This ICCV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

S²M²: Scalable Stereo Matching Model for Reliable Depth Estimation

Junhong Min[1, †], Youngpil Jeon[1], Jimin Kim[1], Minyong Choi[1]
[1]Samsung Electronics, †Corresponding author
{junhong1.min, yfeel.jeon, ji.min.kim, my312choi}@samsung.com

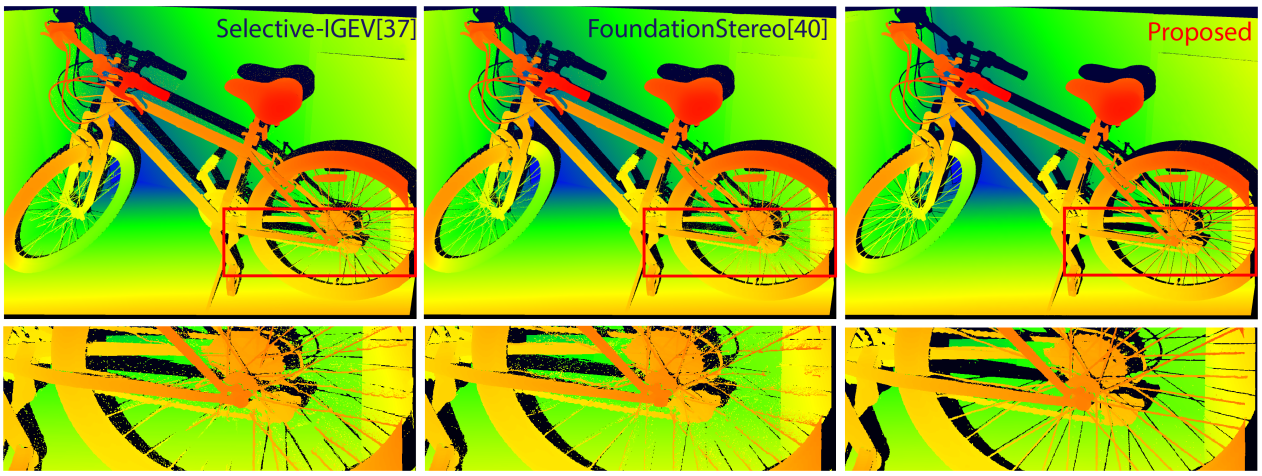


Figure 1. Qualitative comparison of 3D point clouds: Enhanced reliability in fine structures. 3D point cloud visualizations comparing our model against state-of-the-art methods on a challenging sample from the Middlebury v3 test set [28]. Our model yields more reliable reconstructions, especially in delicate structures like bicycle spokes, and reduces over-smoothing effects near edges.

Abstract

The pursuit of a generalizable stereo matching model, capable of performing well across varying resolutions and disparity ranges without dataset-specific fine-tuning, has revealed a fundamental trade-off. Iterative local search methods achieve high scores on constrained benchmarks, but their core mechanism inherently limits the global consistency required for true generalization. However, global matching architectures, while theoretically more robust, have historically been rendered infeasible by prohibitive computational and memory costs. We resolve this dilemma with S²M²: a global matching architecture that achieves state-of-the-art accuracy and high efficiency without relying on cost volume filtering or deep refinement stacks. Our design integrates a multi-resolution transformer for robust long-range correspondence, trained with a novel loss function that concentrates probability on feasible matches. This approach enables a more robust joint estimation of disparity, occlusion, and confidence. S²M² establishes a new state of the art on Middlebury v3 and ETH3D benchmarks, significantly outperforming prior methods in most metrics while reconstructing high-quality details with competitive efficiency.

1. Introduction

Recent advances in deep learning have significantly improved stereo matching accuracy, enabling high-quality depth estimation in well-constrained conditions [1, 4, 18, 20, 23, 31, 32, 37, 42]. However, ensuring reliable performance across diverse environments and image resolutions remains a challenge. A robust stereo matching model should generalize across different conditions (e.g., indoor/outdoor, low/high texture scenes) without requiring dataset-specific tuning. This pursuit of generalization often involves increasing model capacity for improved accuracy, but such gains do not always scale proportionally with size, requiring carefully designed architectures and training strategies. Furthermore, larger models inherently bring higher computational costs, necessitating a balance between efficiency and strong generalization, especially at high resolutions.

The landscape of deep learning-based stereo matching is primarily defined by a trade-off between two dominant

architectural paradigms: global matching and iterative refinement. Transformer-based global matching architectures [19, 23, 38, 39, 45] are particularly well-suited for modeling long-range correspondences to handle occlusions and repetitive patterns. However, they have historically been underexplored because they pose significant efficiency challenges and are difficult to scale, both in processing high-resolution images and in effectively increasing model capacity. This scalability bottleneck led to the dominance of cost volume based iterative refinement methods [18, 20, 42, 49], which demonstrated strong performance on established benchmarks. This paradigm, however, faces two fundamental limitations. First, processing the initial cost volume requires computationally expensive filtering, such as 3D convolutions, which severely hampers scalability. Second, the subsequent iterative refinement, being local in nature, struggles to resolve complex and discontinuous disparity structures, often failing to escape local minima.

To overcome the limitations inherent in both dominant paradigms, we propose the Scalable Stereo Matching Model (S²M²), which revitalizes the global matching approach by directly tackling its well-known scalability issues. S²M² introduces a multi-resolution transformer architecture that enables robust feature interactions both across different scales and within each resolution level, ensuring stable learning as the model scales efficiently. Additionally, an adaptive feature fusion mechanism selectively integrates multi-resolution information, preserving fine details while maintaining global consistency. These combined enhancements allow S²M² to achieve robust depth estimation across varying image resolutions and scene complexities.

In addition to the architecture, we introduce a dedicated loss function designed to guide probability concentration on feasible matches. By explicitly encouraging the network to focus probability mass on valid candidates, our model not only achieves more precise disparity predictions but also jointly produces a corresponding confidence score derived directly from this distribution. This allows for robust filtering of uncertain predictions in ambiguous areas, such as occlusions and textureless surfaces [7, 25], which is crucial for reliable performance in safety-critical applications [10, 17, 24].

The performance of our model was validated on challenging real-world benchmarks, conducting comprehensive comparisons against state-of-the-art methods. As Figures 1 and 2 illustrate, our approach demonstrates strong visual and quantitative performance across diverse real-world datasets [28, 29]. While real-world benchmarks are indispensable for evaluating stereo matching, they inherently present limitations such as restricted disparity ranges or imperfect ground-truth annotations, especially in complex or fine-structured regions. To complement this real-world evaluation, we additionally constructed a high-fidelity synthetic

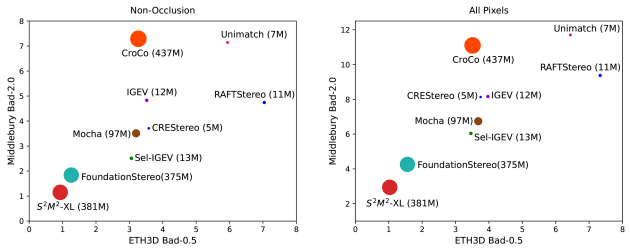


Figure 2. Comprehensive evaluation on two representative benchmarks, ETH3D [29] and Middlebury v3 [28]. Accuracy is measured using the Bad-2.0 metric for Middlebury and Bad-0.5 for ETH3D. Each method is plotted with its corresponding model size (in millions of parameters).

dataset, which enables a more thorough assessment under controlled conditions, including scenes with larger disparity variations and precisely defined ground truth. By achieving state-of-the-art results on both real-world and synthetic benchmarks, our work is the first to conclusively demonstrate that a carefully designed global matching architecture can be scalable in terms of both input resolution and model capacity, overcoming its well-known limitations to achieve robust and state-of-the-art performance.

To summarize, our key contributions include:

- **A Highly Scalable Global Matching Architecture:** A multi-resolution Transformer architecture scalable in terms of both input (high resolutions, large disparities) and model size.
- **Accurate and Reliable Depth Estimation:** A novel loss function that boosts disparity accuracy, while the joint estimation of confidence and occlusion ensures final depth reliability.
- **New SOTA and a Rigorous Validation Framework:** State-of-the-art performance on established real-world benchmarks, and a novel synthetic dataset to validate robustness in scenarios unaddressed by existing benchmarks.

2. Related Works

The era of deep learning in stereo matching began with CNN-based methods [9, 17], which effectively replaced handcrafted components of classical algorithms like SGM [15] with powerful deep feature extractors. While these models significantly improved accuracy by learning robust feature representations, their reliance on local convolutions resulted in limited generalization across diverse domains and resolutions. This limitation became a critical bottleneck, necessitating the exploration of new architectural paradigms.

Modern stereo matching has largely diverged into two parallel pathways: Transformer-based global matching and cost volume-based iterative refinement. The former, Transformer-based methods, leverage self- and cross-attention to estab-

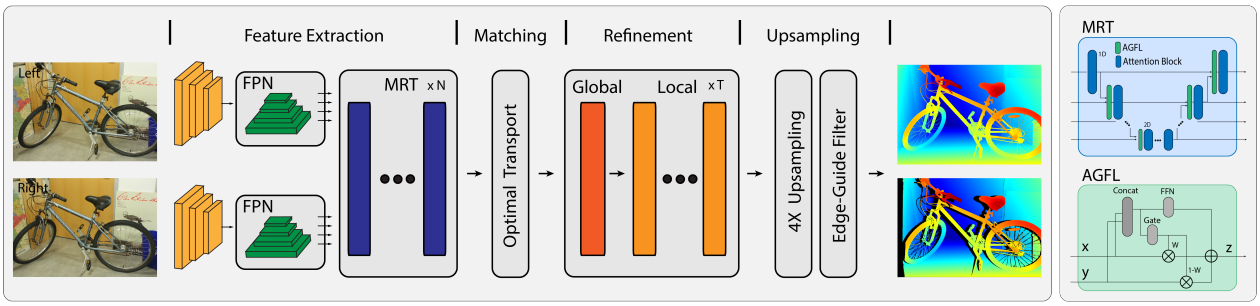


Figure 3. Our proposed architecture, S2M², comprises four main stages: (1) feature extraction, where left and right input images are processed by FPN and MRT to yield multi-resolution feature maps; (2) global matching, which computes all-pairs correlation at 1/4 resolution and estimates disparity, occlusion, and confidence using optimal transport; (3) refinement, iteratively updating these estimates; and (4) upsampling, enhancing depth boundaries to restore full resolution. The feature extraction stage employs a Multi-Resolution Transformer (MRT) for hierarchical feature extraction, incorporating an Adaptive Gated Fusion Layer (AGFL) for efficient multi-scale feature alignment.

lish robust global correspondences, making them particularly effective for handling occlusions and repetitive patterns. An early model, STTR [19], provided foundational insights by framing the task as a sequence-to-sequence problem, though with limited scalability. More recent approaches [23, 38, 39, 45] have refined their global correspondence mechanisms by incorporating techniques from related fields such as dense keypoint matching [27, 30, 35] and point cloud registration [26, 47, 48]. Nonetheless, the quadratic complexity of attention mechanisms remains a critical bottleneck, hindering scalability in terms of both input resolution and model capacity. This presents a key research gap that our work aims to address.

In parallel, cost volume based iterative refinement has emerged as a dominant pathway, demonstrating strong practical performance. This approach was pioneered by RAFT-Stereo [20], which introduced recurrent updates on a learned correlation volume. Subsequent works have advanced this framework through several key strategies. For instance, CREStereo [18] introduced a more robust, hierarchical recurrent refinement to better handle fine details and real-world image distortions. Another line of work, such as IGEV [42], incorporated geometric guidance to improve disparity propagation. More recently, efforts have centered on refining feature representations, with Selective-IGEV [37] selectively integrating frequency information and MoCha-Stereo [5, 6] employing motif-based attention for robustness. Despite these advancements, the performance of these methods hinges on a high-quality correlation volume produced by computationally expensive filtering, and the inherently local nature of this process makes them less effective in globally ambiguous regions with repetitive textures or large discontinuities, reinforcing the need for true global correspondence mechanisms.

Another distinct line of research has focused on computationally efficient methods for real-time applications [2, 41, 43, 44]. These approaches typically achieve efficiency by replacing expensive 3D cost aggregation with lightweight, 2D-based correlation or spatial propagation networks. While they achieve competitive accuracy in low-resolution settings suitable for real-time use, their design often leads to a significant performance degradation on high-resolution imagery. This reveals a fundamental trade-off between efficiency and high-resolution accuracy, a gap that our scalable architecture is designed to bridge.

A recent trend aims to improve generalization by leveraging large-size, pre-trained models. These methods typically build upon the iterative refinement framework but replace the standard feature encoder with a powerful backbone pre-trained on other large-scale tasks. This is often achieved by incorporating a general purpose vision encoder or even a complete monocular depth estimation model as a feature extractor [3, 16, 40, 50]. By transferring rich prior knowledge, these approaches demonstrate impressive generalization capabilities. However, this strategy introduces a significant trade-off: their reliance on massive, often fixed-weight components results in models that are too large and computationally expensive for many practical applications. This highlights the ongoing challenge of achieving both strong generalization and efficiency, a balance our scalable architecture aims to achieve.

3. Method

Our proposed model, S²M², is designed to revitalize the global matching paradigm by addressing its long-standing scalability challenges. To achieve this, our architecture is composed of four main stages, as illustrated in Figure 3: (1) a feature extractor that leverages a multi-resolution architecture to produce a single, powerful feature map; (2) a global matching module that uses optimal transport [8] to establish robust long-range correspondences; (3) a hybrid refinement stage that corrects errors both globally and locally; and (4) a final upsampling process to restore full resolution while

preserving details.

3.1. Feature Extraction

The goal of our feature extraction stage is to produce a single, high-resolution feature map for each image, rich with the contextual information necessary for robust global matching. While using a large-scale Vision Transformer for this task is a common approach, it is computationally prohibitive for high-resolution stereo, creating the primary scalability bottleneck. To address this, we propose an efficient hierarchical architecture. The process begins with a lightweight CNN backbone and a Feature Pyramid Network (FPN) that generate an initial multi-scale feature pyramid: $\{\mathbf{F}_4, \mathbf{F}_8, \mathbf{F}_{16}, \mathbf{F}_{32}\}$. These features are then fed into our novel Multi-Resolution Transformer (MRT) for deep contextual refinement.

3.1.1. Multi-Resolution Transformer (MRT)

The MRT is designed to overcome the efficiency limitations of conventional Transformers. Instead of processing a single high-resolution feature map, it operates on the entire feature pyramid in parallel. Its efficiency stems from a hybrid attention strategy: to align with epipolar geometry, attention is restricted to the horizontal axis (1D) at higher resolutions, while full 2D self-attention is applied only at the coarsest level ($\mathbf{F}_{32}$) to capture global context. This design strikes a critical balance between performance and computational cost.

The internal structure of each MRT block, illustrated in Figure 3, is also carefully designed. Each block is composed of a self-attention module and a cross-attention module, with each respective attention layer being followed by an FFN to produce highly discriminative features. To ensure stable training of a deep stack of MRT blocks, each block is also equipped with cross-scale skip connections.

3.1.2. Adaptive Gated Fusion Layer (AGFL)

We build a high-capacity feature extractor by stacking multiple MRT blocks. A key challenge in such a deep, multi-resolution architecture is ensuring effective information exchange and coherent learning across all scales. Prior architectures like U-Net can suffer from gradient bottlenecks at the highest resolution, hindering the training of coarser-level features [23]. While our parallel MRT architecture avoids this specific bottleneck, it still requires a mechanism for robust cross-scale communication.

To facilitate this, we introduce the Adaptive Gated Fusion Layer (AGFL). The AGFL acts as a dynamic gate within each MRT block, learning to selectively fuse information between scales. This allows global context from coarser levels to effectively inform finer ones and vice versa, ensuring stable training and a powerful, coherent multi-scale representation. The fusion process is formulated as:

$$\mathbf{z} = \mathbf{w} \odot \mathbf{x} + (1 - \mathbf{w}) \odot \mathbf{y} + \text{MLP}_r(\text{Concat}(\mathbf{x}, \mathbf{y})) \quad (1)$$

with the weight $\mathbf{w}$ computed as:

$$\mathbf{w} = \sigma(\text{MLP}_w(\text{Concat}(\mathbf{x}, \mathbf{y}))). \quad (2)$$

3.2. Global Matching

Given the enhanced feature maps from our feature extractor, the next stage establishes dense global correspondences. A common approach in global matching involves first constructing an all-pairs correlation volume $\mathbf{C} \in \mathbb{R}^{H' \times W' \times W'}$ from the 1/4-resolution features ($\mathbf{F}_4$), where $H' = H/4$ and $W' = W/4$. This volume encodes feature similarities along the epipolar lines. Following this standard practice in global matching, we formulate the task as a global assignment problem solved via entropy-regularized optimal transport [8], instead of relying on a simple `argmax` operation on this volume. By enforcing a soft, bidirectionally consistent (one-to-one) matching constraint, the Sinkhorn algorithm finds a globally optimal transport plan $\mathbf{T}$ that is robust to ambiguities where a simple best match from one perspective might not be the best from the other.

$$\mathbf{T} = \text{OT}(\mathbf{C}). \quad (3)$$

To handle occlusions, we incorporate the standard dustbin trick [23, 27].

A key advantage of this formulation is that the resulting transport plan $\mathbf{T}$ provides a rich, probabilistic representation from which all initial estimates can be efficiently derived. The disparity map $\mathbf{D}$ is computed as the expected value over all candidates:

$$\mathbf{D}_{ij} = \frac{\sum_{w=1}^{W'} \mathbf{T}_{ijw} \mathbf{p}_w}{\sum_{w=1}^{W'} \mathbf{T}_{ijw}}, \quad (4)$$

where $\mathbf{p}_w$ represents the candidate disparities. The occlusion map $\mathbf{O}$ is obtained by marginalizing probabilities across the disparity axis, and the confidence map $\mathbf{\Gamma}$ is derived by summing the probability mass within a local window around the peak probability:

$$\mathbf{\Gamma}_{ij} = \sum_{w \in \mathcal{W}(w^*)} \mathbf{T}_{ijw}, \quad \text{where} \quad w^* = \arg\max_w \mathbf{T}_{ijw}. \quad (5)$$

These three estimates—disparity, occlusion, and confidence—serve as a robust starting point for the subsequent refinement stage.

3.3. Refinement

The refinement stage consists of two steps: a global disparity adjustment and an iterative local refinement process. The global step addresses occluded regions by propagating disparity values from high-confidence areas, following a confidence-aware disparity aggregation strategy [23]. Given the initial disparity map $\mathbf{D}_0$ and confidence map $\mathbf{\Gamma}_0$, the

global refinement step updates disparity in unreliable regions as

$$\mathbf{D}_0' = \mathbf{D}_0 + \Phi(\mathbf{F}_4^L, \mathbf{\Gamma}_0, \mathbf{D}_0), \tag{6}$$

where Φ represents the aggregation function that primarily propagates disparity from high-confidence regions to low-confidence regions based on the confidence weights.

For local refinement, we iteratively update disparity, occlusion, and confidence maps using a residual correction strategy inspired by RAFT-style recurrent refinement [20]. At each iteration t, the updates are formulated as

$$\mathbf{D}_{t+1} = \mathbf{D}_t + \Delta\mathbf{D}_t, \tag{7}$$

$$\mathbf{O}_{t+1} = \mathbf{O}_t + \Delta\mathbf{O}_t, \tag{8}$$

$$\mathbf{\Gamma}_{t+1} = \mathbf{\Gamma}_t + \Delta\mathbf{\Gamma}_t. \tag{9}$$

where $\Delta\mathbf{D}_t$, $\Delta\mathbf{O}_t$, and $\Delta\mathbf{\Gamma}_t$ are residual updates computed by the refinement network. These updates are conditioned on local feature correlations, disparity estimates, and the confidence map, ensuring that occlusions and unreliable disparities are iteratively corrected.

3.4. Upsampling

Following refinement, the low-resolution disparity map is upsampled to its original resolution. We first perform an initial $4\times$ upsampling using a weighted combination of neighboring disparity values, where the weights are learned from the feature maps, similar to prior works [18, 32, 42]. To further enhance the quality of object boundaries, we then apply a final edge-guided filter. This filter uses adaptive weights, predicted from the left image and a disparity-warped right image, to refine the upsampled disparity map. This final filtering step improves stereo consistency while preserving fine details at depth edges.

4. Loss Function

We train our model with a composite loss function including disparity, occlusion, confidence (all supervised by L1 loss with ground truths), and a Probabilistic Mode Concentration (PMC) loss. Standard L1 losses on disparity, confidence, and occlusion do not directly guide the Transformer's matching probabilities. To address this, we introduce the PMC loss, which encourages concentrated matching probabilities within valid disparity regions, enhancing feature discriminability and global matching robustness.

4.1. Probabilistic Mode Concentration (PMC) Loss

Global matching is performed on 1/4 downsampled feature maps, resulting in at most 16 disparity candidates at each pixel in this lower resolution space. Let K (with $K \leq 16$) denote the number of these candidates. For each candidate $c_{ij}^{(k)}$, we define its neighborhood as

$$\mathcal{N}(c_{ij}^{(k)}) = \{w \in \{1, \ldots, W\} \mid |w - c_{ij}^{(k)}| \leq \delta\}, \tag{10}$$

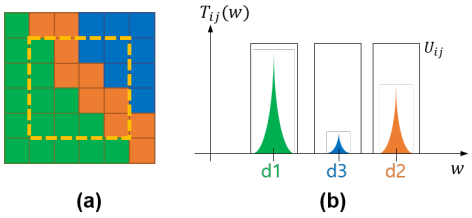


Figure 4. Illustration of Probabilistic Mode Concentration (PMC) Loss. (a) Disparity map with a dashed box highlighting a 4×4 region, which corresponds to a single pixel in the 1/4 downsampled feature space. (b) Matching probability distribution at the selected location, where PMC Loss ensures that the probability mass remains concentrated within valid disparity candidates.

and take the union over all candidates:

$$\mathcal{U}_{ij} = \bigcup_{k=1}^{K} \mathcal{N}(c_{ij}^{(k)}). \tag{11}$$

The aggregated probability mass over this union is then computed as

$$S_{ij} = \sum_{w \in \mathcal{U}_{ij}} T_{ij}(w), \tag{12}$$

where $\mathbf{T}_{ij} = [T_{ij}(1), T_{ij}(2), \ldots, T_{ij}(W')]$ represents the matching probability vector obtained from Eq. 3

The PMC loss enforces that S_{ij} remains close to 1 for non-occluded pixels:

$$\mathcal{L}_{\text{PMC}} = \frac{1}{|\Omega|} \sum_{(i,j)\in\Omega} \max(1 - S_{ij} - \epsilon, 0), \tag{13}$$

where $\Omega = \{(i,j) \mid O(i,j) = 1\}$ represents the set of non-occluded pixels, and ϵ is a small margin.

This additional loss regularizes the cost volume, ensuring that the matching probability remains concentrated within valid candidate regions, thereby complementing the disparity loss and guiding the network toward confident and precise matches.

The final loss function combines disparity, confidence, occlusion, and PMC losses, each weighted by a hyperparameter:

$$\mathcal{L} = \lambda_D\, \mathcal{L}_D + \lambda_{\mathcal{O}}\, \mathcal{L}_{\mathcal{O}} + \lambda_{\mathcal{C}}\, \mathcal{L}_{\mathcal{C}} + \lambda_{\text{PMC}}\, \mathcal{L}_{\text{PMC}}. \tag{14}$$

5. Experiments

5.1. Implementation Details

All models were trained using the AdamW optimizer, an initial learning rate of 1×10^{-5}, a one-cycle schedule, and three local refinement iterations. For the ablation studies, we

Model Configuration			Training Strategy	Disparity Metrics		Auxiliary Metrics		Model Complexity	
Ch	Ntr	AGFL	PMC Loss	EPE	Bad-2.0	Occ AP	Conf AP	Params (M)	GFLOPs
128	1	No	No	0.489	1.974	0.973	0.977	23	645
128	1	Yes	No	0.496	1.601	0.973	0.978	24	647
128	1	Yes	Yes	0.451	1.490	0.974	0.978	24	647
128	3	Yes	Yes	0.388	1.281	0.974	0.978	43	772
256	1	Yes	Yes	0.320	0.928	0.975	0.979	93	2146
256	3	Yes	No	0.299	0.810	0.976	0.980	170	2644
256	3	Yes	Yes	0.254	0.750	0.976	0.980	170	2644

Table 1. Ablation study on model configuration (ch, ntr, Adaptive Gated Fusion Layer (AGFL)) and training strategy (Probabilistic Mode Concentration (PMC) loss). Disparity metrics (EPE, bad-2.0) evaluate stereo matching accuracy, while auxiliary metrics (occ AP, conf AP) measure average precision (AP) for occlusion and confidence estimation. Model complexity is represented by params (M), the number of learnable parameters, and GFLOPs, which quantifies computational cost in giga floating-point operations per second (GFLOPs) for a 960×512 px input. ch: base channel dimension, ntr: number of transformer blocks.

Iterations	GFLOPs	Disparity Metrics		Auxiliary Metrics	
		EPE	Bad-2.0	Occ AP	Conf AP
0	305	0.666	2.768	0.961	0.977
1	427	0.470	1.658	0.974	0.977
2	537	0.479	1.526	0.974	0.977
3	647	0.451	1.490	0.974	0.978

Table 2. Impact of local refinement iterations on computational cost and performance. Model Configuration (Ch=128,Ntr=1). Increasing the number of refinement iterations improves disparity estimation accuracy (EPE, Bad-2.0) and auxiliary metrics (Occ AP, Conf AP), but also increases GFLOPs.

used a 0.5M image subset from several synthetic datasets [18, 21, 36]. Our final models for the benchmark evaluation were trained on an expanded 2M image dataset (denoted '+2Mtr') that includes a diverse mix of data [12, 18, 21, 23, 32–34, 36, 40, 46], and subsequently underwent multi-resolution fine-tuning. All experiments were conducted on an Nvidia H100 GPU, and models were evaluated at their original resolutions. Further details on data processing and augmentations are provided in the supplementary material.

All evaluations are conducted using standard stereo matching metrics to assess disparity accuracy, including End-Point Error (EPE), Root Mean Square Error (RMS), and Bad-p thresholds (Bad-0.5, Bad-1.0, Bad-2.0, Bad-4.0). Each metric is reported for both non-occluded regions and all pixels, where lower values indicate better performance. For confidence and occlusion estimation, we adopt an Average Precision (AP) evaluation approach, where values closer to 1 indicate better performance.

5.2. Ablation Study

We conducted an ablation study on the Sceneflow dataset [21] to validate our architectural design choices, with results summarized in Table 1 and Table 2. First, we confirmed the scalability of our architecture: increasing the channel dimension (Ch) and number of Transformer blocks (Ntr) consistently improves accuracy. Next, we validated the effectiveness of our key components, showing that the Gated Fusion mechanism enhances feature aggregation and the Probabilistic Mode Concentration (PMC) loss boosts accuracy without added computational cost. Finally, we analyzed the refinement process and found that S²M² requires far fewer iterations than prior methods. Unlike approaches reliant on many updates, our model's strong initial disparity estimate allows it to achieve near-optimal performance with just three refinement steps, demonstrating the effectiveness of our global matching foundation.

5.3. Benchmark Evaluation

We evaluate S²M² on diverse benchmarks to verify its effectiveness across realistic and challenging stereo conditions. To thoroughly assess its capabilities and study scaling effects, we introduce several variants of S²M². These differ in model size (S: Ch=128, Ntr=1; L: Ch=256, Ntr=3; XL: Ch=384, Ntr=3) and are trained with datasets of varying scales, including conventional-sized datasets (denoted '+1Mtr') and expanded datasets of approximately 2 million stereo pairs (denoted '+2Mtr'), including the FoundationStereo dataset[40]. This multi-faceted approach allows for a comprehensive analysis of performance scaling with respect to both model capacity and training data volume. Specifically, we conduct comparisons on our synthetic dataset, as well as on established real-world datasets such as ETH3D [29], Middlebury v3 [28], and the KITTI 2015 benchmark [22].

Our Synthetic Benchmark

Our synthetic benchmark is specifically designed to assess performance in challenging high-resolution, large-disparity scenarios that push modern methods to their limits. This setting proves difficult for many existing approaches, as shown in Table 3. For instance, methods relying on explicit cost volumes (e.g., IGEV [42], FoundationStereo [40], Mocha [5]) are constrained by a maximum disparity, which we capped at 960 pixels due to their memory requirements. While this covers over 99% of the disparities in our test set, it highlights a fundamental architectural limitation. Unsurprisingly, these

Model	Non-Occlusion						All Pixels						Mem(G)	TFLOPs	Param(M)
	EPE	RMS	Bad-0.5	Bad-1.0	Bad-2.0	Bad-4.0	EPE	RMS	Bad-0.5	Bad-1.0	Bad-2.0	Bad-4.0			
CoEx [2]	20.17	63.95	38.78	27.82	21.06	16.75	37.55	103.82	46.95	36.78	29.85	24.87	31.31	0.71	2.72
BGNet+ [41]	14.04	42.84	43.97	29.55	20.63	14.82	30.31	88	50.98	37.85	29.11	22.87	22.49	2.42	5.31
RAFT [20]	20.41	48.12	24.41	17.71	13.87	11.32	33.07	76.51	33.31	26.43	21.84	18.31	8.18	30.35	11.11
fastACV [43]	15.09	48.39	36.73	24.77	17.69	13.31	32.50	93.99	45.57	34.61	27.34	22.10	8.01	0.60	3.07
UniMatch [45]	6.46	17.54	55.77	31.11	16.35	10.66	13.96	37.93	61.77	40.05	25.82	18.54	49.3	16.35	7.35
CroCo [39]	15.64	35.66	33.82	23.71	17.43	13.48	24.99	57.23	40.70	30.96	24.31	19.57	2.62	164.7	437
IGEV [42]	7.03	21.16	23.04	13.86	9.66	7.13	15.42	42.85	31.11	22.17	17.30	13.82	10.87	27.63	12.59
MC [11]	35.56	73.77	32.48	26.72	23.08	20.56	56.07	112.67	41.12	35.47	31.46	28.29	7.73	899	24.62
NMRF [14]	40.88	82.83	58.67	43.24	36.66	33.07	62.04	121.41	64.06	50.57	44.28	40.31	9.07	3.28	6.113
Sel-IGEV [37]	9.23	25.92	22.37	14.09	9.79	7.36	18.23	48.79	30.06	21.82	16.86	13.49	13.4	35.77	13.14
MoCha [5]	40.37	102.65	26.9	21.62	18.17	15.73	59.75	141.19	35.34	29.91	25.94	22.82	12.87	22.14	97.14
S²M²-S (+1Mtr)	6.45	16.36	19.10	11.46	7.63	5.39	11.39	31.06	26.94	18.93	14.15	10.78	8.17	5.76	23.91
S²M²-L (+1Mtr)	3.05[3]	10.17[3]	14.92	8.72	5.58	3.74	8.21[3]	25.92	23.48	16.53	12.21	9.12	11.39	25.96	170
FoundationStereo [40] (+2Mtr)	4.60	11.27	10.96[3]	7.23[3]	4.95[3]	3.49[3]	8.26	23.04[3]	18.42[3]	13.98[3]	10.73[3]	8.08[3]	53.78	124.65	375
S²M²-S (+2Mtr)	4.82	10.83	13.13	7.62	5	3.57	9.14	26.32	20.75	14.55	10.83	8.26	8.17	5.76	23.91
S²M²-L (+2Mtr)	2.99[2]	8.72[2]	10.60[2]	6.00[2]	3.80[2]	2.54[2]	7.20[2]	22.40[2]	18.11[2]	12.72[2]	9.39[2]	6.96[2]	11.39	25.96	170
S²M²-XL (+2Mtr)	**2.30**[1]	**7.95**[1]	**9.22**[1]	**5.15**[1]	**3.29**[1]	**2.18**[1]	**6.2**[1]	**20.58**[1]	**16.51**[1]	**11.60**[1]	**8.56**[1]	**6.36**[1]	14.84	56.94	381

Table 3. Quantitative results on our synthetic benchmark, evaluated over the full disparity range. All models process the input images at their original resolution of 2448 × 2048 pixels. The last three columns report peak memory usage, total computational cost (in tera floating-point operations, TFLOPs), and model size (in millions of parameters). The three best-performing models for each metric are marked with blue superscripts.

Model	Non-Occlusion						All Pixels					
	EPE	RMS	Bad-0.5	Bad-1.0	Bad-2.0	Bad-4.0	EPE	RMS	Bad-0.5	Bad-1.0	Bad-2.0	Bad-4.0
RAFTStereo [20]	0.18	0.36	7.04	2.44	0.44	0.15	0.19	0.42	7.33	2.60	0.56	0.22
CREStereo [18]	0.13	0.28	3.58	0.98	0.22	0.10	0.14	0.31[3]	3.75	1.09	0.29	0.12
Unimatch [45]	0.19	0.38	5.94	1.83	0.25	0.08	0.21	0.44	6.44	2.07	0.34	0.14
CroCo [39]	0.14	0.30	3.27	0.99	0.39	0.13	0.15	0.35	3.51	1.14	0.50	0.18
IGEV [42]	0.14	0.34	3.52	1.12	0.21	0.11	0.20	0.86	3.97	1.51	0.54	0.41
Sel-IGEV [37]	0.12	0.29	3.06	1.23	0.22	0.05	0.15	0.57	3.46	1.56	0.51	0.28
CAS [23]	0.14	0.30	3.96	0.87	0.32	0.07	0.15	0.32	4.11	1.09	0.21[3]	0.10[3]
Mocha V2 [6]	0.13[3]	0.36	3.20	1.41	0.33	0.09	0.18	0.74	3.68	1.83	0.71	0.41
S²M²-L (+1Mtr)	0.13[3]	0.24[3]	2.06[3]	0.46[3]	0.13[3]	0.05[2]	0.13[2]	0.27[2]	2.24[3]	0.53[3]	0.17[2]	0.08[2]
FoundationStereo [40] (+2Mtr)	**0.09**[1]	**0.20**[1]	1.26[2]	0.26[2]	0.08[2]	0.05[2]	0.13[2]	0.61	1.56[2]	0.48[2]	0.26	0.21
S²M²-XL (+2Mtr)	0.10[2]	**0.20**[1]	**0.93**[1]	**0.22**[1]	**0.06**[1]	**0.03**[1]	**0.10**[1]	**0.22**[1]	**1.03**[1]	**0.26**[1]	**0.08**[1]	**0.04**[1]

Table 4. Quantitative results on the ETH3D stereo benchmark [29]. The three best-performing models for each metric are marked with blue superscripts.

Model	Non-Occlusion						All Pixels					
	EPE	RMS	Bad-0.5	Bad-1.0	Bad-2.0	Bad-4.0	EPE	RMS	Bad-0.5	Bad-1.0	Bad-2.0	Bad-4.0
RAFTStereo [20]	1.27	8.41	27.7	9.37	4.74	2.75	2.71	12.6	33.0	15.1	9.37	6.42
CREStereo [18]	1.15	7.70	33.3	8.25	3.71	2.04	2.10	10.5	31.3	14.0	8.13	5.05
Unimatch [45]	1.31	6.45[3]	51.5	23.6	7.14	2.96	1.89	8.03[2]	54.8	28.4	11.7	6.07
CroCo [39]	1.76	8.91	40.6	16.9	7.29	4.18	2.36	10.6	44.4	21.6	11.1	6.75
IGEV [42]	2.89	12.8	32.4	9.41	4.83	3.33	3.64	15.1	31.3	13.8	8.16	5.79
Sel-IGEV [37]	0.91	7.26	24.6[3]	6.53	2.51	1.36	1.54	9.26	36.6	11.4	6.04	3.74
CAS [23]	1.01	6.74	31.3	8.71	3.33	1.77	1.52	8.54	35.3	13.0	6.29	3.75
Mocha V2 [6]	1.10	7.69	24.8	6.81	3.51	2.34	1.83	9.91	29.4[3]	11.4	6.74	4.54
S²M²-L (+1Mtr)	0.80[3]	6.36[2]	25.5	6.31[3]	2.08[3]	0.97[2]	1.30[3]	8.31[3]	29.5	10.2[3]	4.76[3]	2.85[3]
FoundationStereo [40] (+2Mtr)	0.78[2]	6.48	22.5[2]	4.39[2]	1.84[2]	1.04[3]	1.24[2]	8.39	26.4[2]	8.14[2]	4.26[2]	2.72[2]
S²M²-XL (+2Mtr)	**0.69**[1]	**6.01**[1]	**22.1**[1]	**3.57**[1]	**1.15**[1]	**0.54**[1]	**1.00**[1]	**7.39**[1]	**25.3**[1]	**6.60**[1]	**2.94**[1]	**1.63**[1]

Table 5. Quantitative results on the Middlebury v3 Benchmark [28]. The three best-performing models for each metric are marked with blue superscripts.

models, as well as lightweight approaches [2, 11, 14, 41, 43], struggle to generalize in this demanding environment.

In stark contrast, our S²M² family demonstrates exceptional robustness and a superior accuracy-efficiency trade-off. Our largest model, S²M²-XL (+2Mtr), establishes a new state-of-the-art across all metrics. While FoundationStereo [40] achieves the best performance among prior methods, it is computationally prohibitive. This trade-off is visualized in the scatter plot of Figure 5, which plots accuracy against model size. The figure clearly shows FoundationStereo [40] occupying a high-accuracy but high-cost region, whereas our S²M² family forms a compelling Pareto front, offering significantly better performance at every level of computational budget and validating the scalability of our architecture.

Model	Fine-tune	Non-Occlusion			All Pixels		
		D1-bg	D1-fg	D1-all	D1-bg	D1-fg	D1-all
Sel-IGEV [37]	Yes	1.22	2.55	1.44	1.33	2.61	1.55
Sel-IGEV [37]	No	2.91	12.9	4.57	3.06	13.4	4.79
FoundationStereo [40]	No	2.48	6.13	3.08	2.59	6.25	3.20
S²M²-XL	No	2.51	6.22	3.12	2.61	6.31	3.23

Table 6. Quantitative results on the KITTI 2015 Benchmark [22]. The results highlight the dramatic performance gap between models with and without dataset-specific fine-tuning, suggesting that top leaderboard scores are heavily influenced by adaptation to dataset biases rather than true generalization.

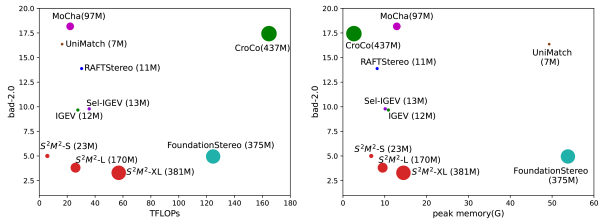


Figure 5. Scalability Analysis

ETH3D Benchmark [29]

The ETH3D benchmark [29] is characterized by real-world scenes with predominantly small disparity values, making it an excellent testbed for assessing precision and robustness in a different regime from our large-disparity synthetic data. On this benchmark, S²M² once again demonstrates its state-of-the-art capabilities, as shown in Table 4.

Our S²M²-XL model surpasses prior methods like FoundationStereo [40] across the majority of metrics. Achieving top performance on both our large-disparity synthetic dataset and the small-disparity ETH3D benchmark validates the exceptional versatility of our global matching architecture across a wide spectrum of disparity ranges. The model's ability to maintain high accuracy in ETH3D's varied conditions, including challenging low-texture surfaces and noisy scenes, further highlights its practical robustness.

The Middlebury v3 Benchmark [28]

The Middlebury v3 benchmark [28], with its high-resolution images and precise ground-truth, serves as the ultimate test for a model's ability to preserve fine-grained details. On this demanding benchmark, S²M² again demonstrates the superiority of our approach, establishing a new state-of-the-art as detailed in Table 5.

Our S²M²-XL model not only outperforms strong baselines like FoundationStereo [40], but does so with particularly large margins on the most stringent metrics (Bad-1.0 and Bad-2.0). This numerical superiority translates to tangible qualitative improvements. For instance, as seen in Figures 1, our model faithfully reconstructs highly intricate structures like bicycle spokes—a common failure case for many other methods. This result indicates that our global matching framework is effective at preserving fine details by establishing accurate initial correspondences for subsequent refinement.

Critical Re-evaluation of the KITTI Benchmark [13, 22]

While the KITTI benchmark [22] is a standard for stereo evaluation, we argue its leaderboard scores are an unreliable indicator of true generalization due to inherent noise in its LiDAR-based ground truth. Our analysis in Table 6 highlights this discrepancy. Powerful generalist models like FoundationStereo [40] and our S²M²-XL, despite their reliable visual quality and top performance on other benchmarks, exhibit error rates on KITTI that are more than double those of the top fine-tuned models on the leaderboard. As seen in Table 6, this vast performance gap suggests that top KITTI scores are less a measure of a model's inherent capability and more a reflection of its ability to adapt to dataset-specific biases and noises.

Our supplementary material provides extensive evidence supporting this claim. We demonstrate that while fine-tuning improves error metrics, it simultaneously degrades photometric similarity measured via left-right warping. Furthermore, we provide 3D visualizations that show fine-tuned models adapting to noise patterns, and present manual inspection results identifying consistent ground-truth biases of up to 2 pixels. Therefore, we recommend that KITTI leaderboard scores be interpreted with caution, as they may disproportionately reward models that fit to flawed annotations rather than those that achieve genuine real-world accuracy.

6. Conclusion

This paper presents S²M², a multi-resolution Transformer framework that overcomes the long-standing challenges of global matching, positioning it as a practical, state-of-the-art methodology. The proposed framework is highly scalable, robustly handling challenging high resolutions and wide disparity ranges, while its performance improves consistently with model size. It produces reliable depth estimates by jointly modeling disparity, confidence, and occlusion. Our extensive experiments not only establish a new state-of-the-art on real-world benchmarks but also demonstrate the framework's robustness in demanding scenarios, thereby opening new avenues for developing highly accurate and generalizable stereo systems.

References

[1] Abhishek Badki, Alejandro Troccoli, Kihwan Kim, Jan Kautz, Pradeep Sen, and Orazio Gallo. Bi3d: Stereo depth estimation via binary classifications. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1600–1608, 2020.

[2] Antyanta Bangunharcana, Jae Won Cho, Seokju Lee, In So Kweon, Kyung-Soo Kim, and Soohyun Kim. Correlate-and-excite: Real-time stereo matching via guided cost volume excitation. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3542–3548. IEEE, 2021.

[3] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1013–1027, 2025.

[4] Jia-Ren Chang and Yong-Sheng Chen. Pyramid stereo matching network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5410–5418, 2018.

[5] Ziyang Chen, Wei Long, He Yao, Yongjun Zhang, Bingshu Wang, Yongbin Qin, and Jia Wu. Mocha-stereo: Motif channel attention network for stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27768–27777, 2024.

[6] Ziyang Chen, Yongjun Zhang, Wenting Li, Bingshu Wang, Yong Zhao, and CL Chen. Motif channel opened in a white-box: Stereo matching via motif correlation graph. *arXiv preprint arXiv:2411.12426*, 2024.

[7] Alex Costanzino, Pierluigi Zama Ramirez, Matteo Poggi, Fabio Tosi, Stefano Mattoccia, and Luigi Di Stefano. Learning depth estimation for transparent and mirror surfaces. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9244–9255, 2023.

[8] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013.

[9] Shivam Duggal, Shenlong Wang, Wei-Chiu Ma, Rui Hu, and Raquel Urtasun. Deeppruner: Learning efficient stereo matching via differentiable patchmatch. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4384–4393, 2019.

[10] Andreas Eitel, Jost Tobias Springenberg, Luciano Spinello, Martin Riedmiller, and Wolfram Burgard. Multimodal deep learning for robust rgb-d object recognition. *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 681–687, 2015.

[11] Miaojie Feng, Junda Cheng, Hao Jia, Longliang Liu, Gangwei Xu, and Xin Yang. Mc-stereo: Multi-peak lookup and cascade search range for stereo matching. In *International Conference on 3D Vision (3DV)*, pages 344–353. IEEE, 2024.

[12] Adrien Gaidon, Qiao Wang, Yohann Cabon, and Eleonora Vig. Virtual worlds as proxy for multi-object tracking analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4340–4349, 2016.

[13] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2012.

[14] Tongfan Guan, Chen Wang, and Yun-Hui Liu. Neural markov random field for stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5459–5469, 2024.

[15] Heiko Hirschmuller. Stereo processing by semi-global matching and mutual information. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(2):328–341, 2008.

[16] Hualie Jiang, Zhiqiang Lou, Laiyan Ding, Rui Xu, Minglang Tan, Wenjie Jiang, and Rui Huang. Defom-stereo: Depth foundation model based stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21857–21867, 2025.

[17] 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 *Proceedings of the IEEE international conference on computer vision*, pages 66–75, 2017.

[18] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16263–16272, 2022.

[19] Zhaoshuo Li, Xingtong Liu, Nathan Drenkow, Andy Ding, Francis X Creighton, Russell H Taylor, and Mathias Unberath. Revisiting stereo depth estimation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6197–6206, 2021.

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

[21] Nikolaus Mayer, Eddy Ilg, Philip Hausser, 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4040–4048, 2016.

[22] Moritz Menze and Andreas Geiger. Object scene flow for autonomous vehicles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3061–3070, 2015.

[23] Junhong Min and Youngpil Jeon. Confidence aware stereo matching for realistic cluttered scenario. In *2024 IEEE International Conference on Image Processing (ICIP)*, pages 3491–3497. IEEE, 2024.

[24] Adithyavairavan Murali, Alex Dobson, Dhiraj Gandhi, Lerrel Pinto, and Abhinav Gupta. Learning to grasp and regrasp using vision and touch. In *Proceedings of the 2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2537–2544, 2020.

[25] Matteo Poggi, Seungryong Kim, Fabio Tosi, Sunok Kim, Filippo Aleotti, Dongbo Min, Kwanghoon Sohn, and Stefano

Mattoccia. On the confidence of stereo matching in a deep-learning era: a quantitative evaluation. *IEEE transactions on pattern analysis and machine intelligence*, 44(9):5293–5313, 2021.

[26] Zi Jian Yew Qin, Hao Yu, Changjian Xu, Yulan Wang, Yew-Soon Liu, Feng Lu, and Pheng-Ann Heng. Geometric transformer for fast and robust point cloud registration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11143–11152, 2022.

[27] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4938–4947, 2020.

[28] Daniel Scharstein, Heiko Hirschmüller, York Kitajima, Greg Krathwohl, Nera Nešić, Xi Wang, and Porter Westling. 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*, pages 31–42. Springer, 2014.

[29] Thomas Schops, Johannes L Schonberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3260–3269, 2017.

[30] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. Loftr: Detector-free local feature matching with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8922–8931, 2021.

[31] Vladimir Tankovich, Christian Hane, Yinda Zhang, Adarsh Kowdle, Sean Fanello, and Sofien Bouaziz. Hitnet: Hierarchical iterative tile refinement network for real-time stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14362–14372, 2021.

[32] Fabio Tosi, Yiyi Liao, Carolin Schmitt, and Andreas Geiger. Smd-nets: Stereo mixture density networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8942–8952, 2021.

[33] Jonathan Tremblay, Thang To, and Stan Birchfield. Falling things: A synthetic dataset for 3d object detection and pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 2038–2041, 2018.

[34] Qiang Wang, Shizhen Zheng, Qingsong Yan, Fei Deng, Kaiyong Zhao, and Xiaowen Chu. 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)*, pages 1–6. IEEE, 2021.

[35] Qing Wang, Jiaming Zhang, Kailun Yang, Kunyu Peng, and Rainer Stiefelhagen. Matchformer: Interleaving attention in transformers for feature matching. In *Proceedings of the Asian Conference on Computer Vision*, pages 2746–2762, 2022.

[36] Wenshan Wang, Delong Zhu, Xiangwei Wang, Yaoyu Hu, Yuheng Qiu, Chen Wang, Yafei Hu, Ashish Kapoor, and Sebastian Scherer. Tartanair: A dataset to push the limits of visual slam. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4909–4916. IEEE, 2020.

[37] Xianqi Wang, Gangwei Xu, Hao Jia, and Xin Yang. Selective-stereo: Adaptive frequency information selection for stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19701–19710, 2024.

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

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

[40] Bowen Wen, Matthew Trepte, Joseph Aribido, Jan Kautz, Orazio Gallo, and Stan Birchfield. Foundationstereo: Zero-shot stereo matching. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5249–5260, 2025.

[41] Bin Xu, Yuhua Xu, Xiaoli Yang, Wei Jia, and Yulan Guo. Bilateral grid learning for stereo matching networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12497–12506, 2021.

[42] Gangwei Xu, Xianqi Wang, Xiaohuan Ding, and Xin Yang. Iterative geometry encoding volume for stereo matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 21919–21928, 2023.

[43] Gangwei Xu, Yun Wang, Junda Cheng, Jinhui Tang, and Xin Yang. Accurate and efficient stereo matching via attention concatenation volume. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(4):2461–2474, 2023.

[44] Haofei Xu and Juyong Zhang. Aanet: Adaptive aggregation network for efficient stereo matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1959–1968, 2020.

[45] Haofei Xu, Jing Zhang, Jianfei Cai, Hamid Rezatofighi, Fisher Yu, Dacheng Tao, and Andreas Geiger. Unifying flow, stereo and depth estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[46] Guorun Yang, Xiao Song, Chaoqin Huang, Zhidong Deng, Jianping Shi, and Bolei Zhou. Drivingstereo: A large-scale dataset for stereo matching in autonomous driving scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 899–908, 2019.

[47] Zi Jian Yew and Gim Hee Lee. Regtr: End-to-end point cloud correspondences with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6677–6686, 2022.

[48] Hao Yu, Zheng Qin, Ji Hou, Mahdi Saleh, Dongsheng Li, Benjamin Busam, and Slobodan Ilic. Rotation-invariant

transformer for point cloud matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5384–5393, 2023.

[49] Haoliang Zhao, Huizhou Zhou, Yongjun Zhang, Yong Zhao, Yitong Yang, and Ting Ouyang. Eai-stereo: Error aware iterative network for stereo matching. In *Proceedings of the Asian Conference on Computer Vision*, pages 315–332, 2022.

[50] Jingyi Zhou, Haoyu Zhang, Jiakang Yuan, Peng Ye, Tao Chen, Hao Jiang, Meiya Chen, and Yangyang Zhang. All-in-one: Transferring vision foundation models into stereo matching. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 10797–10805, 2025.