MODepth: Benchmarking Multi-frame Monocular Depth Estimation with Optical Image Stabilization

ANONYMOUS AUTHOR(S)

This paper presents MODepth, a multi-frame monocular depth estimation system based on the controlled motion of an optical image stabilization (OIS) module. By actively injecting acoustic signals, we induce regular translational movements of the OIS lens, resulting in controllable camera pose changes and simplifying inter-frame pose estimation. Leveraging multi-frame images captured under OIS-controlled lens movements, we design a high-precision depth estimation network, MODNet, and introduce the principal point offset estimation module and pose estimation modules to fully exploit geometric information across frames. To validate the effectiveness of our approach, we collect a new dataset MODdata with 1100 samples in nearly 220 indoor scenarios and benchmark our model as an OIS-based multi-frame depth estimation method, comparing it to ground truth obtained from a depth sensor and other state-of-the-art monocular depth estimation algorithms. Our method achieves competitive or superior performance compared to fully supervised baselines, reaching an RMSE of 0.439, which outperforms all evaluated methods, demonstrating that self-supervised fine-tuning with OIS-induced parallax is a viable alternative to ground-truth supervision.

The relevant code and dataset will be released upon acceptance of this paper.

Additional Key Words and Phrases: Monocular Depth Estimation, Optical Image Stabilization

1 Introduction

Monocular depth estimation has attracted considerable attention due to its wide range of applications and low hardware requirements (i.e., requiring only a single camera) [Bian et al. 2021a; Godard et al. 2019; Yu et al. 2020]. In particular, single-frame monocular depth estimation aims to recover the 3D geometric structure from a single RGB image [Zhao et al. 2023, 2020; Zhou et al. 2019]. However, due to the absence of disparity or multi-view information, this type of method typically relies on supervised training with dense ground-truth (GT) depth maps captured by depth sensors [Bhat et al. 2021; Ranftl et al. 2021]. Moreover, these approaches tend to over-rely on the trained semantic priors, which severely limits their generalization ability in novel or complex scenes and often necessitates the collection of new datasets [Piccinelli et al. 2023].

Multi-frame monocular depth estimation leverages video sequences or multiple consecutive images captured by a moving monocular camera, and estimates depth by enforcing inter-frame 3D geometric consistency [Yao et al. 2018, 2019]. This approach effectively alleviates the strong reliance on GT-based supervised learning inherent in single-frame methods. Accurate relative pose estimation between frames is essential to guarantee the quality of geometric supervision. In scenarios with constrained camera motion, such as autonomous driving, pose estimation networks perform reliably: large inter-frame translations and small rotations simplify the pose estimation task and yield high accuracy, thereby improving depth estimation performance [Bian et al. 2021b; Li et al. 2021]. In contrast, hand-held camera scenarios introduce complex and unstable motion patterns, making pose prediction more challenging. Inaccurate pose estimates in such cases degrade the effectiveness of geometric supervision and adversely affect depth estimation accuracy [Jiang et al. 2021; Yu et al. 2020; Zhao et al. 2023].

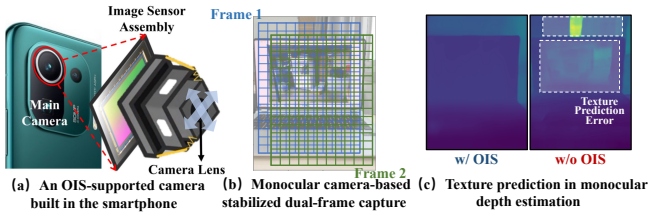


Fig. 1. MODepth leverages OIS mechanism lens control for stable parallax acquisition under static equipment conditions.

This raises an important question: *can we design multi-frame capture protocols for smartphones that induce regular and stable camera motion patterns similar to those in autonomous driving scenarios?* Achieving such controlled camera poses would simplify the pose estimation problem, enhance the effectiveness of geometric supervision in multi-frame monocular depth estimation, and ultimately lead to higher-quality depth maps.

Inspired by the optical image stabilization (OIS)-based vision enhancement techniques [Lu et al. 2024; Pan et al. 2022b; Trippel et al. 2017], we explore leveraging the controlled motion of camera lenses induced by OIS modules to generate regular camera pose changes. In related works, researchers inject high-frequency and inaudible acoustic signals to actively control the lens motion in the OIS module, which means the acoustic signals can cause the camera lens to shift in a controlled manner. This approach allows for capturing sequential images with regular motion patterns while keeping the camera body stationary. Building upon this principle, in this paper, we propose MODepth, a novel multi-frame monocular depth estimation framework and benchmark built upon the OIS-induced stereoscopic image acquisition paradigm.

Our contributions span both dataset construction and model design. First, we construct a new dataset consisting of 1,100 OIS-driven image pairs captured in diverse indoor environments. In each pair, a reference frame is captured with the lens at rest, followed by a target frame obtained while the lens is actively moved using inaudible acoustic signals that actuate the OIS module. This results in stable and repeatable parallax without requiring external camera motion. Second, we introduce a two-stage learning framework tailored to indoor depth estimation. Inspired by Croco and other works utilizing synthetic data for pretraining, we first develop a synthetic dataset that mimics the parallax characteristics induced by OIS in real-world settings. Using this dataset, we perform supervised pretraining to initialize a ViT-based depth estimation backbone with strong inductive priors over indoor 3D structures. Subsequently, we fine-tune the model via self-supervised learning on real OIS image pairs. This phase incorporates a structure-aware photometric consistency loss that leverages the known characteristics of the

OIS-induced lens motion, ensuring geometric coherence during optimization. Through this hybrid strategy, our model learns to produce dense and geometrically consistent depth maps even under challenging indoor conditions. Together, our benchmark and method demonstrate that exploiting hardware-induced lens motion offers a practical and accurate solution to monocular depth estimation, particularly in scenarios where conventional stereo or motion cues are unavailable. Our contributions are threefold:

- We propose MODepth, an OIS-based multi-frame monocular depth estimation framework with a two-stage pipeline: supervised pretraining on synthetic data and self-supervised fine-tuning on real OIS image pairs.
- We build a real-world dataset of 1,100 indoor OIS image pairs with ground-truth depth maps for accurate benchmarking.
- Our hybrid training strategy achieves an RMSE of 0.439, surpassing several fully supervised methods without requiring ground-truth labels during fine-tuning.
- MODepth establishes a strong benchmark for high-precision depth estimation under stationary camera settings enabled by OIS-induced parallax.

2 Background and Related Work

2.1 Monocular Depth Estimation

Supervised Monocular Depth Estimation: With the advent of deep learning, numerous supervised methods have been proposed to improve estimation accuracy and generalization. Eigen et al.[Eigen and Fergus 2015] introduced multi-scale networks to capture global-to-local structures. Laina et al.[Laina et al. 2016] employed Fully Convolutional Residual Networks (FCRNs) to refine prediction resolution. Subsequent works extended this line using dual-stream architectures [Li et al. 2017], multi-scale feature fusion [Hu et al. 2019], and encoder-decoder frameworks [Fang et al. 2020]. To better represent the continuous nature of depth, DORN [Fu et al. 2018] proposed a Spacing-Increasing Discretization (SID) strategy, while VNL [Yin et al. 2019] introduced virtual normal loss as geometric supervision. More recently, AdaBins [Bhat et al. 2021] adaptively discretized depth ranges, and DPT [Ranftl et al. 2021] utilized Vision Transformers (ViTs) to enhance global context aggregation.

Self-supervised Monocular Depth Estimation: Self-supervised methods often optimize a photometric consistency loss between adjacent frames, sometimes enhanced by masking or regularization strategies. Godard et al.[Godard et al. 2019] proposed a minimum reprojection loss and auto-masking to handle occlusions. Zhou et al.[Zhou et al. 2019] introduced a sparse-to-dense optical flow network to improve depth-flow consistency. Other works enhance geometric reasoning through patch-level correspondence [Yu et al. 2020], two-view triangulation and optical flow matching [Zhao et al. 2020], or auto-rectification modules [Bian et al. 2021a].

Further refinements include depth factorization and residual pose estimation [Ji et al. 2021], consistency constraints in planar and linear structures [Jiang et al. 2021], and Manhattan-world assumptions such as coplanar and normal vector constraints [Li et al. 2021]. Some methods leverage classical structure-from-motion (SfM) priors [Zhao et al. 2023] to guide learning without ground truth.

2.2 OIS Control via Acoustic Injection

Optical Image Stabilization (OIS) is a common hardware module integrated into modern smartphone and camera lenses to reduce motion blur. As shown in Fig. 1(a), by physically shifting internal lens elements to counteract camera shake, OIS provides stabilized imaging without modifying the position of the image sensor. Traditionally, OIS is passively driven by internal gyroscopic feedback to compensate for hand tremors in real time.

Recent research has revealed that the OIS mechanism can also be actively controlled via external acoustic stimulation. Specifically, studies such as OISSR and DoCam [Pan et al. 2022a,b] demonstrate that injecting high-frequency sound signals can perturb the readings of three-axis MEMS gyroscopes, which serve as the core sensors driving OIS behavior. By emitting sine wave signals near the gyroscope's resonance frequency—typically in the 18–30 kHz range, which is inaudible to humans and considered biologically safe [Gao et al. 2020]—the sensed angular velocity can be artificially manipulated. As a result, the OIS actuator interprets these perturbed signals as motion and correspondingly drives the lens to move in a stable and repeatable pattern, while the CMOS image sensor and the device body remain static.

This phenomenon enables a novel form of internal lens actuation without mechanical intervention or external calibration, effectively producing structured intra-camera motion. Such controlled oscillation introduces predictable and repeatable parallax between captured frames. Since only the lens group moves while the sensor remains fixed, this method yields optical parallax akin to that generated by real stereo or ego-motion setups—yet with much simpler hardware constraints. Our work leverages this principle to create OIS-driven image pairs for depth estimation, forming a geometry-consistent, scalable supervision signal in indoor environments.

3 System Design

3.1 SfOLM: Structure from OIS-controlled Lens Motion

In classical Structure-from-Motion (SfM) systems, depth supervision is obtained from a set of temporally adjacent frames captured under different camera poses. Given the camera intrinsic matrix $\mathbf{K}$, the known relative pose $T_{r\to t} = [\mathbf{R}|\mathbf{t}]$ between reference frame I_r and target frame I_t, and the depth Z_{ij} at pixel $\mathbf{p}^r_{ij}$ in the reference frame, its corresponding pixel $\mathbf{p}^t_{ij}$ in the target frame can be computed as:

$$\mathbf{p}^t_{ij} \sim K T_{r\to t} Z_{ij} K^{-1} \mathbf{p}^r_{ij} \tag{1}$$

Unlike conventional SfM methods that rely on large-scale camera movements, SfOLM investigates whether the subtle motions induced by internal lens shifts in optical image stabilization (OIS) can serve as supervisory signals for monocular depth learning. Under OIS control, where lens motion is internally actuated, the projection geometry governed by camera intrinsics and extrinsics remains valid. However, the relative transformations between frames are no longer caused by global camera motion. Although the physical camera remains stationary—i.e., $[R_{real}|t_{real}] = [\mathcal{I}|\vec{0}]$, OIS introduces micro-scale disturbances through slight displacements and rotations of internal lens elements. These disturbances can be equivalently modeled as a virtual camera motion with a small transformation $[R_{OIS}|t_{OIS}]$. Therefore, the final relative pose $T_{r\to t} = [R_{OIS}|t_{OIS}]$. In

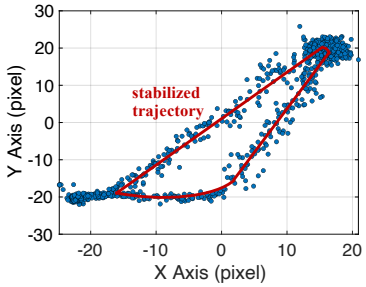


Fig. 2. Principal Point Shift Trajectory under OIS-Controlled Motion.

addition, lens-only motion induced by OIS also leads to shifts in the principal point, effectively altering the camera intrinsics. If the reference frame adopts the original intrinsic matrix $K_r = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix}$, then under the influence of OIS, the target frame is associated with a perturbed intrinsic matrix $K_t = \begin{bmatrix} f_x & 0 & c_x + \delta c_x \\ 0 & f_y & c_y + \delta c_y \\ 0 & 0 & 1 \end{bmatrix}$. Therefore, under the structure from OIS-controlled lens motion (SFOLM), the pixel reprojection relationship between frames becomes:

$$\mathbf{p}_{ij}^t \sim K_t T_{r\to t} Z_{ij} K_r^{-1} \mathbf{p}_{ij}^r \tag{2}$$

A view synthesis loss can be employed to supervise depth estimation by enforcing photometric consistency between the reference frame and the reprojected target frame:

$$\mathcal{L}_{vs} = \Psi(\tilde{I}_t, I_r) \tag{3}$$

$$\tilde{I}_t = proj(I_t, K_r, Z_r, T_{r\to t}, K_t) \tag{4}$$

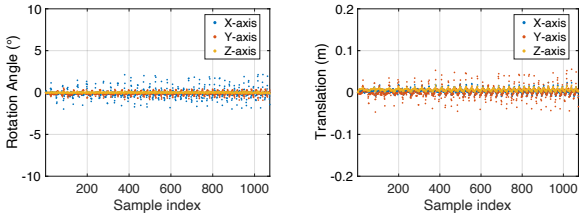
where Ψ denotes the photometric reconstruction loss, formulated as a weighted combination of pixel-wise l_1 distance and structural similarity (SSIM) [Wang et al. 2004], to jointly capture low-level intensity differences and perceptual structural alignment:

$$\Psi(\tilde{I}_t, I_r) = (1-\alpha)||\tilde{I}_t - I_t||_1 + \frac{\alpha}{2}(1 - SSIM(\tilde{I}_t, I_t)) \tag{5}$$

where α is a hyperparameter and defaults to 0.85. The function $proj(\cdot)$ constructs a sampling grid based on the reference frame's camera intrinsics K_r, the estimated depth map Z_r, the relative pose transformation $T_{r\to t}$, and the target frame's intrinsics K_t. This grid is then used to perform differentiable resampling of the input image. According to Eqs. 3 and 4, minimizing the view synthesis loss requires not only accurate depth estimation, but also precise estimation of the camera intrinsics and the relative pose between frames.

3.2 Intrinsic and Extrinsic Parameter Variability Under SfOLM

In the working principle of optical image stabilization (OIS) [Cardani 2006], the lens is physically shifted within a limited range to compensate for unintended hand tremors or device vibrations. Due to mechanical constraints and design tolerances, the lens displacement is inherently bounded—usually within ±0.5 to ±1 mm in consumer-grade camera modules.



(a) Rotation Angles Variation (b) Translation Vector Variation

Fig. 3. Analysis of Camera Pose Changes under OIS-Controlled Motion.

As described in Section 2.2, injecting a cosine acoustic signal perturbs the IMU module and induces controlled, predictable lens motion via the OIS actuator. Relative to the static reference frame, this motion results in limited and stable perturbations to the camera's intrinsic and extrinsic parameters.

To empirically validate the range of intrinsic and extrinsic parameter variations induced by OIS, we conducted a controlled preliminary experiment using a smartphone camera and a standard checkerboard calibration target. The physical distance d between the camera and the checkerboard was pre-measured and fixed throughout the experiment. We first captured a reference image of the calibration board with the camera held stationary, ensuring that no OIS actuation was present. Next, we activated the smartphone's internal speaker and injected a sinusoidal acoustic signal, which coupled with the IMU module and triggered periodic OIS lens motion. While this induced regular micro-movements of the internal lens group, the camera body remained physically static. During this period, we continuously captured 1000 frames under OIS actuation, which served as the target frames for our analysis.

For both the reference frame and each of the 1000 target frames, we first detect the 2D corner positions of the calibration board using a standard checkerboard detection algorithm. Let $\mathbf{C}_{ref}$ and $\mathbf{C}_{t_j}$ denote the detected corner sets in the reference frame and the j-th target frame, respectively, where each $\mathbf{C} = \{p^i = (x_i, y_i) | i \in [1, N]\}$ contains N ordered 2D image coordinates of checkerboard corners.

Given the detected 2D checkerboard corners $\mathbf{C}_{ref}$ and $\mathbf{C}_{t_j}$, and the known depth d at each corner location, we formulate an optimization framework based on the proposed SfOLM formulation to estimate the camera intrinsic matrix $\mathbf{K}_{t_j}$ and extrinsic pose $[\mathbf{R}_{t_j}|\mathbf{t}_{t_j}]$ for each target frame:

$$\underset{K_{t_j}, T_{ref\to t_j}}{argmin} \sum_{i=0}^{N} ||p_{ref}^i - \pi(K_{t_j} T_{ref\to t_j} d K_{ref}^{-1} p_{t_j}^i)||_2^2 \tag{6}$$

where $K_{t_j} = \begin{bmatrix} f_x & 0 & c_x + \delta c_{t_j}^x \\ 0 & f_y & c_y + \delta c_{t_j}^y \\ 0 & 0 & 1 \end{bmatrix}$, $K_{ref} = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix}$ and $\pi(\cdot)$ denotes the standard perspective projection: $\pi([x, y, z]^T) = [x/z, y/z]^T$. Moreover, (f_x, f_y) denote the focal lengths, and (c_x, c_y) the nominal principal point coordinates which can all be obtained through standard camera calibration procedures using the checkerboard pattern. And the terms $\delta c_{t_j}^x$ and $\delta c_{t_j}^y$ represent deviations of the principal point caused by dynamic lens shifts under OIS actuation.

Based on the physical constraints of OIS-controlled lens motion, we impose prior bounds on both the intrinsic perturbations and the extrinsic transformations during optimization. Given that the Xiaomi Mi 11 Pro smartphone in our experiments features a primary camera equipped with a CMOS sensor of size 1/1.12 inches and a pixel pitch of 1.4μm, such lens displacements can result in principal point shifts of up to 35–40 pixels. Accordingly, we constrain the principal point deviations as follows:

$$\delta c_{t_j}^x, \delta c_{t_j}^y \in [-\epsilon_p, \epsilon_p], with\ \epsilon_p \approx 40\ pixels \tag{7}$$

Similarly, considering the mechanical limits of the OIS actuator, we constrain the rotational and translational components of the extrinsic transformation $T_{ref\to t_j} = [\mathbf{R}_{t_j} | \mathbf{t}_{t_j}] = [R_z^{t_j}(\psi) R_y^{t_j}(\theta) R_z^{t_j}(\phi) | \mathbf{t}_{t_j}]$ (ψ, θ, ϕ are the rotation angles around Z, Y, X-axes) as follows:

$$-\epsilon_r \le \psi, \theta, \phi \le \epsilon_r\ and\ ||t_{t_j}|| < \epsilon_t,\ with\ \epsilon_r \approx 1^\circ, \epsilon_t \approx 2cm \tag{8}$$

These constraints serve as physically informed priors in our optimization framework, ensuring that the estimated intrinsic and extrinsic parameters remain within the feasible operating range dictated by the hardware characteristics of the Xiaomi Mi 11 Pro's OIS mechanism.

Subsequently, we obtain the optimized results as illustrated in Fig. 2 and Fig. 3. It can be observed that the displacement of the lens principal point exhibits a consistent and structured pattern. Meanwhile, the relative pose matrices between target frames and reference frame also demonstrate regularity and coherence, indicating that the lens motion is highly modelable.

To enable accurate depth estimation, we jointly optimize camera intrinsics, relative poses, and depth predictions using view synthesis loss. Benefiting from the inherently consistent scale of our camera setup, the predicted depths align well with real-world metrics, eliminating the need for median scaling commonly used in SfM-based methods [Godard et al. 2019; Yu et al. 2020; Zhao et al. 2023].

3.3 Depth Estimation Network

In this section, we present the proposed depth estimation network, MODNet. The network takes as input a pair of RGB images: a reference frame captured with the lens in a static state, and a target frame captured under lens motion induced by optical image stabilization (OIS). The network outputs a dense depth map corresponding to the reference frame, representing the scene geometry from the reference viewpoint. We construct our depth estimation network using the Croco-based [Weinzaepfel et al. 2022, 2023] framework, which is designed based on a ViT [Alexey 2020] (Vision Transformer) backbone. This framework consists of an encoder, a decoder, and an output head.

Details of Encoder. We denote the shared-weight encoder as $\mathcal{E}$, which is based on a Vision Transformer (ViT)[Alexey 2020]. It encodes both input images $I \in \mathbb{R}^{3\times H\times W}$ into patch-level features $\mathcal{E}(I) \in \mathbb{R}^{N\times C_P}$ (where $N = \frac{H}{16}\times\frac{W}{16}$ and C_p is set to 1024). Notably, we replace the standard sinusoidal positional embedding with Rotary Positional Embedding (RoPE)[Su et al. 2023] to inject positional information, which has shown improved performance in modeling spatial relationships.

Details of Decoder. As illustrated in Figure 4, the attention module in the decoder consists of three main components: multi-head self-attention, multi-head cross-attention, and a multi-layer perceptron (MLP). These components collaboratively enable effective fusion of features from the two input frames.

Details of Depth estimation head module. As shown in Fig. 4, our Head module takes the encoder output $\mathcal{E}(P_1)$ and the intermediate features (D_1, D_2, D_3) from selected attention-based blocks in the decoder as inputs to generate the depth map for the reference frame. These patch embedding features are first reconstructed into image-like representations via a PRCNN module. Subsequently, they are fused using an attention-based feature fusion block (AttentionFF, as shown in Fig. 6). Finally, a lightweight head block produces the final dense depth prediction. For more architectural details of the depth estimation network, please refer to the Appendix.

3.4 Principal Point Offset Estimation Module

To explicitly model the principal point shifts induced by OIS-controlled lens motion, we introduce a raft-based [Teed and Deng 2020] principal point offset estimation Module that predicts the displacement of the imaging center $(\delta c_{t_j}^x, \delta c_{t_j}^y)$ in the target frame $I_{t_j} \in \mathcal{R}^{3\times H\times W}$ relative to the reference frame $I_r \in \mathcal{R}^{3\times H\times W}$. The RAFT-based module is first employed to predict the disparity of the target frame I_{t_j} relative to the reference frame I_r. As shown in Fig. 7, a convolutional encoder is used to extract dense features $E(I) \in \mathcal{R}^{L\times\frac{H}{8}\times\frac{W}{8}}$ from the input image. It consists of six blocks: two at $\frac{1}{2}$, two at $\frac{1}{4}$, and two at $\frac{1}{8}$ resolution. This hierarchical structure captures both local details and global context for downstream depth estimation. Then, a comprehensive correlation volume for the pairs is generated to assess the visual similarity:

$$\begin{aligned} &C(E(I_r), E(I_{t_j})) \in \mathcal{R}^{\frac{H}{8}\times\frac{W}{8}\times\frac{H}{8}\times\frac{W}{8}} \\ &C_{h_r,w_r,h_{t_j},w_{t_j}} = \sum_{l}^{L} E(I_r)_{l,h_r,w_r} \cdot E(I_{t_j})_{l,h_{t_j},w_{t_j}}, \end{aligned} \tag{9}$$

Next, a correction module iteratively updates the initial disparity $\mathbf{di}_{init}$ (initialized as zero) into a refined disparity $\mathbf{di}_{vis}$ by integrating visual features $C(E(I_r), E(I_{t_j}))$ and $E(I_r)$ respectively. To remove invalid disparity values near image boundaries, a mask M is applied to obtain the masked disparity $\hat{\mathbf{di}}_{vis} = M \odot \mathbf{di}_{vis}$. We then compute the average disparity $mean(\hat{\mathbf{di}}_{vis}) \in \mathcal{R}^2$ along the x and y axes. Finally, this mean displacement is passed through a ResNet-style MLP block to regress the predicted principal point offset $(\delta c_{t_j}^x, \delta c_{t_j}^y)$.

3.5 Pose Estimation Network

We employ a pose estimation network to estimate the camera pose between two input frames. Similar to works [Fan et al. 2023; Guo et al. 2018; Kuznietsov et al. 2017; Zhao et al. 2023], our PoseNet is based on the widely-used U-Net architecture [Ronneberger et al. 2015], which consists of an encoder-decoder network with skip connections. This structure allows the network to capture both high-level semantic information and low-level details, essential for accurately estimating the relative pose between images. We use ResNet18 [He et al. 2016] as the encoder, which consists of 11 million parameters and has been pre-trained on ImageNet [Deng et al. 2009].

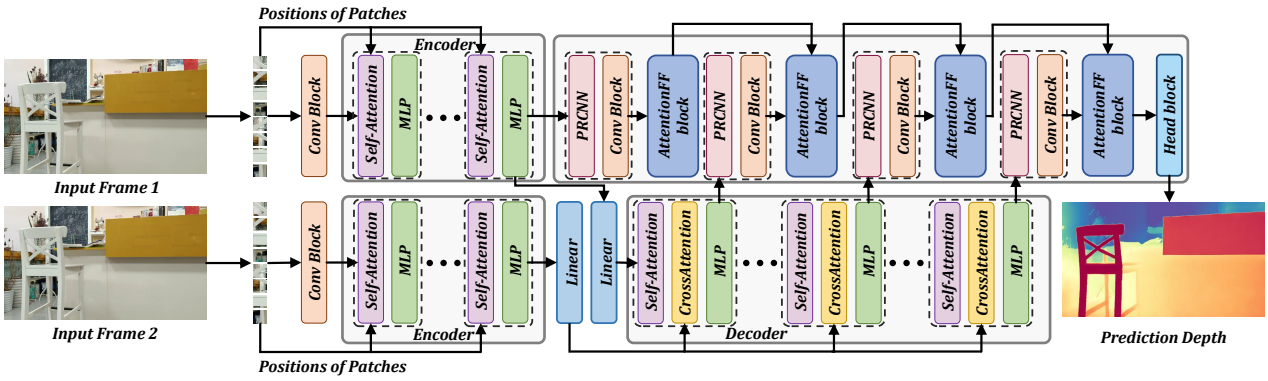


Fig. 4. Architecture of Our Depth Estimation Network MODNet.

3.6 Pre-training on simulated datasets

We adopt the self-supervised pretraining strategy from CroCo v2 [Weinzaepfel et al. 2023] to initialize our encoder and decoder. Using the original CroCo head [Weinzaepfel et al. 2022], the model learns to reconstruct the reference frame from partially masked inputs across both views, encouraging spatial correspondence and 3D-aware feature learning without explicit depth supervision.

To effectively pretrain the network under micro-parallax conditions, we construct a synthetic dataset by simulating OIS-induced lens motion as virtual extrinsic perturbations. By matching focal length and FOV to our Xiaomi 11 Pro setup, we generate stereo-like image pairs with realistic parallax. The simulator also provides clean, high-resolution ground-truth depth maps, offering high-quality supervision unavailable from noisy real-world sensors.

Once the image pairs and corresponding depth maps are obtained, the network is trained using a weighted combination of three loss terms: 1) Mean Squared Error (MSE) Loss ensures overall depth accuracy by minimizing pixel-wise squared differences:

$$\mathcal{L}_{mse} = \frac{1}{N}\sum_{i=1}^{N}(Z_i - \hat{Z}_i)^2 \tag{10}$$

Where $N = H \times W$ is the total number of pixels. 2) Edge-Aware Smoothness Loss encourages spatial smoothness while preserving edges, modulated by image gradients:

$$\mathcal{L}_{smooth} = \sum_{i,j}\left(\left|\partial_x \hat{Z}_{i,j}\right| e^{-\left|\partial_x I_{i,j}\right|} + \left|\partial_y \hat{Z}_{i,j}\right| e^{-\left|\partial_y I_{i,j}\right|}\right) \tag{11}$$

where ∂_x and ∂_y represent horizontal and vertical gradients of the predicted depth map. 3) Gradient Loss enforces consistency between predicted and ground truth depth gradients to preserve structural details:

$$\mathcal{L}_{grad} = \sum_{i,j}\left|\partial_x \hat{Z}_{i,j} - \partial_x Z_{i,j}\right| + \left|\partial_y \hat{Z}_{i,j} - \partial_y Z_{i,j}\right| \tag{12}$$

The final training objective is:

$$\mathcal{L}_{total} = \lambda_1 \mathcal{L}_{mse} + \lambda_2 \mathcal{L}_{smooth} + \lambda_3 \mathcal{L}_{grad} \tag{13}$$

Where λ_1, λ_2 and λ_3 are weighting factors that balance the influence of each loss component and are set to $1, 0.1, 0.1$ by default.

3.7 Self-Supervised training on real datasets

Given the practical challenges of collecting large-scale paired RGB-D data in real-world settings, we leverage the fact that capturing RGB image pairs with micro-parallax using OIS-controlled lens motion is significantly more feasible. Using a single handheld smartphone, we can efficiently record large volumes of OIS-induced image pairs without requiring specialized hardware or depth sensors. Therefore, we adopt a self-supervised training framework based on SfOLM to fine-tune the pretrained model on real-world OIS RGB data. This approach enables the model to adapt to real image distributions while still benefiting from the 3D geometric priors learned during synthetic pretraining.

Given a reference frame I_r and a target frame I_{t_j}, our model predicts dense depth maps Z_r and Z_{t_j} for both images using the depth estimation network. Simultaneously, the principal point offset module estimates the target frame's offset $(\delta c_{t_j}^x, \delta c_{t_j}^y)$ relative to the reference frame, and a pose estimation network regresses the relative camera pose $T_{r\to t_j}$ between the two frames. These outputs are jointly optimized by minimizing a set of self-supervised losses as described below. 1) Photometric Loss $\mathcal{L}_{vs}$. As described in Sec. 3.1, we implement the photometric reconstruction loss based on Eq. 3, 4, and 5, which model the inter-frame reprojection under OIS-controlled motion and visibility constraints. 2) Geometry Consistency Loss L_{gc}. To further improve prediction accuracy, we impose a geometric consistency constraint between the predicted depth maps of the reference and target frames. Specifically, we require that the depth predictions Z_r and Z_{t_j} describe the same underlying 3D scene structure and minimize their mutual discrepancy. Using the inter-frame projection relationship defined in Eq. 1, we first warp the target frame's depth map Z_{t_j} into the reference view to obtain a reconstructed depth $\hat{Z}_{t_j}$. The geometric consistency loss is then defined as:

$$\mathcal{L}_{gc} = \frac{|Z_r - \hat{Z}_{t_j}|}{Z_r + \hat{Z}_{t_j}} \tag{14}$$

3) Edge-Aware Smoothness Loss $\mathcal{L}_{smooth}$. Consistent with the pretraining stage, we also adopt the smoothness loss defined in Equation 11, which encourages spatial continuity in the predicted depth while preserving edge details aligned with the RGB image structure. 4) Inverse Loss $\mathcal{L}_{in}$. While both the photometric loss and geometric

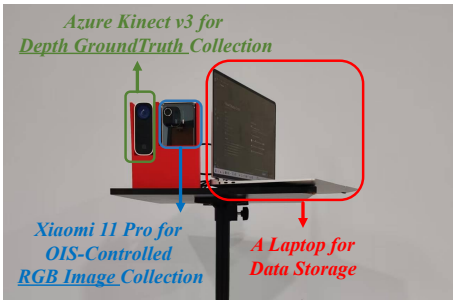


Fig. 5. Experiment setup: Xiaomi 11 Pro and Azure Kinect v3.

consistency loss project the target frame onto the reference frame, the inverse loss $\mathcal{L}_{in} = \mu_1 \mathcal{L}'_{gc} + \mu_2 \mathcal{L}'_{vs}$ measures the discrepancy by projecting the reference frame onto the target frame (μ_1 and μ_2 are hyper-parameters that balance the loss.). This bidirectional formulation enhances consistency and regularizes the depth prediction across both views. Thus, the overall objective of our self-supervised training framework is as follows:

$$\mathcal{L}_{all} = \theta_1 \mathcal{L}_{vs} + \theta_2 \mathcal{L}_{gc} + \theta_3 \mathcal{L}_{smooth} + \theta_4 \mathcal{L}_{in} \tag{15}$$

Where θ_1,θ_2, θ_3, and θ_4 are weighting factors that balance the influence of each loss component.

4 Evalution

4.1 Implementation Details

We use a Xiaomi 11 Pro smartphone camera to capture RGB images, while injecting an acoustic signal at approximately 20,150 Hz to induce regular lens motion via its OIS module. To obtain ground-truth depth, we additionally employ a Kinect v3 depth sensor to record aligned RGB-D data. Moreover, we implement our network using the PyTorch [Paszke et al. 2019] framework and train it using the AdamW [Loshchilov and Hutter 2017] optimizer for efficient and stable optimization. And our model is trained on an NVIDIA A800 GPU with 80GB memory. Additionally, we set the input image and output disparity map size of the depth estimation network to 480×352.

4.2 Datasets and Metrics

Simulated Datasets. We construct a large-scale synthetic dataset using the Habitat simulator [Savva et al. 2019], rendering RGB and depth image pairs at 720p resolution. A total of 81,600 image pairs are collected from 816 diverse indoor scenes drawn from HM3D [Ramakrishnan et al. 2021], ScanNet [Dai et al. 2017], Replica [Straub et al. 2019], and ReplicaCAD [Szot et al. 2021]. Each pair consists of a reference frame and a target frame with aligned RGB and depth information. For each scene, 100 pairs are randomly sampled to ensure diversity. The resulting dataset is split into training, validation, and test sets in an 8:1:1 ratio for use in the pretraining stage.

Real-world Datasets. To construct our dataset, we use a Kinect v3 depth sensor to capture high-quality RGB-D image pairs with accurate depth ground truth. Specifically, as shown in Fig. 5, we mount a Xiaomi 11 Pro smartphone and the Kinect v3 on a custom rig with a fixed relative pose. We then perform extrinsic calibration between the two devices, allowing the depth maps obtained from the Kinect to be projected onto the RGB frame of the smartphone. For each sample, we first record a reference frame using the smartphone while the OIS module remains inactive, alongside the corresponding RGB-D pair from the Kinect v3. The projected depth map onto the reference frame serves as the ground truth. We then activate the built-in speaker to emit a cosine signal at approximately 20,150 Hz, which triggers periodic lens motion via the OIS module, and capture a target frame during this induced motion. This completes one RGB image pair with motion-induced parallax and its associated depth supervision. In total, we collect 1,100 such samples across diverse real-world indoor scenes for evaluating depth estimation performance.

Metrics. Following [Fan et al. 2023; Wu et al. 2022; Zhao et al. 2023], we use standard metrics, including error-based metrics: Mean Absolute Relative Error (Abs Rel), Mean Log10 Error (Log10), and Root Mean Squared Error (RMSE). For accuracy-based metrics, we compute the percentage of pixels $max(\frac{Z_p}{Z_g}, \frac{Z_g}{Z_p}) = \delta < threshold$, where $threshold \in [1.25^1, 1.25^2, 1.25^3]$ and where Z_p and Z_g represent the predicted and ground truth depth map.

4.3 Ablation Studies

Methods	Error Metric ↓			Accuracy Metric (%) ↑		
	Abs Rel ↓	Log10 ↓	RMSE ↓	δ_1 ↑	δ_2 ↑	δ_3 ↑
case (a)	0.247	0.102	0.788	61.1	84.2	93.5
case (b)	0.232	0.098	0.784	59.8	87.1	95.9
case (c)	0.224	0.094	0.721	64.5	88.4	96.5
case (d)	0.215	0.085	0.640	68.5	88.7	96.3
case (e)	0.183	0.068	0.523	77.4	92,8	97.9
case (f)	0.178	0.069	0.458	76.3	**93.3**	98.1
case (g)	**0.165**	**0.065**	**0.439**	**79.4**	93.2	**98.5**

Table 1. Ablation Studies of MODepth on MOData.

To assess the contributions of different components in MODepth, we conduct an ablation study across nine configurations (a–g), focusing on supervised pretraining, self-supervised fine-tuning, and the impact of individual loss terms. Cases (a–c) examine models trained without simulated data pretraining, progressively incorporating photometric loss $\mathcal{L}_{vs}$ and smoothness regularization (a), geometric consistency loss $\mathcal{L}_{gc}$ (b), and inverse projection loss $\mathcal{L}_{in}$(c). Case (d) evaluates the effect of supervised pretraining alone. Cases (e–g) investigate models that are first pretrained on simulated OIS-style data and then fine-tuned using self-supervised objectives: (e) using $\mathcal{L}_{vs}$ and $\mathcal{L}_{smooth}$, (f) adding geometric consistency loss $\mathcal{L}_{gc}$, and (g) further incorporating inverse loss $\mathcal{L}_{in}$, which corresponds to our final model. This study demonstrates that combining synthetic pretraining with comprehensive bidirectional self-supervision significantly enhances depth prediction performance in indoor scenes.

As shown in Tab. 1, our ablation study reveals several key insights. First, comparing cases (a) to (c), we observe that progressively adding geometric consistency loss and reverse loss improves both accuracy and error metrics, with the reverse projection loss in (c) reducing Abs Rel from 0.247 to 0.224 and increasing δ_1 from 61.1% to 64.5%, highlighting its complementary effect in enforcing bidirectional

Methods	Error Metric ↓			Accuracy Metric (%) ↑		
	Abs Rel ↓	Log10 ↓	RMSE ↓	δ_1 ↑	δ_2 ↑	δ_3 ↑
AdaBins [Bhat et al. 2021]	0.216	0.096	0.704	61.9	87.5	96.5
DPT [Ranftl et al. 2021]	0.175	0.074	0.516	75.3	93.7	96.8
idisc [Piccinelli et al. 2023]	**0.151**	**0.065**	0.479	76.8	**95.6**	**98.9**
MODepth(Ours)	0.165	**0.065**	**0.439**	**79.4**	93.2	98.5

Table 2. Quantitative comparison of our MODepth to other SOTA supervised monocular depth estimation methods on MODdata.

consistency. Case (d), which only uses supervised pretraining without self-supervised fine-tuning, achieves better performance than (a–c), demonstrating the importance of strong inductive priors from synthetic data. However, the best results are obtained in cases (e–g), where self-supervised fine-tuning is applied on top of supervised pretraining. Notably, case (g), which integrates photometric, geometric, and reverse losses, achieves the lowest error (Abs Rel = 0.165) and the highest accuracy (δ_1 = 79.4%), confirming the efficacy of our full pipeline. This progression demonstrates that the combination of synthetic pretraining and comprehensive self-supervised objectives is crucial for high-quality indoor depth estimation.

4.4 Comprehensive Comparison

We compare MODepth with several state-of-the-art supervised monocular depth estimation models, including AdaBins [Bhat et al. 2021], DPT [Ranftl et al. 2021], and idisc [Piccinelli et al. 2023]. To ensure fairness, all baseline models are fine-tuned on our OIS-based dataset using their official pre-trained weights and default settings. As shown in Tab. 2 and Fig. 8, MODepth achieves highly competitive performance. It reports the lowest RMSE (0.439) and the highest δ_1 accuracy (79.4%), demonstrating its superiority in estimating geometrically consistent and high-precision depth, particularly in near-range indoor scenes. While iDisc achieves slightly better Abs Rel (0.151 vs. 0.165), our method performs better in key accuracy metrics and produces more stable overall results. Remarkably, MODepth achieves these results without relying on any ground-truth depth supervision during fine-tuning. Instead, it benefits from a hybrid training strategy that combines synthetic-data-based supervised pretraining with self-supervised fine-tuning on OIS-induced image pairs. This shows that our approach not only matches but in some cases outperforms fully supervised methods, highlighting the effectiveness of hardware-induced parallax as a natural and scalable supervisory signal for depth estimation in indoor environments.

4.5 Impact of Self-Supervision Framework

We evaluate the impact of our self-supervised framework on the final depth estimation accuracy by comparing it with alternative self-supervision strategies. To ensure fairness, all methods are pretrained on the same synthetic dataset using supervised learning, followed by fine-tuning on real-world OIS RGB data using their respective self-supervised pipelines. As shown in Tab. 3, methods such as MonoDepth [Godard et al. 2019], IndoorDepth [Fan et al. 2023], and GASMono [Zhao et al. 2023]—which are primarily designed around structure-from-motion (SfM)-based modeling—achieve significantly inferior results under our OIS-based setting. These approaches typically assume large ego-motion or stereo baselines, which are not present in our micro-parallax data. In contrast, our method leverages the SfOLM framework tailored to OIS-induced image pairs and achieves substantially better performance across all metrics, including a notably low RMSE of 0.439 and high δ_1 accuracy of 79.4%. These results demonstrate the importance of designing self-supervision objectives aligned with the physical characteristics of the data, and validate the effectiveness of our geometry-aware training framework in micro-parallax scenarios.

Methods	Error Metric ↓			Accuracy Metric (%) ↑		
	Abs Rel ↓	Log10 ↓	RMSE ↓	δ_1 ↑	δ_2 ↑	δ_3 ↑
MonoDepth [Godard et al. 2019]	0.609	0.193	1.611	14.1	60.1	84.7
IndoorDepth [Fan et al. 2023]	0.869	0.247	2.404	22.6	46.2	66.1
Gasmono [Zhao et al. 2023]	1.086	0.286	2.233	14.8	32.6	54.1
MODepth(Ours)	**0.165**	**0.065**	**0.439**	**79.4**	**93.2**	**98.5**

Table 3. Evaluation on MODdata of different self-supervised monocular depth estimation methods.

Methods	Error Metric ↓			Accuracy Metric (%) ↑		
	Abs Rel ↓	Log10 ↓	RMSE ↓	δ_1 ↑	δ_2 ↑	δ_3 ↑
RAFT-stereo [Lipson et al. 2021]	1.004	0.213	1.777	42.6	65.5	77.8
LEAStereo [Cheng et al. 2020]	0.967	0.279	2.461	3.24	16.7	61.8
MODepth(Ours)	**0.165**	**0.065**	**0.439**	**79.4**	93.2	**98.5**

Table 4. Performance Comparison with Stereo Matching Methods on OIS Image Pairs

4.6 Evaluation of Using Stereo Algorithms

We further evaluate the applicability of existing stereo matching algorithms on our dataset, using RAFT-Stereo [Lipson et al. 2021] and LEAStereo [Cheng et al. 2020] as representative baselines. As shown in Tab. 4, both models are applied to our OIS-induced image pairs without modification. Due to the absence of known baseline distances between the reference and target frames, we cannot directly convert disparity to metric depth. Instead, we adopt the median scaling strategy: disparity is inverted to approximate depth and then scaled by the median ratio for evaluation. The results demonstrate a significant performance gap between stereo methods and our approach. This is primarily because lens-induced motion in OIS not only alters the camera's relative pose but also introduces small but non-negligible changes to the intrinsic parameters—particularly principal point shifts. These deviations violate the assumptions of stereo algorithms, which typically rely on fixed intrinsics and rectified epipolar geometry. As a result, existing stereo methods struggle to produce reliable depth estimates in our setting, confirming that they are not well-suited for OIS-induced micro-parallax data.

5 Conclusion

We present MODepth, a monocular depth estimation framework that leverages OIS-controlled lens motion via acoustic injection to enable stable inter-frame parallax. By incorporating pose and principal point offset estimation, our network MODNet effectively utilizes multi-frame geometric cues. Evaluated on the MODdata dataset, our method outperforms existing monocular approaches and even fully supervised baselines, achieving an RMSE of 0.439, demonstrating the strength of OIS-driven self-supervision.

References

Dosovitskiy Alexey. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv: 2010.11929* (2020).

Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. 2021. Adabins: Depth estimation using adaptive bins. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 4009–4018.

Jia-Wang Bian, Huangying Zhan, Naiyan Wang, Tat-Jun Chin, Chunhua Shen, and Ian Reid. 2021a. Auto-rectify network for unsupervised indoor depth estimation. *IEEE transactions on pattern analysis and machine intelligence* 44, 12 (2021), 9802–9813.

Jia-Wang Bian, Huangying Zhan, Naiyan Wang, Zhichao Li, Le Zhang, Chunhua Shen, Ming-Ming Cheng, and Ian Reid. 2021b. Unsupervised scale-consistent depth learning from video. *International Journal of Computer Vision* 129, 9 (2021), 2548–2564.

Brent Cardani. 2006. Optical image stabilization for digital cameras. *IEEE Control Systems Magazine* 26, 2 (2006), 21–22.

Xuelian Cheng, Yiran Zhong, Mehrtash Harandi, Yuchao Dai, Xiaojun Chang, Hongdong Li, Tom Drummond, and Zongyuan Ge. 2020. Hierarchical neural architecture search for deep stereo matching. *Advances in neural information processing systems* 33 (2020), 22158–22169.

Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. 2017. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5828–5839.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 248–255.

David Eigen and Rob Fergus. 2015. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *Proceedings of the IEEE international conference on computer vision*. 2650–2658.

Chao Fan, Zhenyu Yin, Yue Li, and Feiqing Zhang. 2023. Deeper into Self-Supervised Monocular Indoor Depth Estimation. *arXiv preprint arXiv:2312.01283* (2023).

Zhicheng Fang, Xiaoran Chen, Yuhua Chen, and Luc Van Gool. 2020. Towards good practice for CNN-based monocular depth estimation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*. 1091–1100.

Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. 2018. Deep ordinal regression network for monocular depth estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2002–2011.

Ming Gao, Feng Lin, Weiye Xu, Muertikepu Nuermaimaiti, Jinsong Han, Wenyao Xu, and Kui Ren. 2020. Deaf-aid: mobile IoT communication exploiting stealthy speaker-to-gyroscope channel. In *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking*. 1–13.

Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow. 2019. Digging into self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*. 3828–3838.

Xiaoyang Guo, Hongsheng Li, Shuai Yi, Jimmy Ren, and Xiaogang Wang. 2018. Learning monocular depth by distilling cross-domain stereo networks. In *Proceedings of the European conference on computer vision (ECCV)*. 484–500.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

Junjie Hu, Mete Ozay, Yan Zhang, and Takayuki Okatani. 2019. Revisiting single image depth estimation: Toward higher resolution maps with accurate object boundaries. In *2019 IEEE winter conference on applications of computer vision (WACV)*. IEEE, 1043–1051.

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

Hualie Jiang, Laiyan Ding, Junjie Hu, and Rui Huang. 2021. PLNet: Plane and line priors for unsupervised indoor depth estimation. In *2021 International Conference on 3D Vision (3DV)*. IEEE, 741–750.

Yevhen Kuznietsov, Jorg Stuckler, and Bastian Leibe. 2017. Semi-supervised deep learning for monocular depth map prediction. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 6647–6655.

Iro Laina, Christian Rupprecht, Vasileios Belagiannis, Federico Tombari, and Nassir Navab. 2016. Deeper depth prediction with fully convolutional residual networks. In *2016 Fourth international conference on 3D vision (3DV)*. IEEE, 239–248.

Boying Li, Yuan Huang, Zeyu Liu, Danping Zou, and Wenxian Yu. 2021. StructDepth: Leveraging the structural regularities for self-supervised indoor depth estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*. 12663–12673.

Jun Li, Reinhard Klein, and Angela Yao. 2017. A two-streamed network for estimating fine-scaled depth maps from single rgb images. In *Proceedings of the IEEE international conference on computer vision*. 3372–3380.

Lahav Lipson, Zachary Teed, and Jia Deng. 2021. Raft-stereo: Multilevel recurrent field transforms for stereo matching. In *2021 International Conference on 3D Vision (3DV)*. IEEE, 218–227.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101* (2017).

Yu Lu, Dian Ding, Hao Pan, Yongjian Fu, Liyun Zhang, Feitong Tan, Ran Wang, Yi-Chao Chen, Guangtao Xue, and Ju Ren. 2024. M3Cam: Extreme Super-resolution via Multi-Modal Optical Flow for Mobile Cameras. In *Proceedings of the 22nd ACM Conference on Embedded Networked Sensor Systems*. 744–756.

Hao Pan, Feitong Tan, Yi-Chao Chen, Gaoang Huang, Qingyang Li, Wenhao Li, Guangtao Xue, Lili Qiu, and Xiaoyu Ji. 2022a. DoCam: depth sensing with an optical image stabilization supported RGB camera. In *Proceedings of the 28th Annual International Conference on Mobile Computing and Networking*. 405–418.

Hao Pan, Feitong Tan, Wenhao Li, Yi-Chao Chen, and Guangtao Xue. 2022b. OISSR: Optical Image Stabilization Based Super Resolution on Smartphone Cameras. In *Proceedings of the 30th ACM International Conference on Multimedia*. 2978–2986.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems* 32 (2019).

Luigi Piccinelli, Christos Sakaridis, and Fisher Yu. 2023. idisc: Internal discretization for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 21477–21487.

Santhosh K Ramakrishnan, Aaron Gokaslan, Erik Wijmans, Oleksandr Maksymets, Alex Clegg, John Turner, Eric Undersander, Wojciech Galuba, Andrew Westbury, Angel X Chang, et al. 2021. Habitat-matterport 3d dataset (hm3d): 1000 large-scale 3d environments for embodied ai. *arXiv preprint arXiv:2109.08238* (2021).

René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. 2021. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*. 12179–12188.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*. Springer, 234–241.

Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, et al. 2019. Habitat: A platform for embodied ai research. In *Proceedings of the IEEE/CVF international conference on computer vision*. 9339–9347.

Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, et al. 2019. The Replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797* (2019).

J Su, Y Lu, S Pan, A Murtadha, B Wen, and Y Liu Roformer. 2023. Enhanced transformer with rotary position embedding., 2021. *DOI: https://doi. org/10.1016/j. neucom* (2023).

Andrew Szot, Alexander Clegg, Eric Undersander, Erik Wijmans, Yili Zhao, John Turner, Noah Maestre, Mustafa Mukadam, Devendra Singh Chaplot, Oleksandr Maksymets, et al. 2021. Habitat 2.0: Training home assistants to rearrange their habitat. *Advances in neural information processing systems* 34 (2021), 251–266.

Zachary Teed and Jia Deng. 2020. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*. Springer, 402–419.

Timothy Trippel, Ofir Weisse, Wenyuan Xu, Peter Honeyman, and Kevin Fu. 2017. WALNUT: Waging doubt on the integrity of MEMS accelerometers with acoustic injection attacks. In *2017 IEEE European symposium on security and privacy (EuroS&P)*. IEEE, 3–18.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing* 13, 4 (2004), 600–612.

Philippe Weinzaepfel, Vincent Leroy, Thomas Lucas, Romain Brégier, Yohann Cabon, Vaibhav Arora, Leonid Antsfeld, Boris Chidlovskii, Gabriela Csurka, and Jérôme Revaud. 2022. Croco: Self-supervised pre-training for 3d vision tasks by cross-view completion. *Advances in Neural Information Processing Systems* 35 (2022), 3502–3516.

Philippe Weinzaepfel, Thomas Lucas, Vincent Leroy, Yohann Cabon, Vaibhav Arora, Romain Brégier, Gabriela Csurka, Leonid Antsfeld, Boris Chidlovskii, and Jérôme Revaud. 2023. CroCo v2: Improved cross-view completion pre-training for stereo matching and optical flow. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 17969–17980.

Cho-Ying Wu, Jialiang Wang, Michael Hall, Ulrich Neumann, and Shuochen Su. 2022. Toward practical monocular indoor depth estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 3814–3824.

Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. 2018. MVSNet: Depth Inference for Unstructured Multi-view Stereo. *European Conference on Computer Vision (ECCV)* (2018).

Yao Yao, Zixin Luo, Shiwei Li, Tianwei Shen, Tian Fang, and Long Quan. 2019. Recurrent MVSNet for High-resolution Multi-view Stereo Depth Inference. *Computer Vision and Pattern Recognition (CVPR)* (2019).

Wei Yin, Yifan Liu, Chunhua Shen, and Youliang Yan. 2019. Enforcing geometric constraints of virtual normal for depth prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*. 5684–5693.

Zehao Yu, Lei Jin, and Shenghua Gao. 2020. P 2 net: Patch-match and plane-regularization for unsupervised indoor depth estimation. In *European conference on computer vision*. Springer, 206–222.

Chaoqiang Zhao, Matteo Poggi, Fabio Tosi, Lei Zhou, Qiyu Sun, Yang Tang, and Stefano Mattoccia. 2023. GasMono: Geometry-Aided Self-Supervised Monocular Depth Estimation for Indoor Scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 16209–16220.

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

Junsheng Zhou, Yuwang Wang, Kaihuai Qin, and Wenjun Zeng. 2019. Moving indoor: Unsupervised video depth learning in challenging environments. In *Proceedings of the IEEE/CVF international conference on computer vision*. 8618–8627.

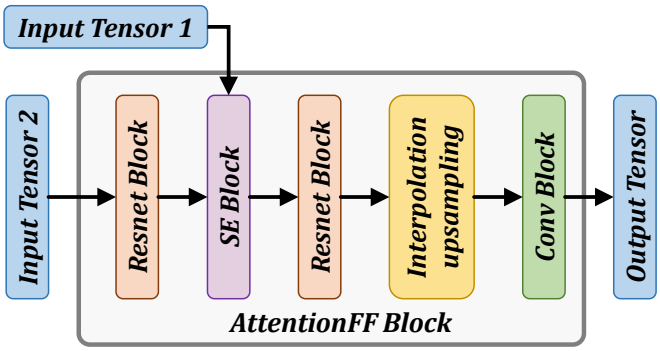


Fig. 6. Architecture of Our Attention Fusion Feature Block.

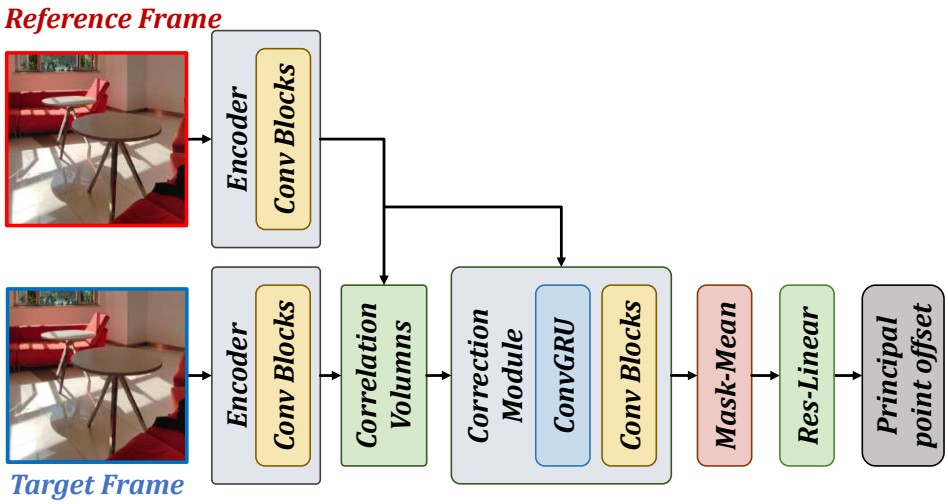


Fig. 7. Architecture of Our Principal Point Offset Estimation Module.

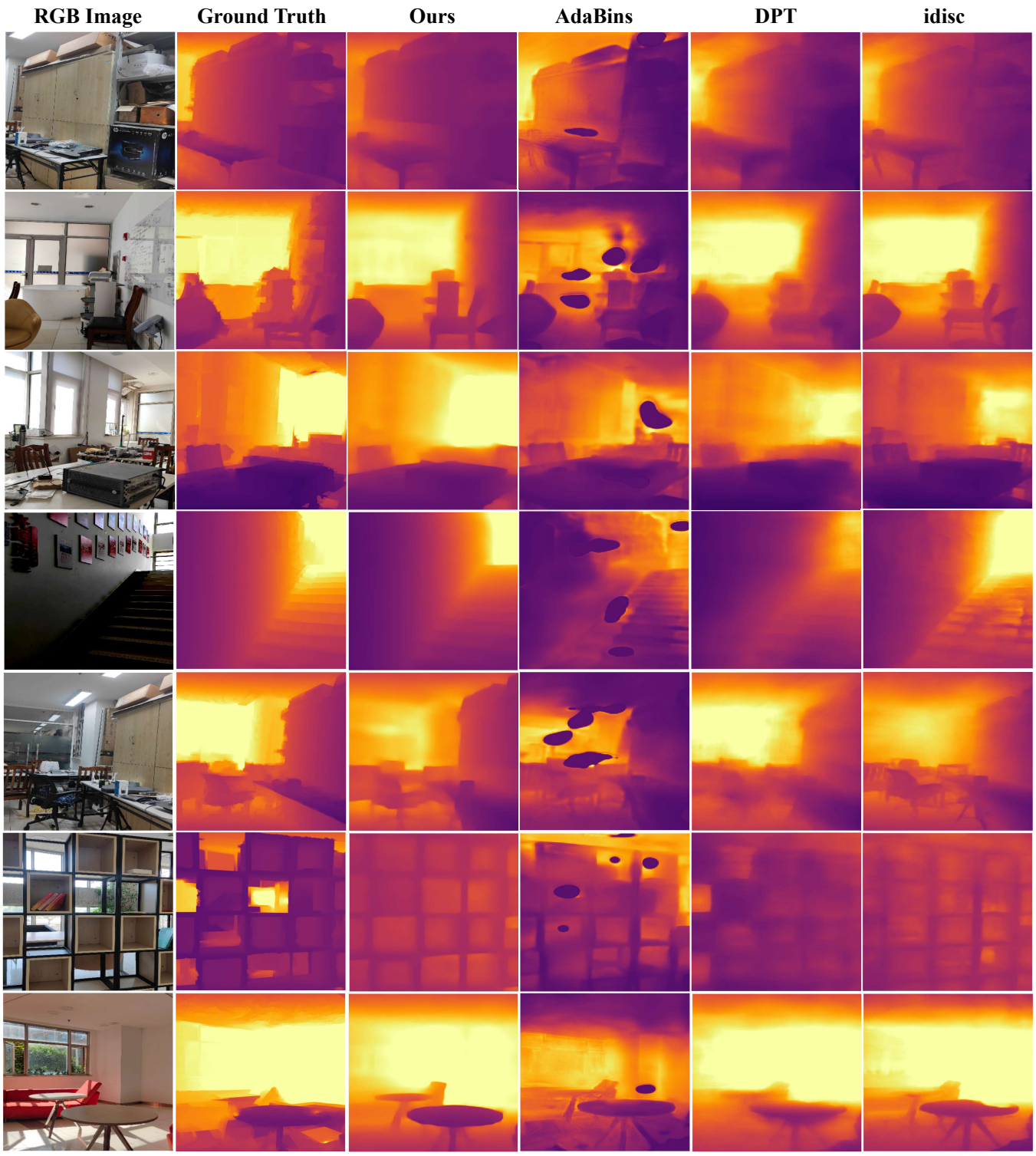


Fig. 8. Qualitative Comparison of our MODepth to other SOTA supervised monocular depth estimation methods on MODdata.