FoundDP: Revisiting Weak Disparity Observability in Dual-Pixel Depth Estimation

Fengchen He*, Hao Xu*, Dayang Zhao, Tingwei Quan, and Shaoqun Zeng

Huazhong University of Science and Technology, Wuhan, China
*Equal contribution
linyark, xuhao, dayangzhao, quantingwei, sqzeng@hust.edu.cn

Abstract. Dual-pixel (DP) imaging enables metric depth estimation from a single camera using sub-aperture disparity. However, the extremely small effective baseline limits disparity observability, leading to structural degradation and depth failure in textureless, low-contrast, or downsampled regions. Existing DP-based methods rely primarily on local disparity cues and therefore become unreliable when disparity signals are weak or ambiguous. To address this limitation, we propose *FoundDP*, a unified framework that integrates metric DP depth with global structural priors from a monocular depth foundation model. Our method preserves metric scale through DP-derived depth and leverages Vision Transformer (ViT) features to restore structural consistency in weak-disparity regions. To ensure reliable metric guidance under DP imaging conditions, we identify and mitigate ViT representation degradation induced by DP defocus blur via ViT feature alignment, enabling stable metric-guided depth estimation. Extensive experiments on synthetic and real-world DP benchmarks show that FoundDP delivers superior performance, with consistent gains in structural fidelity and metric accuracy, especially under reduced disparity observability. Code will be available at: https://github.com/EchoLighting/FoundDP

Keywords: Dual-Pixel · Metric Depth · Depth Guidance

1 Introduction

Dual-pixel (DP) imaging [16,29] equips a single camera with implicit stereo capability by splitting each pixel into left and right sub-pixels that capture light from slightly different viewpoints. Originally introduced for phase-detection autofocus [4, 31], this hardware design enables disparity acquisition without requiring additional sensors or explicit stereo baselines [16]. Due to its minimal manufacturing overhead and widespread deployment in consumer devices, DP technology has become a practical hardware primitive for computational imaging, supporting applications such as image synthesis [12,17,22], image deblurring [1,19,40,43], reflection removal [26, 47], and rain-drop removal [18]. More recently, learning-based approaches have leveraged the implicit sub-aperture disparity in DP images for metric depth estimation [9,12,22,38] and disparity recovery [15,21,46], enabling physically grounded depth inference from a single exposure.

Although DP imaging can be modeled as a stereo system with a narrow effective baseline, its extremely small baseline fundamentally limits disparity observability. Existing DP-based methods predominantly regress depth from local sub-pixel correspondence [22, 38], which performs reliably in texture-rich and edge-distinct regions where disparity cues are observable. However, in textureless, low-light, planar, or distant scenes, correspondence signals weaken toward the noise floor [12, 32], leading to unreliable disparity estimation and unstable depth predictions, as illustrated in Fig. 1(a). Furthermore, spatial downsampling reduces sub-pixel disparity precision, further weakening disparity observability and making depth estimation more unstable. Since metric DP depth estimation depends on observable disparity [9], reduced observability inevitably causes structural degradation and depth failure, constituting a fundamental bottleneck for existing DP-based methods.

Fig. 1: (a) Qualitative comparison. Our method yields sharper boundaries and more accurate depth in weak-disparity regions (highlighted in red). (b) Complementary failure modes: DPNet preserves metric scale but degrades structurally, while DAV2 recovers structure but lacks metric scale. Best viewed by zooming in.

In contrast, recent monocular depth foundation models based on Vision Transformers (ViT) [24, 27, 34, 41] demonstrate remarkable structural reasoning and cross-scene generalization through large-scale pretraining. Models such as Depth Anything V2 (DAV2) [42] and MoGe [34] recover globally coherent scene geometry even in textureless or low-contrast regions by modeling long-range contextual dependencies. However, monocular predictions are statistically inferred rather than physically constrained, and remain ambiguous up to an

affine transformation due to the absence of metric observability [27]. As shown in Fig. 1(b), although foundation models maintain structural consistency, they lack reliable metric scale.

These observations expose a fundamental tension between physical observability and structural reasoning. DP imaging provides metric grounding but relies heavily on locally observable disparity, whereas foundation models offer strong global structural priors without physical scale constraints. This raises a key question: *can metric depth estimation benefit from foundation-scale structural priors without sacrificing physical consistency?*

Motivated by this, we propose *FoundDP*, a framework that leverages disparity observability to reconcile physically grounded DP cues with foundation-based structural priors for metric depth estimation. Our approach anchors metric scale through DP-derived depth while leveraging ViT-extracted global representations to restore structural continuity in weak-disparity regions. Specifically, we first obtain an initial metric depth from a DP network, adopt a ViT encoder to extract globally consistent structural features, and then apply depth guidance to obtain the final predictions. This design enhances structural fidelity under weak-disparity conditions while preserving metric consistency.

However, directly applying foundation models to DP images is non-trivial. Due to optical defocus, DP images exhibit significant blur in out-of-focus regions [2]. Compared with the sharp natural images used for foundation model pretraining, defocus introduces frequency attenuation and distribution shift in DP images, which systematically affects ViT representations and leads to inconsistent global attention responses during depth guidance. We identify this as a previously underexplored source of guidance instability and introduce a ViT feature alignment strategy. By enforcing feature consistency between paired clear images and their defocus-degraded counterparts, we align ViT features toward the feature distribution of ideal sharp inputs, thereby improving the stability of depth guidance.

We conduct extensive evaluations on multiple public synthetic [30] and real-world DP datasets [9, 17, 25], and additionally capture a dataset to analyze robustness under disparity attenuation. Results show that FoundDP consistently outperforms prior DP methods [9, 10, 15, 22] in both normal and weak-disparity regions, validating the effectiveness of metric-guided depth estimation with foundation-based structural priors.

The main contributions of this work are summarized as follows:

- We propose *FoundDP*, a disparity-observability-aware framework that integrates metric DP depth cues with monocular foundation models to improve depth estimation under weak-disparity conditions.
- We identify ViT feature degradation under DP defocus blur as a key barrier to depth guidance and employ a feature alignment strategy to restore consistency for reliable metric-guided depth estimation.
- Extensive experiments on multiple public DP benchmarks demonstrate that FoundDP achieves superior performance, with consistent and often large gains in weak-disparity and reduced-observability conditions.

2 Related Work

2.1 DP-based Depth Estimation

Since the introduction of DP sensors in consumer cameras, their potential for single-exposure depth recovery has been widely explored. DP imaging records signals from left and right sub-apertures at each pixel location, which can be interpreted as an extremely small-baseline stereo system [22]. Wadhwa et al. [32] explicitly separated DP images into sub-views and applied conventional stereo matching, validating the implicit stereo nature of DP imaging and its feasibility for depth estimation. These physically grounded approaches reveal the intrinsic coupling among DP disparity, focus depth, and defocus blur, but remain highly sensitive to imaging conditions.

To improve robustness, learning-based methods were introduced. Garg et al. [9] proposed an inverse depth estimation framework with affine-invariant constraints. Zhang et al. [48] leveraged dual DP cameras for supervised depth learning. Pan et al. [22] developed a DP simulator and jointly addressed depth estimation and deblurring. Xin et al. [38] formulated depth inference from defocus maps under unsupervised optimization. Kim et al. [15] enforced bidirectional disparity consistency across spatial and focal domains, achieving improved performance on synthetic and real datasets. Yu et al. [46] proposed all-directional disparity estimation for real-world QPD images, emphasizing the influence of noise, blur, and directional effects. Ghanekar et al. [10] further explored joint optical-computational design via coded apertures to enhance disparity observability. Lee et al. [5] explored incorporating foundation-model features to improve the robustness of DP disparity estimation.

Despite these advances, the extremely small baseline of DP imaging fundamentally limits disparity observability. In textureless, planar, or low-light scenarios, disparity signals become unstable, leading to degradation in weak-disparity regions. This intrinsic constraint suggests that relying primarily on DP cues—even with improved modeling or optical design—remains insufficient for stable and reliable metric depth estimation in complex real-world scenes.

2.2 Monocular Depth Estimation Foundation Models

Monocular depth estimation (MDE) seeks to recover dense geometry from a single RGB image and is inherently ill-posed. Early learning-based methods relied primarily on convolutional neural networks (CNNs). Eigen et al. [7] introduced a multi-scale architecture that predicts global structure followed by local refinement. Subsequent works improved stability through loss design and distribution modeling, including LRC [11] and DORN [8]. Multi-task frameworks such as PAD-Net [39] incorporated auxiliary geometric cues, while MiDaS [28] and LeReS [45] improved cross-dataset generalization via scale-invariant training and refinement strategies. Although effective under controlled settings, CNN-based models are limited in capturing long-range dependencies due to their local receptive fields.

The adoption of ViT has shifted MDE toward the foundation model paradigm. ViT [6] models global dependencies through self-attention, and DPT [27] first integrated ViT into dense depth prediction, marking a transition from CNN-based to Transformer-based architectures. Building on this paradigm, subsequent works have scaled model capacity and data diversity to enhance generalization.

Depth Anything [41] employed a vision foundation encoder with teacher-student distillation, leveraging unlabeled real-world images for robust zero-shot depth prediction. DAV2 [42] further improved structural fidelity by training a large teacher on high-quality synthetic depth data before distillation to real images. MoGe [34] and UniDepth [24] improved geometric consistency and metric depth estimation via supervision design and scale modeling, respectively. These ViT-based models demonstrate strong structural reasoning and cross-domain generalization through large-scale training and global attention mechanisms.

Diffusion models have also been explored for MDE. Marigold [14] reformulated denoising generation as conditional depth prediction by fine-tuning Stable Diffusion. However, diffusion-based approaches require iterative inference and incur higher computational cost. In contrast, discriminative ViT-based foundation models achieve competitive structural modeling with significantly higher efficiency, making them more practical for real-world deployment.

3 Method

3.1 Overall Framework

DP imaging provides metric depth cues grounded in explicit physical disparity induced by sub-aperture separation. However, the extremely small effective baseline limits disparity observability, particularly in textureless, low-contrast, or downsampled regions, leading to structural discontinuities and depth collapse. To address this limitation, we reconcile *metric observability* from DP imaging with *foundation-scale structural reasoning* from monocular depth models. DP depth serves as a metric anchor, while global structural priors are introduced to restore geometric coherence in weak-disparity regions.

As illustrated in Fig. 2(a), FoundDP consists of three stages: (1) DP Depth Estimation Module (DDE) producing initial metric depth D_{dp}; (2) Structure Refinement Module (SR) generating D_{metric}; (3) Depth Guidance Module (DG) using metric DP depth to guide ViT-based structural features, producing the final prediction D_{guide}.

To further analyze the complementary properties of DP depth and foundation-based depth priors, Fig. 2(b) presents depth histogram comparisons. As shown in Fig. 2(b)(1), foundation depth D_{base} captures coherent global structure but exhibits shifted peak locations relative to ground truth, indicating metric scale inconsistency, whereas DP depth D_{dp} preserves approximate scale but contains substantial noise and structural instability. Fig. 2(b)(2) further illustrates that metric DP depth guides foundation priors to suppress noise and improve distribution consistency over DP depth. As a result, the final guided depth D_{guide}

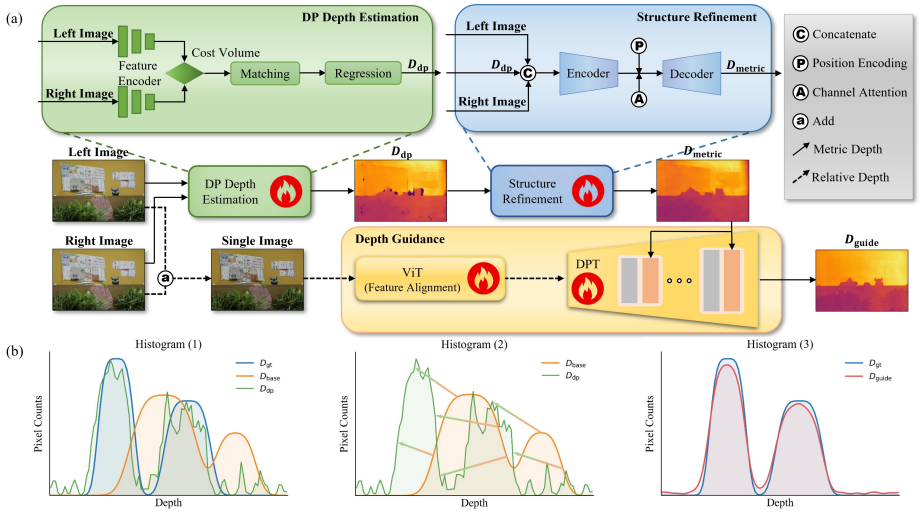

Fig. 2: (a) Overall framework of FoundDP. DP sub-aperture images are used to estimate metric depth D_{dp}, which is refined to D_{metric} for improved geometric consistency. In parallel, a monocular depth foundation model extracts structural priors. A depth guidance module then uses metric depth to guide the structural features and produce the final depth D_{guide}. (b) Illustration of depth histogram analysis. (1) Compared to ground truth D_{gt}, foundation depth D_{base} exhibits peak shifts, while DP depth D_{dp} contains significant noise. (2) Metric depth guides foundation priors to correct structural errors over noisy DP depth. (3) The final depth D_{guide} closely aligns with the ground-truth distribution, demonstrating improved structural and metric accuracy.

closely matches the ground-truth depth distribution, as shown in Fig. 2(b)(3), demonstrating that our framework effectively combines metric observability with global structural consistency.

The overall formulation is:

$$\begin{cases} D_{\mathrm{dp}} = \mathcal{F}_{\mathrm{DDE}}(I_L, I_R), \\ D_{\mathrm{metric}} = \mathcal{F}_{\mathrm{SR}}(I_L, I_R, D_{\mathrm{dp}}), \\ D_{\mathrm{guide}} = \mathcal{F}_{\mathrm{DG}}\left(D_{\mathrm{metric}}, \Phi_{\mathrm{ViT}}\left(\frac{I_L + I_R}{2}\right)\right), \end{cases} \tag{1}$$

where Φ_{ViT} denotes structural features extracted by the foundation encoder, and $\mathcal{F}_{\mathrm{DDE}}$, $\mathcal{F}_{\mathrm{SR}}$, and $\mathcal{F}_{\mathrm{DG}}$ represent the DDE, SR, and DG respectively.

3.2 DP Depth Estimation Module

We first construct a DP-based network to generate an initial metric depth map D_{dp} with explicit physical scale. Given left and right DP images $I_L, I_R \in \mathbb{R}^{3\times H\times W}$, a shared-weight encoder extracts multi-scale representations. The encoder integrates strided convolutions for receptive field enlargement with dilated convolutions and pooling branches to enhance contextual aggregation in textureless regions. Multi-scale features are fused at a unified resolution and compressed

via 1×1 convolutions, producing matching features $F_L, F_R \in \mathbb{R}^{C \times H' \times W'}$, where $H' = H/4$ and $W' = W/4$.

A 3D cost volume [12, 22] is constructed through explicit displacement concatenation. Instead of conventional single-sided search, we define a symmetric displacement hypothesis set centered at zero disparity, $d \in [-D/2, D/2]$. For each hypothesis, aligned left and right features are concatenated along the channel dimension, forming a volume of size $2C \times D \times H' \times W'$. This symmetric design aligns with the optical symmetry of DP sub-apertures and reduces systematic bias during depth learning. The volume is regularized using a 3D convolutional network to jointly model spatial and disparity consistency.

To obtain continuous predictions, we adopt Softmin-based regression. The regularized volume is upsampled to the target resolution, and Softmin is applied along the disparity dimension to obtain a probability distribution. The final depth is computed as

$$D_{\mathrm{dp}}(x, y) = \mathbb{E}_{d \sim P(d|x,y)}[d], \tag{2}$$

where

$$P(d \mid x, y) = \frac{\exp(-C(x, y, d))}{\sum_{d'} \exp(-C(x, y, d'))}, \tag{3}$$

and $\mathbb{E}$ denotes expectation over valid pixels. This strategy avoids discrete quantization artifacts and produces smoother predictions in textureless and low-contrast regions. Nevertheless, because this module relies primarily on local disparity cues, structural degradation may persist in weak-disparity areas, motivating subsequent refinement and guidance.

3.3 Structure Refinement Module

Although the DP module provides physically grounded depth D_{dp}, its reliance on local disparity cues makes it vulnerable to structural degradation in regions with weak disparity observability, such as textureless or planar areas. To address this limitation while preserving metric scale, we introduce a structure refinement network that predicts a residual structural correction conditioned on both image appearance and the initial DP depth.

Specifically, the refinement module estimates a correction field:

$$\Delta D = \mathcal{SR}\left(I_L, I_R, D_{\mathrm{dp}}\right), \qquad D_{\mathrm{metric}} = D_{\mathrm{dp}} + \Delta D, \tag{4}$$

where $\mathcal{SR}(\cdot)$ denotes the refinement network and ΔD represents the learned structural correction. This residual formulation preserves the physically grounded metric scale while compensating for structural errors caused by insufficient disparity observability.

To implement $\mathcal{SR}(\cdot)$, the network jointly processes I_L, I_R, and D_{dp} using a hierarchical encoder-decoder architecture [23]. The inputs are concatenated along the channel dimension and passed through multi-scale residual convolutional blocks, enabling progressive aggregation of local image evidence and global

structural context. A channel attention [3, 33] and 2D positional encoding [37] mechanism is applied at high-resolution stages to emphasize geometrically informative responses and suppress unreliable regions. The decoder reconstructs the refined depth in a residual manner, ensuring structural continuity while maintaining metric consistency.

The resulting depth D_{metric} provides structurally refined metric depth and serves as the geometric foundation for the subsequent guidance.

3.4 Depth Guidance Module

ViT Alignment under DP Degradation. After obtaining refined metric depth, we introduce monocular structural priors from a pretrained ViT [6]. Unlike prior fusion approaches that assume clean representations, DP defocus and optical blur attenuate high-frequency content and induce structural bias in ViT features, as illustrated in Fig. 3. Directly fusing such degraded features can propagate structural inconsistencies and compromise depth reliability.

Fig. 3: Effect of ViT alignment under defocus degradation. Aligned ViT features yield more structurally consistent predictions in blurred regions.

To mitigate this degradation, we apply explicit ViT feature alignment prior to restoring structural consistency. Specifically, paired clear RGB images and their DP-degraded counterparts are processed by a shared ViT encoder, and a feature-space alignment constraint is imposed to reduce representation discrepancies caused by defocus blur. This alignment encourages degraded features to recover structurally consistent representations, providing more reliable global priors.

Formally, we supervise the feature alignment with the following objective:

$$\mathcal{L}_{\text{align}} = \sum_{l} \|\Phi_l(I_{\text{clear}}) - \Phi_l(I_{\text{blur}})\|_2^2 , \tag{5}$$

where Φ_l denotes the ViT feature at level l. This objective explicitly enforces representation consistency between defocused and clean inputs, mitigating defocus-induced structural bias.

Depth Guidance with Metric Conditioning. Inspired by depth-prompt strategies [13, 20, 35, 36, 44], we treat the refined DP metric depth as an explicit geometric condition during depth guidance, preserving metric scale while leveraging foundation-level structural priors.

The RGB image is first encoded by the ViT into multi-level token features, which are reshaped into multi-scale feature maps $\{F_1, F_2, F_3, F_4\}$ [42]. These features are progressively decoded in a DPT-style architecture. At each decoding stage, let X denote the intermediate feature map at the current scale. The

aligned and normalized metric depth $\hat{D}$ is injected as a spatial condition to guide features.

Specifically, feature updates are formulated as:

$$X' = X + \phi(X) + g(\hat{D}) \odot \psi(\hat{D}), \tag{6}$$

where $\phi(\cdot)$ denotes a residual feature transformation applied to X, $\psi(\cdot)$ extracts metric-conditioned modulation signals from $\hat{D}$, and $g(\cdot)$ is an adaptive gating function that controls the spatial influence of the metric condition. The element-wise modulation $\odot$ enables selective enhancement of geometrically reliable regions. This design encourages the network to rely on DP-derived metric cues in high-confidence areas while allowing foundation-based structural priors to dominate in weak-disparity regions, thereby reducing error propagation.

3.5 Loss Function

We adopt a unified Smooth L1 loss in logarithmic depth space to supervise all three stages:

$$\mathcal{L} = \frac{1}{N} \sum_{i} \mathrm{SmoothL1}\big(\log_{10} D^{(i)}, \log_{10} D^{(i)}_{\mathrm{gt}}\big), \tag{7}$$

where $D^{(i)}$ denotes predictions from DP estimation, refinement, and guidance, and D_{gt} is the ground truth. We do not adopt scale-invariant log losses (e.g., SiLog [7]), as dual-pixel has the capability to recover metric depth, where preserving absolute scale is necessary for physical consistency.

4 Experiments

4.1 Implementation Details

Our framework uses DAV2 as the foundation model and is implemented in PyTorch, trained on a single NVIDIA RTX 4090D GPU. All trainable modules are optimized using Adam with an initial learning rate of 1×10^{-4} and trained for 100 epochs per stage. Following common practice, we resize all images to a fixed resolution to ensure computational efficiency and consistent evaluation across datasets. In our experiments, images are resized to 512×768 for both training and evaluation.

Training Strategy. Given the cascaded design and heterogeneous objectives of different modules, we adopt a stage-wise optimization scheme for stable convergence. First, the Dual-Pixel Depth Estimation Module is trained independently to establish reliable metric-scale predictions. The Structure Refinement Module is then trained with DDE frozen to enhance geometric continuity in weak-disparity regions. Next, the pretrained ViT encoder of Depth Guidance Module is introduced and refined under DP imaging conditions to mitigate representation degradation caused by defocus blur. Finally, DDE, SR, and the ViT encoder are fixed, and only the DPT is optimized. This staged training preserves metric consistency while enabling stable integration of global structural priors.

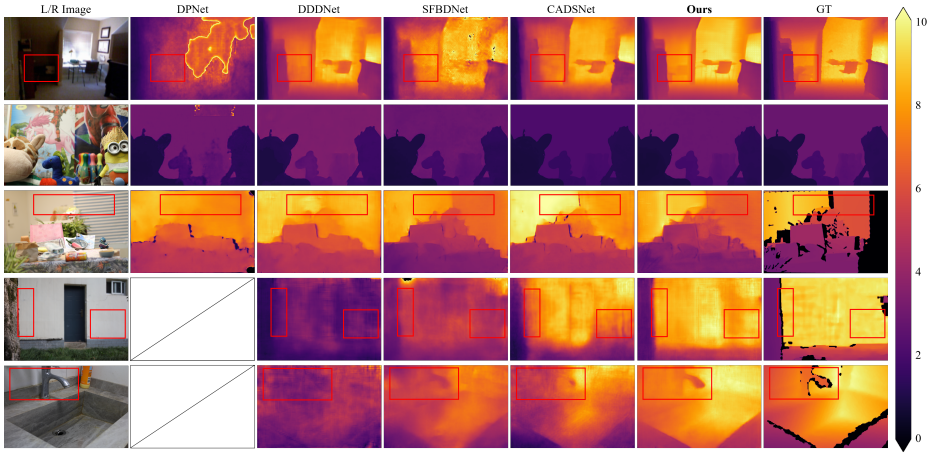

Fig. 4: Qualitative comparison. **Rows 1-3:** Results on representative DP datasets. Our method preserves sharper structural boundaries and more coherent geometry, especially in weak-disparity regions (highlighted in red). **Rows 4-5:** Results on the downsampling dataset. Our model remains robust under disparity attenuation, whereas competing methods exhibit structural degradation.

4.2 Datasets and Evaluation Metrics

Datasets. We evaluate on both synthetic and real-world DP datasets. For each dataset, the training and test sets are generated or captured under the same imaging model or camera setup.

Synthetic: NYUData [30]. Since the synthetic dataset contains only single-view RGB images and depth maps, we generate DP image pairs using a ray-tracing-based DP simulator [12], which produces pixel-wise left/right point spread functions and corresponding DP observations.

Real-world: DP2020 [25], DP5K [17], and DP2019 [9]. The real datasets provide DP left-right image pairs with corresponding depth annotations.

To study robustness against disparity attenuation, we additionally construct a downsampling dataset named DPDown70 using a Canon EOS R6 Mark II camera equipped with an RF 50mm lens at an aperture of f/8, including ∼1k training images and ∼70 test samples. Original 4000×6000 images are resized to 512×768.

To ensure consistent metric evaluation, all valid depths are linearly mapped to a unified physical range of 1-10 meters. Values outside this range are ignored. For datasets providing disparity only, depth is computed as:

$$D = \frac{1}{a + bd}, \quad a = \frac{1}{D_{\max}}, \quad b = \frac{1}{D_{\min}} - \frac{1}{D_{\max}}, \tag{8}$$

with $D_{\min} = 1.0$ m and $D_{\max} = 10.0$ m.

Metrics. We evaluate depth quality from three perspectives:

Affine-Invariant Error (AI) [9,15,22,46]: AI(1) and AI(2) measure mean absolute and squared errors after affine alignment.

Rank Consistency [15,46]: Spearman correlation ρ_s.

Threshold Accuracy [12,22]: $\delta < 1.25$ (Acc-1) and $\delta < 1.25^2$ (Acc-2).

These complementary metrics jointly assess structural fidelity, relative ordering, and metric accuracy.

Table 1: Quantitative test on NYUData [30], DP2020 [25], DP5K [17] and DP2019 [9].

Method	NYUData [30]					DP2020 [25]				
	AI(1)↓	AI(2)↓	$1-\lvert\rho_s\rvert$↓	Acc-1↑	Acc-2↑	AI(1)↓	AI(2)↓	$1-\lvert\rho_s\rvert$↓	Acc-1↑	Acc-2↑
DPNet [9]	0.9271	1.2896	0.2179	0.2925	0.5777	0.3509	0.5629	0.1649	0.7521	0.8180
SFBDNet [15]	0.5330	0.8576	0.1489	0.6579	0.9174	0.0633	0.2095	0.0241	0.9963	0.9987
DDDNet [22]	0.2271	0.3333	0.0299	0.9520	0.9953	0.0851	0.2564	0.1138	0.6934	0.9757
CADSNet [10]	0.2018	0.2931	0.0248	0.9271	0.9933	0.0162	0.0424	0.0467	0.8153	0.9997
Ours	**0.1475**	**0.2368**	**0.0178**	**0.9661**	**0.9988**	**0.0110**	**0.0280**	**0.0232**	**0.9998**	**1.0000**
Method	DP5K [17]					DP2019 [9]				
	AI(1)↓	AI(2)↓	$1-\lvert\rho_s\rvert$↓	Acc-1↑	Acc-2↑	AI(1)↓	AI(2)↓	$1-\lvert\rho_s\rvert$↓	Acc-1↑	Acc-2↑
DPNet [9]	0.4564	0.9296	0.3318	0.6675	0.8903	0.1613	0.3211	0.5884	0.4311	0.5807
SFBDNet [15]	0.3406	0.8228	0.1131	0.7757	0.8973	0.1549	0.3087	0.5223	0.6312	0.7974
DDDNet [22]	0.3617	0.8279	0.1706	0.7849	0.9088	0.1473	0.2935	0.4539	0.6820	0.9031
CADSNet [10]	0.2680	0.7854	**0.0977**	0.6716	0.8773	0.1313	0.2807	0.3491	0.7335	0.9277
Ours	**0.2579**	**0.7562**	0.0989	**0.7976**	**0.9251**	**0.1168**	**0.2622**	**0.3224**	**0.8571**	**0.9404**

Table 2: Quantitative evaluation *in weak regions* on NYUData [30], DP2020 [25], DP5K [17] and DP2019 [9].

Method	NYUData [30]					DP2020 [25]				
	AI(1)↓	AI(2)↓	$1-\lvert\rho_s\rvert$↓	Acc-1↑	Acc-2↑	AI(1)↓	AI(2)↓	$1-\lvert\rho_s\rvert$↓	Acc-1↑	Acc-2↑
DPNet [9]	0.8512	1.1677	0.2340	0.2544	0.5225	0.3389	0.5938	0.1931	0.8087	0.8706
SFBDNet [15]	0.5038	0.7853	0.1698	0.6579	0.9243	0.0622	0.2155	0.0258	0.9942	0.9979
DDDNet [22]	0.2504	0.3471	0.0438	0.9391	0.9940	0.0717	0.2938	0.1397	0.7220	0.9933
CADSNet [10]	0.1949	0.2725	0.0320	0.9276	0.9947	0.0113	0.0416	0.0437	0.7472	0.9995
Ours	**0.1280**	**0.1872**	**0.0180**	**0.9675**	**0.9995**	**0.0044**	**0.0157**	**0.0113**	**1.0000**	**1.0000**
Method	DP5K [17]					DP2019 [9]				
	AI(1)↓	AI(2)↓	$1-\lvert\rho_s\rvert$↓	Acc-1↑	Acc-2↑	AI(1)↓	AI(2)↓	$1-\lvert\rho_s\rvert$↓	Acc-1↑	Acc-2↑
DPNet [9]	0.3877	0.8228	0.3746	0.6879	0.8962	0.1955	0.3685	0.6234	0.4107	0.6061
SFBDNet [15]	0.3147	0.7439	0.2986	0.7592	0.8996	0.1874	0.3542	0.5490	0.5740	0.7291
DDDNet [22]	0.3393	0.7736	0.3079	0.7551	0.8936	0.1894	0.3553	0.6033	0.5176	0.6693
CADSNet [10]	0.2781	0.7283	0.2526	0.6965	0.8738	0.1834	0.3489	0.5226	0.6143	0.7098
Ours	**0.2441**	**0.7019**	**0.2286**	**0.7777**	**0.9264**	**0.1795**	**0.3432**	**0.5067**	**0.6229**	**0.7552**

4.3 Comparison with DP Methods

We compare against representative DP-based methods: DPNet [9], DDDNet [22], SFBDNet [15], and CADSNet [10]. All methods are evaluated under identical preprocessing and metric protocols. The best results are highlighted in bold and the second-best results are underlined in the table.

General Comparison. As shown in Table 1, our method achieves the best or second-best results. Notably, we consistently obtain the lowest affine-invariant errors and the highest threshold accuracies. Compared to earlier DPNet and DDDNet, our approach substantially reduces overall error. Against recent methods such as SFBDNet and CADSNet, our framework further improves metric accuracy while preserving structural consistency, demonstrating that global structural priors effectively compensate for DP observability limitations. Visual comparisons in Fig. 4 (Rows 1-3) confirm that our method produces smoother planar surfaces, more stable boundaries, and fewer depth holes.

Performance in Weak-Disparity Regions. Weak-disparity regions are identified using gradient-based analysis. Sobel responses are smoothed, thresholded, and refined with morphological operations to obtain stable masks. More details are listed in supplementary materials.

Table 2 shows that competing DP methods degrade significantly in weak regions, while our method maintains substantially lower AI errors and higher threshold accuracies. The improvement gap is more pronounced than in global evaluation, indicating that foundation-based structural priors effectively compensate when local DP correspondence fails.

Table 3: Quantitative evaluation on our *Downsampling Dataset* DPDown70.

Method	DPDown70				
	AI(1)↓	AI(2)↓	$1-\lvert\rho_s\rvert$↓	Acc-1↑	Acc-2↑
SFBDNet [15]	0.6458	1.0073	0.2478	0.4113	0.7309
DDDNet [22]	0.6379	0.9120	0.2215	0.3524	0.6746
CADSNet [10]	<u>0.4003</u>	<u>0.6281</u>	<u>0.1238</u>	<u>0.5047</u>	<u>0.8074</u>
Ours	**0.2482**	**0.4078**	**0.0443**	**0.7035**	**0.9163**

Downsampling Robustness. Under resolution reduction (Table 3), most DP methods suffer severe performance degradation. In contrast, our method maintains strong metric accuracy and structural coherence. The qualitative results in Fig. 4 (Rows 4-5) further demonstrate that our approach preserves geometric contours despite disparity attenuation.

Disparity Observability Analysis. The weak-disparity region analysis above evaluates performance under naturally occurring observability variations. To further investigate the impact of disparity observability in a controlled manner, we progressively reduce disparity observability via downsampling.

In DP imaging, disparity cues arise from subpixel correspondence between dual sub-aperture views. Downsampling suppresses fine-scale intensity variations and reduces effective gradient strength, thereby weakening correspondence signals required for depth estimation. As a result, disparity observability is inversely related to the downsampling factor, allowing resolution reduction to serve as a continuous observability proxy.

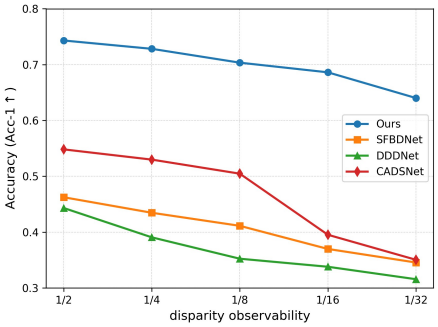

Fig. 5: Depth accuracy versus disparity observability. Our method shows improved robustness under reduced observability.

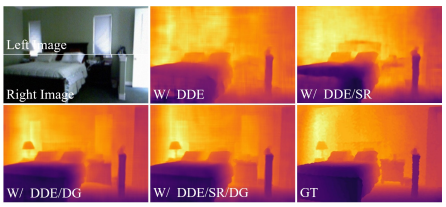

Fig. 6: Ablation results. Progressive integration of SR and DG improves boundary sharpness, structural continuity, and weak-region stability.

Table 4: Ablation study on three modules.

Components			NYUData				
DDE	SR	DG	AI(1)↓	AI(2)↓	$1-\lvert\rho_s\rvert$↓	Acc-1↑	Acc-2↑
✓			0.2463	0.3764	0.0332	0.9297	0.9610
✓	✓		0.2312	0.3486	0.0298	0.9491	0.9715
✓		✓	0.1865	0.2871	0.0220	0.9348	0.9883
✓	✓	✓	**0.1475**	**0.2368**	**0.0178**	**0.9661**	**0.9988**

Table 5: Ablation study on ViT feature alignment.

Components		DP5K				
Base	ViT-Ali.	AI(1)↓	AI(2)↓	$1-\lvert\rho_s\rvert$↓	Acc-1↑	Acc-2↑
✓		0.2741	0.7993	0.1042	0.7839	0.9197
✓	✓	**0.2579**	**0.7562**	**0.0989**	**0.7976**	**0.9251**

We evaluate depth accuracy under the Acc-1 criterion as a function of disparity observability. As shown in Fig. 5, all methods exhibit performance degradation as observability decreases, confirming the fundamental dependence of DP depth estimation on reliable disparity cues. However, existing DP methods degrade rapidly under reduced observability, indicating their strong reliance on local correspondence signals. In contrast, our method maintains consistently higher accuracy and exhibits a more gradual degradation trend, demonstrating improved robustness when disparity observability becomes limited.

4.4 Ablation Studies

Module Contribution Analysis. Table 4 reports progressive integration of core modules. Using DDE alone provides stable metric depth but leaves structural artifacts. Adding SR reduces local discontinuities and improves affine-invariant accuracy, confirming its role in repairing depth holes. Incorporating DG further yields the largest performance gain, demonstrating that global structural priors effectively stabilize geometry when disparity observability deteriorates. Qualitative results in Fig. 6 match the quantitative trends: depth discontinuities are progressively suppressed, and structural coherence improves as modules are added.

Effect of ViT feature alignment. Table 5 evaluates the impact of the proposed ViT feature alignment strategy. Directly incorporating foundation features

improves performance, confirming the benefit of global structural priors. However, performance remains constrained due to representation degradation caused by DP blur, which introduces bias in the extracted ViT features; introducing feature alignment mitigates this effect and improves performance.

Table 6: ViT feature alignment analysis.

ViT Feature	Cosine Similarity↑
Origin	0.8737
Refined	0.9454

To further analyze this effect, Table 6 measures features before and after alignment using cosine similarity to clean-image features as reference. DP blur degrades ViT feature representations, while the proposed alignment strategy restores feature alignment. By correcting this degradation-induced bias, the alignment enables more effective metric-guided integration with DP metric cues, leading to consistent improvements across all metrics in Table 5. This confirms that mitigating ViT representation degradation is essential for stable depth estimation under DP imaging.

Overall, the ablation results confirm that (1) SR mitigates local geometric degradation, (2) DG injects global structural reasoning, and (3) ViT feature alignment stabilizes representations under DP-induced blur. Together, these components address the core limitation of weak disparity observability in DP depth estimation. Lastly, complexity and runtime analysis further demonstrate that the proposed framework achieves these improvements with practical inference efficiency (see supplementary for details).

5 Conclusion and Discussion

DP depth estimation is fundamentally constrained by weak disparity observability, causing structural degradation and unreliable predictions in weak-disparity regions. We address this limitation by reconciling physically grounded metric cues from DP imaging with global structural priors from monocular depth foundation models. Our FoundDP preserves metric scale through DP disparity while leveraging foundation representations to restore coherence where disparity becomes unreliable. Extensive experiments on synthetic and real-world benchmarks demonstrate consistent gains in structural fidelity and metric accuracy, particularly under weak-disparity and downsampling conditions.

Despite these advances, the framework inherently depends on observable disparity signals; performance may degrade when DP measurements are heavily corrupted by noise or extreme optical degradation. While foundation-based priors improve robustness, their adaptation to DP-specific imaging characteristics remains imperfect. Moreover, integrating a ViT-based encoder introduces additional computational overhead, limiting real-time deployment on resource-constrained devices. Nevertheless, incorporating foundation-level structural reasoning is essential for stabilizing metric DP depth under extremely small-baseline and low-disparity conditions, establishing a principled bridge between physical imaging constraints and large-scale representation learning.

Acknowledgments

This work was supported by National Natural Science Foundation of China (Grant No. 32471146) and the project N20240194. The authors thank Echossom, Miya, and Xinge for valuable discussions and assistance.

References

1. Abuolaim, A., Afifi, M., Brown, M.S.: Improving single-image defocus deblurring: How dual-pixel images help through multi-task learning. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1231–1239 (2022)
2. Abuolaim, A., Brown, M.S.: Defocus deblurring using dual-pixel data. In: European conference on computer vision. pp. 111–126. Springer (2020)
3. Bastidas, A.A., Tang, H.: Channel attention networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops. pp. 0–0 (2019)
4. Choi, M., Lee, H., Lee, H.e.: Exploring positional characteristics of dual-pixel data for camera autofocus. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13158–13168 (2023)
5. Doehyung, L., Zhuofeng, W., Yusuke, M., Masatoshi, O.: Fmdp: Leveraging a foundation model for dual-pixel disparity estimation. IEICE Proceeding Series **93**, O1–1 (2025)
6. Dosovitskiy, A.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
7. Eigen, D., Puhrsch, C., Fergus, R.: Depth map prediction from a single image using a multi-scale deep network. Advances in neural information processing systems **27** (2014)
8. Fu, H., Gong, M., Wang, C., Batmanghelich, K., Tao, D.: Deep ordinal regression network for monocular depth estimation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2002–2011 (2018)
9. Garg, R., Wadhwa, N., Ansari, S., Barron, J.T.: Learning single camera depth estimation using dual-pixels. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 7628–7637 (2019)
10. Ghanekar, B., Khan, S.S., Sharma, P., Singh, S., Boominathan, V., Mitra, K., Veeraraghavan, A.: Passive snapshot coded aperture dual-pixel rgb-d imaging. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 25348–25357 (2024)
11. Godard, C., Mac Aodha, O., Brostow, G.J.: Unsupervised monocular depth estimation with left-right consistency. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 270–279 (2017)
12. He, F., Zhao, D., Xu, H., Quan, T., Zeng, S.: Simulating dual-pixel images from ray tracing for depth estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 26106–26115 (2025)
13. Jiang, H., Lou, Z., Ding, L., Xu, R., Tan, M., Jiang, W., Huang, R.: Defom-stereo: Depth foundation model based stereo matching. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 21857–21867 (2025)

14. Ke, B., Obukhov, A., Huang, S., Metzger, N., Daudt, R.C., Schindler, K.: Repurposing diffusion-based image generators for monocular depth estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9492–9502 (2024)
15. Kim, D., Jang, H., Kim, I., Kim, M.H.: Spatio-focal bidirectional disparity estimation from a dual-pixel image. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5023–5032 (2023)
16. Kobayashi, M., Johnson, M., Wada, Y., Tsuboi, H., Takada, H., Togo, K., Kishi, T., Takahashi, H., Ichikawa, T., Inoue, S.: A low noise and high sensitivity image sensor with imaging and phase-difference detection af in all pixels. ITE Transactions on Media Technology and Applications **4**(2), 123–128 (2016)
17. Li, F., Guo, H., Santo, H., Okura, F., Matsushita, Y.: Learning to synthesize photorealistic dual-pixel images from rgbd frames. In: 2023 IEEE International Conference on Computational Photography (ICCP). pp. 1–11. IEEE (2023)
18. Li, Y., Monno, Y., Okutomi, M.: Dual-pixel raindrop removal. IEEE Transactions on Pattern Analysis and Machine Intelligence (2024)
19. Li, Y., Yi, Y., Shu, X., Ren, D., Li, Q., Zuo, W.: Learning dual-pixel alignment for defocus deblurring. Neurocomputing **616**, 128880 (2025)
20. Lin, H., Peng, S., Chen, J., Peng, S., Sun, J., Liu, M., Bao, H., Feng, J., Zhou, X., Kang, B.: Prompting depth anything for 4k resolution accurate metric depth estimation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 17070–17080 (2025)
21. Monin, S., Katz, S., Evangelidis, G.: Continuous cost aggregation for dual-pixel disparity extraction. In: 2024 International Conference on 3D Vision (3DV). pp. 675–684. IEEE (2024)
22. Pan, L., Chowdhury, S., Hartley, R., Liu, M., Zhang, H., Li, H.: Dual pixel exploration: Simultaneous depth estimation and image restoration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4340–4349 (2021)
23. Pan, L., Hartley, R., Liu, L., Xu, Z., Chowdhury, S., Yang, Y., Zhang, H., Li, H., Liu, M.: Weakly-supervised depth estimation and image deblurring via dual-pixel sensors. IEEE Transactions on Pattern Analysis and Machine Intelligence (2024)
24. Piccinelli, L., Yang, Y.H., Sakaridis, C., Segu, M., Li, S., Van Gool, L., Yu, F.: Unidepth: Universal monocular metric depth estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10106–10116 (2024)
25. Punnappurath, A., Abuolaim, A., Afifi, M., Brown, M.S.: Modeling defocus-disparity in dual-pixel sensors. In: 2020 IEEE International Conference on Computational Photography (ICCP). pp. 1–12. IEEE (2020)
26. Punnappurath, A., Brown, M.S.: Reflection removal using a dual-pixel sensor. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1556–1565 (2019)
27. Ranftl, R., Bochkovskiy, A., Koltun, V.: Vision transformers for dense prediction. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 12179–12188 (2021)
28. Ranftl, R., Lasinger, K., Hafner, D., Schindler, K., Koltun, V.: Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. IEEE transactions on pattern analysis and machine intelligence **44**(3), 1623–1637 (2020)

29. Shi, Z., Chugunov, I., Bijelic, M., Côté, G., Yeom, J., Fu, Q., Amata, H., Heidrich, W., Heide, F.: Split-aperture 2-in-1 computational cameras. ACM Transactions on Graphics (TOG) **43**(4), 1–19 (2024)
30. Silberman, N., Hoiem, D., Kohli, P., Fergus, R.: Indoor segmentation and support inference from rgbd images. In: European conference on computer vision. pp. 746–760. Springer (2012)
31. Śliwiński, P., Wachel, P.: A simple model for on-sensor phase-detection autofocusing algorithm. Journal of Computer and Communications **1**(06), 11 (2013)
32. Wadhwa, N., Garg, R., Jacobs, D.E., Feldman, B.E., Kanazawa, N., Carroll, R., Movshovitz-Attias, Y., Barron, J.T., Pritch, Y., Levoy, M.: Synthetic depth-of-field with a single-camera mobile phone. ACM Transactions on Graphics (ToG) **37**(4), 1–13 (2018)
33. Wang, Q., Wu, B., Zhu, P., Li, P., Zuo, W., Hu, Q.: Eca-net: Efficient channel attention for deep convolutional neural networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11534–11542 (2020)
34. Wang, R., Xu, S., Dai, C., Xiang, J., Deng, Y., Tong, X., Yang, J.: Moge: Unlocking accurate monocular geometry estimation for open-domain images with optimal training supervision. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 5261–5271 (2025)
35. Wang, Z., Chen, S., Yang, L., Wang, J., Zhang, Z., Zhao, H., Zhao, Z.: Depth anything with any prior. arXiv preprint arXiv:2505.10565 (2025)
36. Wen, B., Trepte, M., Aribido, J., Kautz, J., Gallo, O., Birchfield, S.: Foundationstereo: Zero-shot stereo matching. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5249–5260 (2025)
37. Wu, K., Peng, H., Chen, M., Fu, J., Chao, H.: Rethinking and improving relative position encoding for vision transformer. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10033–10041 (2021)
38. Xin, S., Wadhwa, N., Xue, T., Barron, J.T., Srinivasan, P.P., Chen, J., Gkioulekas, I., Garg, R.: Defocus map estimation and deblurring from a single dual-pixel image. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2228–2238 (2021)
39. Xu, D., Ouyang, W., Wang, X., Sebe, N.: Pad-net: Multi-tasks guided prediction-and-distillation network for simultaneous depth estimation and scene parsing. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 675–684 (2018)
40. Yang, H., Pan, L., Yang, Y., Hartley, R., Liu, M.: Ldp: Language-driven dual-pixel image defocus deblurring network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 24078–24087 (2024)
41. Yang, L., Kang, B., Huang, Z., Xu, X., Feng, J., Zhao, H.: Depth anything: Unleashing the power of large-scale unlabeled data. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10371–10381 (2024)
42. Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J., Zhao, H.: Depth anything v2. Advances in Neural Information Processing Systems **37**, 21875–21911 (2024)
43. Yang, Y., Pan, L., Liu, L., Liu, M.: K3dn: Disparity-aware kernel estimation for dual-pixel defocus deblurring. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13263–13272 (2023)
44. Ye, T., Chen, S., Chen, H., Chai, W., Ren, J., Xing, Z., Li, W., Zhu, L.: Prompthaze: Prompting real-world dehazing via depth anything model. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 9454–9462 (2025)

45. Yin, W., Zhang, J., Wang, O., Niklaus, S., Mai, L., Chen, S., Shen, C.: Learning to recover 3d scene shape from a single image. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 204–213 (2021)
46. Yu, H., Song, S., Sun, L., Su, W., Yang, X., Liu, C.: All-directional disparity estimation for real-world qpd images. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 21836–21846 (2025)
47. Yu, K., Pan, L., Liu, L., Liang, W.: Enhanced dual-pixel image reflection removal via gaussian splatting. In: Proceedings of the 33rd ACM International Conference on Multimedia. pp. 7766–7775 (2025)
48. Zhang, Y., Wadhwa, N., Orts-Escolano, S., Häne, C., Fanello, S., Garg, R.: Du2net: Learning depth estimation from dual-cameras and dual-pixels. In: European Conference on Computer Vision. pp. 582–598. Springer (2020)