EGODEPTH: A RESIDUAL-BASED STEREO VIDEO DEPTH PREDICTION METHOD WITH 4D SPATIOTEMPORAL FEATURE FUSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Accurate and temporally consistent stereo matching is essential for robotic perception. Frame-wise stereo foundation models provide accurate disparity estimates but lack explicit temporal modeling. We present EgoDepth, a method for temporally consistent stereo matching built on Stereo Any Video. Disparity estimates from a frozen FoundationStereo model initialize single-scale recurrent refinement. During training, confidence-weighted residual regularization constrains deviations from the prior, with confidence derived from left-right photometric consistency. With its trainable components optimized on SceneFlow, EgoDepth achieves the lowest end-point error (EPE) on three of four cross-domain benchmarks and the lowest temporal end-point error (TEPE) on all four among the evaluated methods. After mixed-data training, EgoDepth also outperforms Stereo Any Video in both EPE and TEPE on Sintel Final and KITTI Depth. Real-world evaluations further support the generalization of EgoDepth across different scenes and capture settings.

1 INTRODUCTION

Stereo matching estimates disparity to recover metric depth for robotic navigation, manipulation, and 3D reconstruction. In stereo video, disparity estimates must be accurate and temporally consistent under camera and object motion (Karaev et al., 2023; Jing et al., 2025). Real-world deployment further requires reliable predictions across diverse scenes and capture conditions (Wen et al., 2025b).

Learning-based methods employing cost-volume aggregation (Kendall et al., 2017; Chang & Chen, 2018; Guo et al., 2019; Zhang et al., 2019) or recurrent disparity refinement (Lipson et al., 2021; Li et al., 2022; Xu et al., 2023; Wang et al., 2024) have achieved high disparity accuracy. More recent approaches leverage priors from monocular depth foundation models (Yang et al., 2024) to improve cross-domain generalization (Jiang et al., 2025; Cheng et al., 2025; Bartolomei et al., 2025). FoundationStereo (Wen et al., 2025b) also demonstrates strong cross-domain performance, but estimates disparity independently for each stereo pair without explicit temporal modeling. Video stereo methods exploit cross-frame information to improve temporal consistency (Karaev et al., 2023; Zhang et al., 2023; Jing et al., 2024; Zeng et al., 2024; Jing et al., 2025; Wang et al., 2025). However, temporal modeling alone does not ensure accurate predictions under domain shifts. A key challenge is therefore to improve temporal consistency while retaining the cross-domain disparity accuracy of stereo foundation models.

To address this challenge, we present EgoDepth, a method for temporally consistent stereo matching. EgoDepth uses disparity predictions from a frozen FoundationStereo model (Wen et al., 2025b) to initialize a single-scale recurrent refiner built on Stereo Any Video (SAV) (Jing et al., 2025). The refiner aggregates cross-frame information and iteratively predicts residual disparity updates. During training, confidence-weighted residual regularization penalizes deviations from the FoundationStereo estimates, with confidence derived from left-right photometric consistency.

EgoDepth achieves the lowest EPE on three of four cross-domain benchmarks and the lowest TEPE on all four in zero-shot evaluation (Table 1). On Sintel Final, it reduces EPE and TEPE by 76.0% and 57.9%, respectively, relative to SAV. With mixed-data post-training, EgoDepth also surpasses SAV in both metrics on all three evaluated benchmarks (Table 2). EgoDepth-Fast reduces average inference time from 0.56 to 0.29 seconds per frame while maintaining comparable disparity accuracy

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

and temporal consistency. Real-world results further support its generalization across diverse scenes and capture settings.

Our contributions are summarized as follows:

- We present EgoDepth, a method for temporally consistent stereo matching that combines a frozen stereo foundation prior with single-scale recurrent refinement and confidence-weighted residual regularization.
- We demonstrate through extensive experiments that EgoDepth achieves strong cross-domain generalization, attaining the lowest EPE on three of four cross-domain benchmarks and the lowest TEPE on all four among the evaluated methods.
- We present EgoDepth-Fast, an accelerated variant that substantially reduces inference time while retaining competitive disparity accuracy and temporal consistency.

2 Related Work

2.1 Stereo Matching and Foundation Priors

Learning-based stereo methods estimate disparity through cost-volume aggregation (Kendall et al., 2017; Chang & Chen, 2018; Guo et al., 2019; Zhang et al., 2019) or correlation-based recurrent refinement (Lipson et al., 2021; Li et al., 2022; Xu et al., 2023; Wang et al., 2024). Efficient variants reduce computation in these stages (Guo et al., 2025; Wen et al., 2025a).

Monocular depth foundation models (Yang et al., 2024) supply transferable priors for stereo feature extraction, initialization, and refinement (Jiang et al., 2025; Cheng et al., 2025; Bartolomei et al., 2025). FoundationStereo (Wen et al., 2025b) combines large-scale synthetic training with monocular feature priors for zero-shot stereo matching. These pairwise methods do not explicitly model temporal consistency.

2.2 Temporally Consistent Stereo Matching

Video stereo methods exploit motion compensation, feature reuse, and prediction fusion (Li et al., 2023; Zhang et al., 2023; Cheng et al., 2024); temporal completion and recurrent state fusion provide initialization for refinement (Zeng et al., 2024). Other approaches use spatiotemporal attention and bidirectional alignment (Karaev et al., 2023; Jing et al., 2024), or memory-based selection and reweighting of past observations (Wang et al., 2025).

SAV (Jing et al., 2025) integrates monocular video depth features into recurrent stereo refinement without camera poses or optical flow. Building on SAV, EgoDepth initializes single-scale temporal refinement with frozen FoundationStereo disparities and regularizes deviations from this prior using confidence derived from left-right photometric consistency during training.

3 Method

3.1 Overview

Given a rectified stereo sequence $\{(I_L^t, I_R^t)\}_{t=1}^T$, where $I_L^t, I_R^t \in \mathbb{R}^{H\times W\times 3}$, our method estimates a temporally consistent sequence of left-view disparity maps $\{d^t \in \mathbb{R}^{H\times W}\}_{t=1}^T$. As illustrated in Fig. 1, it first extracts matching and context features and obtains a per-frame disparity initialization from a frozen FoundationStereo model (Sec. 3.2); these representations are then processed by a single-scale temporal recurrent module to produce the final disparity sequence (Sec. 3.3).

3.2 Feature Extraction and Disparity Initialization

We adopt the feature representation of SAV (Jing et al., 2025), which combines trainable stereo encoders with temporally stable features from a pretrained Video Depth Anything (VDA) encoder (Chen et al., 2025). The VDA encoder remains frozen, while its shallow adapter and the stereo encoders are optimized during training.

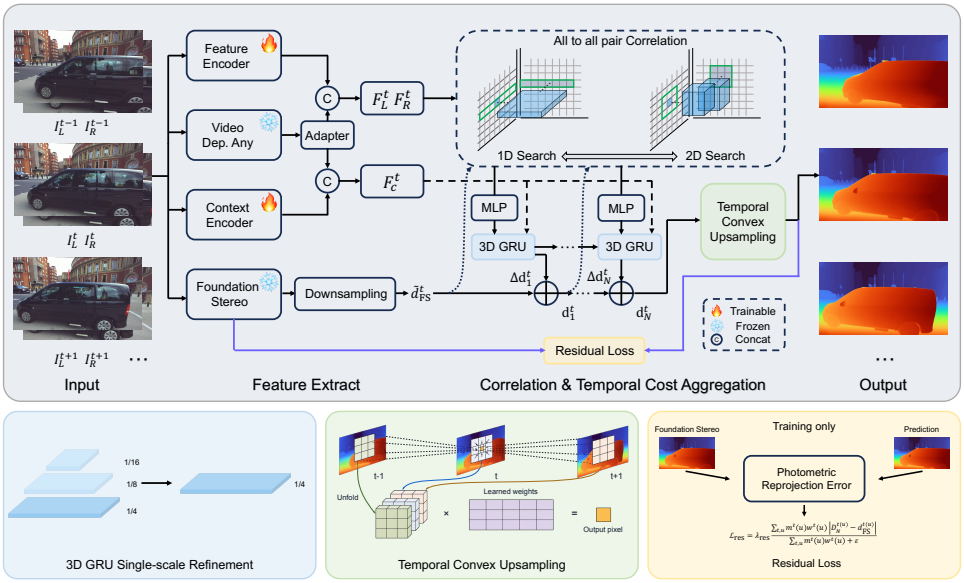


Figure 1: Overview of EgoDepth. A frozen FoundationStereo model provides the initial disparity, which is refined by a single-scale temporal recurrent module. Photometric-consistency-weighted residual regularization is applied during training.

Let $\mathcal{D}$ denote the frozen VDA encoder, $\mathcal{A}$ its trainable adapter, and $\mathcal{E}_f$ and $\mathcal{E}_c$ the stereo feature and context encoders, respectively. For frame t, the features are constructed as

$$V_v^t = \mathcal{A}\big(\mathcal{D}(I_v^{1:T})_t\big)\,, \quad F_v^t = [V_v^t; \mathcal{E}_f(I_v^t)]\ (v \in \{L, R\}), \quad F_c^t = [V_L^t; \mathcal{E}_c(I_L^t)]. \tag{1}$$

Here, $[\cdot;\cdot]$ denotes channel-wise concatenation. The matching features F_L^t and F_R^t provide the stereo correspondence representation, while F_c^t conditions the recurrent update.

In parallel, each stereo pair (I_L^t, I_R^t) is processed by a frozen FoundationStereo model (Wen et al., 2025b), yielding a full-resolution disparity estimate d_{FS}^t. We bilinearly resize the estimate to $1/4$ resolution, rescale its disparity magnitude, and use it as the initial disparity estimate for recurrent refinement:

$$\tilde{d}_{\mathrm{FS}}^t = \frac{1}{4}\,\mathcal{I}_{H/4,W/4}\big(d_{\mathrm{FS}}^t\big)\,, \qquad d_0^t = \tilde{d}_{\mathrm{FS}}^t. \tag{2}$$

Here, $\mathcal{I}_{H/4,W/4}$ denotes bilinear resizing to spatial resolution $H/4 \times W/4$.

3.3 Correlation and Temporal Recurrent Refinement

Starting from d_0^t, the disparity is refined iteratively at $1/4$ resolution. At iteration n, the right-view feature F_R^t is warped toward F_L^t according to d_{n-1}^t, producing $\hat{F}_R^t$. The resulting feature pair is processed by the all-to-all-pair correlation module introduced in SAV (Jing et al., 2025), which evaluates correlations within alternating 1D and 2D local search windows to construct the cost volume C_n^t. A multilayer perceptron then converts C_n^t into a compact representation for recurrent refinement.

The encoded correlation features, the context feature F_c^t, and the current disparity d_{n-1}^t are passed to the recurrent update block. We retain the temporal and spatial attention and separable 3D convolutional GRU used by SAV (Karaev et al., 2023; Jing et al., 2025), but perform recurrent refinement at a single $1/4$-resolution scale. The update block aggregates information across space and time and predicts a residual disparity Δd_n^t:

$$d_n^t = d_{n-1}^t + \Delta d_n^t. \tag{3}$$

Temporal convex upsampling (Jing et al., 2025) converts each recurrent estimate d_n^t into a full-resolution disparity map D_n^t. After N updates, the final output is $d^t = D_N^t$. Each output value is obtained as a learned convex combination over a $3 \times 3 \times 3$ spatiotemporal neighborhood.

3.4 Training Objective

We retain the iteration-weighted sequence loss of SAV (Jing et al., 2025) to supervise every full-resolution recurrent prediction. Let D_n^t denote the prediction at iteration n, D_{gt}^t the ground-truth disparity, and m^t the training validity mask shared by both losses. For N recurrent updates,

$$\mathcal{L}_{\mathrm{seq}} = \frac{1}{T}\sum_{t=1}^{T}\sum_{n=1}^{N}\alpha_n \frac{\sum_u m^t(u)\left|D_n^t(u) - D_{\mathrm{gt}}^t(u)\right|}{\sum_u m^t(u)}. \tag{4}$$

where α_n assigns larger weights to later iterations.

We regularize the final disparity prediction against the FoundationStereo estimate, using confidence weights derived from left-right photometric consistency. With RGB values scaled to $[0, 1]$, we define

$$\begin{aligned} e^t(u) &= \min_{s\in\{-1,+1\}} \frac{1}{3}\sum_{c=1}^{3}\Big|I_{L,c}^t(u) - I_{R,c}^t\big(u_x + s\, d_{\mathrm{FS}}^t(u), u_y\big)\Big|, \\ w^t(u) &= \mathrm{clamp}\Big(\exp\big[-\beta e^t(u)\big], c_{\min}, 1\Big). \end{aligned} \tag{5}$$

Here, β controls the decay rate and $c_{\min}$ sets the lower confidence bound. Non-integer image coordinates are sampled bilinearly, and the smaller photometric error of the two horizontal warping directions is used at each pixel.

The confidence-weighted residual term and the complete training objective are

$$\mathcal{L}_{\mathrm{res}} = \lambda_{\mathrm{res}} \frac{\sum_{t,u} m^t(u)\, w^t(u)\left|D_N^t(u) - d_{\mathrm{FS}}^t(u)\right|}{\sum_{t,u} m^t(u)\, w^t(u) + \epsilon}, \qquad \mathcal{L} = \mathcal{L}_{\mathrm{seq}} + \mathcal{L}_{\mathrm{res}}. \tag{6}$$

Here, λ_{res} controls the residual strength, and ϵ ensures numerical stability. The FoundationStereo estimate is detached when evaluating $\mathcal{L}_{\mathrm{res}}$.

4 Experiments

4.1 Training and Inference Settings

We follow the optimizer, learning-rate scheduler, data augmentation, and sequence-loss settings of SAV (Jing et al., 2025). We optimize the trainable components on SceneFlow (Mayer et al., 2016) for 120,000 iterations with a maximum learning rate of 10^{-4}, then continue training on Dynamic Replica (Karaev et al., 2023), Infinigen-StereoVideo (Jing et al., 2026), and Virtual KITTI 2 (Cabon et al., 2020) for another 120,000 iterations with a maximum learning rate of 5×10^{-5}. Training uses a batch size of two five-frame stereo clips. For confidence-weighted residual regularization, we set $\lambda_{\mathrm{res}} = 0.05$, $\beta = 15$, and $c_{\min} = 0.05$. For efficient inference, EgoDepth-Fast applies mixed precision, TensorRT compilation (NVIDIA Corporation, 2026), and Triton-based operator fusion (Tillet et al., 2019) to accelerate the frozen FoundationStereo prior and temporal refiner.

4.2 Benchmark Evaluation

4.2.1 Benchmarks and Metrics

We evaluate EgoDepth on Sintel Final (Butler et al., 2012), Dynamic Replica (Karaev et al., 2023), Infinigen-StereoVideo (Jing et al., 2026), Virtual KITTI 2 (Cabon et al., 2020), KITTI Depth (Uhrig et al., 2017), and UnrealStereo4K (Tosi et al., 2021). Using the metric implementation of SAV (Jing et al., 2025), we report end-point error (EPE) to evaluate frame-wise disparity accuracy, and temporal end-point error (TEPE) and the 1-pixel temporal outlier rate $\delta_{1\mathrm{px}}^t$ to assess temporal consistency. EPE and TEPE are reported in pixels, and $\delta_{1\mathrm{px}}^t$ as a percentage. Lower values indicate better performance for all three metrics.

4.2.2 ZERO-SHOT GENERALIZATION

We evaluate the checkpoint from the SceneFlow training stage on Sintel Final, Dynamic Replica, Infinigen-StereoVideo, and Virtual KITTI 2. Following the evaluation protocol of SAV (Jing et al., 2025), we compare EgoDepth with image-based and video-based methods using the baseline results reported in that work.

EgoDepth achieves the lowest TEPE on all four benchmarks and the lowest EPE on three among the compared methods (Table 1). Improvements over SAV are particularly pronounced on Sintel Final and Dynamic Replica, with substantial reductions in both frame-wise and temporal disparity errors. On Virtual KITTI 2, the advantage lies in temporal accuracy rather than frame-wise accuracy.

Table 1: Zero-shot generalization on four benchmarks. EgoDepth uses 4 recurrent updates and 12 FoundationStereo iterations; iteration budgets vary across methods. Bold and underlined values indicate the best and second-best results, respectively.

Method	Sintel Final			Dynamic Replica			Infinigen SV			Virtual KITTI2		
	TEPE↓	δ^t_{1px}↓	EPE↓	TEPE↓	δ^t_{1px}↓	EPE↓	TEPE↓	δ^t_{1px}↓	EPE↓	TEPE↓	δ^t_{1px}↓	EPE↓
Image-based												
RAFT-Stereo (Lipson et al., 2021)	2.09	13.70	4.89	0.145	2.02	0.38	2.33	19.80	3.77	1.25	18.39	1.45
IGEV-Stereo (Xu et al., 2023)	1.74	12.65	2.44	0.081	0.80	0.20	2.35	19.33	3.81	1.42	18.88	1.39
Selective-IGEV (Wang et al., 2024)	1.96	12.95	2.82	0.133	1.29	0.29	2.85	21.83	5.69	1.41	19.10	1.38
DEFOM-Stereo (Jiang et al., 2025)	2.33	12.43	2.38	0.080	0.75	<u>0.13</u>	2.01	18.91	3.24	1.07	16.63	1.34
MonSter (Cheng et al., 2025)	1.70	11.06	3.30	0.157	0.86	0.45	<u>1.65</u>	15.73	<u>2.27</u>	<u>0.73</u>	**10.93**	**0.67**
Video-based												
DynamicStereo (Karaev et al., 2023)	1.43	11.80	4.26	0.083	0.73	0.24	1.98	15.68	4.67	1.06	15.63	1.47
BiDAStereo (Jing et al., 2024)	1.26	11.65	1.99	0.091	1.04	0.65	1.99	16.50	6.52	1.02	14.82	1.43
Stereo Any Video (Jing et al., 2025)	<u>1.07</u>	<u>9.98</u>	<u>1.75</u>	<u>0.057</u>	<u>0.67</u>	0.25	<u>1.65</u>	<u>14.27</u>	3.16	0.74	12.45	<u>1.01</u>
EgoDepth	**0.45**	**5.28**	**0.42**	**0.038**	**0.527**	**0.083**	**1.48**	**13.42**	**2.15**	**0.63**	<u>10.94</u>	1.04

4.2.3 MIXED-DATA EVALUATION

We evaluate the checkpoint from the mixed-data training stage and its accelerated variant, EgoDepth-Fast, on UnrealStereo4K, Sintel Final, and KITTI Depth, using four recurrent updates. On Sintel Final and KITTI Depth, external baseline results are taken from SAV (Jing et al., 2025), except for Fast-FoundationStereo, which we evaluate using the respective dataset protocols. For UnrealStereo4K, we adapt the SAV evaluation pipeline.

Both EgoDepth and EgoDepth-Fast achieve lower EPE and TEPE than SAV across all three benchmarks (Table 2). The two variants rank first and second across all three error metrics on UnrealStereo4K and Sintel Final. On KITTI Depth, they jointly achieve the lowest TEPE, while their EPE remains higher than that of the best image-based method. EgoDepth-Fast retains comparable prediction quality.

4.2.4 QUALITATIVE COMPARISON

Figures 2 and 3 compare disparity predictions on indoor and outdoor sequences from South Kensington Stereo Video (Jing et al., 2026). In the indoor sequence, EgoDepth preserves more coherent disparity structure in the highlighted reflective regions across the displayed frames, while several baselines exhibit persistent artifacts. In the outdoor sequence, some baselines produce inconsistent predictions in the highlighted region as the viewpoint changes, whereas EgoDepth maintains a more consistent disparity structure across frames.

Table 2: Mixed-data evaluation. SF: SceneFlow; M: Middlebury; DR: Dynamic Replica; ISV: Infinigen-StereoVideo; KD: KITTI Depth; VK: Virtual KITTI 2. Bold training-data labels mark evaluation datasets; dashes indicate test sets used in training. Runtime is averaged over the three benchmarks on a single H200. Best and second-best results are bold and underlined.

Method	Training Data	UnrealStereo4K			Sintel Final			KITTI Depth			Runtime↓
		TEPE↓	δ^t_{1px}↓	EPE↓	TEPE↓	δ^t_{1px}↓	EPE↓	TEPE↓	δ^t_{1px}↓	EPE↓	(s/frame)
Image-based											
RAFT-Stereo (Lipson et al., 2021)	10 mixed datasets	2.72	22.65	1.89	–	–	–	0.41	7.96	<u>0.51</u>	0.27
IGEV-Stereo (Xu et al., 2023)	SF+M	2.44	20.50	1.67	1.53	13.49	2.23	0.46	9.12	0.60	<u>0.12</u>
Selective-IGEV (Wang et al., 2024)	SF+M	2.12	18.00	1.44	1.64	13.78	2.14	0.48	9.66	0.60	0.38
DEFOM-Stereo (Jiang et al., 2025)	15 mixed datasets	1.94	17.01	1.39	–	–	–	0.39	7.08	**0.48**	0.13
MonSter (Cheng et al., 2025)	10 mixed datasets	2.47	16.04	2.25	–	–	–	0.41	7.73	0.60	0.32
FoundationStereo (Wen et al., 2025b)	7 mixed datasets	1.83	13.25	1.15	–	–	–	0.40	7.20	0.57	0.60
Fast-FoundationStereo (Wen et al., 2025a)	Synthetic+ Stereo4D	2.47	17.65	1.76	1.44	11.90	12.25	<u>0.35</u>	5.58	0.57	**0.05**
Video-based											
DynamicStereo (Karaev et al., 2023)	SF+DR+ ISV+**KD**	2.79	23.66	1.98	1.38	11.84	3.38	0.42	8.21	0.57	0.15
BiDAStereo (Jing et al., 2024)	SF+DR+ ISV+**KD**	2.59	22.28	1.85	1.33	11.34	1.97	0.42	8.33	0.58	0.19
Stereo Any Video (Jing et al., 2025)	SF+DR+ ISV+VK	1.84	17.13	1.27	0.99	9.09	1.48	<u>0.35</u>	6.05	0.61	0.58
EgoDepth	SF+DR+ ISV+VK	**1.60**	<u>12.78</u>	**1.05**	**0.83**	**8.49**	<u>1.42</u>	**0.33**	<u>5.36</u>	0.57	0.56
EgoDepth-Fast	SF+DR+ ISV+VK	<u>1.65</u>	**12.76**	<u>1.08</u>	<u>0.86</u>	<u>8.57</u>	**1.36**	**0.33**	**5.34**	0.57	0.29

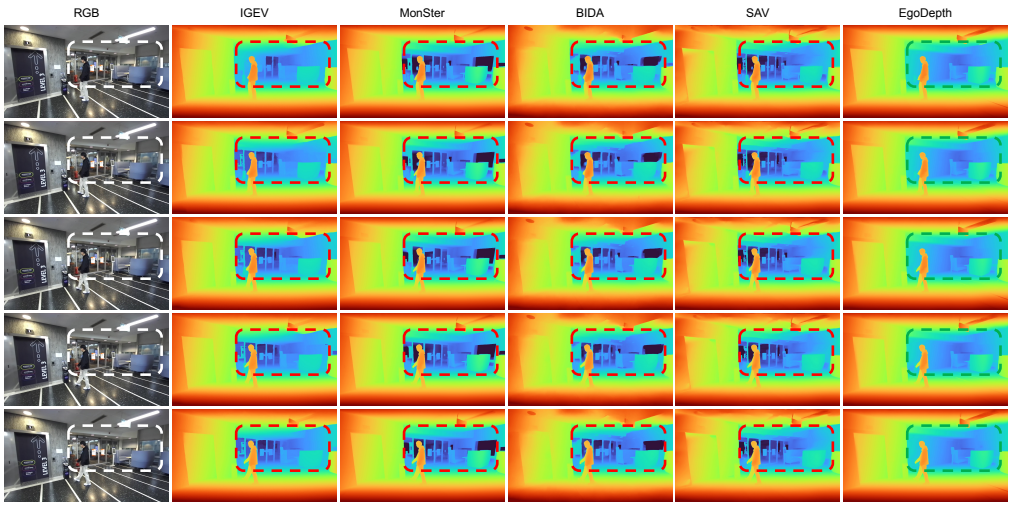


Figure 2: Qualitative comparison on an indoor scene from South Kensington Stereo Video. White dashed boxes in the RGB images indicate regions of interest; red and green dashed boxes in the disparity maps highlight artifacts and more coherent structures, respectively.

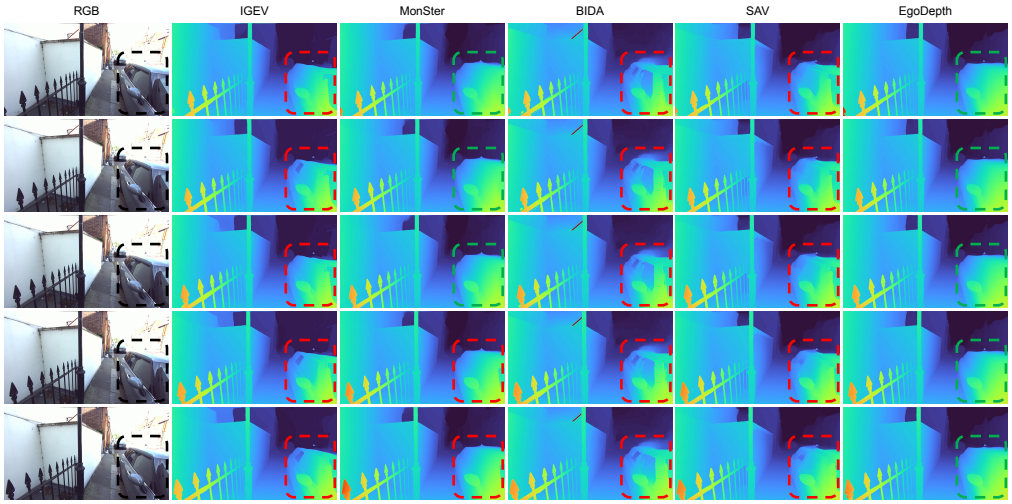


Figure 3: Qualitative comparison on an outdoor scene from South Kensington Stereo Video. Black dashed boxes in the RGB images indicate regions of interest; red and green dashed boxes in the disparity maps highlight artifacts and more coherent structures, respectively.

Table 3: Component ablation on the FlyingThings3D test split. MS and FS denote multi-scale refinement and FoundationStereo, respectively.

Variant	GRU	FS prior	Residual reg.	EPE↓	TEPE↓	$\delta^t_{1\text{px}}$ ↓
Baseline	MS	–	–	0.857	1.224	12.81
+ FS initialization	MS	✓	–	0.672	0.992	10.61
+ single-scale GRU	1/4	✓	–	<u>0.576</u>	<u>0.875</u>	<u>8.70</u>
+ residual reg.	MS	✓	✓	0.673	0.993	10.59
EgoDepth	1/4	✓	✓	**0.555**	**0.845**	**8.48**

4.3 Ablation Study

We conduct component ablations on the FlyingThings3D test split using checkpoints from the SceneFlow training stage described in Section 4.1. Table 3 compares a retrained SAV baseline (Jing et al., 2025) with four FoundationStereo-initialized variants, combining multi-scale or single-scale refinement with and without residual regularization. All variants perform 24 recurrent updates at $1/4$ resolution; the multi-scale variants additionally perform 12 updates at each of the $1/16$ and $1/8$ scales.

FoundationStereo initialization reduces all three error metrics relative to the retrained baseline. Single-scale refinement further improves frame-wise disparity accuracy and temporal consistency, both without and with residual regularization, despite requiring fewer recurrent updates. Residual regularization has little effect in the multi-scale setting, but improves all three metrics when combined with single-scale refinement. The full model achieves the lowest errors across all three metrics.

5 Application

5.1 Egocentric Depth Estimation

We examine generalization across egocentric datasets and capture devices using RGB stereo videos from the open datasets Gen-EgoData (Genrobot.AI, 2026), EgoPro (Lightwheel, 2026), and Ego-Live (Li et al., 2026), together with Orbbec infrared (IR) recordings (Figures 4 and 5). EgoDepth recovers the relative depth of hands, manipulated objects, and surrounding surfaces under these different capture conditions. Across the displayed frames, foreground objects remain distinguishable from the background, with coherent contours and depth ordering as the viewpoint changes. EgoDepth-Fast produces visually similar scene structures, preserving these qualitative properties across both RGB and IR inputs.

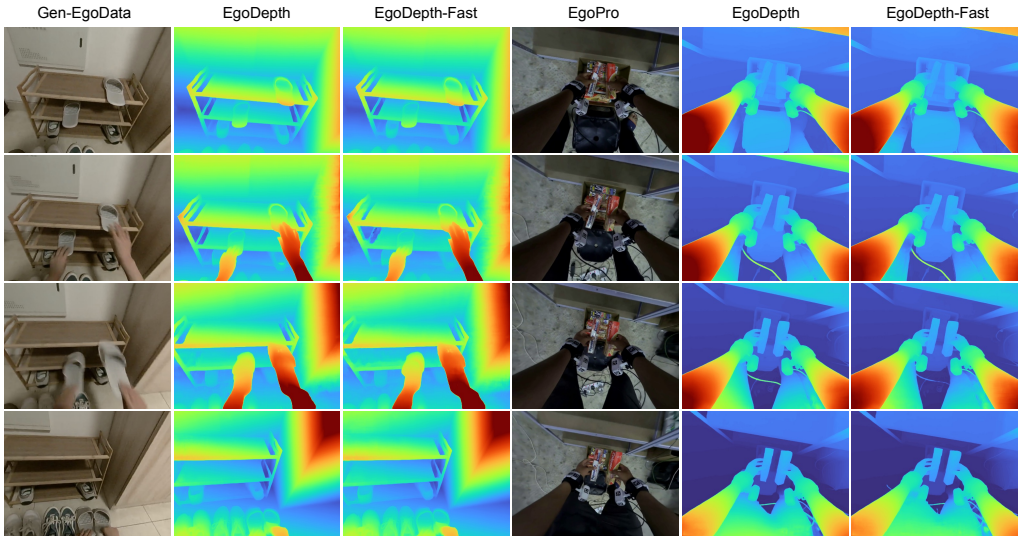


Figure 4: Qualitative results of EgoDepth and EgoDepth-Fast on egocentric stereo videos from Gen-EgoData (left) and EgoPro (right). Each block shows input images and the two depth predictions, with time progressing from top to bottom.

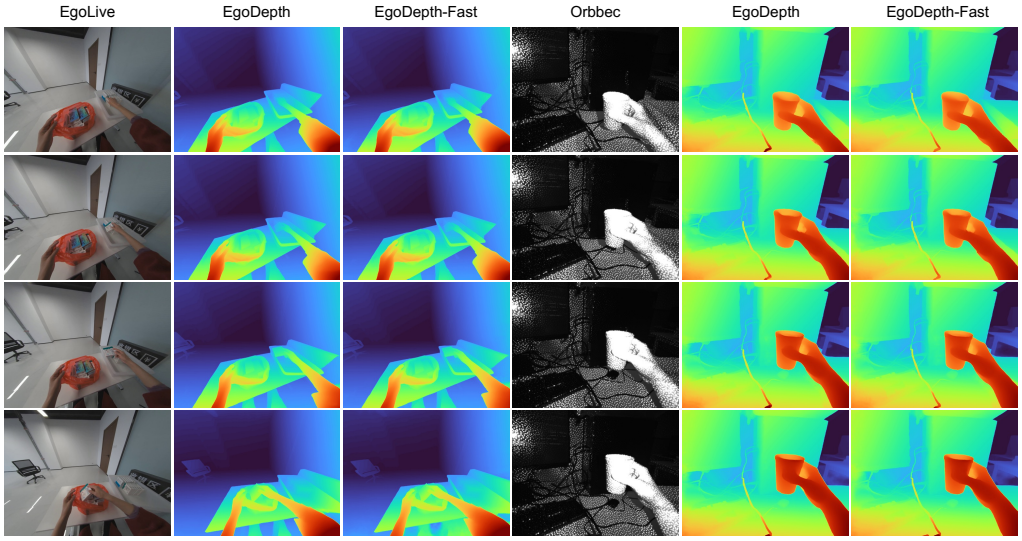


Figure 5: Qualitative results of EgoDepth and EgoDepth-Fast on egocentric stereo videos from EgoLive (left) and Orbbec (right), using RGB and infrared (IR) inputs, respectively. Each block shows input images and the two depth predictions, with time progressing from top to bottom.

5.2 Quantitative Evaluation

We evaluate relative geometry using an Orbbec camera and a stepped calibration target at 23 distances from 0.8 to 3.0 m. For raw depth and EgoDepth, we compare pairwise height differences along the target coordinate system's z-axis with the calibrated reference. We report mean absolute error (MAE) and the percentages of point pairs with errors within 5 and 10 mm.

EgoDepth lowers MAE and increases the percentages within both error thresholds at all 23 distances (Table 4). At 3.0 m, MAE decreases from 42.11 to 15.20 mm, a 63.9% reduction. Figure 6 also shows more regular boundaries around the target's frame with EgoDepth than with raw depth.

Table 4: Pairwise height-difference accuracy at 23 camera-to-target distances. MAE is in millimeters; threshold columns give the percentages of pairs with errors within 5 and 10 mm.

	raw			EgoDepth		
Dist. (mm)	MAE	≤ 5 (%)	≤ 10 (%)	MAE	≤ 5 (%)	≤ 10 (%)
800	2.430	89.89	99.97	**1.951**	**96.47**	**100.00**
830	2.627	87.09	99.93	**2.007**	**96.31**	**100.00**
860	2.821	84.26	99.70	**1.918**	**96.84**	**100.00**
890	2.719	85.61	99.71	**1.873**	**97.34**	**100.00**
920	3.257	77.36	98.90	**2.022**	**95.79**	**100.00**
950	2.861	83.41	99.54	**2.379**	**90.89**	**100.00**
980	2.705	85.94	99.87	**1.836**	**96.79**	**100.00**
1010	2.829	84.28	99.89	**1.901**	**97.11**	**100.00**
1040	3.356	76.78	97.54	**1.846**	**97.20**	**100.00**
1070	2.987	81.72	99.37	**1.885**	**97.68**	**100.00**
1100	3.294	77.47	98.19	**2.164**	**93.73**	**100.00**
1130	3.285	77.38	98.71	**2.237**	**93.42**	**100.00**
1160	3.397	76.55	97.28	**2.035**	**96.19**	**100.00**
1200	3.622	72.40	97.83	**2.107**	**93.91**	**100.00**
1400	5.587	52.00	84.49	**2.702**	**86.00**	**100.00**
1600	7.311	40.56	72.63	**3.690**	**70.73**	**97.01**
1800	7.366	40.71	70.01	**5.018**	**56.11**	**88.16**
2000	12.481	29.98	53.70	**8.965**	**37.28**	**65.59**
2200	22.588	15.81	29.59	**12.235**	**25.61**	**48.54**
2400	27.267	13.26	24.31	**13.864**	**22.90**	**42.84**
2600	38.587	8.81	16.94	**15.775**	**19.81**	**38.12**
2800	26.969	12.70	22.64	**16.112**	**19.74**	**37.44**
3000	42.106	8.34	16.29	**15.197**	**22.20**	**38.37**

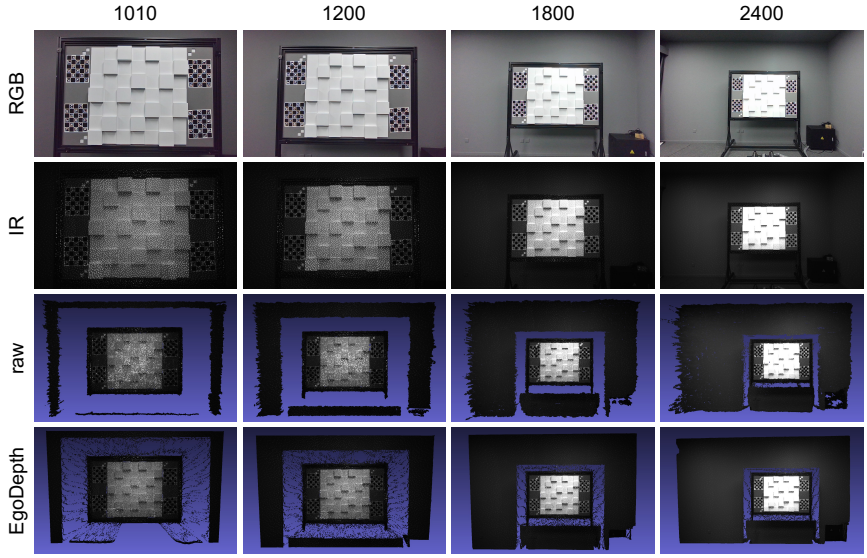


Figure 6: Point-cloud comparison on the calibration target. Columns indicate distance (mm); IR and raw denote the left infrared view and raw sensor depth, respectively.

6 Conclusion

We presented EgoDepth, a stereo video framework that combines a frozen FoundationStereo prior with single-scale temporal refinement and confidence-weighted residual regularization during training. Benchmark evaluations show gains in frame-wise disparity accuracy and temporal consistency across domains, while EgoDepth-Fast reduces inference latency with comparable prediction quality. Real-world relative-geometry measurements and qualitative comparisons on egocentric videos further support generalization across scenes and imaging modalities. Together, these results highlight the value of using stereo foundation predictions as both an initialization and a training reference for accurate, temporally consistent video depth estimation.

AI USE STATEMENT

Generative AI tools assisted with checking methodological descriptions, interpreting and summarizing experimental results, and drafting and revising the manuscript. They also supported literature searches and reference checks, LaTeX formatting, and figure and table editing. We have reviewed all AI-assisted work. These activities were guided by the authors, who made the final editorial decisions and take responsibility for the scientific claims, reported results, and all AI-assisted content.

REFERENCES

Luca Bartolomei, Fabio Tosi, Matteo Poggi, and Stefano Mattoccia. Stereo anywhere: Robust zero-shot deep stereo matching even where either stereo or mono fail. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 1013–1027. Computer Vision Foundation / IEEE, 2025. doi: 10.1109/CVPR52734.2025.00103. URL https://openaccess.thecvf.com/content/CVPR2025/html/Bartolomei_Stereo_Anywhere_Robust_Zero-Shot_Deep_Stereo_Matching_Even_Where_Either_CVPR_2025_paper.html.

Daniel J. Butler, Jonas Wulff, Garrett B. Stanley, and Michael J. Black. A naturalistic open source movie for optical flow evaluation. In Andrew W. Fitzgibbon, Svetlana Lazebnik, Pietro Perona, Yoichi Sato, and Cordelia Schmid (eds.), *Computer Vision - ECCV 2012 - 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part VI*, volume 7577 of *Lecture Notes in Computer Science*, pp. 611–625. Springer, 2012. doi: 10.1007/978-3-642-33783-3_44. URL https://doi.org/10.1007/978-3-642-33783-3_44.

Yohann Cabon, Naila Murray, and Martin Humenberger. Virtual KITTI 2. *CoRR*, abs/2001.10773, 2020. URL https://arxiv.org/abs/2001.10773.

Jia-Ren Chang and Yong-Sheng Chen. Pyramid stereo matching network. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pp. 5410–5418. Computer Vision Foundation / IEEE Computer Society, 2018. doi: 10.1109/CVPR.2018.00567. URL http://openaccess.thecvf.com/content_cvpr_2018/html/Chang_Pyramid_Stereo_Matching_CVPR_2018_paper.html.

Sili Chen, Hengkai Guo, Shengnan Zhu, Feihu Zhang, Zilong Huang, Jiashi Feng, and Bingyi Kang. Video depth anything: Consistent depth estimation for super-long videos. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 22831–22840. Computer Vision Foundation / IEEE, 2025. doi: 10.1109/CVPR52734.2025.02126. URL https://openaccess.thecvf.com/content/CVPR2025/html/Chen_Video_Depth_Anything_Consistent_Depth_Estimation_for_Super-Long_Videos_CVPR_2025_paper.html.

Junda Cheng, Longliang Liu, Gangwei Xu, Xianqi Wang, Zhaoxing Zhang, Yong Deng, Jinliang Zang, Yurui Chen, Zhipeng Cai, and Xin Yang. Monster: Marry monodepth to stereo unleashes power. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 6273–6282. Computer Vision Foundation / IEEE, 2025. doi: 10.1109/CVPR52734.2025.00588. URL https://openaccess.thecvf.com/content/CVPR2025/html/Cheng_MonSter_Marry_Monodepth_to_Stereo_Unleashes_Power_CVPR_2025_paper.html.

Ziang Cheng, Jiayu Yang, and Hongdong Li. Stereo matching in time: 100+ FPS video stereo matching for extended reality. In *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2024, Waikoloa, HI, USA, January 3-8, 2024*, pp. 8704–8713. IEEE, 2024. doi: 10.1109/WACV57701.2024.00852. URL https://doi.org/10.1109/WACV57701.2024.00852.

Genrobot.AI. Gen-EgoData. Hugging Face dataset, 2026. URL https://huggingface.co/datasets/genrobot2025/Gen-EgoData. Accessed September 26, 2026.

Xianda Guo, Chenming Zhang, Youmin Zhang, Wenzhao Zheng, Dujun Nie, Matteo Poggi, and Long Chen. Lightstereo: Channel boost is all you need for efficient 2d cost aggregation. In *IEEE International Conference on Robotics and Automation, ICRA 2025, Atlanta, GA, USA, May 19-23, 2025*, pp. 8738–8744. IEEE, 2025. doi: 10.1109/ICRA55743.2025.11127711. URL https://doi.org/10.1109/ICRA55743.2025.11127711.

Xiaoyang Guo, Kai Yang, Wukui Yang, Xiaogang Wang, and Hongsheng Li. Group-wise correlation stereo network. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 3273–3282. Computer Vision Foundation / IEEE, 2019. doi: 10.1109/CVPR.2019.00339. URL http://openaccess.thecvf.com/content_CVPR_2019/html/Guo_Group-Wise_Correlation_Stereo_Network_CVPR_2019_paper.html.

Hualie Jiang, Zhiqiang Lou, Laiyan Ding, Rui Xu, Minglang Tan, Wenjie Jiang, and Rui Huang. Defom-stereo: Depth foundation model based stereo matching. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 21857–21867. Computer Vision Foundation / IEEE, 2025. doi: 10.1109/CVPR52734.2025.02036. URL https://openaccess.thecvf.com/content/CVPR2025/html/Jiang_DEFOM-Stereo_Depth_Foundation_Model_Based_Stereo_Matching_CVPR_2025_paper.html.

Junpeng Jing, Ye Mao, and Krystian Mikolajczyk. Match-stereo-videos: Bidirectional alignment for consistent dynamic stereo matching. In Ales Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part LX*, volume 15118 of *Lecture Notes in Computer Science*, pp. 415–432. Springer, 2024. doi: 10.1007/978-3-031-73027-6_24. URL https://doi.org/10.1007/978-3-031-73027-6_24.

Junpeng Jing, Weixun Luo, Ye Mao, and Krystian Mikolajczyk. Stereo any video: Temporally consistent stereo matching. In *IEEE/CVF International Conference on Computer Vision, ICCV 2025, Honolulu, HI, USA, October 19-25, 2025*, pp. 20836–20846. IEEE, 2025. doi: 10.1109/ICCV51701.2025.01937. URL https://doi.org/10.1109/ICCV51701.2025.01937.

Junpeng Jing, Ye Mao, Anlan Qiu, and Krystian Mikolajczyk. Match stereo videos via bidirectional alignment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(8):9337–9353, 2026. doi: 10.1109/TPAMI.2026.3679033.

Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Dynamicstereo: Consistent dynamic depth from stereo videos. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pp. 13229–13239. IEEE, 2023. doi: 10.1109/CVPR52729.2023.01271. URL https://doi.org/10.1109/CVPR52729.2023.01271.

Alex Kendall, Hayk Martirosyan, Saumitro Dasgupta, Peter Henry, Ryan Kennedy, Abraham Bachrach, and Adam Bry. End-to-end learning of geometry and context for deep stereo regression. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 66–75, 2017. doi: 10.1109/ICCV.2017.17.

Jiankun Li, Peisen Wang, Pengfei Xiong, Tao Cai, Ziwei Yan, Lei Yang, Jiangyu Liu, Haoqiang Fan, and Shuaicheng Liu. Practical stereo matching via cascaded recurrent network with adaptive correlation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp. 16242–16251. IEEE, 2022. doi: 10.1109/CVPR52688.2022.01578. URL https://doi.org/10.1109/CVPR52688.2022.01578.

Yihang Li, Xuelong Wei, Jingzhou Luo, Yingjing Xiao, Yibo Bai, Guangyuan Zhou, Teng Zou, Chenguang Gui, Jiajun Wen, He Zhang, Kangliang Chen, Xing Pan, Shuaiyan Liu, Daming Wang, Tao An, Jiayi Li, Shibo Jin, Wanwan Zhang, Tianyu Wang, Boren Wei, Zhixuan Huang, Fangsheng Liu, Ruodai Li, Hui Zhang, Anson Li, Yicheng Gong, Peng Cao, Jiaming Liang, and

Liang Lin. Egolive: A large-scale egocentric dataset from real-world human tasks, 2026. URL https://arxiv.org/abs/2604.23570.

Zhaoshuo Li, Wei Ye, Dilin Wang, Francis X. Creighton, Russell H. Taylor, Ganesh Venkatesh, and Mathias Unberath. Temporally consistent online depth estimation in dynamic scenes. In *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2023, Waikoloa, HI, USA, January 2-7, 2023*, pp. 3017–3026. IEEE, 2023. doi: 10.1109/WACV56688.2023.00303. URL https://doi.org/10.1109/WACV56688.2023.00303.

Lightwheel. EgoPro. Hugging Face dataset, 2026. URL https://huggingface.co/datasets/LightwheelAI/EgoPro. Accessed September 26, 2026.

Lahav Lipson, Zachary Teed, and Jia Deng. Raft-stereo: Multilevel recurrent field transforms for stereo matching. In *International Conference on 3D Vision, 3DV 2021, London, United Kingdom, December 1-3, 2021*, pp. 218–227. IEEE, 2021. doi: 10.1109/3DV53792.2021.00032. URL https://doi.org/10.1109/3DV53792.2021.00032.

Nikolaus Mayer, Eddy Ilg, Philip Häusser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pp. 4040–4048. IEEE Computer Society, 2016. doi: 10.1109/CVPR.2016.438. URL https://doi.org/10.1109/CVPR.2016.438.

NVIDIA Corporation. NVIDIA TensorRT Documentation, 2026. URL https://docs.nvidia.com/deeplearning/tensorrt/. Accessed: 2026-09-01.

Philippe Tillet, Hsiang-Tsung Kung, and David D. Cox. Triton: an intermediate language and compiler for tiled neural network computations. In Tim Mattson, Abdullah Muzahid, and Armando Solar-Lezama (eds.), *Proceedings of the 3rd ACM SIGPLAN International Workshop on Machine Learning and Programming Languages, MAPL@PLDI 2019, Phoenix, AZ, USA, June 22, 2019*, pp. 10–19. ACM, 2019. doi: 10.1145/3315508.3329973. URL https://doi.org/10.1145/3315508.3329973.

Fabio Tosi, Yiyi Liao, Carolin Schmitt, and Andreas Geiger. Smd-nets: Stereo mixture density networks. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pp. 8942–8952. Computer Vision Foundation / IEEE, 2021. doi: 10.1109/CVPR46437.2021.00883. URL https://openaccess.thecvf.com/content/CVPR2021/html/Tosi_SMD-Nets_Stereo_Mixture_Density_Networks_CVPR_2021_paper.html.

Jonas Uhrig, Nick Schneider, Lukas Schneider, Uwe Franke, Thomas Brox, and Andreas Geiger. Sparsity invariant cnns. In *2017 International Conference on 3D Vision, 3DV 2017, Qingdao, China, October 10-12, 2017*, pp. 11–20. IEEE Computer Society, 2017. doi: 10.1109/3DV.2017.00012. URL https://doi.org/10.1109/3DV.2017.00012.

Xianqi Wang, Gangwei Xu, Hao Jia, and Xin Yang. Selective-stereo: Adaptive frequency information selection for stereo matching. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 19701–19710. IEEE, 2024. doi: 10.1109/CVPR52733.2024.01863. URL https://doi.org/10.1109/CVPR52733.2024.01863.

Yun Wang, Junjie Hu, Qiaole Dong, Yongjian Zhang, Yanwei Fu, Tin Lun Lam, and Dapeng Wu. Ppmstereo: Pick-and-play memory construction for consistent dynamic stereo matching. In Danielle Belgrave, Cheng Zhang, Laura N. Montoya, Hsuan-Tien Lin, Razvan Pascanu, Piotr Koniusz, Marzyeh Ghassemi, Nancy Chen, Iván Vladimir Meza Ruíz, and Arturo Loaiza-Bonilla (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025. URL http://papers.nips.cc/paper_files/paper/2025/hash/8b2fc235787852ead92da2268cd9e90c-Abstract-Conference.html.

Bowen Wen, Shaurya Dewan, and Stan Birchfield. Fast-foundationstereo: Real-time zero-shot stereo matching. *CoRR*, abs/2512.11130, 2025a. doi: 10.48550/ARXIV.2512.11130. URL https://doi.org/10.48550/arXiv.2512.11130.

Bowen Wen, Matthew Trepte, Joseph Aribido, Jan Kautz, Orazio Gallo, and Stan Birchfield. Foundationstereo: Zero-shot stereo matching. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 5249–5260. Computer Vision Foundation / IEEE, 2025b. doi: 10.1109/CVPR52734.2025.00495. URL https://openaccess.thecvf.com/content/CVPR2025/html/Wen_FoundationStereo_Zero-Shot_Stereo_Matching_CVPR_2025_paper.html.

Gangwei Xu, Xianqi Wang, Xiaohuan Ding, and Xin Yang. Iterative geometry encoding volume for stereo matching. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pp. 21919–21928. IEEE, 2023. doi: 10.1109/CVPR52729.2023.02099. URL https://doi.org/10.1109/CVPR52729.2023.02099.

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything V2. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/26cfdcd8fe6fd75cc53e92963a656c58-Abstract-Conference.html.

Jiaxi Zeng, Chengtang Yao, Yuwei Wu, and Yunde Jia. Temporally consistent stereo matching. In Ales Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part XXXI*, volume 15089 of *Lecture Notes in Computer Science*, pp. 341–359. Springer, 2024. doi: 10.1007/978-3-031-72751-1_20. URL https://doi.org/10.1007/978-3-031-72751-1_20.

Feihu Zhang, Victor Adrian Prisacariu, Ruigang Yang, and Philip H. S. Torr. Ga-net: Guided aggregation net for end-to-end stereo matching. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 185–194. Computer Vision Foundation / IEEE, 2019. doi: 10.1109/CVPR.2019.00027. URL http://openaccess.thecvf.com/content_CVPR_2019/html/Zhang_GA-Net_Guided_Aggregation_Net_for_End-To-End_Stereo_Matching_CVPR_2019_paper.html.

Youmin Zhang, Matteo Poggi, and Stefano Mattoccia. Temporalstereo: Efficient spatial-temporal stereo matching network. In *IROS*, pp. 9528–9535, 2023. doi: 10.1109/IROS55552.2023.10341598. URL https://doi.org/10.1109/IROS55552.2023.10341598.