

# WAF-T-Stereo: Warping-Alone Field Transforms for Stereo Matching

Yihan Wang and Jia Deng

Department of Computer Science, Princeton University  
{yw7685, jiadeng}@princeton.edu

**Abstract.** We introduce WAF-T-Stereo, a simple and effective warping-based method for stereo matching. WAF-T-Stereo demonstrates that cost volumes, a common design used in many leading methods, are not necessary for strong performance and can be replaced by warping with improved efficiency. WAF-T-Stereo ranks first on ETH3D (BP-0.5), Middlebury (RMSE), and KITTI (all metrics), reducing the zero-shot error by 81% on ETH3D, while being  $1.8\text{--}6.7\times$  faster than competitive methods. Code and model weights are available at <https://github.com/princeton-vl/WAF-T-Stereo>.

**Keywords:** Stereo Matching · Warping · Dense Correspondences

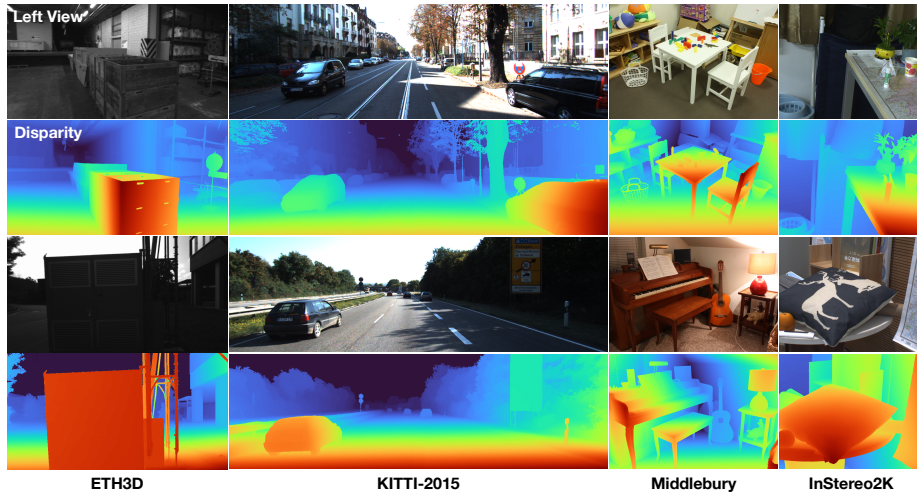
## 1 Introduction

Stereo matching is the problem of estimating the horizontal motion of each pixel given two rectified images from two calibrated cameras. The horizontal motion, or disparity, can be directly converted to depth. Stereo matching has many direct applications, including autonomous driving [10, 29] and augmented reality [9].

Most leading stereo matching methods [7, 11, 20, 23, 30, 57, 59] rely on cost volumes [45, 67], which are constructed by pairwise comparisons across the two input views. While widely adopted, cost volumes incur substantial memory overhead that scales linearly with the disparity range (full cost volume) or look-up radius (partial cost volume), and are therefore typically constructed and processed at low resolution (e.g.,  $1/4$  of the original resolution). However, low resolution can hurt accuracy, especially for highly detailed image regions.

Stereo matching can be understood as a special case of optical flow that is restricted to the horizontal scan line. Recently, WAF-T [54], a new state-of-art optical flow estimator, demonstrates that cost volumes can be replaced by high-resolution feature-space warping, leading to a substantially simpler design and improved accuracy and efficiency. This raises a natural question: *can cost volumes be replaced by warping also for stereo matching?*

In this work we answer this question in the affirmative. We introduce WAF-T-Stereo, a warping-based stereo matching method that achieves a new state of the



**Fig. 1:** WAFT-Stereo achieves strong sim-to-real generalization [1, 29, 40, 42].

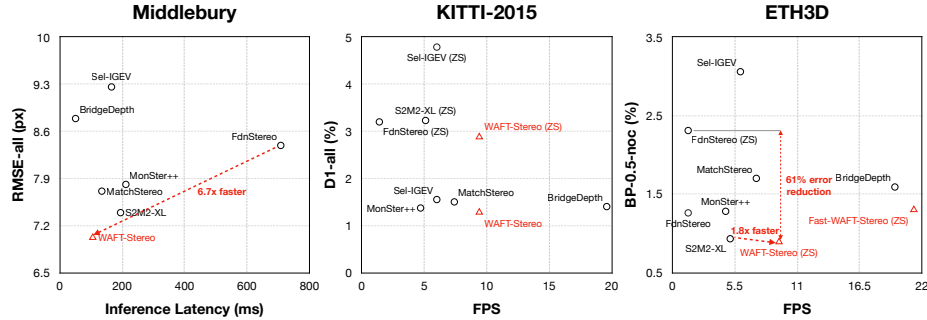
art. WAFT-Stereo demonstrates that strong performance does not require cost-volume-specific designs. Instead, high-resolution warping paired with iterative updates [23, 47, 54] are sufficient.

WAFT-stereo is based on WAFT, but involves non-trivial modifications. WAFT can be directly applied to stereo matching—just changing the prediction to 1D flow—but this trivial adaptation does not work well in practice. WAFT uses recurrent iterative updates which regress updates to flow vectors. This works well for small displacements between adjacent frames of continuous video, but can struggle to converge for very large displacements (hundreds of pixels) typical for high resolution stereo pairs.

WAFT-Stereo addresses the issue of large displacements by introducing a classification module before recurrent updates. The classification module predicts disparity discretized into a predefined set of bins. The classification module uses the same architecture as the recurrent updater. This is a small change to the prediction head and extremely simple to implement, yet it significantly improves accuracy at negligible additional cost.

In addition, WAFT-Stereo introduces several architectural improvements over WAFT. WAFT includes a small U-Net that serves as a lightweight adaptation layer for a pre-trained input encoder. We remove this U-Net by instead fine-tune the pre-trained encoder with LoRA [17], which reduces latency. Also, we replace the high-resolution skip connection with several ResNet blocks in the recurrent update module, which significantly improves accuracy.

WAFT-Stereo achieves state of the art accuracy on standard benchmarks. This was achieved using the same DAv2-L backbone [66] used by leading prior stereo matching approaches (e.g., FoundationStereo [57], DEFOM-Stereo [20], and Monster++ [7]). WAFT-Stereo reduces BP-0.5 by 6% on ETH3D [42], BP-



**Fig. 2:** WAFt-Stereo achieves state-of-the-art performance on Middlebury [40], KITTI-2015 [29], and ETH3D [42] public benchmarks. Our best-performing model reduces the zero-shot error on ETH3D by at least 61%, while being 1.8 – 6.7 $\times$  faster than leading methods [30, 57]. Our real-time model can process 540p stereo pairs at 21 FPS, while maintaining competitive performance. ‘ZS’ denotes zero-shot submissions.

2 by 13% on KITTI-2012 [10], D1 by 6% on KITTI-2015 [29], and RMSE by 6% on Middlebury [40].

WAFt-Stereo is highly efficient. It processes 540p stereo pairs at 10 FPS on an NVIDIA L40 GPU, representing a 1.8–6.7 $\times$  speedup over leading methods (e.g., 1.8 $\times$  over S2M2-XL [30] and 6.7 $\times$  over FoundationStereo [57]). With a smaller backbone, DAv2-S [66], WAFt-Stereo reaches 21 FPS on 540p inputs while maintaining accuracy comparable to prior state-of-the-art methods.

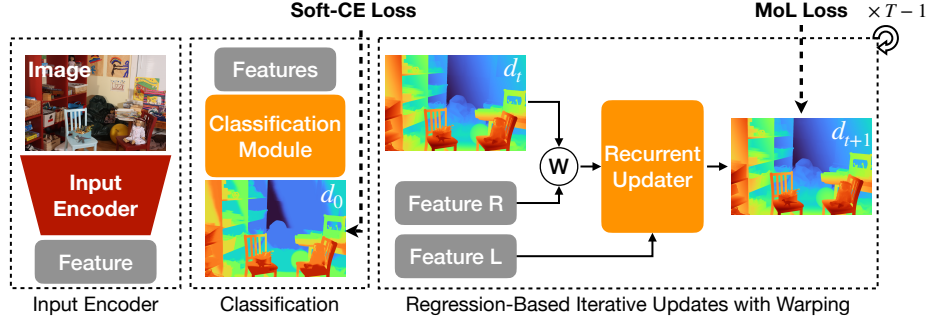
WAFt-Stereo also exhibits strong generalization, when evaluated in the zero-shot setting. Trained exclusively on synthetic data, it achieves the best BP-0.5 on the ETH3D benchmark among all existing submissions, corresponding to an 81% error reduction over the strongest established zero-shot baseline [57]. On KITTI-2015, WAFt-Stereo reduces D1 by 9% relative to leading zero-shot methods [30, 57].

Our main contributions are two-fold: (1) we show that cost volumes are not necessary for strong performance in stereo matching, and (2) we introduce WAFt-Stereo, a fully warping-based architecture with state-of-the-art accuracy, high efficiency, and strong generalization.

## 2 Related Work

**Deep Stereo Matching** Modern stereo matching methods are largely driven by deep learning [6, 7, 11–13, 20, 21, 23, 25, 30, 31, 52, 56, 57, 59–61, 67]. While several methods [25, 56] formulate stereo matching as a direct dense prediction problem, most leading methods [7, 11, 30, 57] adopt the RAFT paradigm [23, 47], iteratively refining predictions by regressing disparity updates.

WAFt-Stereo follows the iterative refinement framework. However, unlike many existing iterative methods [7, 23, 57], WAFt-Stereo performs one-step classification of discretized disparity before regression-based iterative updates. Prior



**Fig. 3:** WAFT-Stereo consists of three parts: (1) an input encoder that extracts features from images; (2) a classification step that estimates probabilities over preset disparity bins, supervised by a soft-cross-entropy loss; and (3) a recurrent updater that takes backward-warped right view features as input and regresses disparity updates, supervised by a Mixture-of-Laplace loss [55] for  $T - 1$  steps.

work has used cost volumes to perform classification because cost volume values can naturally represent visual similarities and therefore matching probabilities after softmax normalization [11, 30, 35, 46, 68]. However, WAFT-Stereo demonstrates that the benefit of classification can be realized just by the classification formulation itself, without using cost volumes or cost-volume-specific designs.

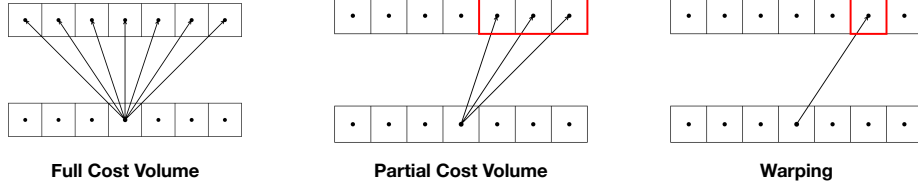
**Cost Volume & Warping** Cost volumes originated from classical stereo matching methods [16, 41], and have become a standard component in modern optical flow and stereo matching methods [7, 23, 30, 45, 57, 67]. Many recent works [7, 11, 23, 30, 47, 57] focus on stronger cost-volume-specific designs for stereo matching.

Warping, in contrast, was primarily studied in the context of optical flow [2, 3, 28]. Warping is similar, although not identical, to a special case of partial cost volumes [45] with look-up radius 1. Despite its simplicity, warping-based designs received much less attention in the past several years, largely due to the strong empirical performance of cost volumes. Recently, WAFT [54] revisited the warping formulation in optical flow. It demonstrated that purely warping-based methods can surpass cost-volume-based methods in both accuracy and efficiency.

WAFT-Stereo builds on the warping framework introduced in WAFT [54]. It adopts a fully warping-based design, removing the reliance on cost-volume-specific network designs used in leading stereo matching methods [7, 11, 20, 23, 30, 57]. This simple warping framework uses standard off-the-shelf architecture components and provides strong performance along with high efficiency.

**Stereo Matching Data** High-quality training data is critical for modern stereo matching methods. However, annotating real-world stereo pairs with accurate ground-truth disparity is expensive and technically challenging. As a result, the total size of publicly available real-world datasets [1, 10, 29, 38, 40] is three orders of magnitude smaller than synthetic datasets [4, 5, 22, 26, 27, 34, 36, 37, 48, 49, 51,





**Fig. 4:** Full cost volumes compute matching costs for all disparity candidates; partial cost volumes compute costs only in a small window around the current disparity estimate; warping aligns the target feature using the current estimate and concatenates aligned and reference features, without computing the matching cost.

[57, 62, 65]. To mitigate the sim-to-real gap, most leading methods adopt mixed training strategies that combine synthetic and real data.

In contrast, WAFT-Stereo is able to outperform leading approaches on public leaderboards when trained exclusively on synthetic data, highlighting its strong generalization capability.

### 3 Method

In this section, we first review the warping-based iterative refinement in WAFT [54], explain why it is advantageous compared to the cost-volume counterpart, and describe its implementation in WAFT-Stereo. We then introduce additional technical components, including the proposed classification step.

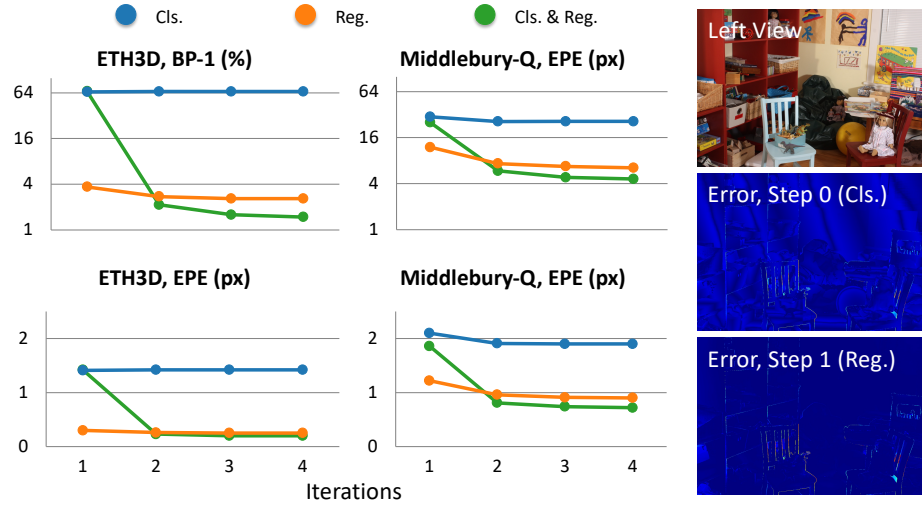
#### 3.1 Iterative Refinement with Warping

Given a rectified stereo pair  $I_L, I_R \in \mathbb{R}^{H \times W \times 3}$ , stereo matching aims to estimate a dense disparity field  $d \in \mathbb{R}^{H \times W}$  that maps each pixel in  $I_L$  to its corresponding location in  $I_R$  along the horizontal epipolar line. Modern iterative methods [7, 30, 57] typically extract image features, construct a cost volume at reduced resolution (commonly 1/4 scale), and recurrently refine the disparity by indexing into the precomputed cost volume.

In optical flow, WAFT [54] replaces cost-volume indexing with feature-space warping: at each iteration, it backward-warps the target-view feature map using the current flow estimate and feeds the aligned features into the recurrent update. This mechanism transfers naturally to stereo matching, since disparity can be viewed as the 1D horizontal component of optical flow.

Let  $F(I_L), F(I_R) \in \mathbb{R}^{h \times w \times c}$  denote the extracted feature maps,  $d_{\text{cur}} \in \mathbb{R}^{h \times w}$  be the current disparity estimate at the same resolution. The backward warping operator  $\text{Warp} : \mathbb{R}^{h \times w} \rightarrow \mathbb{R}^{h \times w \times c}$  applied to  $F(I_R)$  is defined as

$$\text{Warp}(d_{\text{cur}})_p = F(I_R)_{(p_h, p_w - (d_{\text{cur}})_p)},$$



**Fig. 5:** Left: combining classification and regression achieves better performance than using either one alone. Right: the classification step provides a rough estimate, which is later refined by regressions.

where  $p = (p_h, p_w)$  is a pixel in the left view. In practice,  $\text{Warp}(d_{\text{cur}})$  is computed via bilinear sampling and concatenated with  $F(I_L)$  as input to the next refinement iteration.

Warping offers two key advantages over cost volumes. First, its computation and memory scale linearly with the spatial resolution with no dependency on the disparity range, enabling high-resolution indexing [54] that improves accuracy. Second, warping eliminates the need for cost-volume-specific designs, which are typically computationally expensive [7, 57, 59] in practice. Benefiting from a simpler and more standard network design, WAFT-Stereo can process 540p input at 10 FPS, representing a  $6.7\times$  speedup over FoundationStereo [57] and a  $1.8\times$  speedup over S2M2-XL [30].

### 3.2 Classification before Regression

In optical flow, many iterative methods [18, 19, 32, 54, 55] start from an all-zero flow field and recurrently regress residual updates. This strategy works well because typical flow magnitudes are small between consequent frames. In stereo matching, however, large displacements that span hundreds of pixels are common, making purely regression-based iterative updates harder to predict and slower to converge. This is consistent with our observation that regression-only iterative stereo methods [7, 57, 60] often require a relatively large number of refinement iterations at inference time (e.g., 32), which degrades efficiency.

Classification is known to be more effective than regression in many vision tasks. In stereo matching, cost volumes naturally represent matching probabil-

ities over disparity candidates by computing the feature similarity. Several existing methods [11, 30, 68] have performed this classification using cost volumes: BridgeDepth [11] directly supervises the softmax-normalized cost volume with a cross-entropy loss, and derives the initial disparity estimate via soft argmax. S2M2 [30] applies an optimal transport algorithm based on cost volumes to obtain an initial disparity estimate, supervised by the Probabilistic Mode Concentration (PMC) loss. These methods typically require fewer iterations (e.g., 4) in inference than regression-only iterative methods.

WAF-T-Stereo shows that the key reason for faster convergence is the classification formulation itself rather than the use of cost volumes or the associated specialized designs. In particular, our classifier can share the same network design as the regression counterpart (a standard vision transformer [8]), while operating on warped features instead of cost volumes.

In the following, we introduce our implementation of the classification step. Denote the maximum disparity as  $D_{max}$ , the number of disparity bins as  $B$ , the ground-truth disparity as  $d_{gt} \in \mathbb{R}^{H \times W}$ . We define the soft target distribution  $P_{gt}$  for each pixel  $p$  in  $I_L$  as:

$$(P_{gt})_p = \text{Softmax}\left(-\left|(d_{gt})_p - i \times \frac{D_{max}}{B-1}\right|, \quad i = 0, \dots, B-1\right)$$

WAF-T-Stereo predicts the distribution  $P$  in its first step and supervises it with the soft cross-entropy loss (equivalently, KL divergence up to a constant):

$$(L_{cls})_p = - \sum_{i=0}^{B-1} (P_{gt})_{p,i} \log P_{p,i}$$

As shown in Table 5, in the same 4-iteration setting, replacing the first regression-based iteration with classification substantially improves accuracy with negligible additional cost. Figure 5 further illustrates the mechanism: the initial classification provides a coarse but stable estimate that is subsequently refined by the regression-based updates, yielding better final accuracy than the regression-only variant. Although regression can have a better initialization on average, it gets some pixels (those with large displacements) very wrong and will be stuck for them. Classification plus regression solves this issue.

We show additional results on ETH3D (train) to support the claims above. The first step of ‘‘Cls.& Reg’’ achieves 0.89 EPE for pixels with 40px+ disparity and 1.80 EPE for 50px+, better than 1.11 EPE and 2.88 EPE from the first step of ‘‘Reg’’.

### 3.3 Implementation

As shown in Figure 3, the overall architecture of WAF-T-Stereo largely follows WAF-T [54], with an additional decoder for the initial classification step. Below we describe each component, highlighting the simplifications and improvements over the original WAF-T design.

**Input Encoder** The input encoder  $F$  extracts features from the rectified stereo pair  $(I_L, I_R)$ . Given a pretrained backbone, WAFT-Stereo freezes the backbone weights and fine-tunes the model using low-rank adapters (LoRA) [17], instead of attaching a side-tunable small U-Net as in WAFT. We use a fully trainable DPT head [39] to upsample low-resolution features. We ablate different backbone choices [43, 53, 66] in Tables 4 and 5. Following leading stereo methods [7, 11, 20, 57], we use DepthAnythingV2-L [66] for all benchmark submissions.

**Regression-Based Recurrent Updater** At iteration  $t$ , the recurrent updater takes the left-view feature  $F(I_L)$ , the warped right-view feature  $\text{Warp}(d_t)$ , and the hidden state  $h_t \in \mathbb{R}^{h \times w \times c}$  as input, and outputs the next hidden state  $h_{t+1}$ . We adopt the same overall design as WAFT: a ViT-small [8] followed by a DPT [39] feature upsampler. Compared to WAFT, we replace the high-resolution skip connection between hidden states with ResNet blocks, which substantially improves accuracy (see Table 5).

**Classification Module** The classification module shares the same architecture as the recurrent updater (ViT-S + DPT). Given the predicted disparity distribution  $P$ , we compute the initial disparity estimate via soft-argmax and use it to perform the first warping operation in the regression stage:

$$d_0 = \sum_{i=0}^{B-1} \left( i \cdot \frac{D_{\max}}{B-1} \right) P_i.$$

**Prediction Head & Loss** We use the Mixture-of-Laplace (MoL) loss [55] for the regression-based iterative updates. The initial distribution  $P$  and the MoL parameters are predicted from the hidden state  $h_t$  using an MLP, and are then upsampled to the input resolution via convex upsampling [47]. The total training objective sums the classification loss and the discounted MoL losses across iterations:

$$L_{\text{total}} = L_{\text{cls}} + \sum_{i=1}^{T-1} \gamma^{T-i} L_{\text{MoL}}^i,$$

where  $T$  is the total number of iterations (including the initial classification step),  $\gamma$  is a decay factor, and  $L_{\text{MoL}}^i$  is the MoL loss at regression iteration  $i$ .

**Improvements over Existing Methods** WAFT-Stereo can be viewed as a meta-architecture that simplifies the network design for stereo matching. It combines a lightweight warping operator with standard building blocks (e.g., ViTs and ResNet-style modules), allowing it to leverage well-optimized kernels while avoiding the substantial memory overhead of 3D cost volumes [57, 59]. Moreover, the accuracy of WAFT-Stereo consistently improves with larger backbones and increased training data, highlighting its strong scalability.

| Method                | Zero-Shot | ETH3D (test) |             |             |             |             |             |
|-----------------------|-----------|--------------|-------------|-------------|-------------|-------------|-------------|
|                       |           | BP-0.5-noc↓  | BP-1-noc↓   | BP-2-noc↓   | BP-0.5-all↓ | BP-1-all↓   | BP-2-all↓   |
| RAFT-Stereo [47]      | ✗         | 7.04         | 2.44        | 0.44        | 7.33        | 2.60        | 0.56        |
| CREStereo [22]        | ✗         | 3.58         | 0.98        | 0.22        | 3.75        | 1.09        | 0.29        |
| CroCo-Stereo [56]     | ✗         | 3.27         | 0.99        | 0.39        | 3.51        | 1.14        | 0.50        |
| Selective-IGEV [52]   | ✗         | 3.06         | 1.23        | 0.22        | 3.46        | 1.56        | 0.51        |
| BridgeDepth [11]      | ✗         | 1.59         | 0.38        | 0.11        | 1.76        | 0.50        | 0.16        |
| MonSter++ [7]         | ✗         | 1.28         | 0.25        | 0.09        | 1.59        | 0.45        | 0.23        |
| MatchStereo [63]      | ✗         | 1.70         | 0.70        | 0.17        | 1.90        | 0.79        | 0.20        |
| DEFOM-Stereo [20]     | ✗         | 2.21         | 0.70        | 0.14        | 2.43        | 0.78        | 0.19        |
| FoundationStereo [57] | ✗         | 1.26         | 0.26        | 0.08        | 1.56        | 0.48        | 0.26        |
| S2M2-XL [30]          | ✗         | 0.93         | <b>0.22</b> | <b>0.06</b> | 1.03        | <b>0.26</b> | 0.08        |
| FoundationStereo [57] | ✓         | 2.31         | 1.52        | -           | -           | -           | -           |
| Ours (DAv2S, 4)       | ✓         | 1.30         | 0.31        | 0.11        | 1.42        | 0.36        | 0.13        |
| Ours (DAv2L, 5)       | ✓         | <b>0.89</b>  | <b>0.28</b> | 0.07        | <b>0.97</b> | 0.31        | <b>0.07</b> |

**Table 1:** Trained on synthetic data, WAFT-Stereo achieves the best BP-0.5 and BP-2-all on ETH3D benchmark, demonstrating its strong sim-to-real generalization. This also represents a 61%–81% error reduction on the best established zero-shot results.

## 4 Experiments

In this section, we describe the training pipeline of WAFT-Stereo, analyze the results of benchmark submissions, and ablate our design choices, including the training data we used.

**Training Pipeline & Datasets** WAFT-Stereo is first trained on synthetic data for zero-shot evaluation and submissions, then fine-tuned on real-world data for final submissions. The synthetic datasets we used are SceneFlow [26], FallingThings [49], FSD [57], TartanAir [51], TartanGround [34], Spring [27], CREStereo [22], Sintel [4], Virtual KITTI 2 [5], UnrealStereo4K [48], WMGStereo [62], and HR-VS [64]. The real-world datasets we used for fine-tuning are Booster [38], InStereo2K [1], KITTI [10, 29], and Middlebury [40].

We report results on ETH3D [42], KITTI-2012/2015 [10, 29], and Middlebury [40] public benchmarks, and conduct ablations on the training splits of ETH3D [42], KITTI [10, 29], and Middlebury [40].

**Metrics** We report commonly used evaluation metrics, including (1) **BP-X**: the percentage of pixels whose error exceeds X pixels, (2) **D1**: the percentage of pixels whose error exceeds 3 pixels and 5% of the ground-truth disparity, and (3) **RMSE**: the root mean square error.

**Implementation Details** Following WAFT [54], we perform warping at half resolution, i.e.,  $(h, w) = (\frac{H}{2}, \frac{W}{2})$ , and apply an  $8 \times 8$  patchifier before the ViT blocks in the recurrent updater. We use 4 ResNet blocks for high-resolution processing within the updater. For the input encoder, we set the LoRA [17] rank to 8. We use  $B = 40$  preset disparity bins for classification with a maximum disparity  $D_{max} = 800$ .

For brevity, we denote a WAFT-Stereo configuration by its pretrained backbone and the number of iterations, e.g., (DAv2-L, 5).

#### 4.1 Training on Synthetic Data

In the first stage, we train WAFT-Stereo exclusively on a mixture of synthetic stereo datasets [4, 5, 22, 26, 27, 34, 48, 49, 51, 57, 62, 64]. This training mixture (denoted as **SynLarge**) contains approximately 7.7 million stereo pairs in total. We train on 480p random crops with batch size 32 and learning rate  $5 \times 10^{-4}$  for 400k steps, using AdamW optimizer [24] with OneCycle scheduler [44].

Unless otherwise specified, we use this synthetic-only checkpoint for all zero-shot evaluations and benchmark submissions. As we will show later, synthetic-only training already yields strong performance and strong sim-to-real generalization for WAFT-Stereo.

#### 4.2 ETH3D

We report zero-shot results on ETH3D public benchmark using the model trained exclusively on synthetic data (see Section 4.1). We follow the benchmark protocols and report BP-0.5, BP-1, and BP-2 on both all pixels and non-occluded pixels (denoted as ‘all’ and ‘noc’ suffixes).

As shown in Table 1, WAFT-Stereo ranks first on BP-0.5 and BP-2-all among all public submissions, demonstrating its high accuracy. WAFT-Stereo is also more efficient than leading methods [7, 30, 57]. Compared to FoundationStereo [57], WAFT-Stereo achieves a  $6.7\times$  lower latency and uses  $5.2\times$  fewer MACs. Compared to S2M2-XL [30], WAFT-Stereo achieves  $1.8\times$  lower latency and uses  $2.7\times$  fewer MACs. Compared to MonSter++ [7], WAFT-Stereo achieves  $2.0\times$  lower latency and uses  $2.1\times$  fewer MACs.

WAFT-Stereo also demonstrates strong sim-to-real generalization. Compared to the strongest established zero-shot baseline, FoundationStereo [57], our best-performing model reduces BP-0.5 by 61% and BP-1 by 81% on non-occluded pixels. Moreover, our fastest variant, built on the DepthAnythingV2-S backbone [66], achieves 44% lower BP-0.5 and 80% lower BP-1 on non-occluded pixels, while maintaining real-time speed, processing 540p stereo pairs at more than 21 FPS.

#### 4.3 KITTI-2012/2015

We report both fine-tuned and zero-shot results on KITTI. For fine-tuned submissions, we further train the synthetic-pretrained model on the KITTI training splits using  $372 \times 1240$  crops for 3k steps, with batch size 16 and learning rate  $10^{-4}$ . We follow the standard evaluation protocols, report D1 for KITTI-2015 and BP-2 for KITTI-2012, each on both all pixels and non-occluded pixels.

As shown in Table 2, WAFT-Stereo achieves the best BP-2 on KITTI-2012 and the best D1 on KITTI-2015, representing a 13% and a 6% error reduction



| Method                | Zero-Shot | KITTI-12 (test) |             | KITTI-15 (test) |             |             |             |
|-----------------------|-----------|-----------------|-------------|-----------------|-------------|-------------|-------------|
|                       |           | BP-2-noc↓       | BP-2-all↓   | D1-bg-noc ↓     | D1-noc ↓    | D1-bg-all ↓ | D1-all↓     |
| GWCNet [14]           | ✗         | 2.16            | 2.71        | 1.61            | 1.92        | 1.74        | 2.11        |
| RAFT-Stereo [23]      | ✗         | 1.44            | 1.69        | 1.58            | 1.82        | 1.92        | 2.42        |
| CREStereo [22]        | ✗         | 1.72            | 2.18        | 1.33            | 1.54        | 1.45        | 1.69        |
| DLNR [68]             | ✗         | -               | -           | 1.42            | 1.61        | 1.60        | 1.76        |
| MoCha-Stereo [6]      | ✗         | 1.51            | 1.91        | 1.17            | 1.40        | 1.27        | 1.49        |
| Selective-IGEV [52]   | ✗         | 1.59            | 2.05        | 1.22            | 1.44        | 1.33        | 1.55        |
| Croco-Stereo [56]     | ✗         | -               | -           | 1.30            | 1.51        | 1.38        | 1.59        |
| NMRF [68]             | ✗         | 1.59            | 2.07        | 1.18            | 1.46        | 1.28        | 1.57        |
| BridgeDepth [11]      | ✗         | 1.32            | 1.65        | 1.05            | 1.31        | 1.13        | 1.40        |
| DEFOM-Stereo [20]     | ✗         | 1.43            | 1.79        | 1.15            | 1.33        | 1.25        | 1.41        |
| MatchStereo [63]      | ✗         | 1.63            | 2.08        | 1.23            | 1.40        | 1.34        | 1.50        |
| MonSter++ [7]         | ✗         | 1.30            | 1.70        | 1.02            | 1.29        | 1.12        | 1.37        |
| Ours (DAv2L, 5)       | ✗         | <b>1.18</b>     | <b>1.47</b> | <b>0.98</b>     | <b>1.21</b> | <b>1.05</b> | <b>1.28</b> |
| Selective-IGEV [52]   | ✓         | -               | -           | 2.91            | 4.57        | 3.06        | 4.79        |
| FoundationStereo [57] | ✓         | -               | -           | 2.48            | 3.08        | 2.59        | 3.20        |
| S2M2-XL [30]          | ✓         | -               | -           | 2.51            | 3.12        | 2.61        | 3.23        |
| Ours (DAv2L, 5)       | ✓         | -               | -           | <b>2.39</b>     | <b>2.79</b> | <b>2.47</b> | <b>2.88</b> |

**Table 2:** WAFT-Stereo ranks first on KITTI public benchmarks, reducing BP-2 by 13% on KITTI-2012 and D1 by 6% on KITTI-2015. It also outperforms leading methods when not fine-tuned on the training split, reducing D1 by 9%.

on previous state-of-the-art method, MonSter++ [7], respectively. We achieve this with a  $2.0\times$  speed-up and the same DepthAnythingV2-L [66] backbone.

Since prior work has observed strong dataset bias on KITTI [30], we additionally report zero-shot performance on KITTI-2015. Trained exclusively on synthetic data, WAFT-Stereo achieves the best zero-shot result, reducing D1 by at least 10% compared to previous leading methods [30, 52, 57].

#### 4.4 Middlebury

We report fine-tuned results on this benchmark. Starting from the synthetic-pretrained checkpoint, we fine-tune WAFT-Stereo on a mixture of real-world and synthetic data [1, 22, 27, 38, 40, 48, 49, 57, 64] using  $540 \times 960$  crops for 50k steps, with batch size 16 and learning rate  $2 \times 10^{-4}$ . We follow the standard protocol and report BP-2, BP-4, and RMSE on both all pixels and non-occluded pixels.

As shown in Table 3, WAFT-Stereo achieves the best RMSE, improving over previous methods by 5%. With the same DepthAnythingV2-L [66] backbone, it matches or surpasses FoundationStereo [57], MonSter++ [7], and DEFOM-Stereo [20] on BP-4 while being  $6.7\times$ ,  $2.0\times$ , and  $1.5\times$  faster, respectively.

However, WAFT-Stereo falls behind the leading methods [30, 57, 63] in terms of the averaged BP-2 metric across all 15 scenes. We observe that this performance gap is largely dominated by a single particularly challenging scene, Classroom2E, where the stereo pair exhibits substantial illumination differences. Excluding this scene significantly narrows the gap.

| Method                | Middlebury (test) |             |             |             |             |             |
|-----------------------|-------------------|-------------|-------------|-------------|-------------|-------------|
|                       | BP-2-noc↓         | BP-2-all↓   | BP-4-noc↓   | BP-4-all↓   | RMSE-noc↓   | RMSE-all↓   |
| RAFT-Stereo [23]      | 4.74              | 9.37        | 2.75        | 6.42        | 8.41        | 12.6        |
| CREStereo [22]        | 3.71              | 8.13        | 2.04        | 5.05        | 7.70        | 10.5        |
| CroCo-Stereo [56]     | 7.29              | 11.1        | 4.18        | 6.75        | 8.91        | 10.6        |
| GMStereo [61]         | 7.14              | 11.7        | 2.96        | 6.07        | 6.45        | 8.03        |
| MoCha-Stereo [6]      | 3.51              | 6.74        | 2.34        | 4.54        | 7.69        | 9.91        |
| Selective-IGEV [52]   | 2.51              | 6.04        | 1.36        | 3.74        | 7.26        | 9.26        |
| MonSter++ [7]         | 2.60              | 6.40        | 1.18        | 3.52        | 6.03        | 7.81        |
| DEFOM-Stereo [20]     | 2.39              | 5.02        | 1.22        | 3.02        | 5.81        | 7.73        |
| BridgeDepth [11]      | 3.78              | 6.92        | 1.61        | 3.48        | 7.16        | 8.79        |
| MatchStereo [63]      | 1.85              | 4.90        | 0.91        | 3.02        | 5.94        | 7.71        |
| FoundationStereo [57] | 1.84              | 4.26        | 1.04        | 2.72        | 6.48        | 8.39        |
| S2M2-XL [30]          | <b>1.15</b>       | <b>2.94</b> | <b>0.54</b> | <b>1.63</b> | 6.01        | 7.39        |
| Ours (DAv2L, 5)       | 2.53              | 4.39        | 1.05        | 2.15        | <b>5.61</b> | <b>7.02</b> |

**Table 3:** WRAFT-Stereo achieves the best RMSE on Middlebury public benchmark, reducing the error by 6% on non-occluded pixels and 5% on all pixels.

Despite this, WRAFT-Stereo remains on the Pareto frontier when considering accuracy–efficiency trade-offs. Moreover, when performing scene-wise analysis, WRAFT-Stereo achieves the best BP-2-all on four, the best BP-2-noc on three, the best BP-4-all on six, and the best BP-4-noc on three out of the total 15 scenes, demonstrating its strong competitiveness.

#### 4.5 Ablations on Training Data

The strong performance of WRAFT-Stereo does not stem merely from larger-scale training data. In Table 4, we ablate the training data and demonstrate that the WRAFT-Stereo architecture achieves comparable or superior accuracy to leading methods [7, 11, 57] under the same or even fewer training data. We report standard metrics on the training splits of KITTI [10, 29], ETH3D [42], and Middlebury [40], following existing work [7, 11, 57, 59].

In the first block, we follow the zero-shot setting of FoundationStereo [57] (denoted as “FSD-ZS”), restricting the training data to a mixture of synthetic datasets [4, 5, 22, 26, 49] and a single real-world dataset [1]. Under this setting, WRAFT-Stereo achieves performance comparable to FoundationStereo while delivering a  $6.7\times$  speed-up and  $5.3\times$  fewer MACs.

The second block corresponds to the standard zero-shot setting of MonSter++ [7], which uses a larger collection of training datasets [1, 4, 5, 21, 22, 26, 27, 33, 38, 48–51, 57, 58, 64, 69] compared to the FSD-ZS setting. Even with fewer data, WRAFT-Stereo achieves significantly better performance across all metrics, together with a  $2.0\times$  speed-up and  $2.1\times$  fewer MACs.

The third block compares WRAFT-Stereo with BridgeDepth [11], the best zero-shot method under the commonly reported SceneFlow-only training setting [26]. Using the same DepthAnythingV2-L backbone [66], WRAFT-Stereo significantly outperforms BridgeDepth when trained on 7.7 million diverse synthetic

| Data      | Method                        | KITTI-12 (train) |         | KITTI-15 (train) |         | ETH3D (train) |           | Middlebury-Q (train) |           | Latency |
|-----------|-------------------------------|------------------|---------|------------------|---------|---------------|-----------|----------------------|-----------|---------|
|           |                               | D1-all↓          | D1-noc↓ | D1-all↓          | D1-noc↓ | BP-1-all↓     | BP-1-noc↓ | BP-2-all↓            | BP-2-noc↓ |         |
| FSD-ZS    | FoundationStereo [57]         | -                | 2.30    | -                | 2.80    | -             | 0.50      | -                    | 1.30      | 708ms   |
|           | Ours (DAv2L, 5)               | 2.54             | 2.36    | 2.93             | 2.85    | 0.60          | 0.52      | 2.29                 | 1.38      | 106ms   |
| FSD-ZS+   | MonSter++ [7]                 | 2.95             | 2.69    | 3.25             | 3.16    | 1.08          | 0.89      | 4.19                 | 2.93      | 212ms   |
| SynLarge  | BridgeDepth [11] <sup>†</sup> | 2.81             | 2.57    | 3.15             | 3.06    | 0.79          | 0.71      | 3.29                 | 1.90      | 51ms    |
|           | Ours (DAv2S, 4)               | 2.80             | 2.58    | 3.20             | 3.08    | 0.86          | 0.76      | 4.14                 | 2.66      | 47ms    |
|           | Ours (DAv2B, 4)               | 2.58             | 2.38    | 2.88             | 2.80    | 0.71          | 0.63      | 2.76                 | 1.67      | 66ms    |
|           | Ours (DAv2L, 5)               | 2.67             | 2.46    | 2.89             | 2.80    | 0.53          | 0.47      | 2.19                 | 1.27      | 106ms   |
| SceneFlow | BridgeDepth [11] <sup>†</sup> | 3.11             | 2.80    | 4.44             | 4.27    | 1.21          | 1.08      | 4.52                 | 3.04      | 51ms    |
|           | Ours (DAv2S, 4)               | 10.1             | 9.78    | 16.8             | 16.6    | 71.0          | 71.1      | 8.86                 | 6.61      | 47ms    |

**Table 4:** We report the zero-shot ablation results on the training splits of KITTI [29], ETH3D [42] and Middlebury-Q [40]. The training data is identified in the first column. WAFT-Stereo architecture achieves comparable or better performance than leading methods with the same or fewer training data, while being  $2.0 - 6.7\times$  faster. <sup>†</sup> denotes our re-implementation of BridgeDepth [11], which provides better accuracy than reported in the original paper.

pairs (see Section 4.1). Consistently, WAFT-Stereo surpasses BridgeDepth across all public benchmarks [10, 29, 40, 42] (see Tables 1 to 3)

However, as shown in the last block, training WAFT-Stereo solely on SceneFlow (35k pairs) leads to severe overfitting to the dataset bias and consequently degrades generalization. These results suggest that performance under the widely adopted SceneFlow-only zero-shot setting may not faithfully reflect the actual capability of a model.

#### 4.6 Ablations and Analysis on Architecture Design

We conduct extensive ablations on our architectural design choices in Table 5. Unless otherwise specified, all models are trained with 480p image crops, batch size 32, and learning rate  $5 \times 10^{-4}$  for 100k iterations.

We use a smaller synthetic subset [22, 26, 51, 57] for training in ablation experiments, containing approximately 1.6 million stereo pairs. We report zero-shot performance on the training splits of ETH3D and Middlebury. The maximum disparity is set to  $D_{\max} = 320$ .

*Number of Disparity Bins* In the initial classification stage, the number of predefined disparity bins should not be too small. We observe a significant performance drop when only 5 bins are used. Considering the trade-off between accuracy and memory efficiency, we choose  $B = 40$  as the default setting.

*Classification & Regression* Combining classification and regression yields the best performance in practice, as shown in Figure 5 and Table 5. Compared to the regression-only variant, the hybrid design consistently achieves better results on both Middlebury and ETH3D.

In contrast, classification primarily provides coarse disparity estimates. The classification-only variant benefits little from iterative refinement and significantly underperforms the other two variants on BP-X metrics.

| Experiment  | Feat.   | #Iters |      | #HR blks | #Bins | ETH3D     | Middlebury-Q | Reg. Loss | #MACs |
|-------------|---------|--------|------|----------|-------|-----------|--------------|-----------|-------|
|             |         | Cls.   | Reg. |          |       | BP-1-all↓ | BP-2-all↓    |           |       |
| Baseline    | DAv2S   | 1      | 3    | 4        | 40    | 1.48      | 4.62         | MoL       | 560G  |
| Disp. bins  | DAv2S   | 1      | 3    | 4        | 5     | 2.46      | 6.45         | MoL       | 560G  |
|             |         |        |      |          | 20    | 1.70      | 5.04         |           |       |
|             |         |        |      |          | 80    | 1.27      | 5.39         |           |       |
| Cls. & Reg. | DAv2S   | 0      | 4    | 4        | 40    | 2.59      | 6.45         | MoL       | 561G  |
|             |         | 4      | 0    |          |       | 66.8      | 26.1         | MoL       | 560G  |
|             |         | C.V.   | 3    |          |       | 1.53      | 4.74         | MoL       | 561G  |
| HR blks     | DAv2S   | 1      | 3    | 0        | 40    | 1.99      | 6.20         | MoL       | 494G  |
| Diff. Feat. | DAv2L   | 1      | 3    | 4        | 40    | 0.80      | 2.63         | MoL       | 2172G |
|             | DINOv3L | 1      | 3    |          |       | 0.87      | 2.80         | MoL       | 2175G |
| Reg. Loss   | DAv2S   | 1      | 3    | 4        | 40    | 2.44      | 6.39         | $L_1$     | 560G  |

**Table 5:** We report the zero-shot ablation results on the training splits of ETH3D and Middlebury-Q. The effect of changes can be identified through comparisons to the first row. See Section 4.6 for more details.

We also implement a classification variant based on cost volumes, following prior work [11, 68]. However, introducing cost volumes does not provide noticeable performance improvements. This observation suggests that the gains mainly stem from the classification formulation itself, rather than from cost-volume-specific design choices.

**High-Resolution Blocks** The high-resolution ResNet-blocks [15] remarkably improve accuracy. This is reasonable since the low-resolution processing after patchification may lose information about high-resolution details.

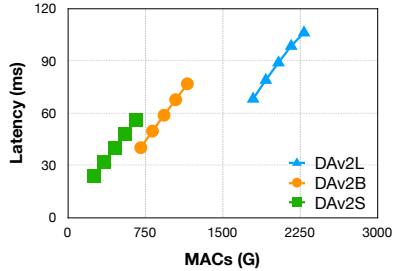
**Different Backbones** WAFT-Stereo can be easily integrated with various pre-trained backbones [43, 66]. For fair comparison with prior work, we adopt the DAv2 series [66], following existing methods [7, 11, 20, 57]. We leave more explorations on backbone choices for future work.

**Regression Loss** As in optical flow, the Mixture-of-Laplace (MoL) loss [55] yields significantly better performance for iterative refinement than the  $L_1$  loss widely adopted in prior work [7, 11, 23, 30, 57].

**Inference Cost** We provide a detailed cost breakdown in Figure 6 and Table 6. All profiling results use 540p inputs, BF16 precision, and an NVIDIA L40 GPU.

WAFT-Stereo is faster than most leading methods, with BridgeDepth [11] as the only faster baseline in our comparison. However, WAFT-Stereo remains on the Pareto frontier of the speed–accuracy trade-off (see Tables 1 to 4).

We further analyze how inference latency grows with an increasing number of iterations. Figure 6 shows that more iterations reduces parallelism, though WAFT-Stereo already requires substantially fewer iterations than prior iterative methods [23, 57, 59]. We leave better designs for future work.



**Fig. 6:** The latency grows almost linearly with the number of iterative updates, reflecting reduced parallelism as the iteration count increases.

| Method           | Latency | #MACs | #Params |
|------------------|---------|-------|---------|
| MonSter++ [7]    | 212ms   | 4.92T | 0.39B   |
| MatchStereo [63] | 135ms   | 0.39T | 0.08B   |
| BridgeDepth [11] | 51ms    | 1.34T | 0.35B   |
| DEFOM [20]       | 163ms   | 4.91T | 0.38B   |
| S2M2-XL [30]     | 195ms   | 6.26T | 0.41B   |
| FdnStereo [57]   | 708ms   | 12.1T | 0.38B   |
| Ours (DAv2S, 4)  | 47ms    | 0.56T | 0.08B   |
| Ours (DAv2B, 4)  | 66ms    | 1.05T | 0.15B   |
| Ours (DAv2L, 5)  | 106ms   | 2.29T | 0.38B   |

**Table 6:** More details on the inference cost of WAF-T-Stereo and leading methods, profiled with 540p input, BF16 precision on an L40.

## Acknowledgements

This work was partially supported by the National Science Foundation.

## References

- Bao, W., Wang, W., Xu, Y., Guo, Y., Hong, S., Zhang, X.: InStereo2k: a large real dataset for stereo matching in indoor scenes. *Science China Information Sciences* **63**, 1–11 (2020) 2, 4, 9, 11, 12
- Black, M.J., Anandan, P.: The robust estimation of multiple motions: Parametric and piecewise-smooth flow fields. *Computer vision and image understanding* **63**(1), 75–104 (1996) 4
- Brox, T., Bruhn, A., Papenberg, N., Weickert, J.: High accuracy optical flow estimation based on a theory for warping. In: *European conference on computer vision*. pp. 25–36. Springer (2004) 4
- Butler, D.J., Wulff, J., Stanley, G.B., Black, M.J.: A naturalistic open source movie for optical flow evaluation. pp. 611–625 (2012) 4, 9, 10, 12
- Cabon, Y., Murray, N., Humenberger, M.: Virtual KITTI 2. *arXiv preprint arXiv:2001.10773* (2020) 4, 9, 10, 12
- Chen, Z., Long, W., Yao, H., Zhang, Y., Wang, B., Qin, Y., Wu, J.: Mocha-stereo: Motif channel attention network for stereo matching. pp. 27768–27777 (2024) 3, 11, 12
- Cheng, J., Liao, W., Cai, Z., Liu, L., Xu, G., Wang, X., Wang, Y., Yuan, Z., Deng, Y., Zang, J., et al.: Monster++: Unified stereo matching, multi-view stereo, and real-time stereo with monodepth priors. *arXiv preprint arXiv:2501.08643* (2025) 1, 2, 3, 4, 5, 6, 8, 9, 10, 11, 12, 13, 14, 15
- Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: *International Conference on Learning Representations (ICLR)* (2021) 7, 8

9. Engel, J., Somasundaram, K., Goesele, M., Sun, A., Gamino, A., Turner, A., Tatlatof, A., Yuan, A., Souti, B., Meredith, B., et al.: Project aria: A new tool for egocentric multi-modal ai research. arXiv preprint arXiv:2308.13561 (2023) [1](#)
10. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the KITTI vision benchmark suite. pp. 3354–3361 (2012) [1](#), [3](#), [4](#), [9](#), [12](#), [13](#)
11. Guan, T., Guo, J., Wang, C., Liu, Y.H.: Bridgedepth: Bridging monocular and stereo reasoning with latent alignment. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 27681–27691 (2025) [1](#), [3](#), [4](#), [7](#), [8](#), [9](#), [11](#), [12](#), [13](#), [14](#), [15](#)
12. Guan, T., Wang, C., Liu, Y.H.: Neural markov random field for stereo matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5459–5469 (2024) [3](#)
13. Guo, X., Zhang, C., Zhang, Y., Wang, R., Nie, D., Zheng, W., Poggi, M., Zhao, H., Ye, M., Zou, Q., et al.: Stereo anything: Unifying zero-shot stereo matching with large-scale mixed data. arXiv preprint arXiv:2411.14053 (2024) [3](#)
14. Guo, X., Yang, K., Yang, W., Wang, X., Li, H.: Group-wise correlation stereo network. pp. 3273–3282 (2019) [11](#)
15. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778 (2016) [14](#)
16. Hosni, A., Rhemann, C., Bleyer, M., Rother, C., Gelautz, M.: Fast cost-volume filtering for visual correspondence and beyond. IEEE transactions on pattern analysis and machine intelligence **35**(2), 504–511 (2012) [4](#)
17. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. Iclr **1**(2), 3 (2022) [2](#), [8](#), [9](#)
18. Huang, Z., Shi, X., Zhang, C., Wang, Q., Cheung, K.C., Qin, H., Dai, J., Li, H.: Flowformer: A transformer architecture for optical flow. In: European conference on computer vision. pp. 668–685. Springer (2022) [6](#)
19. Jahedi, A., Luz, M., Rivinius, M., Bruhn, A.: Ccmr: High resolution optical flow estimation via coarse-to-fine context-guided motion reasoning. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 6899–6908 (2024) [6](#)
20. Jiang, H., Lou, Z., Ding, L., Xu, R., Tan, M., Jiang, W., Huang, R.: Defom-stereo: Depth foundation model based stereo matching. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 21857–21867 (2025) [1](#), [2](#), [3](#), [4](#), [8](#), [9](#), [11](#), [12](#), [14](#), [15](#)
21. Karaev, N., Rocco, I., Graham, B., Neverova, N., Vedaldi, A., Ruppert, C.: DynamicStereo: Consistent dynamic depth from stereo videos. pp. 13229–13239 (2023) [3](#), [12](#)
22. Li, J., Wang, P., Xiong, P., Cai, T., Yan, Z., Yang, L., Liu, J., Fan, H., Liu, S.: Practical stereo matching via cascaded recurrent network with adaptive correlation. pp. 16263–16272 (2022) [4](#), [9](#), [10](#), [11](#), [12](#), [13](#)
23. Lipson, L., Teed, Z., Deng, J.: RAFT-Stereo: Multilevel recurrent field transforms for stereo matching. In: International Conference on 3D Vision (3DV). pp. 218–227 (2021) [1](#), [2](#), [3](#), [4](#), [11](#), [12](#), [14](#)
24. Loshchilov, I.: Decoupled weight decay regularization. ICLR (2019) [10](#)
25. Mayer, N., Ilg, E., Haussler, P., Fischer, P., Cremers, D., Dosovitskiy, A., Brox, T.: A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4040–4048 (2016) [3](#)

26. Mayer, N., Ilg, E., Hausser, P., Fischer, P., Cremers, D., Dosovitskiy, A., Brox, T.: A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. pp. 4040–4048 (2016) [4](#), [9](#), [10](#), [12](#), [13](#)
27. Mehl, L., Schmalfluss, J., Jahedi, A., Nalivayko, Y., Bruhn, A.: Spring: A high-resolution high-detail dataset and benchmark for scene flow, optical flow and stereo. In: Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2023) [4](#), [9](#), [10](#), [11](#), [12](#)
28. Memin, E., Perez, P.: A multigrid approach for hierarchical motion estimation. In: ICCV. vol. 98, pp. 933–938 (1998) [4](#)
29. Menze, M., Geiger, A.: Object scene flow for autonomous vehicles. pp. 3061–3070 (2015) [1](#), [2](#), [3](#), [4](#), [9](#), [12](#), [13](#)
30. Min, J., Jeon, Y., Kim, J., Choi, M.: S2m2: Scalable stereo matching model for reliable depth estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 26729–26739 (2025) [1](#), [3](#), [4](#), [5](#), [6](#), [7](#), [9](#), [10](#), [11](#), [12](#), [14](#), [15](#)
31. Min, J., Kim, J., Min, C.H., Kim, M., Jeon, Y., Choi, M.: Depthfocus: Controllable depth estimation for see-through scenes. arXiv preprint arXiv:2511.16993 (2025) [3](#)
32. Morimitsu, H., Zhu, X., Cesar, R.M., Ji, X., Yin, X.C.: Dpflow: Adaptive optical flow estimation with a dual-pyramid framework. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 17810–17820 (2025) [6](#)
33. Niklaus, S., Mai, L., Yang, J., Liu, F.: 3d ken burns effect from a single image. ACM Transactions on Graphics (ToG) **38**(6), 1–15 (2019) [12](#)
34. Patel, M., Yang, F., Qiu, Y., Cadena, C., Scherer, S., Hutter, M., Wang, W.: Tartanground: A large-scale dataset for ground robot perception and navigation. In: 2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 20524–20531. IEEE (2025) [4](#), [9](#), [10](#)
35. Peng, R., Wang, R., Wang, Z., Lai, Y., Wang, R.: Rethinking depth estimation for multi-view stereo: A unified representation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8645–8654 (2022) [4](#)
36. Raistrick, A., Lipson, L., Ma, Z., Mei, L., Wang, M., Zuo, Y., Kayan, K., Wen, H., Han, B., Wang, Y., et al.: Infinite photorealistic worlds using procedural generation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12630–12641 (2023) [4](#)
37. Raistrick, A., Mei, L., Kayan, K., Yan, D., Zuo, Y., Han, B., Wen, H., Parakh, M., Alexandropoulos, S., Lipson, L., et al.: Infinigen indoors: Photorealistic indoor scenes using procedural generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21783–21794 (2024) [4](#)
38. Ramirez, P.Z., Costanzino, A., Tosi, F., Poggi, M., Salti, S., Mattoccia, S., Di Stefano, L.: Booster: A benchmark for depth from images of specular and transparent surfaces. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) (2023) [4](#), [9](#), [11](#), [12](#)
39. Ranftl, R., Bochkovskiy, A., Koltun, V.: Vision transformers for dense prediction. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 12179–12188 (2021) [8](#)
40. Scharstein, D., Hirschmüller, H., Kitajima, Y., Krathwohl, G., Nešić, N., Wang, X., Westling, P.: High-resolution stereo datasets with subpixel-accurate ground truth. In: Pattern Recognition: 36th German Conference, GCPR 2014, Münster, Germany, September 2-5, 2014, Proceedings 36. pp. 31–42. Springer (2014) [2](#), [3](#), [4](#), [9](#), [11](#), [12](#), [13](#)



41. Scharstein, D., Szeliski, R.: A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International journal of computer vision* **47**(1), 7–42 (2002) [4](#)
42. Schops, T., Schonberger, J.L., Galliani, S., Sattler, T., Schindler, K., Pollefeys, M., Geiger, A.: A multi-view stereo benchmark with high-resolution images and multi-camera videos. pp. 3260–3269 (2017) [2](#), [3](#), [9](#), [12](#), [13](#)
43. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khali-dov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., et al.: Dinov3. *arXiv preprint arXiv:2508.10104* (2025) [8](#), [14](#)
44. Smith, L.N., Topin, N.: Super-convergence: Very fast training of neural networks using large learning rates. In: *Artificial intelligence and machine learning for multi-domain operations applications*. vol. 11006, pp. 369–386. SPIE (2019) [10](#)
45. Sun, D., Yang, X., Liu, M.Y., Kautz, J.: Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 8934–8943 (2018) [1](#), [4](#)
46. Tankovich, V., Hane, C., Zhang, Y., Kowdle, A., Fanello, S., Bouaziz, S.: HITNet: Hierarchical iterative tile refinement network for real-time stereo matching. pp. 14362–14372 (2021) [4](#)
47. Teed, Z., Deng, J.: RAFT: Recurrent all-pairs field transforms for optical flow. pp. 402–419 (2020) [2](#), [3](#), [4](#), [8](#), [9](#)
48. Tosi, F., Liao, Y., Schmitt, C., Geiger, A.: Smd-nets: Stereo mixture density networks. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 8942–8952 (2021) [4](#), [9](#), [10](#), [11](#), [12](#)
49. Tremblay, J., To, T., Birchfield, S.: Falling things: A synthetic dataset for 3d object detection and pose estimation. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. pp. 2038–2041 (2018) [4](#), [9](#), [10](#), [11](#), [12](#)
50. Wang, Q., Zheng, S., Yan, Q., Deng, F., Zhao, K., Chu, X.: IRS: A large naturalistic indoor robotics stereo dataset to train deep models for disparity and surface normal estimation. In: *IEEE International Conference on Multimedia and Expo (ICME)* (2021) [12](#)
51. Wang, W., Zhu, D., Wang, X., Hu, Y., Qiu, Y., Wang, C., Hu, Y., Kapoor, A., Scherer, S.: TartanAir: A dataset to push the limits of visual slam. In: *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. pp. 4909–4916 (2020) [4](#), [9](#), [10](#), [12](#), [13](#)
52. Wang, X., Xu, G., Jia, H., Yang, X.: Selective-Stereo: Adaptive frequency information selection for stereo matching. pp. 19701–19710 (2024) [3](#), [9](#), [11](#), [12](#)
53. Wang, Y., Zhou, J., Zhu, H., Chang, W., Zhou, Y., Li, Z., Chen, J., Pang, J., Shen, C., He, T.: Pi3: Permutation-equivariant visual geometry learning. *arXiv preprint arXiv:2507.13347* (2025) [8](#)
54. Wang, Y., Deng, J.: Waft: Warping-alone field transforms for optical flow. *arXiv preprint arXiv:2506.21526* (2025) [1](#), [2](#), [4](#), [5](#), [6](#), [7](#), [9](#)
55. Wang, Y., Lipson, L., Deng, J.: Sea-raft: Simple, efficient, accurate raft for optical flow. In: *European Conference on Computer Vision*. pp. 36–54. Springer (2024) [4](#), [6](#), [8](#), [14](#)
56. Weinzaepfel, P., Lucas, T., Leroy, V., Cabon, Y., Arora, V., Brégier, R., Csurka, G., Antsfeld, L., Chidlovskii, B., Revaud, J.: CroCo v2: Improved cross-view completion pre-training for stereo matching and optical flow. pp. 17969–17980 (2023) [3](#), [9](#), [11](#), [12](#)

57. Wen, B., Trepte, M., Aribido, J., Kautz, J., Gallo, O., Birchfield, S.: Foundation-stereo: Zero-shot stereo matching. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5249–5260 (2025) [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [8](#), [9](#), [10](#), [11](#), [12](#), [13](#), [14](#), [15](#)
58. Wu, C.Y., Wang, J., Hall, M., Neumann, U., Su, S.: Toward practical monocular indoor depth estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3814–3824 (2022) [12](#)
59. Xu, G., Wang, X., Ding, X., Yang, X.: Iterative geometry encoding volume for stereo matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21919–21928 (2023) [1](#), [3](#), [6](#), [8](#), [12](#), [14](#)
60. Xu, G., Wang, X., Zhang, Z., Cheng, J., Liao, C., Yang, X.: IGEV++: Iterative multi-range geometry encoding volumes for stereo matching. arXiv preprint arXiv:2409.00638 (2024) [3](#), [6](#)
61. Xu, H., Zhang, J., Cai, J., Rezatofghi, H., Yu, F., Tao, D., Geiger, A.: Unifying flow, stereo and depth estimation. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) (2023) [3](#), [12](#)
62. Yan, D., Raistrick, A., Deng, J.: What makes good synthetic training data for zero-shot stereo matching? arXiv preprint arXiv:2504.16930 (2025) [4](#), [9](#), [10](#)
63. Yan, T., Liu, T., Yang, X., Zhao, Q., Xia, Z.: Matchattention: Matching the relative positions for high-resolution cross-view matching. arXiv preprint arXiv:2510.14260 (2025) [9](#), [11](#), [12](#), [15](#)
64. Yang, G., Manela, J., Happold, M., Ramanan, D.: Hierarchical deep stereo matching on high-resolution images. pp. 5515–5524 (2019) [9](#), [10](#), [11](#), [12](#)
65. Yang, G., Song, X., Huang, C., Deng, Z., Shi, J., Zhou, B.: DrivingStereo: A large-scale dataset for stereo matching in autonomous driving scenarios. pp. 899–908 (2019) [4](#)
66. Yang, L., Kang, B., Huang, Z., Xu, X., Feng, J., Zhao, H.: Depth anything: Unleashing the power of large-scale unlabeled data. pp. 10371–10381 (2024) [2](#), [3](#), [8](#), [10](#), [11](#), [12](#), [14](#)
67. Zbontar, J., LeCun, Y.: Computing the stereo matching cost with a convolutional neural network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1592–1599 (2015) [1](#), [3](#), [4](#)
68. Zhao, H., Zhou, H., Zhang, Y., Chen, J., Yang, Y., Zhao, Y.: High-frequency stereo matching network. pp. 1327–1336 (2023) [4](#), [7](#), [11](#), [14](#)
69. Zhou, S., Zhang, J., Zuo, W., Xie, H., Pan, J., Ren, J.S.: Davanet: Stereo deblurring with view aggregation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10996–11005 (2019) [12](#)