IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

Scale-invariant and View-relational Representation Learning for Full Surround Monocular Depth

Kyumin Hwang[1,*], Wonhyeok Choi[1,*], Kiljoon Han[1], Wonjoon Choi[1], Minwoo Choi[1], Yongcheon Na[2], Minwoo Park[2], and Sunghoon Im[1,†]

***Abstract*— Recent foundation models demonstrate strong generalization capabilities in monocular depth estimation. However, directly applying these models to Full Surround Monocular Depth Estimation (FSMDE) presents two major challenges: (1) high computational cost, which limits real-time performance, and (2) difficulty in estimating metric-scale depth, as these models are typically trained to predict only relative depth. To address these limitations, we propose a novel knowledge distillation strategy that transfers robust depth knowledge from a foundation model to a lightweight FSMDE network. Our approach leverages a hybrid regression framework combining the knowledge distillation scheme–traditionally used in classification–with a depth binning module to enhance scale consistency. Specifically, we introduce a cross-interaction knowledge distillation scheme that distills the scale-invariant depth bin probabilities of a foundation model into the student network while guiding it to infer metric-scale depth bin centers from ground-truth depth. Furthermore, we propose view-relational knowledge distillation, which encodes structural relationships among adjacent camera views and transfers them to enhance cross-view depth consistency. Experiments on DDAD and nuScenes demonstrate the effectiveness of our method compared to conventional supervised methods and existing knowledge distillation approaches. Moreover, our method achieves a favorable trade-off between performance and efficiency, meeting real-time requirements.**

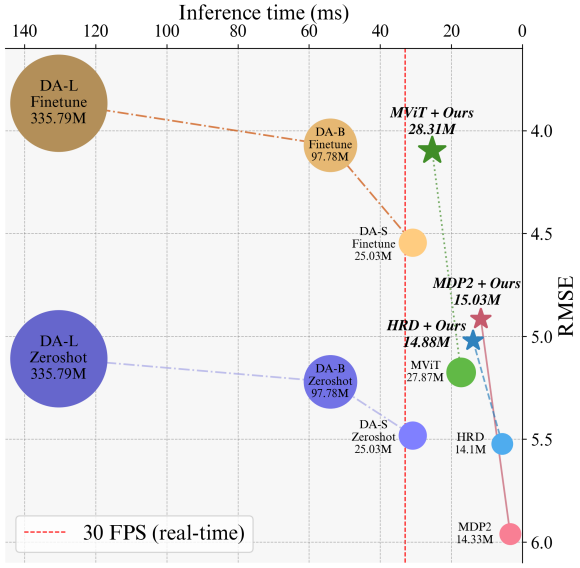

Fig. 1. Inference speed vs. RMSE trade-off curve for **nuScenes** dataset (upper-right is optimal). Each circle represents a model size, and the number inside each circle indicates the number of model parameters. (DA: DepthAnything [1], MDP2: Monodepth2 [2], HRD: HRDepth [3], MViT: MonoViT [4])

I. Introduction

Full surround camera systems have become increasingly popular in autonomous vehicles, providing a cost-effective alternative to LiDAR-based solutions by capturing a comprehensive view of the environment. In this context, Full Surround Monocular Depth Estimation (FSMDE) [5], [6], [7] has emerged as a practical and affordable solution, attracting significant research attention. Recent FSMDE methods have focused on developing lightweight network architectures that effectively balance computational efficiency and depth estimation accuracy, enabling reliable real-time decision-making for autonomous vehicles.

To this end, prior FSMDE approaches have employed lightweight networks trained in either self-supervised [5], [6], [8], [9], [7] or supervised [10] settings to efficiently learn depth by incorporating inter-view geometric consistency or spatio-temporal cues. Meanwhile, recent advancements in foundation models [11], [12], [1], [13] demonstrate remarkable generalization capabilities across various tasks, including monocular depth estimation [1], [13]. These large-scale models, trained on vast datasets, exhibit strong robustness in estimating relative depth across various environments. However, applying foundation models to FSMDE for autonomous driving presents two critical challenges:

(1) Computational cost: Foundation models are inherently large and computationally expensive, making real-time inference infeasible for autonomous driving.
(2) Metric-scale depth estimation: Foundation models,

Manuscript received: August 18, 2025; Revised: October 17, 2025; Accepted: November 9, 2025. This paper was recommended for publication by Editor Abhinav Valada upon evaluation of the Associate Editor and Reviewers' comments. This research was supported by Hyundai Motor Groups, Institute of Information & Communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (No. RS-2025-02219277, AI Star Fellowship Support Project(DGIST)), Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (RS-2025-25420118) and LG AI STAR Talent Development Program for Leading Large-Scale Generative AI Models in the Physical AI Domain (RS-2025-25442149) (*: Equal Contribution, †: Corresponding Author).

[1]Kyumin Hwang, Wonhyeok Choi, Kiljoon Han, Wonjoon Choi, Minwoo Choi, and Sunghoon Im are with the Department of Electrical Engineering & Computer Sciences, Daegu Gyeongbuk Institute of Science and Technology (DGIST), Daegu 42988, South Korea (kyumin@dgist.ac.kr; smu06117@dgist.ac.kr; kiljoon_h@dgist.ac.kr; wjchoi@dgist.ac.kr; subminu@dgist.ac.kr; sunghoonim@dgist.ac.kr).

[2]Yongcheon Na and Minwoo Park are with the Department of Autonomous Driving Perception Technology Vanguard Team, Hyundai Motor Company, Gyeonggi 13529, South Korea (ycna@hyundai.com, minwoo.park@hyundai.com)

Digital Object Identifier (DOI): see top of this page.

©2026 IEEE

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

trained on diverse datasets with varying camera intrinsics, typically produce only relative depth, making it difficult to ensure consistent metric depth across views.

To address these limitations, we propose a *Cross-interaction Knowledge Distillation* (CKD) scheme that transfers robust depth information from a foundation model to a lightweight FSMDE student network. Our method builds on the widely used hybrid regression paradigm [14], [15], [16], where depth binning techniques effectively improve scale consistency in supervised Monocular Depth Estimation (MDE). The depth binning module serves two primary functions: regressing depth bin centers to represent scale-variant depth distributions, and predicting scale-invariant depth probabilities for each pixel. The proposed CKD distills the foundation model's bin probabilities into the student network, ensuring that it captures the foundation model's scale-invariant and generalized representation as illustrated in Fig. 2-(a).

Additionally, we propose a *View-relational Knowledge Distillation* (VRKD), which transfers the inter-camera relationships learned by the teacher model to the student model. It allows the student to leverage spatial information from adjacent camera views in FSMDE. Conventional FSMDE methods typically refine metric depth estimation by enforcing geometric constraints or incorporating spatio-temporal cues across the full surround camera system. In contrast, our approach distills the structural relationships between cameras into the student network at the probability level as shown in Fig. 2-(b). Inspired by relational knowledge distillation [17], our framework encodes depth distribution relationships between adjacent views using a potential function that measures the relational energy across the N-camera-view system. This enables our framework to transfer the teacher's view-relational structure to the student network, enhancing multi-view depth consistency.

Through extensive experiments, we demonstrate that our method achieves an average improvement of 5.88% on DDAD and 11.87% on nuScenes compared to conventional supervised MDE. Additionally, our approach offers 5.13 – 11.14× faster inference speed compared to the teacher foundation model, enabling real-time inference on practical autonomous driving applications as shown in Fig. 1. We also compare our method with existing knowledge distillation methods [18], [19], [20], [21], showing that our method significantly outperforms these techniques in the FSMDE training scenario.

Our main contributions are summarized as follows:

- We propose a *Cross-interaction Knowledge Distillation* (CKD) scheme that transfers a scale-invariant depth knowledge from a foundation model to a lightweight student network.
- We present a *View-relational Knowledge Distillation* (VRKD) that effectively distills inter-camera relationships from the teacher to the student network.
- We validate the effectiveness of our method by surpassing both depth-supervised approaches and existing distillation methods across two FSMDE datasets.

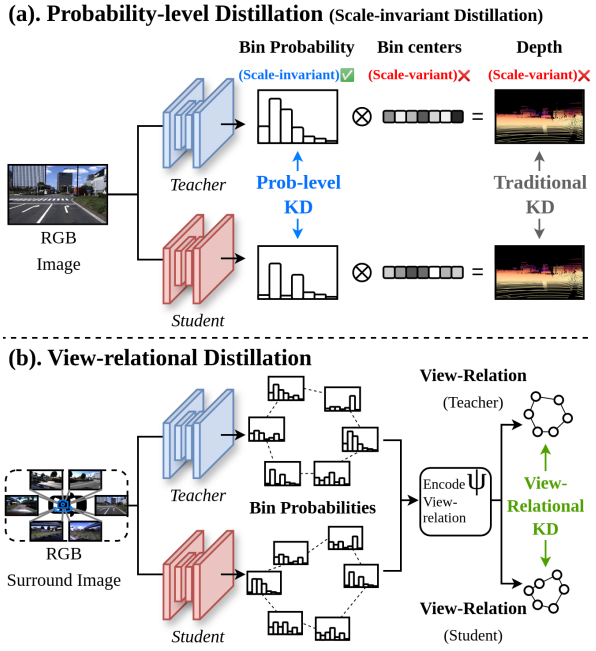

Fig. 2. Conceptual illustration of our method. (a) Leveraging an effective depth binning module from supervised methods, we perform scale-invariant distillation at the probability level, avoiding the scale sensitivity of output-level distillation. (b) We use a potential function between adjacent views to distill relational information.

II. Related Work

A. Monocular Depth Estimation

Since [22] introduced deep learning-based monocular depth estimation, the field has seen active progress [23], [24], [25]. Subsequent advancements include ViT-based architectures for effective global context embedding [26], [27] and methods leveraging geometric priors, such as ground planes or normal maps [28], [29]. Pioneering supervised methods like DORN [30] discretized depth estimation into a classification task with fixed intervals. Subsequent works advanced this by introducing adaptive binning, where methods like AdaBins [14] and LocalBins [15] dynamically predict depth distributions at the global and per-pixel levels, respectively, to better suit scene-specific content. More recently, ZoeDepth [31] introduced the MetricBins module to enable the joint estimation of relative and metric depth by iteratively refining bin centers. Moreover, depth foundation models such as Depth Anything [1] have emerged, achieving robust generalization across diverse environments.

B. Full Surround Monocular Depth Estimation

Numerous self-supervised monocular depth estimation methods [2], [3], [4], [32], [33] have achieved impressive performance without relying on explicit ground-truth supervision. Building upon these advancements, FSM [5] pioneers Full Surround Monocular Depth Estimation (FSMDE), offering a cost-efficient approach to multi-camera depth estimation for autonomous driving by incorporating multi-camera spatio-temporal context to reconstruct scale-aware depth. Following FSM, several methods [8], [6] leverage cross-view

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

and temporal information to enhance depth estimation. For instance, these works employ diverse strategies such as cross-view self-attention [6] and volumetric feature fusion [8].

C. Knowledge Distillation

Knowledge Distillation (KD) [18] has emerged as an effective technique for model compression and knowledge transfer. The core idea is to guide the student to mimic the teacher's knowledge through various mechanisms, such as aligning predictions [34], transferring intermediate features [19], [20], or capturing structural relationships between features [17], [21]. In computer vision, KD has been widely applied to tasks such as image classification [35], [36], object detection [34]. For depth estimation, [37] enhanced unsupervised monocular methods using KD with an error correction mechanism, while [38] introduced a feature-based KD framework for mobile devices.

III. Method

A. Problem Definition

The task of full surround monocular depth estimation aims to predict the metric depth maps from given a set of input surrounding samples $\mathbf{I} = \{I_i\}_{i=1}^{N}$ and the corresponding ground truth depth maps $\mathbf{D} = \{D_i\}_{i=1}^{N}$, where the N is the number of surround cameras. The objective of our method is to improve the depth estimation performance across the surround camera views by distilling the knowledge of a teacher foundation model to a student model.

As aforementioned, we assume that the teacher model $\mathcal{F}_\theta^t$ and the student model $\mathcal{F}_\theta^s$ adopt the same depth binning module structure for hybrid regression [14], [15]. Rather than directly predicting per-pixel depth values, the depth binning method discretizes depth into B bins. Simultaneously, the model $\mathcal{F}$ jointly predicts the pixel-wise depth bin centers C, and the corresponding probabilities P as follows:

$$\begin{aligned} C, P &= \mathcal{F}_\theta(I), \\ \text{where } C = \{c_k\}_{k=1}^{B},\ P &= \{p_k\}_{k=1}^{B}, \end{aligned} \tag{1}$$

where H and W denote height and width of an image I, respectively, and $c_k, p_k \in \mathbb{R}^{H\times W}$. By combining these depth bin centers and probabilities, the final depth estimation $\hat{D}$ for each view is then computed as follows:

$$\begin{aligned} \hat{D}[u,v] &= \sum_{k=1}^{B} c_k[u,v]\cdot p_k[u,v], \\ \forall u \in \{1,\cdots,H\},\ v &\in \{1,\cdots,W\}, \end{aligned} \tag{2}$$

where u and v denote the indices corresponding to the width W and height H of the image, respectively. Note that each depth bin center $c_k \in C$ is adaptively determined for each sample through the network's prediction. In conventional supervised methods, the student model predicts the depth bin centers C^s and the corresponding probabilities P^s, which are then used to compute the student's depth output $\hat{D}^s$ via the above formulation. The error between $\hat{D}^s$ and the ground-truth depth D is minimized using a specific loss term L_{depth} (*e.g.,* L1, SiLog [22]) as follows:

$$\mathcal{L}_{\text{sup}} = L_{\text{depth}}(\hat{D}^s, D). \tag{3}$$

B. Cross-interaction Knowledge Distillation

The network should infer the absolute scale (*i.e.,* metric scale) depth bin centers C based on the depth distribution of each sample. On the other hand, the probability P, which is used to estimate depth through the weighted sum of depth bin centers, represents the probability distribution of relative depth without considering the actual scale. The main idea of our method is to improve overall depth accuracy by distilling the probability P^t of the foundation model–which encapsulates robust and scale-invariant information–into the student's probability P^s, while simultaneously training the student's depth bin centers C^s to align with the actual metric scale of ground truth depth.

To achieve this, we propose a ***cross-interaction knowledge distillation*** scheme that systematically transfers the robust representation of the teacher model's relative depth to the student network by leveraging bin probabilities and bin centers from the teacher and student models, respectively. The overall training mechanism is depicted in Fig. 3. We first obtain the probabilities and depth bin centers of both teacher network $\mathcal{F}_\theta^t$ and student network $\mathcal{F}_\theta^s$ as follows:

$$C^t, P^t = \mathcal{F}_\theta^t(I), \quad C^s, P^s = \mathcal{F}_\theta^s(I). \tag{4}$$

To transfer the probability distribution from the teacher network to the student network, we employ a loss function $\mathcal{L}_{\text{ckd}}$ that reconstructs the teacher's depth prediction using the teacher's bin centers and the student's probability distribution, formulated as follows:

$$\begin{aligned} \mathcal{L}_{\text{ckd}} &= L_{\text{depth}}(\hat{D}^{\text{cross}}, \hat{D}^t), \\ \text{where}\quad \hat{D}^{\text{cross}}[u,v] &= \sum_{k=1}^{B} c_k^t[u,v]\cdot p_k^s[u,v] \end{aligned} \tag{5}$$

where $c_k^t \in C^t$, $p_k^s \in P^s$, and $\hat{D}^t$ indicate the teacher's depth prediction, respectively. The objective of this loss is to encourage the student network to learn the teacher's probability, while student depth bin centers are guided by Eq. 3.

C. View-relational Knowledge Distillation

Existing full surround self-supervised monocular depth estimation approaches [5], [6], [8], [7] commonly leverage the geometric constraint from spatio-temporal information from each camera coordinate. On the other hand, in a supervised scenario where the absolute scale depth (*i.e.,* metric depth for each viewpoint) is given as ground truth, we propose a learning-based method, ***view-relational knowledge distillation***, which allows the structural knowledge between camera views to be distilled from the teacher model to the student model. This method aims to transfer the teacher's structural knowledge using mutual relationships between depth distributions from each adjacent camera view. Similar to [17], given teacher's depth bin probabilities across all cameras $\{\hat{P}_i^t\}_{i=1}^{N}$ and student depth bin probabilities $\{\hat{P}_i^s\}_{i=1}^{N}$ through Eq. 4, we first encode the internal pair-wise relations

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

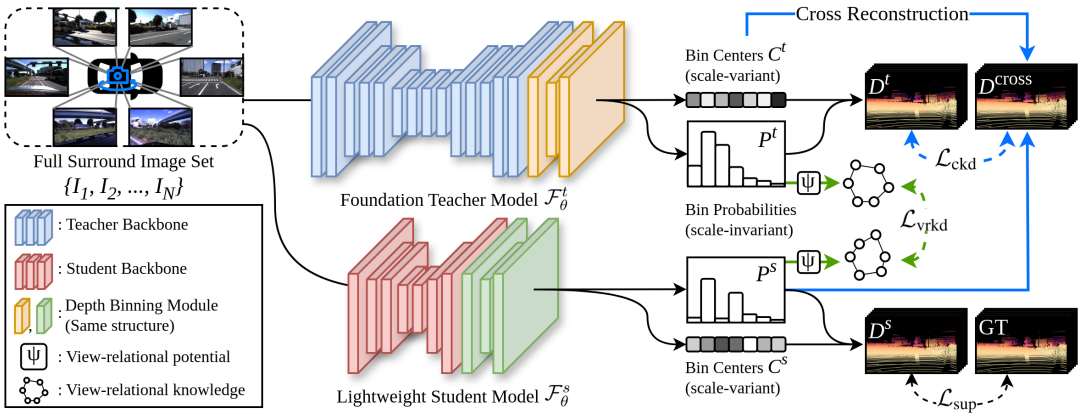

Fig. 3. Illustration of the proposed knowledge distillation schemes. Our method leverages a depth binning module with the same architecture as the teacher model, enabling effective knowledge distillation at the scale-invariant depth bin probability level.

of the teacher's and student's depth bin probabilities using the relational potential function ψ as follows:

$$\Gamma^t_{(i,j)} = \psi(\hat{P}^t_i, \hat{P}^t_j),\ \Gamma^s_{(i,j)} = \psi(\hat{P}^s_i, \hat{P}^s_j),$$
$$\forall \psi(P_i, P_j) = \frac{1}{\mu}\|P_i - P_j\|_2, \tag{6}$$

where i and j are the camera indices, and μ is the normalization factor for distance, respectively.

To pair each camera i with its adjacent camera j, which shares overlapping fields of view, we define the adjacent camera pair set (i,j) within the set $\mathcal{A} = \{(1,2),(2,3),\cdots,(N,1)\}$, defined cyclically. The student network is trained using the view-relational knowledge distillation loss $\mathcal{L}_{\text{vrkd}}$, which encourages the student to mimic the teacher's camera-wise relations. Specifically, the relation between adjacent camera views $\Gamma^s_{i,j}$ in the student network is encouraged to approximate $\Gamma^t_{i,j}$ from the teacher using the Huber loss [39] L_{huber}, as follows:

$$\mathcal{L}_{\text{vrkd}} = \sum_{(i,j)\in\mathcal{A}} L_{\text{huber}}(\Gamma^t_{i,j}, \Gamma^s_{i,j}). \tag{7}$$

This loss distills the relational information of the adjacent cameras' probabilities–including the scale-invariant depth knowledge–to the student, thereby enabling the student to learn the relationships between adjacent camera views. In the end, the final loss term of our method consists of a weighted sum of the supervised loss term $\mathcal{L}_{\text{sup}}$, the cross-interaction knowledge distillation loss $\mathcal{L}_{\text{ckd}}$ introduced in Sec. III-B, and $\mathcal{L}_{\text{vrkd}}$ as follows:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{sup}} + \lambda_{\text{ckd}} \cdot \mathcal{L}_{\text{ckd}} + \lambda_{\text{vrkd}} \cdot \mathcal{L}_{\text{vrkd}}, \tag{8}$$

where λ_{ckd} and λ_{vrkd} are the weight balancing parameters for each loss terms.

IV. Experiments

a) Datasets: We evaluate our method on the *DDAD* [40] and *nuScenes* [41] datasets. DDAD captures diverse urban environments in the US and Japan using six cameras and a Luminar-H2 sensor, covering a full 360-degree field of view. It contains 73,914 training images (12,319 per camera) and 23,700 validation images (3,950 per camera). With a maximum range of 250m and only 20% overlap between adjacent cameras, DDAD simulates realistic full surround autonomous driving conditions. The nuScenes dataset, widely used in autonomous driving research, is collected from urban scenes across the US and Singapore, utilizing six surrounding cameras and a LiDAR sensor. It includes 120,576 training images (20,096 per camera) and 36,114 validation images (6,019 per camera). The camera setup has at most 10% overlap between adjacent views, posing challenges for cross-view consistency and multi-view depth estimation.

b) Baselines: We adopt three representative Monocular Depth Estimation (MDE) frameworks as student networks, including Monodepth2 [2], HRDepth [3], and MonoViT [4], which are widely used in MDE research. HRDepth and MonoViT, with their enhanced encoder-decoder architectures built on Monodepth2, are particularly suitable for evaluating the effectiveness of our proposed knowledge distillation as the student network size increases. Notably, FSM [5], SurroundDepth [6], and CVCDepth [42] share the same Monodepth2 backbone and are trained in a self-supervised manner. Due to their architectural relevance and prevalence in recent literature, we select Monodepth2, HRDepth, and MonoViT as our backbone networks. For the teacher network, we use DepthAnything [1], a highly generalizable model capable of inferring depth maps across diverse environments. As described in Sec. III, we incorporate a MetricBins (MB) module—identical to the teacher's binning module—into each student network to leverage the bin probability-level representation from the teacher. All [baseline + MB] configurations use 64 bins, with a bin embedding dimension of 128.

c) Implementation details: We follow the training and evaluation protocols of state-of-the-art FSMDE methods [5], [6], [8] to ensure fair comparisons between our approach and conventional supervised methods. For the DDAD dataset, we set the maximum depth to 200m to reflect the long-range capability and apply the self-occlusion masks provided by SurroundDepth [6] to exclude regions occluded by the vehicle. For nuScenes, no self-occlusion mask is needed, and

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

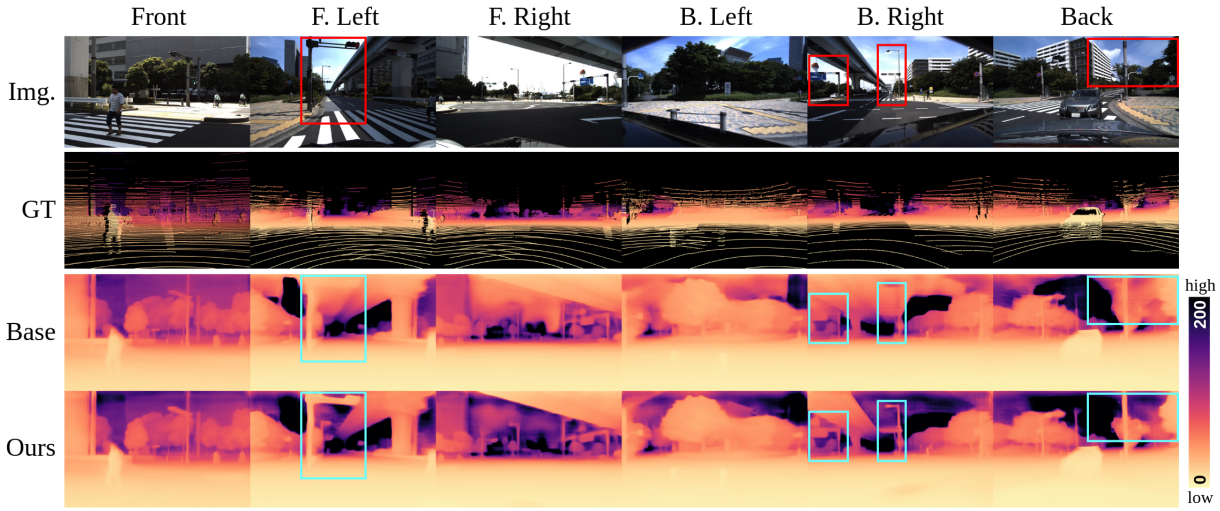

Fig. 4. Qualitative results of fine-tuned Monodepth2 (denoted as Base) and Monodepth2 + Ours (denoted as Ours) on **DDAD** dataset.

TABLE I

EVALUATION RESULTS ON **DDAD** DATASET. FOR THE BINNING MODULE FOR STUDENTS, WE ADOPT THE METRICBINS (MB), WHICH IS THE SAME ARCHITECTURE AS THE TEACHER'S (*i.e., DepthAnything*) BINNING MODULE. $\Delta_{\mathcal{T}}$ DENOTES THE RELATIVE PERFORMANCE INCREMENT FROM EACH BASELINE. UNLIKE SUPERVISED FSMDE, WE APPLY MEDIAN SCALING FOR SELF-SUPERVISED METHODS.

	Methods	Abs Rel ↓	Sq Rel ↓	RMSE ↓	RMSE log ↓	$\delta < 1.25$ ↑	$\delta < 1.25^2$ ↑	$\delta < 1.25^3$ ↑	$\Delta_{\mathcal{T}}(\%)$ ↑	Params. ↓	Latency ↓
	DepthAnything (zero-shot)	*0.270*	*3.291*	*11.866*	*0.621*	*0.601*	*0.826*	*0.904*	-	*335.79* M	*142.26* ms
	DepthAnything (finetuned)	*0.140*	*1.866*	*9.475*	*0.228*	*0.831*	*0.935*	*0.969*	-	*335.79* M	*142.26* ms
Self-supervised	Monodepth2 [2]	0.217	3.641	12.962	0.323	0.699	0.877	0.939	-	14.33 M	3.75 ms
	PackNet-SfM [40]	0.234	3.802	13.253	0.331	0.672	0.860	0.931	-	128.29 M	63.08 ms
	FSM [5]	0.202	-	-	-	-	-	-	-	14.33 M	3.75 ms
	SurroundDepth [6]	0.200	3.392	12.270	0.301	0.740	0.894	0.947	-	59.56 M	11.79 ms
	EGA-Depth-LR [7]	0.195	3.211	12.117	0.297	0.743	0.896	0.947	-	not public	not public
	EGA-Depth-MR [7]	**0.191**	**3.126**	**11.922**	**0.290**	**0.747**	**0.901**	**0.950**	-	not public	not public
Supervised	Monodepth2	0.200	3.087	12.849	0.323	0.679	0.861	0.932	0.00	14.33 M	3.75 ms
	Monodepth2 + MB	0.195	3.016	12.727	0.322	0.686	0.865	0.932	+1.08	15.03 M	12.77 ms
	Monodepth2 + MB + Ours	**0.191**	**2.865**	**12.134**	**0.300**	**0.710**	**0.881**	**0.943**	**+4.64**	15.03 M	12.77 ms
	HRDepth	0.196	2.955	12.428	0.304	0.699	0.875	0.940	0.00	14.10 M	6.19 ms
	HRDepth + MB	0.194	2.961	12.399	0.303	0.700	0.877	0.942	+0.28	14.80 M	15.18 ms
	HRDepth + MB + Ours	**0.183**	**2.684**	**11.578**	**0.283**	**0.736**	**0.896**	**0.950**	**+5.47**	14.80 M	15.18 ms
	MonoViT	0.179	2.602	11.777	0.287	0.725	0.892	0.949	0.00	27.87 M	18.83 ms
	MonoViT + MB	0.178	2.585	11.757	0.286	0.728	0.893	0.950	+0.34	28.31 M	27.71 ms
	MonoViT + MB + Ours	**0.166**	**2.260**	**10.483**	**0.256**	**0.773**	**0.916**	**0.961**	**+7.54**	28.31 M	27.71 ms

the maximum depth is set to 80m. Training image resolutions are set to 640×384 for DDAD and 640×352 for nuScenes, and median scaling is not applied during evaluation. Each model is trained for 5 epochs with a batch size of 12. We follow the hyperparameter setup of supervised training and the MetricBins module of DepthAnything [1]. Weight balancing parameters for our method are set to $\lambda_{\text{ckd}} = 0.1$ and $\lambda_{\text{vrkd}} = 1.0$. The teacher model (DepthAnything) uses pre-trained weights from the outdoor datasets, while all student models are initialized with pre-trained weights from KITTI [43] to accelerate training and improve convergence. We present our results following the standard evaluation metrics introduced by [22]. For evaluating inference time, we conduct measurements using a single NVIDIA RTX A6000.

A. Evaluation Results on FSMDE Datasets

Tab. I summarizes the quantitative evaluation results on the DDAD dataset. We evaluate each baseline (*i.e.,* Monodepth2, HRDepth, and MonoViT) by applying the MetricBins module and our proposed method. Given that the majority of FSMDE research employs self-supervised methods, we additionally reported the performance of self-supervised FSMDE approaches in our table for a general comparison. The results indicate that even with post-processing (i.e., median scaling), self-supervised methodologies generally exhibit lower performance compared to those trained in a supervised manner. When the MetricBins method is added to each baseline, the average improvement ranges from as low as 0.28% to as high as 1.08% compared to the pure baseline. Compared to the improvements achieved by MetricBins, applying the proposed methods (*i.e.*, $\mathcal{L}_{\text{ckd}}$ and $\mathcal{L}_{\text{vrkd}}$) leads to markedly greater improvements, an average improvement ranging from 4.64% to 7.54%. Furthermore, as illustrated in Fig. 4, transferring probability-level knowledge from the teacher network contributes to improved generalization performance in regions lacking LiDAR points, which are problematic for supervised MDE.

On the other hand, applying the MetricBins to Mon-

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

TABLE II

EVALUATION RESULTS ON **NUSCENES** DATASET. $\Delta_{\mathcal{T}}$ DENOTES THE RELATIVE PERFORMANCE INCREMENT FROM EACH BASELINE.

	Methods	Abs Rel ↓	Sq Rel ↓	RMSE ↓	RMSE log ↓	$\delta < 1.25$ ↑	$\delta < 1.25^2$ ↑	$\delta < 1.25^3$ ↑	$\Delta_{\mathcal{T}}(\%)$ ↑	Params. ↓	Latency ↓
	DepthAnything (zero-shot)	*0.208*	*1.324*	*5.108*	*0.280*	*0.735*	*0.901*	*0.951*	-	*335.79* M	*130.41* ms
	DepthAnything (finetuned)	*0.107*	*0.831*	*3.866*	*0.185*	*0.890*	*0.950*	*0.974*	-	*335.79* M	*130.41* ms
Self-supervised	Monodepth2 [2]	0.287	3.349	7.184	0.345	0.641	0.845	0.925	-	14.33 M	3.44 ms
Self-supervised	PackNet-SfM [40]	0.309	2.891	7.994	0.390	0.547	0.796	0.899	-	128.29 M	57.86 ms
Self-supervised	FSM [5]	0.299	-	-	-	-	-	-	-	14.33 M	3.44 ms
Self-supervised	SurroundDepth [6]	0.245	3.067	6.835	0.321	0.719	0.878	0.935	-	59.56 M	10.82 ms
Self-supervised	EGA-Depth-LR [7]	0.239	2.357	6.801	0.317	0.723	0.880	0.936	-	not public	not public
Self-supervised	EGA-Depth-MR [7]	**0.228**	**2.113**	**6.738**	**0.311**	**0.728**	**0.885**	**0.940**	-	not public	not public
Supervised	Monodepth2	0.182	1.593	5.961	0.295	0.744	0.868	0.926	0.00	14.33 M	3.44 ms
Supervised	Monodepth2 + MB	0.188	1.719	6.145	0.311	0.733	0.857	0.917	−3.35	15.03 M	11.71 ms
Supervised	Monodepth2 + MB + Ours	**0.158**	**1.135**	**4.915**	**0.243**	**0.805**	**0.914**	**0.957**	**+13.42**	15.03 M	11.71 ms
Supervised	HRDepth	0.172	1.313	5.523	0.267	0.765	0.893	0.946	0.00	14.10 M	5.67 ms
Supervised	HRDepth + MB	0.174	1.360	5.439	0.278	0.762	0.881	0.934	−1.48	14.80 M	13.92 ms
Supervised	HRDepth + MB + Ours	**0.156**	**1.097**	**5.020**	**0.244**	**0.792**	**0.911**	**0.958**	**+7.18**	14.80 M	13.92 ms
Supervised	MonoViT	0.170	1.643	5.174	0.253	0.769	0.884	0.940	0.00	27.87 M	17.26 ms
Supervised	MonoViT + MB	0.169	1.616	5.121	0.252	0.770	0.884	0.940	+0.54	28.31 M	25.40 ms
Supervised	MonoViT + MB + Ours	**0.134**	**1.072**	**4.096**	**0.214**	**0.830**	**0.914**	**0.954**	**+15.00**	28.31 M	25.40 ms

TABLE III

COMPARISION BETWEEN EXISTING KD APPROACHES AND OUR METHOD ON **DDAD** DATASET. ALL EXPERIMENTS ARE CONDUCTED ON MONODEPTH2 ARCHITECTURES. $\Delta_{\mathcal{T}}$ IMPLIES THE RELATIVE PERFORMANCE INCREMENT FROM MONODEPTH2.

Methods	Abs Rel ↓	Sq Rel ↓	RMSE ↓	RMSE log ↓	$\delta < 1.25$ ↑	$\delta < 1.25^2$ ↑	$\delta < 1.25^3$ ↑	$\Delta_{\mathcal{T}}(\%)$ ↑
Monodepth2	0.200	3.087	12.849	0.323	0.679	0.861	0.932	0.00
KD [18]	0.222	3.444	13.017	0.332	0.662	0.852	0.928	−4.38
FitNets [19]	0.200	2.993	12.623	0.308	0.688	0.872	0.940	+1.84
AT [20]	0.195	2.997	12.505	0.311	0.694	0.875	0.935	+2.28
SP [21]	0.194	2.931	12.331	0.308	0.692	0.876	0.941	+3.05
Ours	**0.191**	**2.865**	**12.134**	**0.300**	**0.710**	**0.881**	**0.943**	**+4.64**

odepth2 and HRDepth results in slight performance degradation on the nuScenes dataset, as shown in Tab. II. Although the performance drop is minor, this indicates that simply incorporating the binning strategy alone does not lead to significant improvements. In contrast, the substantial performance gain achieved by our method, which distills knowledge at the bin probability level, further validates its effectiveness. The performance improvement ranges from 7.18% to 15%, with the largest gain observed in MonoViT, while our method also satisfies real-time latency requirements (*i.e.,* under 33ms), highlighting its practical applicability. These results demonstrate our method's effectiveness and robustness across different datasets and model architectures.

B. Comparison between Our Method and Existing Knowledge Distillation Methods

We compare the existing knowledge distillation approaches with our proposed method to demonstrate our probability-level knowledge distillation method. Due to that, recent distillation methods often leverage class priors or inter-data relationships [44], [45]—particularly function for only classification tasks—we incorporate general distillation methods: conventional output-level KD [18] and three feature-level distillation baselines, including FitNets [19], AT [20], and SP [21]. In the current experimental setup, where accurate metric scale depth estimation is required, a scale-ambiguous teacher model can hinder a student network from learning metric depth properly. To prevent this issue and for a proper comparison, we do not include output-level distillation for all distillation approaches except for KD, which inherently operates at the output level. Furthermore, considering the low transferability in cross-architecture settings (*i.e.*, a transformer teacher model and a CNN student model [46]), we apply feature-level distillation only between the CNN decoders of the teacher and student models.

Tab. III summarizes the evaluation results of the proposed method along with other knowledge distillation approaches. Due to scale ambiguity in the output-level of the teacher foundation model (refer to Sec. I), KD severely degrades performance compared to a fine-tuned student network. Our proposed method, which performs distillation at the scale-invariant bin probability level, achieves the best performance while other feature-level distillation methods show slight improvements. These results further validate the effectiveness of our approach in mitigating scale ambiguity while maximizing the benefits of knowledge distillation.

C. Evaluation Results of Various Binning Methods

To assess the effectiveness of the proposed method in relation to different binning strategies, we integrate various binning methods, including AdaBins [14], LocalBins [15], and MetricBins [31], into our strategy for comparison. Specifically, AdaBins leverages global adaptive depth bin centers $C_{\text{AdaBins}} \in \mathbb{R}^B$ that are shared across all pixels, as opposed to per-pixel adaptive depth bin centers $C_{\text{LocalBins}}, C_{\text{MetricBins}} \in \mathbb{R}^{B\times H\times W}$, where B denotes the number of depth bins. As detailed in Eq. 2, each binning module reconstructs

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

TABLE IV

EVALUATION RESULTS OF VARIOUS BINNING METHODS AND OUR METHOD ON **DDAD** DATASET. ALL EXPERIMENTS ARE CONDUCTED ON MONODEPTH2 ARCHITECTURES. (AB: ADABINS, LB: LOCALBINS, MB: METRICBINS)

Methods	Abs Rel ↓	Sq Rel ↓	RMSE ↓	RMSE log ↓	$\delta < 1.25$ ↑	$\delta < 1.25^2$ ↑	$\delta < 1.25^3$ ↑	$\Delta_{\mathcal{T}}(\%)$ ↑	Params. ↓	Latency ↓
Monodepth2	0.200	3.087	12.849	0.323	0.679	0.861	0.932	0.00	14.33 M	3.75 ms
Monodepth2 + AB	**0.502**	**5.004**	**13.421**	**0.471**	**0.377**	**0.659**	**0.828**	**−48.92**	16.36 M	5.46 ms
Monodepth2 + AB + Ours	1.194	14.074	16.496	0.789	0.167	0.342	0.531	−172.04	16.36 M	5.46 ms
Monodepth2 + LB	0.194	3.011	12.880	0.317	0.699	0.871	0.933	+1.61	18.20 M	10.58 ms
Monodepth2 + LB + Ours	**0.191**	**2.819**	**12.446**	**0.295**	**0.716**	**0.884**	**0.945**	**+4.93**	18.20 M	10.58 ms
Monodepth2 + MB	0.195	3.016	12.727	0.322	0.686	0.865	0.932	+1.08	15.03 M	12.77 ms
Monodepth2 + MB + Ours	**0.191**	**2.865**	**12.134**	**0.300**	**0.710**	**0.881**	**0.943**	**+4.64**	15.03 M	12.77 ms

the depth map by utilizing the depth bin probability $P \in \mathbb{R}^{B\times H\times W}$ along with the corresponding global or local depth bin centers. Due to these methodological differences, AdaBins produces a depth bin probability that reflects the overall depth structure of the scene, whereas LocalBins and MetricBins yield locally adaptive probability distributions.

Tab. IV presents the results of applying each binning method and our method to Monodepth2. The use of AdaBins leads to significant performance degradation because the global binning strategy of AdaBins does not align with the teacher model's locally adaptive nature, leading to improper knowledge distillation of the probability distribution. Meanwhile, LocalBins and MetricBins, which adopt the same pixel-level mechanism as the teacher model, enable effective knowledge distillation, achieving relative improvements of 4.93% and 4.64%, respectively.

D. Ablation Studies of Our Method

We conduct ablation studies on two key components of our method: $\mathcal{L}_{\text{ckd}}$ and $\mathcal{L}_{\text{vrkd}}$, which play a crucial role in distilling knowledge at the bin probability level. To investigate the impact of these losses, we evaluate the performance of Monodepth2, HRDepth, and MonoViT on the DDAD dataset by comparing results with and without each loss term. As shown in Tab. V, employing MB alone yields an average improvement of 0.57% over all models. Integrating $\mathcal{L}_{\text{ckd}}$ with MB further improves the performance, achieving a total gain of 3.92%, 5.12%, and 6.37% for Monodepth2, HRDepth, and MonoViT, respectively, which highlights the effectiveness of transferring the teacher foundation model's bin probability distribution to the student model.

Furthermore, incorporating $\mathcal{L}_{\text{vrkd}}$ results in an additional enhancement, leading to an overall 4.64%, 5.47%, and 7.54% performance gain for each model, and these consistent performance improvements demonstrate the stability and robustness of our proposed method. This result suggests that View-relational Knowledge Distillation (Sec. III-C) is complementary to Cross-interaction Knowledge Distillation (Sec. III-B), further enhancing depth estimation performance by leveraging view relational constraints.

TABLE V

ABLATION STUDIES OF OUR METHODS ON **DDAD** DATASET.

MB	$\mathcal{L}_{\text{ckd}}$	$\mathcal{L}_{\text{vrkd}}$	Monodepth2	HRDepth	MonoViT
			$\Delta_{\mathcal{T}}(\%)$ ↑	$\Delta_{\mathcal{T}}(\%)$ ↑	$\Delta_{\mathcal{T}}(\%)$ ↑
✗	✗	✗	0.00	0.00	0.00
✓	✗	✗	+1.08	+0.28	+0.34
✓	✓	✗	+3.92	+5.12	+6.37
✓	✗	✓	+0.83	+0.83	+1.11
✓	✓	✓	**+4.64**	**+5.47**	**+7.54**

V. CONCLUSION

In this paper, we propose a new knowledge distillation strategy for the FSMDE framework that transfers a robust depth representation from a foundation model to a lightweight FSMDE student model suitable for autonomous driving scenarios. Our method is based on the predominant binning strategy used in supervised MDE methodologies, and can be applied seamlessly. The proposed cross-interaction knowledge distillation can effectively transfer the well-structured relative depth knowledge of the teacher foundation model to the student model, which enables the student model to mimic the teacher's robust representation. View-relational Knowledge Distillation enables the student model to learn the relational knowledge of adjacent camera views by injecting the teacher's structural probabilities into the student model. Experiments on DDAD and nuScenes demonstrate that our method achieves significant performance improvements while maintaining real-time processing at 30 FPS. The results also demonstrate superior performance compared to existing distillation methods, highlighting the effectiveness of our approach in combining a binning strategy with FSMDE.

REFERENCES

[1] L. Yang, B. Kang, Z. Huang, X. Xu, J. Feng, and H. Zhao, "Depth anything: Unleashing the power of large-scale unlabeled data," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 10 371–10 381.

[2] C. Godard, O. Mac Aodha, M. Firman, and G. J. Brostow, "Digging into self-supervised monocular depth estimation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 3828–3838.

[3] X. Lyu, L. Liu, M. Wang, X. Kong, L. Liu, Y. Liu, X. Chen, and Y. Yuan, "Hr-depth: High resolution self-supervised monocular depth estimation," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, no. 3, 2021, pp. 2294–2301.

[4] C. Zhao, Y. Zhang, M. Poggi, F. Tosi, X. Guo, Z. Zhu, G. Huang, Y. Tang, and S. Mattoccia, "Monovit: Self-supervised monocular depth estimation with a vision transformer," in *2022 international conference on 3D vision (3DV)*. IEEE, 2022, pp. 668–678.

[5] V. Guizilini, I. Vasiljevic, R. Ambrus, G. Shakhnarovich, and A. Gaidon, "Full surround monodepth from multiple cameras," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 5397–5404, 2022.

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

[6] Y. Wei, L. Zhao, W. Zheng, Z. Zhu, Y. Rao, G. Huang, J. Lu, and J. Zhou, "Surrounddepth: Entangling surrounding views for self-supervised multi-camera depth estimation," in *Conference on robot learning*. PMLR, 2023, pp. 539–549.

[7] Y. Shi, H. Cai, A. Ansari, and F. Porikli, "Ega-depth: Efficient guided attention for self-supervised multi-camera depth estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 119–129.

[8] J.-H. Kim, J. Hur, T. P. Nguyen, and S.-G. Jeong, "Self-supervised surround-view depth estimation with volumetric feature fusion," *Advances in Neural Information Processing Systems*, vol. 35, pp. 4032–4045, 2022.

[9] A. Schmied, T. Fischer, M. Danelljan, M. Pollefeys, and F. Yu, "R3d3: Dense 3d reconstruction of dynamic scenes from multiple cameras," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 3216–3226.

[10] X. Guo, W. Yuan, Y. Zhang, T. Yang, C. Zhang, Z. Zhu, and L. Chen, "A simple baseline for supervised surround-view depth estimation," *arXiv preprint arXiv:2303.07759*, 2023.

[11] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, *et al.*, "Segment anything," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4015–4026.

[12] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695.

[13] L. Yang, B. Kang, Z. Huang, Z. Zhao, X. Xu, J. Feng, and H. Zhao, "Depth anything v2," *Advances in Neural Information Processing Systems*, vol. 37, pp. 21 875–21 911, 2025.

[14] S. F. Bhat, I. Alhashim, and P. Wonka, "Adabins: Depth estimation using adaptive bins," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 4009–4018.

[15] ——, "Localbins: Improving depth estimation by learning local distributions," in *European Conference on Computer Vision*. Springer, 2022, pp. 480–496.

[16] Z. Li, X. Wang, X. Liu, and J. Jiang, "Binsformer: Revisiting adaptive bins for monocular depth estimation," *IEEE Transactions on Image Processing*, 2024.

[17] W. Park, D. Kim, Y. Lu, and M. Cho, "Relational knowledge distillation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 3967–3976.

[18] G. Hinton, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[19] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, and Y. Bengio, "Fitnets: Hints for thin deep nets," *arXiv preprint arXiv:1412.6550*, 2014.

[20] S. Zagoruyko and N. Komodakis, "Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer," *arXiv preprint arXiv:1612.03928*, 2016.

[21] F. Tung and G. Mori, "Similarity-preserving knowledge distillation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 1365–1374.

[22] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," *Advances in neural information processing systems*, vol. 27, 2014.

[23] D. Xu, W. Wang, H. Tang, H. Liu, N. Sebe, and E. Ricci, "Structured attention guided convolutional neural fields for monocular depth estimation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 3917–3925.

[24] J. H. Lee, M.-K. Han, D. W. Ko, and I. H. Suh, "From big to small: Multi-scale local planar guidance for monocular depth estimation," *arXiv preprint arXiv:1907.10326*, 2019.

[25] W. Choi, M. Shin, and S. Im, "Depth-discriminative metric learning for monocular 3d object detection," *Advances in Neural Information Processing Systems*, vol. 36, pp. 80 165–80 177, 2023.

[26] R. Ranftl, A. Bochkovskiy, and V. Koltun, "Vision transformers for dense prediction," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 12 179–12 188.

[27] D. Kim, W. Ka, P. Ahn, D. Joo, S. Chun, and J. Kim, "Global-local path networks for monocular depth estimation with vertical cutdepth," *arXiv preprint arXiv:2201.07436*, 2022.

[28] S. Shao, Z. Pei, W. Chen, X. Wu, and Z. Li, "Nddepth: Normal-distance assisted monocular depth estimation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 7931–7940.

[29] Y. Zhao, S. Kong, and C. Fowlkes, "Camera pose matters: Improving depth prediction by mitigating pose distribution bias," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 15 759–15 768.

[30] H. Fu, M. Gong, C. Wang, K. Batmanghelich, and D. Tao, "Deep ordinal regression network for monocular depth estimation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 2002–2011.

[31] S. F. Bhat, R. Birkl, D. Wofk, P. Wonka, and M. Müller, "Zoedepth: Zero-shot transfer by combining relative and metric depth," *arXiv preprint arXiv:2302.12288*, 2023.

[32] W. Choi, K. Hwang, W. Peng, M. Choi, and S. Im, "Self-supervised monocular depth estimation robust to reflective surface leveraged by triplet mining," *arXiv preprint arXiv:2502.14573*, 2025.

[33] J. Bae, K. Hwang, and S. Im, "A study on the generality of neural network structures for monocular depth estimation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 4, pp. 2224–2238, 2023.

[34] G. Chen, W. Choi, X. Yu, T. Han, and M. Chandraker, "Learning efficient object detection models with knowledge distillation," *Advances in neural information processing systems*, vol. 30, 2017.

[35] Z. Li and D. Hoiem, "Learning without forgetting," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 12, pp. 2935–2947, 2017.

[36] Z. Peng, Z. Li, J. Zhang, Y. Li, G.-J. Qi, and J. Tang, "Few-shot image recognition with knowledge transfer," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 441–449.

[37] A. Pilzer, S. Lathuiliere, N. Sebe, and E. Ricci, "Refine and distill: Exploiting cycle-inconsistency and knowledge distillation for unsupervised monocular depth estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9768–9777.

[38] Y. Wang, X. Li, M. Shi, K. Xian, and Z. Cao, "Knowledge distillation for fast and accurate monocular depth estimation on mobile devices," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 2457–2465.

[39] P. J. Huber, "Robust estimation of a location parameter," in *Breakthroughs in statistics: Methodology and distribution*. Springer, 1992, pp. 492–518.

[40] V. Guizilini, R. Ambrus, S. Pillai, A. Raventos, and A. Gaidon, "3d packing for self-supervised monocular depth estimation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 2485–2494.

[41] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuscenes: A multimodal dataset for autonomous driving," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 621–11 631.

[42] L. Ding, H. Jiang, J. Li, Y. Chen, and R. Huang, "Towards cross-view-consistent self-supervised surround depth estimation," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2024, pp. 10 043–10 050.

[43] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *The International Journal of Robotics Research*, vol. 32, no. 11, pp. 1231–1237, 2013.

[44] Y. Tian, D. Krishnan, and P. Isola, "Contrastive representation distillation," *arXiv preprint arXiv:1910.10699*, 2019.

[45] D. Chen, J.-P. Mei, H. Zhang, C. Wang, Y. Feng, and C. Chen, "Knowledge distillation with the reused teacher classifier," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 11 933–11 942.

[46] Y. Liu, J. Cao, B. Li, W. Hu, J. Ding, and L. Li, "Cross-architecture knowledge distillation," in *Proceedings of the Asian conference on computer vision*, 2022, pp. 3396–3411.