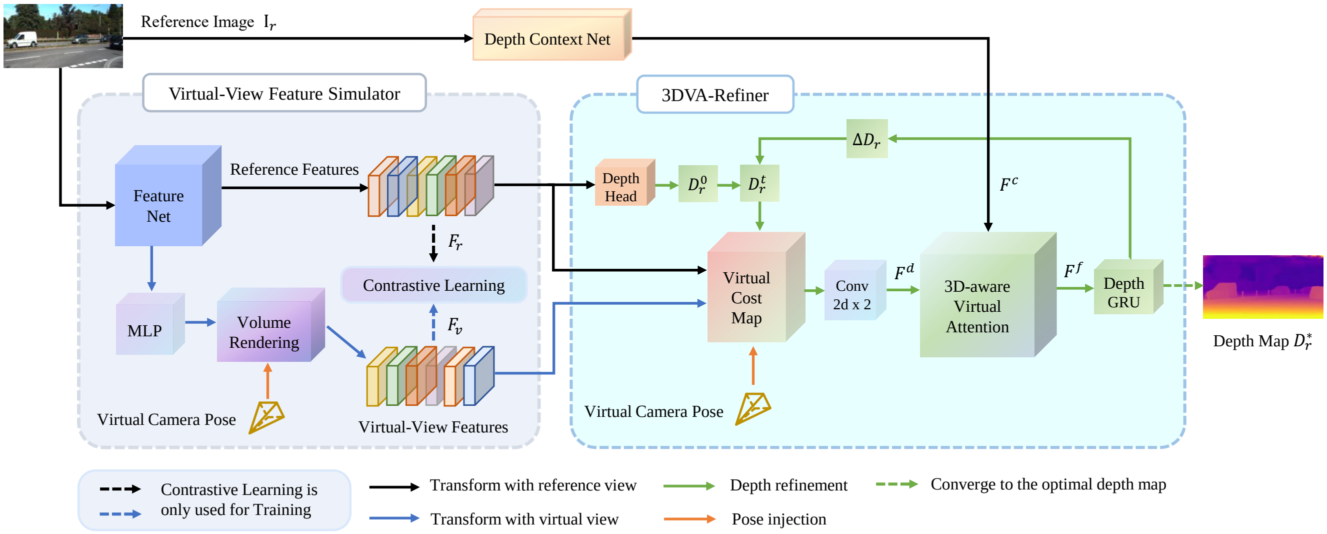


Figure 2: Overview of the framework. Given the reference image I_r, our virtual-view feature simulator extracts the virtual-view feature F_v and the reference feature F_r. In parallel, a depth context network extracts the context feature F^c. Then, a depth head takes in F_r and outputs an initial depth map D_r^0. During the optimization, the proposed 3D-aware virtual attention based refiner (3DVA-Refiner) computes a cost map from the feature pairs with relative pose and initial depth. Finally, D_r^0 is iteratively optimized by the 3DVA-Refiner and gradually converges to the optimal depth D_r^*.

discretization (SID) strategy to discretize depth with the ordinary regression loss. Adabins [6] proposed to divide the depth range into bins for adaptive depth quantization. In addition, other methods introduce auxiliary information to assist the training of the depth network, such as sparse depth [24], mixing datasets [64], or segmentation information [34, 56, 61, 90].

However, it is an inherently ambiguous and ill-posed problem to estimate the 3D world depth from monocular images due to the lack of geometric constraints [58, 89].

Stereo training refers to using stereo image pairs during the training of the depth network. Particularly, left-right consistency allows methods to release the dependence on ground-truth depth for training and to perform self-supervised depth estimation with binocular stereo images [25, 34–36, 60, 83]. Grag et al. [19] first formulated the photometric consistency loss based on left-right images to estimate depth without using LiDAR supervision. Then, MonoDepth [22] and MonoDepth2 [21] performed a stereo-trained version that employed left-right depth reprojection to learn depth from stereo pairs. The reprojection loss was hence computed at a higher resolution [21, 49, 59] to stimulate the depth-awareness of high-level feature maps. Depth Hints [79] was introduced as depth suggestions obtained from simple off-the-shelf stereo algorithms to study the reprojections in depth prediction from stereo-based self-supervision, and the pre-computed semi-global matching (SGM) [29, 30] depth from binocular stereo data was adopted as a supervisory signal. DistDepth [83] adopted stereo images as inputs, and introduced the structure distillation approach to learn knacks from an off-the-shelf relative depth estimator.

Inspired by the leveraging of geometric constraints from multiple cameras for the training of self-supervised depth estimation, we proposed to train a monocular depth estimator based on generated virtual-view semantic features.

2.2 Novel View Synthesis

To provide sufficient 3D geometry constraints and avoid the inconsistency between training and inference strategies, we investigate the potential of novel view synthesis (NVS). The methods based on neural radiance fields [3, 4, 9, 31, 33, 51, 55, 67] (Nerfs) have achieved impressive fidelity on photo-realistic novel-view synthesis. Specifically, the multi-layer perceptron (MLP) was adopted to map spatial coordinates to color (appearance) and density (geometry) values. Novel views were rendered by evaluating the optimized MLP along rays and incorporating the color over the density with the volume rendering [50] approach. In addition, recent approaches also regulate the MLP on feature grids or voxels, resulting in a faster speed of inference [7, 47, 53, 73] and more robust generalizations to novel scenes [52, 70, 88]. Despite these advances, they require multiple images for the rendering of novel scenes.

Considering NVS from a single image, methods also attempted to incorporate the usage of both NeRFs and depth estimation. Earlier works [10, 11, 78] predict layered depth images (LDI) [69] for rendering. Afterwards, methods [72, 77] started learning view synthesis by using deep learning to generate the multiplane image (MPI) [91], which is a camera-centric layered 3D representation with multiple layers of color and alpha values at certain depths. MINE [42] predicted a 4-channel image (volume density and RGB) at arbitrary depth values to reconstruct the camera frustum and synthesize the frustum as novel RGB or depth views with differentiable rendering. Besides, Wiles et al. [80] proposed an end-to-end NVS method by employing a network to predict both a per-pixel depth map and a feature map from a single image. It generated novel-view images through neural rendering and network refinement on the reprojected features. Furthermore, NVS-MonoDepth [5] proposed to improve the monocular depth prediction with novel view synthesis, where the prediction results of a monocular depth network are warped to an additional viewpoint. Recently, Wimbauer et al. [81] generalized the depth prediction formulation to a continuous density field by introducing an encoder-decoder network to predict a dense feature map from a single image. They applied volume rendering to perform both NVS and depth prediction.

Motivated by the above effective neural reconstruction approaches, we hence propose a novel supervised coarse-to-fine framework that incorporates the idea of NVS to construct a virtual-view feature simulator. It feeds the initial results with generated virtual feature pairs to the spatial refinement network, and iteratively optimizes the high-level semantic features for depth estimation.

3 METHOD

3.1 Overview

Given a reference image I_r, our goal is to predict an accurate depth map D_r^* from I_r. We present our novel framework V2Depth in Fig. 2, which mainly consists of three components: a novel virtual-view feature simulator, a depth context net, and a 3D-aware virtual attention based refiner. We use ResNet18 [28] as the feature net and context net to extract features. Specifically, we first regard I_r as the network input. The input image first goes through our proposed virtual-view feature simulator to produce virtual-view feature pairs, which consist of the (real) reference feature F_r and the virtual view feature F_v. In parallel, a depth context network extracts the context feature F^c from I_r. Next, inside our 3DVA-Refiner, a depth head takes in the reference feature F_r, which is extracted from I_r and outputs an initial depth map D_r^0. Furthermore, a cost map is constructed from the current depth map D_r^t, virtual-view and reference view features, and virtual camera pose. Hence, the final feature F^f is generated by a convolution net and the proposed 3D-aware virtual attention. The proposed 3D-aware virtual attention and gated recurrent units (GRU) generate the depth updates $\Delta\mathbf{D}$ and gradually refine it to gain the optimal depth map D_r^*.

3.2 Virtual-View Feature Simulator

We firstly define $\mathbf{I}_{\mathrm{I}} \in [0,1]^{3\times H\times W} = \left(\mathbb{R}^3\right)^{\Omega}$ as the reference image on the lattice $\Omega = \{1,\ldots,H\}\times\{1,\ldots,W\}$. Hence we leverage the volume rendering [50, 81] and feature grid sampling to simulate virtual-view feature maps. During training, we use the left-right stereo image pairs of the widely used datasets [14, 18, 86] as inputs, where the left image is the reference image and the right image is the novel view image ground truth for the training of virtual view synthesis. Under the assumption of homogeneous coordinates [81], a point $\mathbf{x} \in \mathbb{R}^3$ in world coordinates is projected onto the image plane of image i by the following operation: $\pi_i(\mathbf{x}) = K_i T_i \mathbf{x}$, where $T_{\mathrm{i}} \in \mathbb{R}^{4\times 4}$ and $K_{\mathrm{i}} \in \mathbb{R}^{3\times 4}$ denote the corresponding world-to-camera pose matrix and projection matrix, respectively.

We design two components within the virtual-view feature simulator, an image encoder $\mathbf{E}$ that encodes the input image into pixel-aligned feature grids, and a multi-layer perception (MLP) ϕ. Given a spatial location and its corresponding encoded feature, ϕ outputs the volume density. We model the spatial query in the camera space of the input view, rather than a canonical space. It is not only integral for generalization to unseen scenes and object categories, but also for flexibility, since no clear canonical coordinate system exists in scenes with multiple objects or real scenes. The model is trained with the volume rendering method [81]. Given the input image I_r, our feature network predicts a pixel-aligned feature map $\mathbf{F} \in \mathbb{R}^{c\Omega}$. Inspired by the work [81], every feature $f_u = F(u)$ at pixel location $\mathbf{u} \in \Omega$ is able to capture the distribution of local geometry along the ray from the camera origin through the pixel at $\mathbf{u}$.

We adopt MLP ϕ to predict the density of the scene based on the input view. To predict a density value at a 3D coordinate x, we first project $\mathbf{x}$ onto the input image $\mathbf{u}_{\mathrm{r}}' = \pi_{\mathrm{r}}(\mathbf{x})$ and bilinearly sample the feature $f_{\mathbf{u}'} = \mathbf{F}\left(\mathbf{u}'\right)$ at that position. We hence encode the feature $f_{\mathbf{u}'}$, the positional encoding $\gamma(d)$ [51] of the distance d between $\mathbf{x}$ and the camera origin, and the positional encoding $\gamma\left(\mathbf{u}_{\mathrm{r}}'\right)$ of the pixel with the MLP ϕ. We use the feature representation $f_{\mathbf{u}'}$ to describe the density along a ray through the camera center and pixel $\mathbf{u}'$. We then apply ϕ to predict the density $\sigma_{\mathbf{x}}$ at the 3D location $\mathbf{x}$ based on $f_{\mathbf{u}'}$ and a distance to the camera:

$$\sigma_{\mathbf{x}} = \phi\left(f_{\mathbf{u}_{\mathrm{r}}'}, \gamma(d), \gamma\left(\mathbf{u}_{\mathrm{r}}'\right)\right) \tag{1}$$

We perform volume rendering and feature grid sampling to simulate virtual-view feature maps. When performing volume rendering from a novel viewpoint, instead of retrieving feature grids directly from our scene representation, we sample the feature grid from the available feature maps for a point in 3D space. Concretely, we project a 3D point $\mathbf{x}$ into a frame k and then bilinearly sample the feature $f_{\mathbf{x},k} = \mathbf{I}_k\left(\pi_k(\mathbf{x})\right)$ at this position. By combining $sigma_{\mathbf{x}}$ and $f_{x,k}$, we can perform volume rendering for the virtual-view feature simulation. To obtain the feature grid f_k for a point in a novel view, we emit a ray from the camera and integrate the feature grid along the ray over the probability of the ray ending at a certain distance. To approximate this integral, density, feature grids are evaluated at s discrete steps $\mathbf{x}_i$ along the ray. σ_i is regarded as the distance between $\mathbf{x}_i$ and $\mathbf{x}_{i+1}$. We set α_i as the probability of a ray ending between $\mathbf{x}_i$ and $\mathbf{x}_{i+1}$. From the previous α_j, the probability T_i that $\mathbf{x}_i$ is not occluded can be calculated. In other words, the ray does not terminate before $\mathbf{x}_i$.

$$\alpha_i = \exp\left(1 - \sigma_{\mathbf{x}_i}\delta_i\right) \quad T_i = \prod_{j=1}^{i-1}\left(1 - \alpha_j\right) \tag{2}$$

$$\hat{f}_k = \sum_{i=1}^{S} T_i \alpha_i f_{\mathbf{x}_i,k} \tag{3}$$

To train the virtual-view feature simulator based on contrastive representation learning [27, 57], we learn the output feature embedding space to maximize the cosine similarity of the positive image pairs in the training batch while minimizing the cosine similarity of the embedding of the negative image pairs. Inspired by the CLIP method [62], we optimize the same symmetric cross-entropy loss over these similarity scores. This training batch construction technique was proposed in deep metric learning as the multi-class N-pair loss [71], and used for our contrastive representation learning as the InfoNCE loss [57]. In our training batch, all the positive image pairs have two images from different view directions, and all the

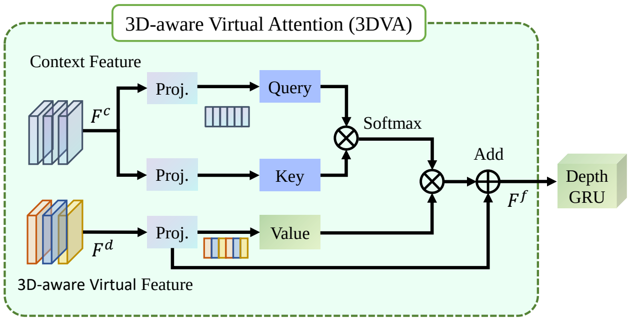


Figure 3: Illustration of the 3D-aware virtual attention (3DVA). The 3D-aware virtual feature F^d and the context features F^c are fed into our 3D-aware virtual attention together. The 3DVA outputs the final feature F^f to the depth GRU for iterative refinement.

negative image pairs have two identical images (from the same view direction). We employ stereo images to train the VVF-Simulator while using only monocular images as input during inference.

3.3 3DVA-Refiner

We adopt the 3D-aware virtual attention based refiner to iteratively refine the initial estimate, so that depth map D_r^* could gradually converges. We also introduced it with the additional 3D-aware guidance from the corporation of the context features F^c from the depth context net and the virtual camera pose vector. The refiner takes in these inputs to produce the prediction offset ΔD_r^t and then updates the depth as follows:

$$D_r^{t+1} \longleftarrow D_r^t + \Delta D_r^t. \tag{4}$$

In detail, we first construct the virtual cost map from D_r^t, P_v, F_r and F_v, as shown in Fig. 2. Notably, the virtual cost map measures the photometric cost in feature space between the reference image I_r and the virtual right image I_v. Given the current depth map D_r^t of the reference image I_r and the virtual camera pose P_v of the virtual image I_v with respect to the real reference image I_r, the virtual cost map is constructed at each pixel x in the reference image I_r:

$$\mathbf{C}_v^t(x) = \left\| \mathcal{F}_v \left(\pi \left(\mathbf{P}_v \circ \pi^{-1}(x, \mathbf{D}_r^t(x)) \right) \right) - \mathcal{F}_r(x) \right\|_2, \tag{5}$$

where $\pi()$ is the projection of 3D points in 3D space onto the image plane and $\pi^{-1}(x, D_r^t(x))$ represents the inverse projection. The transformation converts 3D points from the camera space of I_r to that of I_v. Next, we adopt two convolution layers as a simple feature extractor to obtain the 3D-aware virtual attention features F^d from the virtual cost map, preparing for the following 3DVA. Thus it rectifies the implied content-spatial inconsistency for occluded or moving objects as shown in Fig. 1, Fig. 4 and Fig. 5, as well as effectively promoting information integration between 3D-aware virtual attention features and depth context features.

3D-aware Virtual Attention (3DVA). The construction of the cost volume heavily relies on the static scene assumption, where it supposes that the object points remain static at time t and t^*. Thus, we re-project the features at time t to another plane with pose t^* at time t^*, to match cost values. However, moving objects break this assumption since targets such as cars, trains, or pedestrians with a certain speed could move within the time gap [34, 44]. It thus gives rise to the feature inconsistency deviation, degraded (mismatching) cost values and re-projection loss, and drawbacks to our depth optimization. We discard explicit settings such as the object motion prediction module or disentangle module [16, 39, 82], which brings additional complexity and ignores the potential of complementary context-spatial information. Instead, we deliver our 3D-aware virtual attention (3DVA) to implicitly rectify the mismatching problem, thus achieving feature consistency and estimation integrity. It efficiently cooperates the global 3D-aware virtual features with context features via the attention operation. Quantitative and qualitative results in Section 4 demonstrate the superiority of our method.

Specifically, as shown in Fig. 3, for depth optimization, we first lift the 3DVA feature F^d to value (V) vectors via the mapping function $\sigma(\cdot)$. Meanwhile, we create query (Q) and key (K) vectors by adding the mapping functions $\theta(\cdot)$ and $\phi(\cdot)$ from the context feature F^c, and prepare long-range geometry embedding (LGE). We first allocate the query, key, and value as $Q = \theta(F^c) \oplus LGE$, $K = \phi(F^c) \oplus LGE$, and $V = \sigma(F^s)$, respectively. Subsequently, the 3D-aware virtual attention is denoted as follows:

$$F^d = f_s(Q \otimes K) \otimes V \oplus F^s, \tag{6}$$

where f_s denotes the softmax operation, $\oplus$ denotes the point-wise addition and $\otimes$ denotes matrix multiplication.

Intuitively, compared with directly feeding F^d and F^c for refinement, our 3DVA explicitly aligns the features for occluded and moving objects through the cross-attention mechanism to compensate for the mismatching discrepancy. It thus guarantees 3D-aware feature fusion and seamless depth refinement. It also helps the context feature to fulfill the integrity of targets with occlusion relationship, such as the fourth column (d) of Fig. 1, where our 3DVA successfully rectify the wrong estimation for the contour of cars, pedestrians and bicyclists.

3.4 Supervised Training Loss

Our 3D-aware virtual-view refiner is trained for minimizing the depth error. The depth loss is formulated as the L1 distance between the predicted depth D_r of the reference image I_r and the associated ground truth $\hat{D}_r$:

$$\mathcal{L}_{\text{depth}} = \sum_{s=1}^{m} \gamma^{m-s} \| \mathrm{D_r} - \hat{\mathrm{D}}_\mathrm{r} \|_1, \tag{7}$$

where the discounting factor γ is 0.85 and s is the stage number. There are m optimization stages for depth refinements. At each stage, the 3DVA-Refiner iteratively optimizes the depth n times.

4 EXPERIMENTS AND RESULTS

4.1 Datasets

KITTI is a popular depth estimation benchmark for autonomous driving [20]. It provides over 93,000 depth maps with corresponding raw LiDAR scans and RGB images aligned with the raw data. We follow the widely-used KITTI Eigen split [14] to conduct the monocular depth estimation experiment. It contains 23488 image pairs from 32 scenes for training and 697 images from 29 scenes for

Method	Cap	Input	GT type	Abs Rel ↓	Sq Rel ↓	RMSE ↓	RMSE$_{log}$ ↓	$\delta_1 < 1.25$ ↑	$\delta_2 < 1.25^2$ ↑	$\delta_3 < 1.25^3$ ↑
PackNet-SfM [25]	0-80m	M→S	Velodyne	0.090	0.618	4.220	0.179	0.893	0.962	0.983
DRO [23]	0-80m	Multi-Frame	Velodyne	0.073	0.528	3.888	0.163	0.924	0.969	0.984
V2Depth (ours)	0-80m	Multi-Frame	Velodyne	0.061	0.425	2.788	0.116	0.952	0.985	0.992
BTS [38]	0-80m	Single-Frame	Improved	0.059	0.241	2.756	0.096	0.956	0.993	0.998
GLPDepth [32]	0-80m	Single-Frame	Improved	0.057	–	2.297	0.086	0.967	0.996	0.999
PackNet-SfM [25]	0-80m	M→S	Improved	0.064	0.300	3.089	0.108	0.943	0.989	0.997
BANet [74]	0-80m	Multi-Frame	Improved	0.083	–	3.640	0.134	–	–	–
DeepV2D(2-view) [75]	0-80m	Multi-Frame	Improved	0.064	0.350	2.946	0.120	0.946	0.982	0.991
DRO [23]	0-80m	Multi-Frame	Improved	0.047	0.199	2.629	0.082	0.970	0.994	0.998
Ours	0-80m	Single-Frame	Improved	**0.037**	**0.143**	**1.985**	**0.068**	**0.983**	**0.998**	**0.999**

Table 1: Quantitative results of supervised monocular depth estimation methods on the KITTI Eigen split. Note that the seven widely used metrics are calculated strictly following the baseline [23] and ground-truth median scaling is applied. "M→S" means monocular multiple frame images are used in training while only a single frame image is used for inference. We utilize bold to highlight the best results.

Method	Reference	Input	Abs Rel ↓	Sq Rel ↓	RMSE ↓	RMSE$_{log}$ ↓	$\delta_1 < 1.25$ ↑	$\delta_2 < 1.25^2$ ↑	$\delta_3 < 1.25^3$ ↑	FPS ↑
BANet [74]	ICLR 2019	Multi-Frame	0.083	–	3.640	0.134	–	–	–	–
DeepV2D [75]	ICLR 2020	Multi-Frame	0.064	0.350	2.946	0.120	0.946	0.982	0.991	0.67
DRO [23]	RA-L 2023	Multi-Frame	0.059	0.230	2.799	0.092	0.964	0.994	0.998	6.25
MaGNet [2]	CVPR 2022	Multi-Frame	0.054	0.162	2.158	0.083	0.971	–	–	–
Xu et al. [85]	CVPR 2018	Single-Frame	0.122	0.897	4.677	–	0.818	0.954	0.985	–
DORN [17]	CVPR 2018	Single-Frame	0.072	0.307	2.727	0.120	0.932	0.984	0.995	–
Yin et al. [87]	ICCV 2019	Single-Frame	0.072	–	3.258	0.117	0.938	0.990	0.998	–
PackNet-SAN [24]	CVPR 2021	Single-Frame	0.062	–	2.888	–	0.955	–	–	–
DPT* [63]	ICCV 2021	Single-Frame	0.062	–	2.573	0.092	0.959	0.995	0.999	–
PWA [40]	AAAI 2021	Single-Frame	0.060	0.221	2.604	0.093	0.958	0.994	0.999	–
AdaBins [6]	CVPR 2021	Single-Frame	0.058	0.190	2.360	0.088	0.964	0.995	0.999	2.96
P3Depth [58]	CVPR 2022	Single-Frame	0.071	0.270	2.842	0.103	0.953	0.993	0.998	–
NeWCRFs [89]	CVPR 2022	Single-Frame	0.052	0.155	2.129	0.079	0.974	0.997	0.999	3.48
SIDP [45]	CVPR 2023	Single-Frame	0.050	–	2.020	0.075	0.976	–	–	–
Ours	–	Single-Frame	**0.043**	**0.150**	**1.989**	**0.069**	**0.981**	**0.998**	**0.999**	3.05

Table 2: Quantitative results on KITTI Eigen split with the cap of 0-80m. Note that the seven widely used metrics are calculated strictly following NeWCRFs [89]. "Abs Rel" error occupies the main ranking metric. "*" means using additional data for training. We utilize bold to highlight the best results of single-frame methods and multi-frame methods.

testing. The corresponding depth of each image is sampled sparsely by the rotating LiDAR sensor. The "Improved" ground-truth in Table 1 refers to the improved annotated depth map, which aggregates Lidar points from five successive frames and stereo images.

Virtual KITTI 2 is an updated version of the well-known photo-realistic synthetic video dataset Virtual KITTI [18] and designed to learn and evaluate computer vision models for video understanding tasks such as depth estimation. It consists of 5 sequence clones from the KITTI tracking benchmark and contains 50 monocular videos generated from five different virtual worlds in urban settings under different imaging and weather conditions. These synthetic videos are fully annotated with depth labels.

DrivingStereo is a large-scale stereo dataset that covers a diverse set of outdoor driving scenarios with over 180k images. Credit to the huge amount of available data, deep-learning models usually pre-train on it to improve the robustness to real-world driving scenes [86]. It is designed to learn and evaluate vision models for video understanding tasks such as object detection and depth estimation. High-quality labels of depth maps are generated by a model-guided filtering strategy from multi-frame LiDAR points.

4.2 Implementation Details

We implement our V2Depth in PyTorch and train it for 100 epochs with a mini-batch size of 4. We adopt ResNet18 [28] as the feature network, which is pre-trained on ImageNet [68] and then trained by the contrastive representation learning [57, 62]. The learning rate is 2×10^{-4} for depth refinement, which is decayed by a constant step (gamma=0.5 and step size=30). We set $\beta_1 = 0.9$ and $\beta_2 = 0.999$ in the Adam optimizer. We resize the input images to 320×960 for training, and set the number of sequential images to 2 for 3DVA-Refiner by balancing both computation efficiency and prediction accuracy. We fix $m = 3$ and $n = 4$ in the experiments.

4.3 Evaluation of Our Method

Evaluation on KITTI. We first compare our V2Depth against the state-of-the-art supervised depth estimators on the KITTI dataset; see Tables 1 & 2 for the results. For a fair comparison, all methods are evaluated given the same sequential images. In Table 1, the seven widely-used evaluation metrics are calculated following [23] and the ground-truth median scaling is applied. On the other hand,

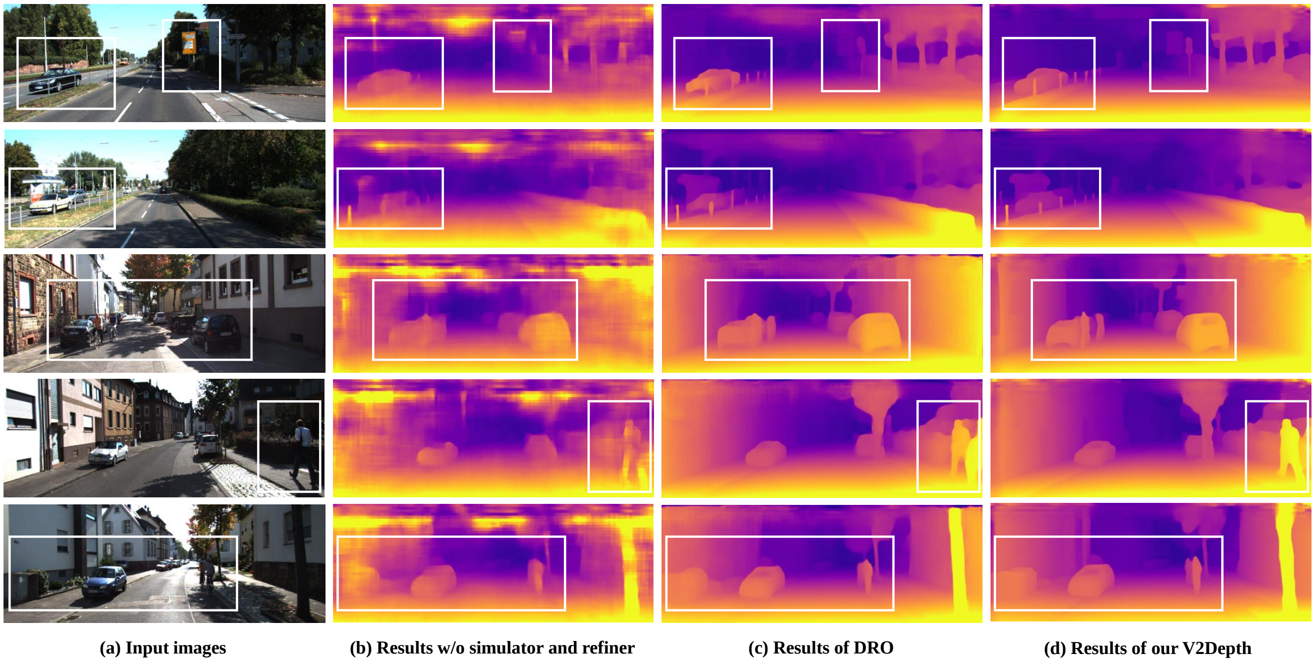

(a) Input images
(b) Results w/o simulator and refiner
(c) Results of DRO
(d) Results of our V2Depth

Figure 4: Qualitative results on the KITTI Eigen split dataset.

Method	Abs Rel ↓	Sq Rel ↓	RMSE ↓	RMSE$_{log}$ ↓
DORN	0.068	0.274	2.693	0.115
Adabins	0.041	0.164	1.981	0.094
DRO	0.040	0.153	1.903	0.092
NeWCRFs	0.039	0.157	1.977	0.085
Ours	**0.032**	**0.125**	**1.683**	**0.081**

Table 3: Quantitative results on the Virtual KITTI 2.

Method	Abs Rel ↓	Sq Rel ↓	RMSE ↓	RMSE$_{log}$ ↓
DORN	0.055	0.126	1.217	0.103
Adabins	0.036	0.083	0.612	0.089
DRO	0.035	0.079	0.595	0.086
NeWCRFs	0.032	0.071	0.548	0.077
Ours	**0.026**	**0.068**	**0.506**	**0.047**

Table 4: Quantitative results on the DrivingStereo.

the seven evaluation metrics in Table 2 are calculated according to NeWCRFs [89]. We also evaluate our method against others in terms of the frames per second (FPS) using the same Nvidia RTX A6000 GPU. Clearly, V2Depth achieves state-of-the-art performance over all the evaluation metrics under both the effectiveness and efficiency considerations.

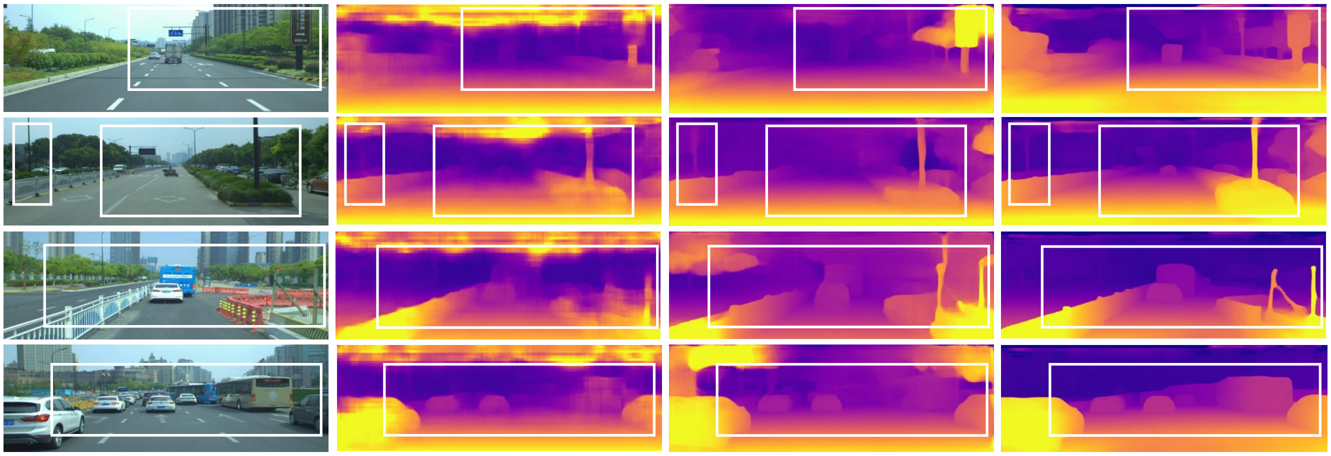
We further show the qualitative comparisons in Fig. 4 by comparing our method (d) with the recent approach DRO [23] (c) and the simplified depth predictor without our simulator and refiner (b). The simplified depth predictor is composed of a feature network and a depth head, and estimates a initial coarse depth map. Then the coarse-to-fine method DRO predicts relatively accurate depth by adding a deep recurrent optimizer as the refiner. Comparing to DRO, our V2Depth has a novel feature simulator and a 3D-aware virtual attention module, and replaces temporal adjacency features with the generated virtual-view features as input to the refiner. As shown in the white boxes of Fig. 4, our method yields more accurate and cleaner depth estimation results for objects with occlusion, for example, the car behind the road railings, bicyclists, and pedestrians. In addition, the results of our V2Depth demonstrate the better depth and clearer contours of objects in some difficult regions such as messy environments and distant traffic signs (please refer to regions marked with white boxes).

Evaluation on Virtual KITTI 2. We further compare our method with recent approaches on the virtual KITTI 2 dataset. We use a subset of the virtual KITTI 2, which contains 1,700 image pairs for training and 193 images for testing. The quantitative results of our method compared with others are shown in Table 3, where four widely used evaluation metrics are calculated for the test set. Notably, our V2Depth achieves significantly outperforms on all evaluation metrics.

Evaluation on DrivingStereo. We further evaluate our approach on the DrivingStereo benchmark. We use a subset of the DrivingStereo dataset, which contains 7251 image pairs for training and 500 images for testing. The quantitative and qualitative results are shown in Table 4 and Fig. 5. Our method outperforms these monocular depth predictors in all evaluation metrics. The qualitative results in Fig. 5 illustrate that our V2Depth is robust to obscured vehicles, distant vehicles, the signage in front of grasses, and the consistency of object boundaries.

4.4 Computation Time Analysis

We evaluate the speed of inference on the Nvidia RTX A6000 GPU with the metric of frames per second (FPS). Credit to our usage of the lightweight ResNet18 [28] backbone, compared to the single-frame method Adabins [6], the inference speed of our V2Depth



(a) Input images (b) Results w/o simulator and refiner (c) Results of DRO (d) Results of our V2Depth

Figure 5: Qualitative results on the DrivingStereo dataset.

is improved by 3%, 3.05 (Ours) vs. 2.96 (Adabins). As shown in Table 2, the inference speed of V2Depth is comparable against the NeWCRFs [89] (3.05 vs. 3.48), while our performance significantly outperforms both the NeWCRFs and the SIDP [45]. In detail, we achieve a 14% reduction in the main metric "Abs Rel" error (0.043 vs. 0.050) compared to the top estimator.

4.5 Ablation Study

IDP	Refiner	TNF	VVF	CL	3DVA	Abs Rel ↓	Sq Rel ↓	δ_1 ↑
✓						0.091	0.473	0.908
✓	✓	✓				0.059	0.230	0.964
✓	✓		✓			0.050	0.186	0.973
✓	✓		✓	✓		0.047	0.171	0.976
✓	✓		✓		✓	0.045	0.159	0.979
✓	✓		✓	✓	✓	**0.043**	**0.150**	**0.981**

Table 5: Ablation study on the KITTI dataset. "IDP" : initial depth predictor, "TNF": temporal neighboring features from the monocular temporal sequences, "VVF": virtual view features generated from our VVF-Simulator, "Refiner": refinement network without our 3DVA, "CL": contrastive learning loss, "3DVA": 3D-aware virtual attention.

To inspect the influence of our designs, we conduct an ablation study on the KITTI [14, 20] benchmark and provide the results in Table 5. The bottom row indicates our full pipeline.

Initial Depth Predictor. We first build an initial depth predictor named "IDP", which includes the same feature network and a depth head but without the simulator or the refiner. As shown in the second column (b) of Fig. 1, Fig. 4, and Fig. 5, the results predicted by IDP are coarser against the full coarse-to-fine framework.

Refiner. We add our refiner to IDP to construct a baseline coarse-to-fine framework. In detail, given the initial depth, virtual camera pose and multiple features, the GRU-based refiner constructs a cost map and iteratively optimizes the depth map by minimizing a feature-metric cost. Note that in contrast to DRO [23], which uses temporal neighboring features from video sequences for their refinement network, our approach generates virtual-view features to provide meaningful 3D geometry for the proposed 3DVA-Refiner.

Virtual-View Features vs. Temporal Neighboring Features We present the results of using virtual-view features and temporal neighboring features in Table 5. The significant improvement in performance, with a reduction in "Abs Rel" error from 0.059 to 0.050, indicates that the virtual view features of our VVF-simulator could provide representative spatial cues and 3D geometry to the model.

Contrastive Learning Loss We replace the MSE loss with the contrastive representation learning loss [57] to learn the virtual-view features. The performance gain demonstrates the effectiveness of contrastive learning in producing more accurate and meaningful virtual-view features by comparing different view feature pairs.

3D-Aware Virtual Attention. We add 3DVA after the cost map to obtain 3D-aware virtual attention by modeling the spatial relation. It learns the 3D geometric constraints between reference view feature maps and virtual-view feature maps. As a result, the learned 3D-aware virtual attention features are fed to the GRU optimizer and yield a noticeable performance gain. As the Table 5, the "Abs Rel" error is reduced from 0.047 to 0.043.

5 CONCLUSION

In this work, we propose V2Depth, a novel monocular depth estimation framework with virtual view feature simulations and 3D-aware refinements. To address the issue of limited spatial cues giving merely a single image, we design a virtual view feature simulator with the contrastive learning strategy. It generates virtual view features to provide representative 3D geometry for depth prediction. Next, we propose a 3D-aware virtual attention refiner to exploit the abundant information of reference and virtual view features. It leverages the cost minimization to iteratively optimize the initial depth estimations. As a result, our V2Depth overcomes the performance bottleneck of the monocular depth estimation on the challenging KITTI benchmark, as well as the Virtual KITTI2 and DrivingStereo datasets. In the future, we would like to apply our virtual-view feature simulator to more 3D scene understanding tasks in autonomous driving.

REFERENCES

[1] Shubhra Aich, Jean Marie Uwabeza Vianney, Md Amirul Islam, and Mannat Kaur Bingbing Liu. 2021. Bidirectional attention network for monocular depth estimation. In *IEEE Int. Conf. on Robotics and Automation (ICRA)*. IEEE, 11746–11752.

[2] Gwangbin Bae, Ignas Budvytis, and Roberto Cipolla. 2022. Multi-View Depth Estimation by Fusing Single-View Depth Probability with Multi-View Geometry. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2842–2851.

[3] Jonathan T Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P Srinivasan. 2021. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *IEEE Int. Conf. on Computer Vision (ICCV)*. 5855–5864.

[4] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. 2022. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 5470–5479.

[5] Zuria Bauer, Zuoyue Li, Sergio Orts-Escolano, Miguel Cazorla, Marc Pollefeys, and Martin R Oswald. 2021. NVS-MonoDepth: Improving monocular depth prediction with novel view synthesis. In *2021 International Conference on 3D Vision (3DV)*. IEEE, 848–858.

[6] Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. 2021. AdaBins: Depth estimation using adaptive bins. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 4009–4018.

[7] Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. 2022. Tensorf: Tensorial radiance fields. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXII*. Springer, 333–350.

[8] Zhi Chen, Xiaoqing Ye, Wei Yang, Zhenbo Xu, Xiao Tan, Zhikang Zou, Errui Ding, Xinming Zhang, and Liusheng Huang. 2021. Revealing the reciprocal relations between self-supervised stereo and monocular depth estimation. In *IEEE Int. Conf. on Computer Vision (ICCV)*. 15529–15538.

[9] Kangle Deng, Andrew Liu, Jun-Yan Zhu, and Deva Ramanan. 2022. Depth-supervised nerf: Fewer views and faster training for free. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 12882–12891.

[10] Helisa Dhamo, Nassir Navab, and Federico Tombari. 2019. Object-driven multi-layer scene decomposition from a single image. In *IEEE Int. Conf. on Computer Vision (ICCV)*. 5369–5378.

[11] Helisa Dhamo, Keisuke Tateno, Iro Laina, Nassir Navab, and Federico Tombari. 2019. Peeking behind objects: Layered depth prediction from a single image. *Pattern Recognition Letters* 125 (2019), 333–340.

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *Int. Conf. on Learning Representations (ICLR)*. OpenReview.net.

[13] Ruofei Du, Eric Turner, Maksym Dzitsiuk, Luca Prasso, Ivo Duarte, Jason Dourgarian, Joao Afonso, Jose Pascoal, Josh Gladstone, Nuno Cruces, et al. 2020. DepthLab: Real-time 3D interaction with depth maps for mobile augmented reality. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology (UIST)*. 829–843.

[14] David Eigen, Christian Puhrsch, and Rob Fergus. 2014. Depth map prediction from a single image using a multi-scale deep network. *Conference and Workshop on Neural Information Processing Systems (NeurIPS)* (2014), 2366–2347.

[15] Fatima El Jamiy and Ronald Marsh. 2019. Survey on depth perception in head mounted displays: distance estimation in virtual reality, augmented reality, and mixed reality. *IET Image Processing* 13, 5 (2019), 707–712.

[16] Ziyue Feng, Liang Yang, Longlong Jing, Haiyan Wang, YingLi Tian, and Bing Li. 2022. Disentangling Object Motion and Occlusion for Unsupervised Multi-frame Monocular Depth. In *European Conf. on Computer Vision (ECCV)*. 228–244.

[17] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. 2018. Deep ordinal regression network for monocular depth estimation. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2002–2011.

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

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

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

[21] Clement Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow. 2019. Digging into self-supervised monocular depth estimation. In *IEEE Int. Conf. on Computer Vision (ICCV)*. 3828–3838.

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

[23] Xiaodong Gu, Weihao Yuan, Zuozhuo Dai, Siyu Zhu, Chengzhou Tang, Zilong Dong, and Ping Tan. 2023. DRO: Deep Recurrent Optimizer for Video to Depth. *IEEE Robotics and Automation Letters (RA-L)* (2023).

[24] Vitor Guizilini, Rares Ambrus, Wolfram Burgard, and Adrien Gaidon. 2021. Sparse auxiliary networks for unified monocular depth prediction and completion. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 11078–11088.

[25] Vitor Guizilini, Rares Ambrus, Sudeep Pillai, Allan Raventos, and Adrien Gaidon. 2020. 3d packing for self-supervised monocular depth estimation. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2485–2494.

[26] Vitor Guizilini, Rareș Ambruș, Dian Chen, Sergey Zakharov, and Adrien Gaidon. 2022. Multi-Frame Self-Supervised Depth with Transformers. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 160–170.

[27] Raia Hadsell, Sumit Chopra, and Yann LeCun. 2006. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, Vol. 2. IEEE, 1735–1742.

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

[29] Heiko Hirschmuller. 2006. Stereo vision in structured environments by consistent semi-global matching. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Vol. 2. IEEE, 2386–2393.

[30] Heiko Hirschmuller. 2007. Stereo processing by semiglobal matching and mutual information. *IEEE Transactions on pattern analysis and machine intelligence* 30, 2 (2007), 328–341.

[31] Ajay Jain, Matthew Tancik, and Pieter Abbeel. 2021. Putting nerf on a diet: Semantically consistent few-shot view synthesis. In *IEEE Int. Conf. on Computer Vision (ICCV)*. 5885–5894.

[32] Doyeon Kim, Woonghyun Ga, Pyungwhan Ahn, Donggyu Joo, Sehwan Chun, and Junmo Kim. 2022. Global-Local Path Networks for Monocular Depth Estimation with Vertical CutDepth. *arXiv preprint arXiv:2201.07436* (2022).

[33] Mijeong Kim, Seonguk Seo, and Bohyung Han. 2022. Infonerf: Ray entropy minimization for few-shot neural volume rendering. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 12912–12921.

[34] Marvin Klingner, Jan-Aike Termöhlen, Jonas Mikolajczyk, and Tim Fingscheidt. 2020. Self-supervised monocular depth estimation: Solving the dynamic object problem by semantic guidance. In *European Conf. on Computer Vision (ECCV)*. Springer, 582–600.

[35] Varun Ravi Kumar, Marvin Klingner, Senthil Yogamani, Stefan Milz, Tim Fingscheidt, and Patrick Mader. 2021. Syndistnet: Self-supervised monocular fisheye camera distance estimation synergized with semantic segmentation for autonomous driving. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*. 61–71.

[36] Varun Ravi Kumar, Senthil Yogamani, Markus Bach, Christian Witt, Stefan Milz, and Patrick Mäder. 2020. Unrectdepthnet: Self-supervised monocular depth estimation using a generic framework for handling common camera distortion models. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 8177–8183.

[37] Tristan Laidlow, Jan Czarnowski, and Stefan Leutenegger. 2019. DeepFusion: Real-time dense 3D reconstruction for monocular SLAM using single-view depth and gradient predictions. In *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 4068–4074.

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

[39] Seokju Lee, Sunghoon Im, Stephen Lin, and In So Kweon. 2021. Learning monocular depth in dynamic scenes via instance-aware projection consistency. In *AAAI Conf. on Artificial Intell. (AAAI)*. 1863–1872.

[40] Sihaeng Lee, Janghyeon Lee, Byungju Kim, Eojindl Yi, and Junmo Kim. 2021. Patch-wise attention network for monocular depth estimation. In *AAAI Conf. on Artificial Intell. (AAAI)*. 1873–1881.

[41] Bo Li, Chunhua Shen, Yuchao Dai, Anton Van Den Hengel, and Mingyi He. 2015. Depth and surface normal estimation from monocular images using regression on deep features and hierarchical crfs. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 1119–1127.

[42] Jiaxin Li, Zijian Feng, Qi She, Henghui Ding, Changhu Wang, and Gim Hee Lee. 2021. Mine: Towards continuous depth mpi with nerf for novel view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 12578–12588.

[43] Yingyan Li, Yuntao Chen, Jiawei He, and Zhaoxiang Zhang. 2022. Densely Constrained Depth Estimator for Monocular 3D Object Detection. In *European Conf. on Computer Vision (ECCV)*. Springer, 718–734.

[44] Zhengqi Li, Tali Dekel, Forrester Cole, Richard Tucker, Noah Snavely, Ce Liu, and William T Freeman. 2019. Learning the depths of moving people by watching frozen people. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 4521–4530.

[45] Ce Liu, Suryansh Kumar, Shuhang Gu, Radu Timofte, and Luc Van Gool. 2023. Single Image Depth Prediction Made Better: A Multivariate Gaussian Take. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*.

[46] Fayao Liu, Chunhua Shen, and Guosheng Lin. 2015. Deep convolutional neural fields for depth estimation from a single image. In *IEEE Conf. on Computer Vision*

and Pattern Recognition (CVPR). 5162–5170.

[47] Lingjie Liu, Jiatao Gu, Kyaw Zaw Lin, Tat-Seng Chua, and Christian Theobalt. 2020. Neural sparse voxel fields. *Advances in Neural Information Processing Systems* 33 (2020), 15651–15663.

[48] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *IEEE Int. Conf. on Computer Vision (ICCV).* 10012–10022.

[49] Chenxu Luo, Zhenheng Yang, Peng Wang, Yang Wang, Wei Xu, Ram Nevatia, and Alan Yuille. 2019. Every pixel counts++: Joint learning of geometry and motion with 3d holistic understanding. *IEEE transactions on pattern analysis and machine intelligence* 42, 10 (2019), 2624–2641.

[50] Nelson Max. 1995. Optical models for direct volume rendering. *IEEE Transactions on Visualization and Computer Graphics* 1, 2 (1995), 99–108.

[51] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. 2021. Nerf: Representing scenes as neural radiance fields for view synthesis. *Commun. ACM* 65, 1 (2021), 99–106.

[52] Norman Müller, Andrea Simonelli, Lorenzo Porzi, Samuel Rota Bulò, Matthias Nießner, and Peter Kontschieder. 2022. Autorf: Learning 3d object radiance fields from single view observations. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR).* 3971–3980.

[53] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. 2022. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics (ToG)* 41, 4 (2022), 1–15.

[54] Fuseini Mumuni and Alhassan Mumuni. 2022. Bayesian cue integration of structure from motion and CNN-based monocular depth estimation for autonomous robot navigation. *International Journal of Intelligent Robotics and Applications* 6, 2 (2022), 191–206.

[55] Michael Niemeyer, Jonathan T Barron, Ben Mildenhall, Mehdi SM Sajjadi, Andreas Geiger, and Noha Radwan. 2022. Regnerf: Regularizing neural radiance fields for view synthesis from sparse inputs. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR).* 5480–5490.

[56] Matthias Ochs, Adrian Kretz, and Rudolf Mester. 2019. Sdnet: Semantically guided depth estimation network. In *Pattern Recognition: 41st DAGM German Conference, DAGM GCPR 2019, Dortmund, Germany, September 10–13, 2019, Proceedings 41.* Springer, 288–302.

[57] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748* (2018).

[58] Vaishakh Patil, Christos Sakaridis, Alexander Liniger, and Luc Van Gool. 2022. P3Depth: Monocular Depth Estimation with a Piecewise Planarity Prior. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR).* 1610–1621.

[59] Sudeep Pillai, Rareş Ambruş, and Adrien Gaidon. 2019. Superdepth: Self-supervised, super-resolved monocular depth estimation. In *2019 International Conference on Robotics and Automation (ICRA).* IEEE, 9250–9256.

[60] Matteo Poggi, Filippo Aleotti, Fabio Tosi, and Stefano Mattoccia. 2020. On the uncertainty of self-supervised monocular depth estimation. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR).* 3227–3237.

[61] Siyuan Qiao, Yukun Zhu, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. 2021. Vip-deeplab: Learning visual perception with depth-aware video panoptic segmentation. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR).* 3997–4008.

[62] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning.* PMLR, 8748–8763.

[63] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. 2021. Vision transformers for dense prediction. In *IEEE Int. Conf. on Computer Vision (ICCV).* 12179–12188.

[64] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. 2020. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE transactions on pattern analysis and machine intelligence* 44, 3 (2020), 1623–1637.

[65] Manuel Rey-Area, Mingze Yuan, and Christian Richardt. 2022. 360MonoDepth: High-Resolution 360deg Monocular Depth Estimation. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR).* 3762–3772.

[66] Elisa Ricci, Wanli Ouyang, Xiaogang Wang, Nicu Sebe, et al. 2018. Monocular depth estimation using multi-scale continuous CRFs as sequential deep networks. *IEEE Trans. Pattern Anal. & Mach. Intell.* 41, 6 (2018), 1426–1440.

[67] Barbara Roessle, Jonathan T Barron, Ben Mildenhall, Pratul P Srinivasan, and Matthias Nießner. 2022. Dense depth priors for neural radiance fields from sparse input views. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR).* 12892–12901.

[68] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision* 115 (2015), 211–252.

[69] Jonathan Shade, Steven Gortler, Li-wei He, and Richard Szeliski. 1998. Layered depth images. In *Proceedings of the 25th annual conference on Computer graphics and interactive techniques.* 231–242.

[70] Prafull Sharma, Ayush Tewari, Yilun Du, Sergey Zakharov, Rares Ambrus, Adrien Gaidon, William T Freeman, Fredo Durand, Joshua B Tenenbaum, and Vincent Sitzmann. 2022. Seeing 3D Objects in a Single Image via Self-Supervised Static-Dynamic Disentanglement. *arXiv preprint arXiv:2207.11232* (2022).

[71] Kihyuk Sohn. 2016. Improved deep metric learning with multi-class n-pair loss objective. *Advances in neural information processing systems* 29 (2016).

[72] Pratul P Srinivasan, Richard Tucker, Jonathan T Barron, Ravi Ramamoorthi, Ren Ng, and Noah Snavely. 2019. Pushing the boundaries of view extrapolation with multiplane images. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR).* 175–184.

[73] Towaki Takikawa, Joey Litalien, Kangxue Yin, Karsten Kreis, Charles Loop, Derek Nowrouzezahrai, Alec Jacobson, Morgan McGuire, and Sanja Fidler. 2021. Neural geometric level of detail: Real-time rendering with implicit 3D shapes. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR).* 11358–11367.

[74] Chengzhou Tang and Ping Tan. 2018. BA-Net: Dense Bundle Adjustment Networks. In *Int. Conf. on Learning Representations (ICLR).*

[75] Zachary Teed and Jia Deng. 2019. DeepV2D: Video to Depth with Differentiable Structure from Motion. In *Int. Conf. on Learning Representations (ICLR).*

[76] Zachary Teed and Jia Deng. 2020. RAFT: Recurrent all-pairs field transforms for optical flow. In *European Conf. on Computer Vision (ECCV).* 402–419.

[77] Richard Tucker and Noah Snavely. 2020. Single-view view synthesis with multiplane images. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR).* 551–560.

[78] Shubham Tulsiani, Richard Tucker, and Noah Snavely. 2018. Layer-structured 3d scene inference via view synthesis. In *Proceedings of the European Conference on Computer Vision (ECCV).* 302–317.

[79] Jamie Watson, Michael Firman, Gabriel J Brostow, and Daniyar Turmukhambetov. 2019. Self-supervised monocular depth hints. In *Proceedings of the IEEE/CVF International Conference on Computer Vision.* 2162–2171.

[80] Olivia Wiles, Georgia Gkioxari, Richard Szeliski, and Justin Johnson. 2020. Synsin: End-to-end view synthesis from a single image. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR).* 7467–7477.

[81] Felix Wimbauer, Nan Yang, Christian Rupprecht, and Daniel Cremers. 2023. Behind the Scenes: Density Fields for Single View Reconstruction. (2023).

[82] Felix Wimbauer, Nan Yang, Lukas Von Stumberg, Niclas Zeller, and Daniel Cremers. 2021. MonoRec: Semi-supervised dense reconstruction in dynamic environments from a single moving camera. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR).* 6112–6122.

[83] Cho-Ying Wu, Jialiang Wang, Michael Hall, Ulrich Neumann, and Shuochen Su. 2022. Toward Practical Monocular Indoor Depth Estimation. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR).* 3814–3824.

[84] Zizhang Wu, Zhi-Gang Fan, Zhuozheng Li, Jizheng Wang, Tianhao Xu, Qiang Tang, Fan Wang, and Zhengbo Luo. 2022. Monocular Fisheye Depth Estimation for Automated Valet Parking: Dataset, Baseline and Deep Optimizers. In *International Conference on Intelligent Transportation Systems (ITSC).* IEEE, 01–07.

[85] Dan Xu, Wei Wang, Hao Tang, Hong Liu, Nicu Sebe, and Elisa Ricci. 2018. Structured attention guided convolutional neural fields for monocular depth estimation. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR).* 3917–3925.

[86] Guorun Yang, Xiao Song, Chaoqin Huang, Zhidong Deng, Jianping Shi, and Bolei Zhou. 2019. Drivingstereo: A large-scale dataset for stereo matching in autonomous driving scenarios. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR).* 899–908.

[87] Wei Yin, Yifan Liu, Chunhua Shen, and Youliang Yan. 2019. Enforcing geometric constraints of virtual normal for depth prediction. In *IEEE Int. Conf. on Computer Vision (ICCV).* 5684–5693.

[88] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. 2021. pixelnerf: Neural radiance fields from one or few images. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR).* 4578–4587.

[89] Weihao Yuan, Xiaodong Gu, Zuozhuo Dai, Siyu Zhu, and Ping Tan. 2022. NeWCRFs: Neural Window Fully-connected CRFs for Monocular Depth Estimation. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR).* 3916–3925.

[90] Zhenyu Zhang, Zhen Cui, Chunyan Xu, Yan Yan, Nicu Sebe, and Jian Yang. 2019. Pattern-affinitive propagation across depth, surface normal and semantic segmentation. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR).* 4106–4115.

[91] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. 2018. Stereo magnification: Learning view synthesis using multiplane images. *arXiv preprint arXiv:1805.09817* (2018).