PPMStereo: Pick-and-Play Memory Construction for Consistent Dynamic Stereo Matching

Yun Wang[1], Junjie Hu[2],* Qiaole Dong[3],† Yongjian Zhang[4]
Yanwei Fu[3], **Tin Lun Lam[2]**, **Dapeng Wu[1]**
[1]City University of Hong Kong, [2]The Chinese University of Hong Kong, Shenzhen
[3] Fudan University,[4] Shenzhen Campus, Sun Yat-sen University
ywang3875-c@my.cityu.edu.hk, dpwu@ieee.org,
{qldong18, yanweifu}@fudan.edu.cn
zhangyj85@mail2.sysu.edu.cn,
{hujunjie,tllam}@cuhk.edu.cn

Abstract

Temporally consistent depth estimation from stereo video is critical for real-world applications such as augmented reality, where inconsistent depth estimation disrupts the immersion of users. Despite its importance, this task remains challenging due to the difficulty in modeling long-term temporal consistency in a computationally efficient manner. Previous methods attempt to address this by aggregating spatio-temporal information but face a fundamental trade-off: limited temporal modeling provides only modest gains, whereas capturing long-range dependencies significantly increases computational cost. To address this limitation, we introduce a memory buffer for modeling long-range spatio-temporal consistency while achieving efficient dynamic stereo matching. Inspired by the two-stage decision-making process in humans, we propose a **P**ick-and-**P**lay **M**emory (PPM) construction module for dynamic **Stereo** matching, dubbed as **PPMStereo**. PPM consists of a 'pick' process that identifies the most relevant frames and a 'play' process that weights the selected frames adaptively for spatio-temporal aggregation. This two-stage collaborative process maintains a compact yet highly informative memory buffer while achieving temporally consistent information aggregation. Extensive experiments validate the effectiveness of PPMStereo, demonstrating state-of-the-art performance in both accuracy and temporal consistency. Codes are available at https://github.com/cocowy1/PPMStereo.

1 Introduction

Stereo matching refers to binocular disparity estimation, which is a fundamental computer vision task focused on estimating the disparity between a pair of rectified stereo images [53, 21, 34]. Deep learning-based stereo matching methods have achieved remarkable progress in terms of accuracy [53, 58, 50, 11], efficiency [46, 59, 52, 2], and robustness [42, 64, 63, 55]. Despite impressive performance for static scenes, these methods exhibit severe temporal inconsistencies when applied to dynamic scenes [24]. This manifests itself as flickering artifacts and blurred disparity maps due to the absence of effective inter-frame temporal information integration. Therefore, the algorithm deployment in dynamic scenarios such as autonomous driving, robotics, and augmented reality platforms is limited, which requires temporally consistent disparity maps.
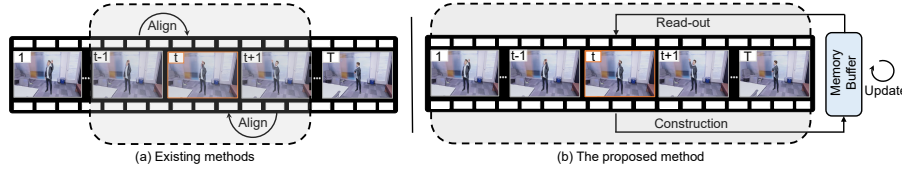
---

*Corresponding author.
†Project Leader

Figure 1: Comparison between prior methods (a) and our method (b). For the t-th frame, prior works process video sequences using small temporal sliding windows with attention or optical flow, restricting cost information propagation. Our method captures long-range spatio-temporal relationships across the input sequence by constructing and updating a compact memory buffer.

To address the task of dynamic stereo matching, recent approaches start to incorporate temporal cues from two main perspectives to achieve temporally consistent estimation. Some methods [30, 61, 12] refine the current disparity with disparity or motion of previous neighbor frame, while achieving limited improvements in temporal consistency due to the narrow temporal context. Secondly, other approaches [24, 22] (Fig. 1 (a)) expand the temporal receptive field by using attention mechanisms to model spatio-temporal relationships [24] within a sliding window while treating all frames equally, which overlooks variations in frame reliability. BiDAStereo [22] further depends on optical flow priors for alignment, may incurring errors from flow inaccuracies and high computational cost. Overall, video-based methods face a core trade-off: narrow context yields marginal improvements, whereas naively using all frames drives up computation without reliability awareness.

Naturally, these considerations lead to a key question: How can we design a model that effectively models long-range temporal relationships while maintaining computational efficiency? To answer this question, we draw inspiration from recent advances in sequence processing and bring a memory buffer into the dynamic stereo matching task. We present **P**ick-and-**P**lay **M**emory for dynamic **Stereo** matching named **PPMStereo** which enables effective and efficient utilization of reference frames for long-range spatio-temporal modeling by dynamically reducing redundant frames while selectively retaining and leveraging the most valuable frames throughout the video sequence to ensure accuracy and efficiency, as illustrated in Fig. 1 (b).

Specifically, our method draws inspiration from human decision-making in complex scenarios, which typically involves the 'pick' process that identifies the most essential elements from a set of candidates and the 'play' process that meticulously balances and leverages the identified elements [5, 18, 38]. In this paper, we propose a novel Pick-and-Play Memory construction method for video stereo matching. Specifically, the 'pick' process identifies the most relevant K frames from T reference frames for the current frame. To facilitate this process, we introduce a novel Quality Assessment Module (QAM), which evaluates each frame's contribution by jointly evaluating confidence, redundancy, and similarity of reference frames. Upon identifying the most relevant K frames, the 'play' process adaptively weights the importance of the features extracted from those K selected frames via a dynamic memory modulation mechanism. Subsequently, we utilize an attention-based memory read-out mechanism that queries the high-quality memory buffer using the current frame's contextual feature, yielding temporally and spatially aggregated cost features. By combining this aggregated cost feature with the current cost and context features, we can use GRU modules to regress the residual disparities.

Extensive experiments show that our method achieves state-of-the-art temporal consistency and accuracy. Specifically, on both the clean and final pass of the Sintel [6] dataset, our model achieves a temporal end-of-point error (TEPE) of 0.62 and 1.11 pixels, with 3-pixel error rates of 5.19% and 7.64%, respectively. Compared to the previous SoTA method, BiDAStereo [22], this represents a 17.3% and 9.02% reduction in TEPE and a 9.74% and 10.32% improvement in 3-pixel error rate, while enjoying lower computational costs. Overall, the contributions of our work can be summarized as follows: (1) We introduce PPMStereo, the first work that successfully builds a memory buffer to tackle dynamic stereo matching, allowing for long-range spatio-temporal modeling in a computationally efficient way. (2) We propose a novel 'Pick-and-Play' memory buffer construction method that first identifies the key subset of reference frames with the pick process and then effectively aggregates them with a play process, enabling highly accurate and temporally consistent disparity estimation. (3) Extensive experiments demonstrate that PPMStereo achieves state-of-the-art performance across multiple dynamic stereo matching benchmarks.

2 Related Work

Deep Stereo Matching. Existing deep stereo matching methods [47] primarily focus on cost volume aggregation for network and representation design. These approaches are generally categorized into regression-based [34, 25, 59, 62, 42, 56, 53] and iterative-based methods [32, 27, 54, 58, 50, 64]. Regression-based methods typically regress a probability volume to estimate disparity maps, which can be further divided into 2D [34, 31, 59, 54] and 3D cost aggregation approaches [19, 62, 42, 33, 43, 53, 52]. These methods either directly regress disparity across a predefined global range [25, 62, 53] or employ a coarse-to-fine refinement strategy to improve accuracy [42, 43, 33]. Recently, iterative-based methods [57, 32, 4, 54, 28, 50, 55, 51, 26] have emerged as the dominant paradigm in stereo matching. These methods leverage multi-level GRU or LSTM modules to iteratively refine disparity maps through recurrent cost volume retrieval, achieving state-of-the-art performance. However, despite their remarkable results, these approaches infer disparities independently for each frame, ignoring temporal correlations across video sequences. As a result, they often suffer from poor temporal consistency, which manifests as flickering artifacts in the disparity outputs.

Dynamic Stereo Matching. A few methods in stereo matching have focused on leveraging temporal cues from dynamic scenes to enhance disparity consistency. These methods can be mainly categorized into two paradigms: (i) **Adjacent-frame Integration**, which propagates disparity or motion fields from the immediately preceding frame to maintain local temporal smoothness. These works [30, 65, 12, 61] typically employ warped disparity or motion estimates for robust initialization, thereby enhancing the temporal consistency. However, these methods are limited by their reliance on only the most recent frame, resulting in a narrow temporal receptive field. (ii) **Multi-frame Integration**, which employs sliding-window aggregation across extended temporal contexts to enforce temporal consistency through attention mechanisms (DynamicStereo) [24] or optical flow priors (BiDAStereo) [22]. Despite their strengths, attention-based methods treat all frames equally without assessing the reliability of reference frames and suffer from high computational costs with a large window. Additionally, flow-based methods are sensitive to optical flow estimation errors and introduce extra computational overheads. In contrast, our method effectively aggregates long-range spatio-temporal information from a compact yet high-quality memory buffer. Thanks to our 'pick' process, PPMStereo remains computationally efficient, even with the enlarged temporal window.

Memory Cues for Video Tasks. Prior works have explored memory model [45] across various video tasks, including optical flow [15], segmentation [37, 66, 9, 10], tracking [60, 17], and video understanding [44, 20], demonstrating its significant effectiveness for video-related tasks. Among them, XMem [9] consolidates memory by selecting prototypes and evicting obsolete features via a least-frequently-used policy, while RMem [66] improves the segmentation accuracy by using a fixed frame memory bank [1]. Prior works have explored memory model [45] across various video tasks, including optical flow [15], segmentation [37, 66, 9, 10], and video understanding [44, 20], demonstrating its significant effectiveness for video-related tasks. Among them, XMem [9] consolidates memory by selecting prototypes and evicting obsolete features via a least-frequently-used policy, while RMem [66] improves the segmentation accuracy by using a fixed frame memory bank [1]. The closest related work is MemFlow [15], which develops an adjacent-frame memory buffer framework to aggregate spatio-temporal motion for optical flow estimation. While effective for optical flow, MemFlow yields limited gains when directly applied to dynamic stereo matching, as it only retains the immediate adjacent frame. Expanding its temporal scope without reliability assessment introduces redundant and noisy cues. In contrast, our method adaptively updates and modulates the most valuable memory cues across the entire sequence, enabling robust long-range spatio-temporal modeling while filtering out inferior ones, leading to significant performance improvements.

3 Methodology

3.1 Overview

Dynamic stereo matching seeks to recover a sequence of temporally consistent disparity maps $\{\mathbf{d}^t\}_{t\in(1,T)} \in \mathbb{R}^{H\times W}$ from stereo video frames $\{\mathbf{I}_L^t, \mathbf{I}_R^t\}_{t\in(1,T)} \in \mathbb{R}^{H\times W\times 3}$, where T is the number of frames, H and W are the height and width dimensions. However, prior approaches struggle to capture long-range temporal dependencies without incurring prohibitive cost. To address this, we introduce **PPMStereo**, which augments the DynamicStereo backbone [24] with a Pick-
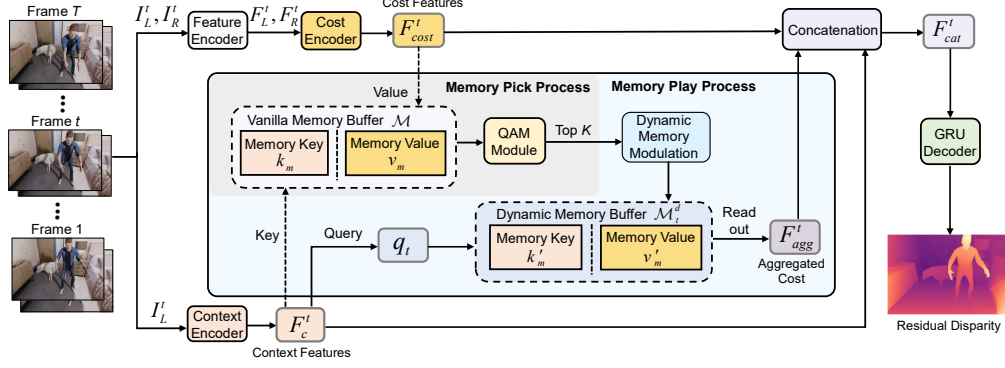
Figure 2: An overview of PPMStereo. The gray part is the memory 'pick' process, and the blue part is the memory play process. Our PPMStereo employs a dynamic memory buffer for modeling long-range spatio-temporal relationships while maintaining computational efficiency.

and-Play Memory (PPM) module that selectively aggregates high-quality references into a compact, query-adaptive buffer, thereby strengthening spatio-temporal modeling while remaining efficient. As illustrated in Fig. 2, the overall pipeline proceeds as follows: *(1) Feature Extraction:* a shared encoder extracts multi-scale features $\{F_L^t, F_R^t\}_{(s)} \in \mathbb{R}^{sH \times sW \times C}$ at scales $s \in \{1/16, 1/8, 1/4\}$, with C channels. These pyramidal representations provide both receptive-field diversity and a convenient substrate for multi-scale matching. *(2) Cost Volume Construction:* at each time step t, we construct a 3D correlation volume from $\{F_L^t, F_R^t\}_{(s)}$ and pass it through a lightweight cost encoder to obtain matching costs F_{cost}^t, subsequently projected to a value embedding v_t. *(3) Context Encoding:* A context encoder operating on the left view produces F_c^t, which are linearly projected to a query q_t and k_t. *(4) Memory Buffer Initialization and Update:* To expose the model to long-range spatio-temporal correlations, we initialize a vanilla memory $\mathcal{M} = \{k_m \in \mathbb{R}^{L \times C}, v_m \in \mathbb{R}^{L \times C}\}$ that stores $k_m = \{k_1, \ldots, k_T\}$ and $v_m = \{v_1, \ldots, v_T\}$ with $L = T \times sH \times sW$. This naive memory buffer stores all reference-frame features, making per-iteration queries prohibitively expensive. To retain accuracy without sacrificing efficiency, we introduce the *Pick-and-Play Memory (PPM)*: driven by a Quality Assessment module (omitting the iteration index n for brevity), PPM first *picks* the most informative references to construct a compact, dynamic buffer $\mathcal{M}_t^d = \{k_m' \in \mathbb{R}^{L' \times C}, v_m' \in \mathbb{R}^{L' \times C}\}$ with $L' = K \times sH \times sW$ and $K \ll T$, and then *plays* by adaptively weighting these entries to produce aggregated cost features that balance contributions across the selected frames. *(6) Iterative Refinement:* following a RAFT-style iterative scheme [32], we alternate GRU-based updates of disparity estimates with PPM-based memory updates, progressively refining $\{d_t\}$ while preserving temporal consistency and computational efficiency .

3.2 Memory Pick Process

Naive heuristic strategies, such as random selection or solely keeping the latest frame, are unreliable. Since the former neglects frame reliability and relevance, while the latter suffers from limited temporal context and knowledge drift [36]. To this end, we introduce a Quality Assessment Module (QAM) that explicitly evaluates the quality of memory elements $\{k_m, v_m\}$ in the vanilla buffer for dynamic stereo matching. To activate QAM, we define two complementary scores that quantify each reference frame's contribution to the final accuracy: a confidence score $\mathbf{S}_t^c$ computed over the value embeddings v_m to prioritize reliable evidence, and a redundancy-aware relevance score $\mathbf{S}_t^r$ computed over the key embeddings k_m to suppress repetitive or low-information entries. The full procedure is summarized in Algorithm 1. $\mathbf{S}_t^c$ and $\mathbf{S}_t^r$ are used together to enable the construction of a compact, high-quality memory $\mathcal{M}_t^d$ that preserves the most informative cross-frame cues.

Confidence Score. Memory values v_m encode pixel-wise horizontal displacements, which are critical for disparity estimation. These features naturally indicate the reliability of its disparity estimation. To this end, we employ a lightweight confidence network[3] that transforms $v_m \in \mathbb{R}^{T \times sH \times sW \times C}$ into confidence maps $u_t \in \mathbb{R}^{T \times sH \times sW}$, quantifying whether memory values v_m corresponding to

---

[3]The confidence network consists of two convolutional layers followed by a sigmoid activation, which ensures efficient and effective confidence estimation.

Algorithm 1 Pseudo code of Pick-and-Play Memory

Input: Video frames sequence $\{I_L^t, I_R^t\}$ of video length T, GRU n-th iterations, $K \ll T$

Intermediates: Vanilla Memory: $\mathcal{M} = \left\{ k_m \in \mathbb{R}^{L \times C}, v_m \in \mathbb{R}^{L \times C)} \right\}$, $L = T \times sH \times sW$

        The query: $q_t \in \mathbb{R}^{1 \times sH \times sW}$, $s \in \{1/16, 1/8, 1/4\}$ is the downsampled scale

        Dynamic Memory: $\mathcal{M}_t^d = \left\{ k_m' \in \mathbb{R}^{L' \times C}, v_m' \in \mathbb{R}^{L' \times C)} \right\}$, $L' = K \times sH \times sW$

Output: The residual disparity map at n-th GRU iteration: Δd_t^n

---

1: **while** $t \leq T$ **do**

       Memory Pick Process:

2:     $\mathbf{S}_t = \mathbf{S}_t^c + \mathbf{S}_t^r$ # QAM, evaluate the quality of memory elements k_m and v_m

3:     $\mathcal{I}_t = \{i \mid \text{rank}\,(\mathbf{S}_t[i]) \leqslant K\}$ # Select top-K reference frames

4:     $\mathcal{M}_t^d = \{k_m' = \text{Cat}\,[\{k_i \mid i \in \mathcal{I}_t\}], v_m' = \text{Cat}\,[\{v_i \mid i \in \mathcal{I}_t\}]\}$

       Memory Play Process:

5:     $\overline{\mathbf{S}}_t[i] = \frac{\mathbf{S}_t[i]}{\sum_i \mathbf{S}_t[i]}, i \in \mathcal{I}_t$ # Balance the contribution of selected memory entries

6:     $q_t = q_t + p_t, \quad k_m' = \overline{\mathbf{S}}_t \cdot k_m' + P_{\mathcal{I}_t}$ # Dynamic memory modulation

7:     $F_{agg}^t = \text{Read-out}(q_t, \mathcal{M}_t^d)$ # Aggregate high-quality spatio-temporal cost information

8:     $\Delta d_t^n = \text{GRU}(F_{agg}^t, F_{cost}^t, F_c^t)$ # Produce the disparity map at the n-th iteration

---

accurate disparity outputs. These confidence maps can provide a frame-level reliability measure by estimating the uncertainty of predicted disparity [42, 49]. During training for N iterations, the confidence maps are supervised using an L_1 loss function to enforce consistency with their ground-truth counterparts. The ground-truth confidence score $\hat{u}_t$ is computed as follows:

$$\hat{u}_t = \exp\left(-\left|\frac{d_t - \hat{d}_t}{\sigma}\right|\right), \tag{1}$$

where d_t and $\hat{d}_t$ represent the predicted and ground-truth disparities for the t-th frame, respectively, and σ is a hyper-parameter empirically set to 5. Over N iterations, we compute the confidence loss L_{conf} across all timesteps $u_{t \in (1,T)}$ as follows:

$$\mathcal{L}_{conf} = \sum_{t=1}^{T} \sum_{n=1}^{N} \gamma^{N-n} \|u_t^n - \hat{u}_t^n\|_1, \tag{2}$$

where n denotes the number of iterations and γ is a decay factor set as 0.9. To obtain a frame-level confidence score $\mathbf{S}_t^c \in \mathbb{R}^{1 \times T}$, we apply average pooling across the spatial dimensions of the confidence maps u_t.

Redundancy-aware Relevance Score. Relying solely on the confidence score is insufficient, as adjacent frames often exhibit strong spatio-temporal correlations, which can result in higher confidence scores. This, in turn, introduces feature redundancy and suppresses contributions from more diverse frames, ultimately limiting the diversity and effectiveness of the memory buffer. To mitigate this issue, we propose a redundancy-aware relevance score to evaluate memory keys k_m, balancing semantic consistency and memory diversity. First, we compute an inter-frame similarity score $\mathbf{Sim}_t \in \mathbb{R}^{1 \times T}$ between the current query q_t and the memory keys k_m, measuring semantic alignment while preserving temporal coherence. For computational efficiency, we employ an attention mechanism combined with spatial downsampling. Specifically, average pooling reduces the spatial resolution of the query and memory keys from $sH \times sW$ to $sH' \times sW'$, followed by L2-normalization
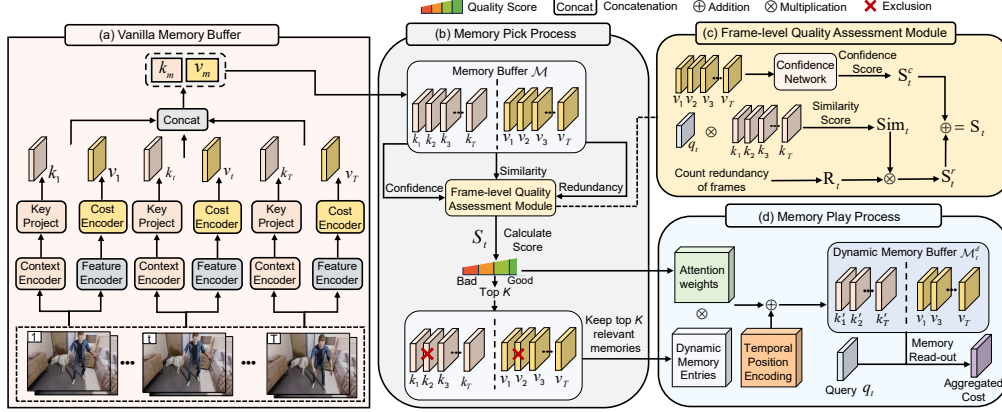
Figure 3: The details of our Pick-and-play Memory Construction Process (PPM).

along the combined feature dimension $f = sH' \times sW' \times C$. The similarity score is computed as:

$$\mathbf{sim}_t = \phi(q_t)\phi(k_m)^T, \text{where } \phi(x) = \frac{\text{AvgPool}(x)}{||\text{AvgPool}(x)||_2} \tag{3}$$

where $\phi(k_m) \in \mathbb{R}^{T \times f}$ and AvgPool($\cdot$) denotes the average pooling operation. However, focusing solely on the most similar regions may overlook occluded areas. Since occluded regions in adjacent frames tend to be highly similar, they can be challenging to reference effectively. To mitigate this, we then introduce a redundancy-aware regularizer $\mathbf{R}_t[k] = e^{-\frac{t_k}{T}}$, where t_k denotes the the cumulative number of times the k-th frame has been selected for the dynamic memory buffer across previous GRU iterations. This term dynamically downweights overused frames while promoting underutilized yet informative references, ensuring a compact yet diverse memory buffer. The final redundancy-aware relevance score $\mathbf{S}_t^r \in \mathbb{R}^{1 \times T}$ combines redundancy and similarity:

$$\mathbf{S}_t^r = \mathbf{R}_t \cdot \mathbf{sim}_t \tag{4}$$

By jointly considering relevance and diversity, our approach enhances feature aggregation while minimizing redundancy, leading to more robust and efficient memory-based processing.

Memory Updating via QAM. We compute the total quality metric for each memory frame as $\mathbf{S}_t = \mathbf{S}_t^c + \mathbf{S}_t^r$ by integrating confidence and redundancy-aware relevance scores. This integrated scoring enables dynamic memory update by retaining the most informative entries via a top-K selection mechanism, ensuring robust adaptation to varying video scenarios while preventing memory overload. Specifically, for the vanilla memory buffer $\mathcal{M} = \{k_m, v_m\}$ with the corresponding quality scores $\mathbf{S}_t \in \mathbb{R}^{1 \times T}$, we sort the quality scores in descending order and only retain the top-K memory features in the vanilla memory buffer as:

$$\mathcal{I}_t = \{i \mid \text{rank}(\mathbf{S}_t[i]) \leqslant K\} \tag{5}$$

$$\mathcal{M}_t^d = \{\text{Cat}\left[\{k_i \mid i \in \mathcal{I}_t\}\right], \text{Cat}\left[\{v_i \mid i \in \mathcal{I}_t\}\right]\}, \tag{6}$$

{where rank($\cdot$) denotes the ranking position in descending order, with rank = 1 corresponding to the highest score, $\mathcal{I}_t$ is the set of selected frames' indices, and Cat denotes the concatenation. The resulting dynamic memory buffer $\mathcal{M}_t^d$ comprises keys $k_m' = \{k_i\}_{(i \in \mathcal{I}_t)}$, and values $v_m' = \{v_i\}_{(i \in \mathcal{I}_t)}$. By enforcing $K \ll T$, this strategy efficiently handles arbitrary video sequences while providing high-quality spatio-temporal cues for dynamic memory aggregation.

3.3 Memory Play Process

After the pick process selects the top-K most relevant memory entries for our dynamic memory buffer $\mathcal{M}_t^d$, we argue that not all selected frames contribute equally to disparity estimation. To further weigh their importance, we introduce a memory play process that dynamically weights the selected memory entries based on learned quality scores. Since dynamic memory construction inherently disrupts temporal ordering, we incorporate temporal position encoding into the framework, ensuring temporal awareness.

Dynamic Memory Modulation. Building on this foundation, we propose a unified dynamic memory modulation strategy that jointly optimizes feature reliability and temporal consistency. Specifically, given the estimated quality score $\mathbf{S}_t$, we first obtain the relative significance of the frames:

$$\overline{\mathbf{S}}_t[i] = \frac{\mathbf{S}_t[i]}{\sum_i \mathbf{S}_t[i]}, i \in \mathcal{I}_t \tag{7}$$

Following [16], we initialize positional encodings (PE) to align with the original memory buffer length T, formalized as $P_{1:T}$. This initialization ensures temporal coherence in feature representation. Therefore, the 'play' process subsequently operates as follows:

$$q_t = q_t + P_t, \qquad k'_m = \overline{\mathbf{S}}_t \cdot k'_m + P_{\mathcal{I}_t} \tag{8}$$

where P_t denotes the positional encoding at timestep t, and $\overline{\mathbf{S}}_t$ represents the aggregated importance weights over the index set $\mathcal{I}_t$. Leveraging the estimated quality scores as reliability indicators, we prioritize more reliable memory entries while maintaining computational efficiency.

Memory Read-out. We aggregate cost features through an attention-based memory read-out mechanism from the dynamic memory buffer $\mathcal{M}_t^d$. Specifically, we first compute soft attention weights by measuring the similarity between the query q_t and modulated memory keys k'_m. The aggregated cost features F_{agg}^t are then obtained by weighting the memory values v'_m through these attention weights:

$$F_{agg}^t = F_{cost}^t + \alpha \cdot \mathrm{Softmax}\left(1/\sqrt{D_k} \times q_t \times {k'_m}^{\mathsf{T}}\right) \times v'_m, \tag{9}$$

where α is a learnable scalar initialized from 0. In this way, we employ the attention to gather additional temporal information. With the context, cost, and aggregated cost features, we can now output a residual disparity map through a GRU unit at the n-th iteration: $\Delta d_n = \mathrm{GRU}(F_{cost}^t, F_{agg}^t, F_c^t)$. After N iterations of PPM and GRU, we can get the final disparity map.

Loss Functions. Our disparity loss functions are inherited from the previous works [24, 22]. Generally, for N iterations, we supervise our network with L_1 distance between our a series of residual flows $\{d_1, \ldots, d_T\}$ and the ground-truth $\hat{d}_t$ with exponentially increasing weights:

$$\mathcal{L}_d = \sum_{t=1}^{T} \sum_{n=1}^{N} \gamma^{N-n} \left\| d_t^n - \hat{d}_t \right\|_1, \tag{10}$$

where γ and N are set as 0.9 and 10, respectively. Therefore, the total loss function is as follows:

$$\mathcal{L}_{total} = \mathcal{L}_d + \mathcal{L}_{conf}. \tag{11}$$

4 Experiments

4.1 Datasets

Our work focuses on videos captured with moving cameras, rendering standard image benchmarks like Middlebury [39], ETH3D [40] unsuitable. For training and evaluation, we employ three synthetic and one real-world stereo video dataset, all featuring dynamic scenes: **SceneFlow (SF)** [34] comprising FlyingThings3D, Driving, and Monkaa, with FlyingThings3D featuring moving 3D objects against varied backgrounds. **Dynamic Replica (DR)** [24], a synthetic indoor dataset with non-rigid objects such as people and animals. **Sintel** [6], a synthetic movie dataset available in clean and final passes. **South Kensington (SV)** [23], a real-world stereo dataset without ground truth data, capturing daily scenarios. We use them for generalization evaluation. Following prior work [24, 22], we train on synthetic datasets (SF and DR + SF) and evaluate the performance on **Sintel**, **DR**, and **SV**.

4.2 Implementation Details

We implement PPMStereo in PyTorch, training on $8\times$ A100 GPUs (batch size = 2) using 320×512 crops from 5-frame sequences, evaluated at full resolution with 20-frame sequences. We use AdamW (lr = 0.0003) with one-cycle scheduling, training for $180k$ iterations (≈ 4.5 days). Data augmentation follows DynamicStereo [24], including random crops and saturation shifts. For efficient memory read-out, we employ FlashAttention [13]. Following prior works [22, 24], we set the number of evaluation

Table 1: Quantitative comparison with SoTA methods. Abbreviations: K - KITTI [35], M - Middlebury [39], ISV–Infinigen SV [23], VK – Virtual KITTI2 [7]. CREStereo utilize 7 datasets for training, including SF [34], Sintel [6], FallingThings [48], InStereo2K [3], Carla [14], AirSim [41], and CREStereo dataset [27]. The best results are in bold, and the second-best are underlined.

| Training data | Method | Sintel Stereo | | | | | | | | Dynamic Replica | | | |
| | | Clean | | | | Final | | | | First 150 frames | | | |
		δ_{3px}	TEPE	δ_{1px}^t	δ_{3px}^t	δ_{3px}	TEPE	δ_{1px}^t	δ_{3px}^t	δ_{1px}	TEPE	δ_{1px}^t	δ_{3px}^t
SF	CODD [30]	8.68	1.44	10.8	5.65	17.46	2.32	18.56	9.79	6.59	0.105	1.04	0.42
	RAFT-Stereo [32]	6.12	0.92	9.33	4.51	10.40	2.10	13.69	7.08	5.51	0.145	2.03	0.65
	DynamicStereo [24]	6.10	0.77	8.41	3.93	8.97	1.45	11.95	5.98	3.44	0.087	0.75	0.24
	BiDAStereo [22]	5.94	0.73	8.29	3.79	8.78	1.26	11.65	5.53	5.17	0.103	1.11	0.40
	PPMStereo (Ours)	5.34	0.64	7.38	3.40	7.87	1.14	10.12	4.99	2.95	0.066	0.67	0.23
	PPMStereo_VDA (Ours)	**4.62**	**0.58**	**6.89**	**3.08**	**7.21**	**1.04**	**9.84**	**4.65**	**2.37**	**0.059**	**0.61**	**0.22**
SF + M + K	CODD [30]	9.11	1.33	12.16	6.23	11.90	2.01	16.16	8.64	10.03	0.152	2.16	0.77
SF + M	RAFT-Stereo [32]	5.86	0.85	8.79	4.13	8.47	1.63	12.40	6.23	3.46	0.114	1.34	0.41
7 datasets (incl. Sintel)	CREStereo [29]	4.58	0.67	6.36	3.26	8.17	1.90	12.29	6.87	1.75	0.088	0.88	0.29
DR + SF	RAFT-Stereo [32]	5.71	0.84	9.15	4.40	9.16	2.27	13.45	7.17	1.89	0.075	0.77	0.25
DR + SF	DynamicStereo [24]	5.77	0.76	8.46	3.93	8.68	1.42	11.93	5.92	3.32	0.075	0.68	0.23
DR + SF	BiDAStereo [22]	5.75	0.75	8.03	3.76	8.52	1.22	11.04	5.30	2.81	0.062	0.62	0.22
DR + SF	**PPMStereo (Ours)**	5.19	0.62	7.21	3.29	7.64	1.11	9.98	4.87	2.52	0.057	0.60	0.20
DR + SF	**PPMStereo_VDA (Ours)**	**4.47**	**0.56**	**6.69**	**2.97**	**7.03**	**1.02**	**9.65**	**4.51**	**1.81**	**0.052**	**0.51**	**0.17**

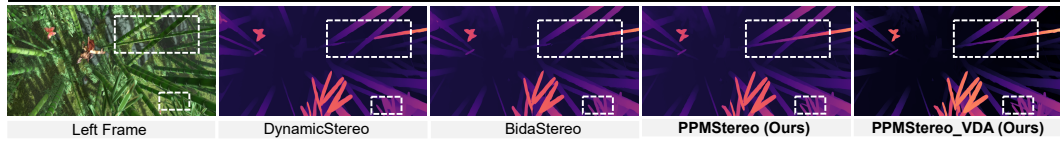| Left Frame | DynamicStereo | BidaStereo | **PPMStereo (Ours)** | **PPMStereo_VDA (Ours)** |

Figure 4: Qualitative comparisons on the Sintel final dataset.

iterations N to 20, while setting $N = 10$ during training. Besides, we adopt n-pixel error rate (δ_{npx}) for accuracy analysis. Additionally, we use the temporal end-point-error (TEPE) to quantify error variation over time, and δ_{npx}^t denotes the percentage of pixels with TEPE exceeding n pixels. Lower values on metrics indicate greater temporal consistency and disparity estimation accuracy. Besides, we replace our original feature extractor with Video Depth Anything (ViT-Small) [8]. This PPMStereo_VDA variant leverages pre-trained representations to further boost performance.

4.3 Comparison with State-of-the-Art Methods

Quantitative Results. As shown in Tab. 1, For the SF version, our PPMStereo achieves state-of-the-art performance, outperforming BiDAStereo [24] by 12.3% & 9.52% and DynamicStereo by 16.8% & 21.3% in TEPE on Sintel clean/final pass. The method also demonstrates strong generalization on Dynamic Replica, surpassing all previous approaches across all metrics. Remarkably, our PPMStereo trained only on synthetic data even largely exceeds the temporal consistency and accuracy of CREStereo [27] on Sintel final pass, despite CREStereo using Sintel data for training. For the SF & DR version, our method achieves superior temporal consistency with a TEPE of 0.057 on Dynamic Replica, significantly outperforming all previous works. Notably, this is achieved with training on only two synthetic datasets, while CREStereo [27] requires seven diverse datasets, demonstrating the efficacy of our long-range temporal modeling. Overall, the results highlight our method's robust performance and generalization ability in both seen and unseen domains. Besides, compared to the previous SoTA method BiDAStereo [22], our method achieves better performance with lower computational costs and memory usage (Please see the appendix for details).

Qualitative Results. Our visual comparisons (Fig. 4) using the DR+SF checkpoint show PPMStereo produces sharper disparity predictions than DynamicStereo [24] and BiDAStereo [22], especially in textureless regions (e.g., glass surfaces) where competing methods exhibit blurring artifacts. Besides, following prior work [22, 24], we validate temporal consistency on static scenes by rendering depth point clouds at 15-degree viewpoint increments (Fig. 5). Our method shows significantly smaller high-

Table 2: Ablations of memory buffer module variants trained on DR+SF. 'OOM' denotes CUDA out of memory. 'Baseline' refers to our backbone model without any memory-related modules.

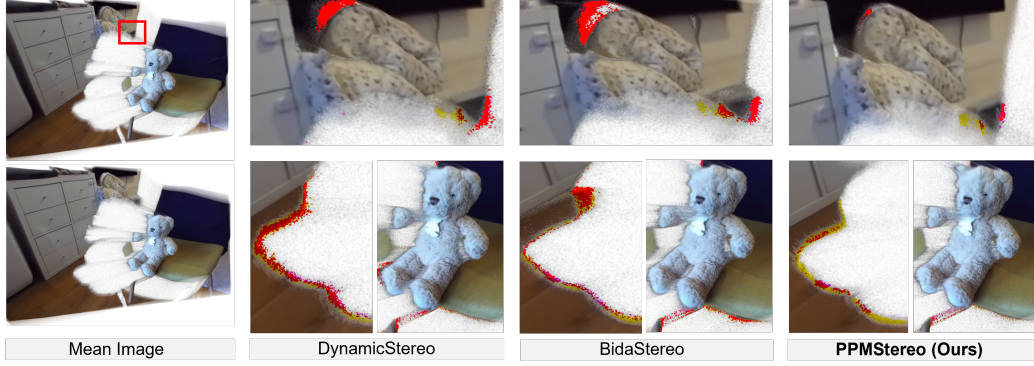| Experiments | Method | Sintel Final | | Dynamic Replica | |
		δ_{3px}	TEPE	δ_{1px}	TEPE
	Baseline	8.65	1.37	3.10	0.074
Memory Buffer	Full	OOM		OOM	
	MemFlow [15]	8.45	1.28	3.11	0.070
	Latest	8.11	1.19	2.89	0.062
	Random	8.42	1.26	2.99	0.064
	XMem [9]	8.04	1.18	2.84	0.061
	RMem [66]	7.93	1.16	2.77	0.061
	Ours	**7.64**	**1.11**	**2.52**	**0.057**
Memory Length	$K = 1$	7.95	1.18	2.70	0.062
	$K = 3$	7.80	1.13	2.58	0.057
	$K = 5$	7.64	1.11	2.52	0.057
	$K = 7$	**7.62**	**1.10**	**2.50**	**0.057**

8

Figure 5: Temporal consistency comparison on 50-frame reconstructed stereo video (all trained on DR + SF). Our method achieves lower variance, demonstrating superior consistency.
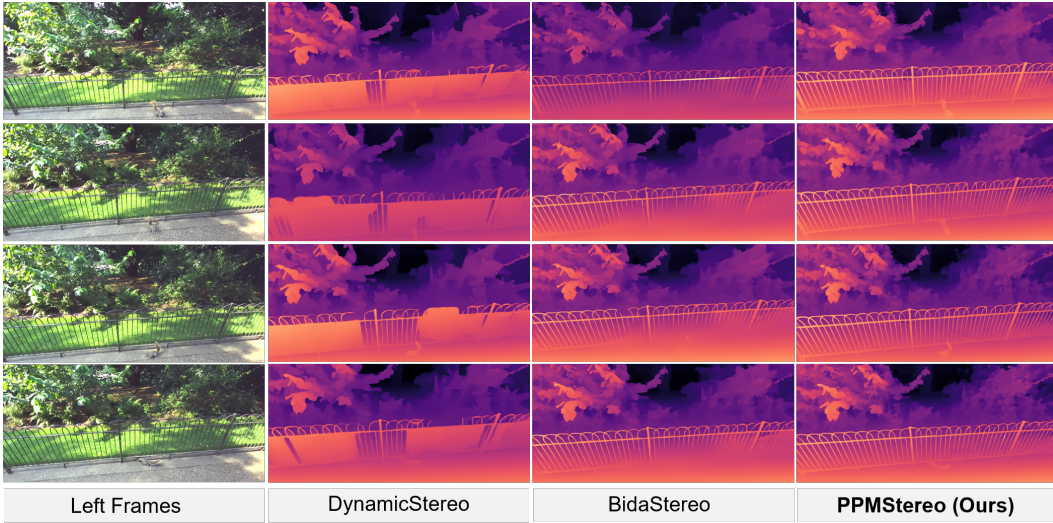

Figure 6: Qualitative generalization comparison on a dynamic outdoor scenario from the SV dataset.

variance regions (> 40 px^2, marked red), confirming superior stability. Furthermore, on the real-world outdoor scenes from the South Kensington dataset [23] (Fig. 6), PPMStereo accurately recovers thin structures such as the fences while maintaining temporal consistency, demonstrating robust generalization to unseen domains. More visualizations are provided in the appendix.

4.4 Ablation Studies

Due to the huge training cost of PPMStereo_VPA, we conduct ablation studies exclusively on PPMStereo below. Besides, all ablated models below are trained on DR + SF.

Memory buffer construction. We train and evaluate 5 different memory buffer variants, namely, keeping frames from (1) full frames (20 frames), (2) MemFlow (1 frame) [15], (3) the latest frames (5 frames), (4) random (5 frames), (5) XMem [9] (distilling all outdated memory features into long-term memory based on attention scores), (6) RMem [66] (5 frames), and (7) ours (5 frames).

Specifically, we replace the memory buffer variants and keep the remaining modules unchanged during training and inference. Table 2 shows three key insights: First, while reference frames improve performance, naive accumulation shows diminishing returns, indicating memory capacity alone is insufficient. Second, frame selection quality critically affects results. The random selection policy underperforms even single-neighbor memory (MemFlow) [15] on Sintel final pass, highlighting selection importance. However, on the DR dataset with minimal inter-frame changes, the random policy performs comparably to advanced variants. Lastly, direct long-term memory integration

9

Table 3: Ablation Study of PPM on Sintel and Dynamic Replica. All models are trained on DR+SF. Note that we directly perform the read-out operation for the ablated model without the 'play' process.

ID	Pick-and-Play Memory		Sintel Final				Dynamic Replica			
	Pick	Play	δ_{3px}	TEPE	δ_{1px}^t	δ_{3px}^t	δ_{1px}	TEPE	δ_{1px}^t	δ_{3px}^t
1	Baseline		8.65	1.37	11.72	5.91	3.10	0.074	0.72	0.23
2	✓		7.81	1.14	10.24	5.07	2.65	0.060	0.64	0.21
3		✓	7.97	1.17	10.36	5.20	2.80	0.062	0.68	0.21
4	✓	✓	**7.64**	**1.11**	**9.98**	**4.87**	**2.52**	**0.057**	**0.60**	**0.20**

Table 4: Ablation study on the 'pick' process. C, Sim, and R denote confidence score, similarity score, and redundancy factor, respectively.

ID	QAM			Sintel Final			Dynamic Replica		
	C	Sim	R	δ_{3px}	TEPE	δ_{3px}^t	δ_{1px}	TEPE	δ_{1px}^t
1	Baseline			7.97	1.17	5.20	2.80	0.062	0.68
2	✓			7.81	1.14	5.06	2.63	0.058	0.65
3	✓	✓		7.74	1.12	4.95	2.57	0.057	0.62
4	✓	✓	✓	**7.64**	**1.11**	**4.87**	**2.52**	**0.057**	**0.60**

Table 5: Ablation study on the 'play' process. Weights and PE denote the weighting operation and the temporal position encoding, respectively.

ID	Play Process		Sintel Final			Dynamic Replica		
	Weights	PE	δ_{3px}	TEPE	δ_{3px}^t	δ_{1px}	TEPE	δ_{1px}^t
1	Baseline		7.81	1.14	5.07	2.65	0.060	0.64
2	✓		7.67	1.12	5.00	2.54	0.060	0.62
3		✓	7.77	1.11	4.93	2.63	0.058	0.61
4	✓	✓	**7.64**	**1.11**	**4.87**	**2.52**	**0.057**	**0.60**

(XMem) shows limited impact, suggesting that simply using all frames may be less effective than the RMem variant. In contrast, our PPM mechanism overcomes these limitations by dynamically identifying and modulating valuable reference frames, achieving significant TEPE improvements on these two datasets (+19.0% TEPE on Sintel and +22.9% TEPE on DR) over the baseline.

Memory length. Table 2 shows the impact of memory length on PPMStereo. Performance improves initially (e.g., +14.8% δ_{1px}^t on Sintel for $K \leq 5$) when trained and evaluated at this memory length, but performance saturates beyond $K = 5$ due to feature redundancy. To balance computational efficiency and model accuracy, we select $K = 5$ as the optimal memory length for our final model.

Contribution of each component. Table 3 shows the proposed PPM module outperforms window-based aggregation through two key processes: (1) The pick process dynamically selects high-quality memory elements from non-adjacent frames, overcoming fixed-window limitations and improving occlusion handling; (2) The play process adaptively weights features by semantic relevance, reducing noise propagation (ID = 3 shows +0.2 on Sintel and +0.017 TEPE improvements on DR compared to the baseline). By combining them, they provide complementary benefits. The pick ensures feature diversity while play suppresses outliers, yielding superior performance in dynamic stereo matching.

QAM. Our QAM module dynamically assesses frame reliability in the memory buffer using a scoring mechanism. We refresh the memory buffer by balancing: (1) cost feature quality (v_m) and (2) redundancy-aware semantic relevance (k_m) (Sec. 3.2). Table 4 shows that our quality score improves both depth accuracy and temporal consistency. Fig. 7 further confirms the confidence map's strong correlation with the error map, validating it as a reliable quality indicator for v_m.

(a) Left Frame   (b) Predicted Disparity Map   (c) Error Map   (d) Confidence Map

Figure 7: Visualization of error map and confidence map. Brighter regions denote higher uncertainty.

Memory modulation. Our proposed memory modulation mechanism (Sec. 3.3) further enhances spatio-temporal modeling, achieving a performance gain with +0.17 δ_{3px} and +0.13 δ_{1px} improvements on the Sintel Final and DR, respectively, as seen in Table 5. The adaptive weighting mechanism dynamically prioritizes the most important spatio-temporal features, highlighting accuracy improvements. Meanwhile, learned positional embeddings endow the model with temporal awareness, improving the overall temporal consistency. Experiments show that these components work together to strengthen the model's ability to capture long-range dependencies and distinguish key spatio-temporal patterns.

5 Conclusion

In this paper, we introduce PPMStereo, the first framework, to our knowledge, to leverage high-quality memory for dynamic stereo matching. By selectively updating and modulating the most valuable memory entries, our proposed pick-and-play memory construction mechanism enables the integration of cost information across long-range spatio-temporal connections, ensuring temporally consistent stereo matching. Extensive experiments demonstrate the effectiveness of our approach across diverse datasets, highlighting its generic applicability.

Acknowledgment

References

[1] Peter Auer. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research (JMLR)*, 3(Nov):397–422, 2002.

[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 *2021 IEEE International Conference on Intelligent Robots and Systems (IROS)*, pages 3542–3548. IEEE, 2021.

[3] Wei Bao, Wei Wang, Yuhua Xu, Yulan Guo, Siyu Hong, and Xiaohu Zhang. Instereo2k: a large real dataset for stereo matching in indoor scenes. *Science China Information Sciences*, 63:1–11, 2020.

[4] 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 Computer Vision and Pattern Recognition Conference (CVPR)*, pages 1013–1027, 2025.

[5] Rajesh Bhargave, Amitav Chakravarti, and Abhijit Guha. Two-stage decisions increase preference for hedonic options. *Organizational Behavior and Human Decision Processes*, 130:123–135, 2015.

[6] Daniel J Butler, Jonas Wulff, Garrett B Stanley, and Michael J Black. A naturalistic open source movie for optical flow evaluation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 611–625. Springer, 2012.

[7] Yohann Cabon, Naila Murray, and Martin Humenberger. Virtual kitti 2. *arXiv preprint arXiv:2001.10773*, 2020.

[8] Sili Chen, Hengkai Guo, Shengnan Zhu, Feihu Zhang, Zilong Huang, Jiashi Feng, and Bingyi Kang. Video depth anything: Consistent depth estimation for super-long videos. 2025.

[9] Ho Kei Cheng and Alexander G Schwing. Xmem: Long-term video object segmentation with an atkinson-shiffrin memory model. In *European Conference on Computer Vision (ECCV)*, pages 640–658. Springer, 2022.

[10] Ho Kei Cheng, Yu-Wing Tai, and Chi-Keung Tang. Rethinking space-time networks with improved memory coverage for efficient video object segmentation. *Advances in Neural Information Processing Systems (NeuralIPS)*, 34:11781–11794, 2021.

[11] Junda Cheng, Longliang Liu, Gangwei Xu, Xianqi Wang, Zhaoxing Zhang, Yong Deng, Jinliang Zang, Yurui Chen, Zhipeng Cai, and Xin Yang. Monster: Marry monodepth to stereo unleashes power. 2025.

[12] Ziang Cheng, Jiayu Yang, and Hongdong Li. Stereo matching in time: 100+ fps video stereo matching for extended reality. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 8719–8728, 2024.

[13] Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in Neural Information Processing Systems (NeuralIPS)*, 35:16344–16359, 2022.

[14] Jean-Emmanuel Deschaud. Kitti-carla: a kitti-like dataset generated by carla simulator. *arXiv preprint arXiv:2109.00892*, 2021.

[15] Qiaole Dong and Yanwei Fu. Memflow: Optical flow estimation and prediction with memory. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19068–19078, 2024.

[16] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[17] Zhihong Fu, Qingjie Liu, Zehua Fu, and Yunhong Wang. Stmtrack: Template-free visual tracking with space-time memory networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13774–13783, 2021.

[18] Herbert Gintis. A framework for the unification of the behavioral sciences. *Behavioral and brain sciences*, 30(1):1–16, 2007.

[19] Xianda Guo, Chenming Zhang, Youmin Zhang, Wenzhao Zheng, Dujun Nie, Matteo Poggi, and Long Chen. Lightstereo: Channel boost is all you need for efficient 2d cost aggregation. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 8738–8744. IEEE, 2025.

[20] Bo He, Hengduo Li, Young Kyun Jang, Menglin Jia, Xuefei Cao, Ashish Shah, Abhinav Shrivastava, and Ser-Nam Lim. Ma-lmm: Memory-augmented large multimodal model for long-term video understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13504–13514, 2024.

[21] Heiko Hirschmuller. Stereo processing by semiglobal matching and mutual information. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 30(2):328–341, 2007.

[22] Junpeng Jing, Ye Mao, and Krystian Mikolajczyk. Match-stereo-videos: Bidirectional alignment for consistent dynamic stereo matching. In *European Conference on Computer Vision (ECCV)*, pages 415–432. Springer, 2024.

[23] Junpeng Jing, Ye Mao, Anlan Qiu, and Krystian Mikolajczyk. Match stereo videos via bidirectional alignment. 2024.

[24] Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Dynamicstereo: Consistent dynamic depth from stereo videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13229–13239, 2023.

[25] 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 (ICCV)*, pages 66–75, 2017.

[26] Jiahao Li, Xinhong Chen, Zhengmin Jiang, Qian Zhou, Yung-Hui Li, and Jianping Wang. Global regulation and excitation via attention tuning for stereo matching. *arXiv preprint arXiv:2509.15891*, 2025.

[27] 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 Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16263–16272, 2022.

[28] Kunhong Li, Longguang Wang, Ye Zhang, Kaiwen Xue, Shunbo Zhou, and Yulan Guo. Los: Local structure guided stereo matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[29] Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. Exploring plain vision transformer backbones for object detection. In *European Conference on Computer Vision (ECCV)*, pages 280–296. Springer, 2022.

[30] Zhaoshuo Li, Wei Ye, Dilin Wang, Francis X Creighton, Russell H Taylor, Ganesh Venkatesh, and Mathias Unberath. Temporally consistent online depth estimation in dynamic scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3018–3027, 2023.

[31] Zhengfa Liang, Yulan Guo, Yiliu Feng, Wei Chen, Linbo Qiao, Li Zhou, Jianfeng Zhang, and Hengzhu Liu. Stereo matching using multi-level cost volume and multi-scale feature constancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2019.

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

[33] Yamin Mao, Zhihua Liu, Weiming Li, Yuchao Dai, Qiang Wang, Yun-Tae Kim, and Hong-Seok Lee. Uasnet: Uncertainty adaptive sampling network for deep stereo matching. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 6311–6319, 2021.

[34] 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 Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4040–4048, 2016.

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

[36] Roy Miles, Mehmet Kerim Yucel, Bruno Manganelli, and Albert Saa-Garriga. Mobilevos: Real-time video object segmentation contrastive learning meets knowledge distillation. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10480–10490, 2023.

[37] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

[38] Laurie R Santos and Alexandra G Rosati. The evolutionary roots of human decision making. *Annual review of psychology*, 66(1):321–347, 2015.

[39] 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 *German conference on pattern recognition (GCPR)*, pages 31–42. Springer, 2014.

[40] Thomas Schöps, Johannes L. Schönberger, S. 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 Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2538–2547, 2017.

[41] Shital Shah, Debadeepta Dey, Chris Lovett, and Ashish Kapoor. Airsim: High-fidelity visual and physical simulation for autonomous vehicles. In *Field and Service Robotics: Results of the 11th International Conference*, pages 621–635. Springer, 2018.

[42] Zhelun Shen, Yuchao Dai, and Zhibo Rao. CFNet: Cascade and fused cost volume for robust stereo matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13906–13915, 2021.

[43] Zhelun Shen, Xibin Song, Yuchao Dai, Dingfu Zhou, Zhibo Rao, and Liangjun Zhang. Digging into uncertainty-based pseudo-label for robust stereo matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 30(2):1–18, 2023.

[44] Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, et al. Moviechat: From dense token to sparse memory for long video understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18221–18232, 2024.

[45] Sainbayar Sukhbaatar, Jason Weston, Rob Fergus, et al. End-to-end memory networks. *Advances in Neural Information Processing Systems (NeuralIPS)*, 28, 2015.

[46] 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 Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14362–14372, 2021.

[47] Fabio Tosi, Luca Bartolomei, and Matteo Poggi. A survey on deep stereo matching in the twenties. *International Journal of Computer Vision (IJCV)*, 133(7):4245–4276, 2025.

[48] Jonathan Tremblay, Thang To, and Stan Birchfield. 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 (CVPRW)*, pages 2038–2041, 2018.

[49] Chen Wang, Xiang Wang, Jiawei Zhang, Liang Zhang, Xiao Bai, Xin Ning, Jun Zhou, and Edwin Hancock. Uncertainty estimation for stereo matching based on evidential deep learning. *Pattern Recognition*, 124:108498, 2022.

[50] Xianqi Wang, Gangwei Xu, Hao Jia, and Xin Yang. Selective-stereo: Adaptive frequency information selection for stereo matching. *arXiv preprint arXiv:2403.00486*, 2024.

[51] Yun Wang, Junjie Hu, Junhui Hou, Chenghao Zhang, Renwei Yang, and Dapeng Oliver* Wu. Rose: Robust self-supervised stereo matching under adverse weather conditions. *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, 2025.

[52] Yun Wang, Kunhong Li, Longguang Wang, Junjie Hu, Dapeng Oliver Wu, and Yulan Guo. Adstereo: Efficient stereo matching with adaptive downsampling and disparity alignment. *IEEE Transactions on Image Processing (TIP)*, 2025.

[53] Yun Wang, Longguang Wang, Kunhong Li, Yongjian Zhang, Dapeng Oliver Wu, and Yulan Guo. Cost volume aggregation in stereo matching revisited: A disparity classification perspective. *IEEE Transactions on Image Processing (TIP)*, 2024.

[54] Yun Wang, Longguang Wang, Hanyun Wang, and Yulan Guo. SPNet: Learning stereo matching with slanted plane aggregation. *IEEE Robotics and Automation Letters*, 2022.

[55] Yun Wang, Longguang Wang, Chenghao Zhang, Yongjian Zhang, Zhanjie Zhang, Ao Ma, Chenyou Fan, Tin Lun Lam, and Junjie Hu. Learning robust stereo matching in the wild with selective mixture-of-experts. *arXiv preprint arXiv:2507.04631*, 2025.

[56] Yun Wang, Jiahao Zheng, Chenghao Zhang, Zhanjie Zhang, Kunhong Li, Yongjian Zhang, and Junjie Hu. Dualnet: Robust self-supervised stereo matching with pseudo-label supervision. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 39, pages 8178–8186, 2025.

[57] 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 (CVPR)*, pages 5249–5260, 2025.

[58] Gangwei Xu, Xianqi Wang, Xiaohuan Ding, and Xin Yang. Iterative geometry encoding volume for stereo matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21919–21928, 2023.

[59] Haofei Xu and Juyong Zhang. AANet: Adaptive aggregation network for efficient stereo matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1959–1968, 2020.

[60] Tianyu Yang and Antoni B Chan. Learning dynamic memory networks for object tracking. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 152–167, 2018.

[61] Jiaxi Zeng, Chengtang Yao, Yuwei Wu, and Yunde Jia. Temporally consistent stereo matching. In *European Conference on Computer Vision (ECCV)*, pages 341–359. Springer, 2024.

[62] Feihu Zhang, Victor Prisacariu, Ruigang Yang, and Philip HS Torr. GA-Net: Guided aggregation net for end-to-end stereo matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 185–194, 2019.

[63] Jiawei Zhang, Xiang Wang, Xiao Bai, Chen Wang, Lei Huang, Yimin Chen, Lin Gu, Jun Zhou, Tatsuya Harada, and Edwin R Hancock. Revisiting domain generalized stereo matching networks from a feature consistency perspective. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13001–13011, 2022.

[64] Yongjian Zhang, Longguang Wang, Kunhong Li, Yun Wang, and Yulan Guo. Learning representations from foundation models for domain generalized stereo matching. In *European Conference on Computer Vision (ECCV)*, pages 146–162. Springer, 2024.

[65] Youmin Zhang, Matteo Poggi, and Stefano Mattoccia. Temporalstereo: Efficient spatial-temporal stereo matching network. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 9528–9535. IEEE, 2023.

[66] Junbao Zhou, Ziqi Pang, and Yu-Xiong Wang. Rmem: Restricted memory banks improve video object segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18602–18611, 2024.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The claims around performance, runtime, and comparison to existing methods are described in the experiments.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: A common limitation among our work and existing works is the inability to proactively distinguish between dynamic and static areas, which is the key to ensuring consistency. Moving forward, our focus lies in exploring how to integrate our method with high-quality memory cues to enhance the performance and in developing a lightweight version of the model.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: No theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We provide the full details necessary to reproduce these results, including architecture details and implementation details of the training datasets and procedures, evaluation protocols for dynamic stereo matching. Details can be seen in the Experiment section.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in

some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: We will release the codes if the paper is accepted.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: We provide details on our architecture in the experiment section. The dataset splits were used and shared in previous work.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: For dynamic stereo matching metrics, we follow prior work and provide TEPE and δ_{1npx}^{t} to evaluate the temporal consistency.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: We provide a comparison of computational resources used in our experiments alongside previous works, with detailed results in the Appendix.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

Answer: [Yes]

Justification: We have reviewed the NeurIPS Code of Ethics.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [NA]

Justification: Our work provides a generic algorithm to learn a high-quality memory buffer for dynamic stereo matching. In the guidelines, it is reported that algorithms to optimize neural networks do not need any societal impact justification.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: The paper poses no such risks.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We cite the original paper that produced the code package or datasets.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [No]

Justification: Our method introduces a novel dynamic stereo matching approach without requiring a new dataset.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: No crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: Paper does not involve crowdsourcing nor resarch with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: We only use LLMs for improving writing.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (https://neurips.cc/Conferences/2025/LLM) for what should or should not be described.