

Temporally Consistent Stereo Matching

Jiaxi Zeng^{1,2}, Chengtang Yao^{1,3}, Yuwei Wu^{1,2*}, and Yunde Jia^{2,1}

¹ Beijing Key Laboratory of Intelligent Information Technology,
School of Computer Science & Technology, Beijing Institute of Technology, China

² Guangdong Laboratory of Machine Perception and Intelligent Computing,
Shenzhen MSU-BIT University, China

³ Horizon Robotics
`{jiaxi,yao.c.t,wuyuwei,jiayunde}@bit.edu.cn`

Abstract. Stereo matching provides depth estimation from binocular images for downstream applications. These applications mostly take video streams as input and require temporally consistent depth maps. However, existing methods mainly focus on the estimation at the single-frame level. This commonly leads to temporally inconsistent results, especially in ill-posed regions. In this paper, we aim to leverage temporal information to improve the temporal consistency, accuracy, and efficiency of stereo matching. To achieve this, we formulate video stereo matching as a process of temporal disparity completion followed by continuous iterative refinements. Specifically, we first project the disparity of the previous timestamp to the current viewpoint, obtaining a semi-dense disparity map. Then, we complete this map through a disparity completion module to obtain a well-initialized disparity map. The state features from the current completion module and from the past refinement are fused together, providing a temporally coherent state for subsequent refinement. Based on this coherent state, we introduce a dual-space refinement module to iteratively refine the initialized result in both disparity and disparity gradient spaces, improving estimations in ill-posed regions. Extensive experiments demonstrate that our method effectively alleviates temporal inconsistency while enhancing both accuracy and efficiency. Currently, our method ranks second on the KITTI 2015 benchmark, while achieving superior efficiency compared to other state-of-the-art methods. The code is available at <https://github.com/jiaxiZeng/Temporally-Consistent-Stereo-Matching>.

Keywords: Stereo Matching · Temporal Consistency · Disparity Completion · Iterative Refinement

1 Introduction

Stereo matching estimates the disparity by finding the horizontal correspondences between the rectified left and right images. It plays a significant role in various fields such as autonomous driving [28, 47], robotics [39], SLAM [3, 29],

*Corresponding author

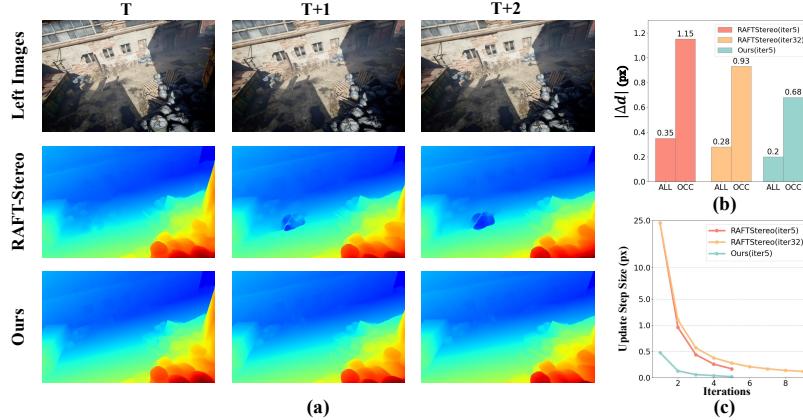


Fig. 1: (a) Visualization of disparity sequences from RAFT-Stereo [22] and our TC-Stereo. The failure cases of RAFT-Stereo lie in the reflective areas on the ground. A small motion can cause severe jitters in disparity predictions, while our method achieves temporally stable outputs. (b) The jitter $|\Delta d|$ between aligned successive disparity maps in all and occluded areas on TartanAir dataset [43]. Our method achieves better temporal consistency than RAFT-Stereo. (c) The update step size changed with the iterations of RAFT-Stereo and our TC-Stereo on TartanAir. Compared to RAFT-Stereo, our method performs a disparity search within a local range.

AR/VR [8, 40]. In these applications, stereo videos are typically taken as input and the output disparity sequences are employed for downstream tasks. However, the majority of existing methods [19, 22, 32, 44, 45] predict disparities independently for each stereo pair, disregarding the coherence between consecutive frames. This typically results in temporally inconsistent results, as shown in the second row of Fig. 1 (a). The inconsistent depth information has been demonstrated to substantially influence the accuracy and consistency of downstream tasks [10, 41, 51]. Although some techniques, such as Extended Kalman Filtering (EKF) [14] or Bundle Adjustment (BA) [38], mitigate the impact of inconsistency, they also have respective limitations. For instance, EKF suffers from non-linear errors, while BA brings intensive computation.

In this work, our goal is to improve the temporal consistency of stereo matching. We analyze the source of temporal inconsistency in stereo matching from two perspectives. On the one hand, most of the existing methods [4, 22, 32] independently infer the disparity map from scratch for each frame. This leads to a global disparity searching range, which introduces greater variability in disparity computation (*e.g.*, larger update step size), increasing the likelihood of inconsistency between consecutive frames. As shown in Fig. 1 (c), our method has a smaller update step size, indicating that it searches for the ground truth within a local disparity range, while RAFT-Stereo regresses the disparity from scratch in a global disparity range. On the other hand, camera or object motion leads to appearance changes in occluded areas, reflective surfaces, or low-texture regions.

The inherent ambiguity in these ill-posed regions causes the model to output unstable results when processing the temporally changing image sequences. Fig. 1 (b) demonstrates that the temporal inconsistency in ill-posed regions, like occlusions, is more serious than that in general regions.

Based on the above analyses, we propose to leverage the temporal information to mitigate temporal inconsistency. To avoid large update steps during the refinement process, we use the semi-dense disparity map from the previous time step for disparity completion, providing a well-initialized disparity for subsequent iterative refinement. This allows the refinement module to focus on local-range disparity updates. Additionally, we conduct a straightforward temporal state fusion to fuse the hidden state from the previous refinement with the current state features from the completion module, thus providing a temporally coherent initial hidden state for further refinement.

Moreover, we find that, for ill-posed areas like occlusions or reflective regions, it is easier to estimate the disparity gradient than the disparity itself. This is primarily because, in the real world, the depth of most regions tends to be smooth and continuous, producing constant or gradually changing gradients in these areas. Based on this observation, we propose a dual-space refinement module. This module takes the temporally initialized disparity and fused states as inputs, and refines the results in both disparity space and disparity gradient space. Through iterative refinement, the local smoothness constraints in the disparity gradient space are progressively extended to more global areas, thereby improving the smoothness of ill-posed regions and resulting in stable outputs.

Extensive experiments show that our method achieves state-of-the-art temporal consistency and accuracy. The contributions of our work can be summarized as follows: (1) We analyze the causes of temporal inconsistency and propose a temporally consistent stereo matching method. (2) We propose a temporal disparity completion and a temporal state fusion module to exploit temporal information, providing a well-initialized disparity and a coherent hidden state for refinement. (3) We propose an iterative dual-space refinement module to refine and stabilize the results in ill-posed regions. (4) Our method improves the temporal consistency of stereo matching and achieves SOTA results on synthetic and real-world datasets.

2 Related Work

Deep Stereo Matching Existing deep stereo matching methods primarily revolve around cost volume to design networks or representations, which is commonly categorized into regression-based methods [4, 11–13, 16, 17, 32, 44, 46] and iterative-based methods [19, 22, 45, 49]. Regression-based methods regress a probability volume to compute the disparity map, which can be classified into 3D volume methods [21, 27, 37, 46] and 4D volume methods [4, 11, 16, 26, 32, 44]. They either directly regress the disparity in the global disparity range [4, 16, 44] or perform a global-to-local regression [11, 26, 32]. However, the motion between frames may cause variation in the global disparity probability distribution, leading to in-

consistencies between sequential disparity maps. Iterative-based methods regard stereo matching as a continuous optimization process in the disparity space, iteratively retrieving the cost volume to refine the disparity map. RAFT-Stereo [22] leverages the optical flow method RAFT [36] for stereo matching and introduces a multi-level GRU to enlarge the receptive field. CREStereo [19] introduces a cascade network and adaptive group correlation layer to address challenges such as thin structures and imperfect rectification. DLNR [52] employs LSTM to alleviate the data coupling problem in the iteration of GRU. PCVNet [49] proposed a parameterized cost volume to accelerate the convergence of iterations by predicting larger update steps. IGEV-Stereo [45] regresses a geometry encoding volume to provide an accurate initial disparity for the GRU module, accelerating the convergence of disparity optimization. Although these methods have achieved excellent results, independently inferring the disparities for each frame leads to poor temporal consistency. In contrast, our method integrates the disparity and state information of the previous frame to ensure coherence between the consecutive outputs.

Video Stereo Matching Recently, there has been increasing attention on video stereo matching. Dynamic-Stereo [15] and CODD [20] focus on consistent stereo matching in dynamic scenes. Dynamic-Stereo jointly processes multiple frames, improving the temporal consistency of the results. However, the latency and computational overhead introduced by multi-frame inference limit its application in online scenarios. CODD predicts SE(3) transformations for each pixel to align successive disparity maps and fuses them together. Different from their post-fusion strategy, we integrate temporal information before our refinement, providing a robust local disparity searching range. TemporalStereo [50] mitigates the adverse effects of occlusions and reflections by augmenting current cost volume with the costs from the previous frame. Nevertheless, it still relies on coarse-to-fine regression in the global disparity range. XR-Stereo [8] extends RAFT-Stereo [22] over time, reducing iterations by reusing results from the previous frame, achieving over 100 FPS in a low-resolution mode. However, it simply takes the previous disparity and hidden state as the initialization of the current frame, disregarding the initial disparity and hidden state are semi-dense due to non-overlapping regions between frames. For these regions, it is still necessary to regress the disparity from scratch. Instead, we suggest completing the disparity map and incorporating the current state information to obtain a robust initialization for local refinement. The importance of these designs for improving temporal consistency was demonstrated in our ablation experiments.

Depth Completion Depth completion relies on a sparse depth map to generate a dense depth map [5, 6, 25, 34]. Some methods [1, 9] formulate depth completion as a stereo matching task, while some stereo matching methods [2, 42] utilizing sparse depth from LiDAR to guide the matching process. Different from these works, our disparity completion module does not rely on additional LiDAR data or complicated architectures. It utilizes a simple encoder-decoder to complete a semi-dense disparity map projected from the previous time stamp and provide the state features of the completed disparity map for further refinement.

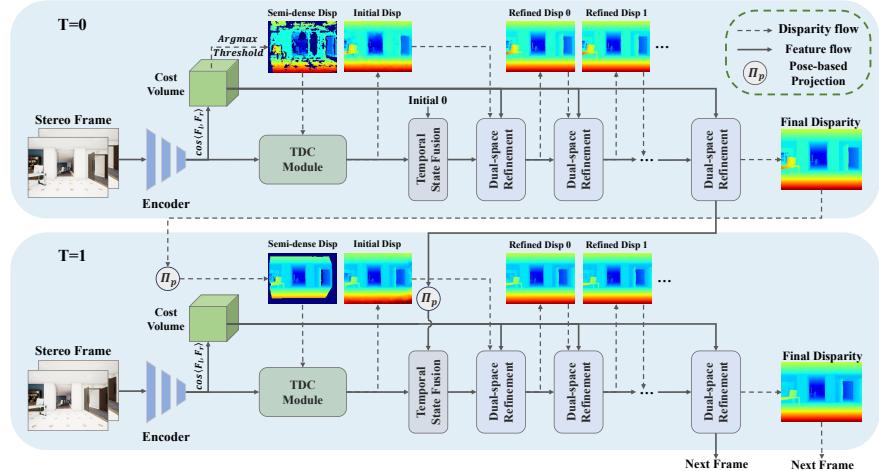


Fig. 2: Pipeline of TC-Stereo. We first use an encoder to extract left and right features for the current stereo frame. These features are then used to construct a cost volume. A semi-dense disparity map, derived from the cost volume (for the first frame) or projected from the previous timestamp (for subsequent frames), is fed into the Temporal Disparity Completion (TDC) module to obtain an initial dense disparity map. The output state of the TDC module is fused with the state from the past to provide an initial hidden state for refinement. The dual-space refinement module iteratively retrieves the cost volume and alternately refines the disparity map in the disparity and disparity gradient spaces. The final disparity map and hidden state are projected into the viewpoint of the next frame, serving as the temporal information for continuous disparity estimation.

Normal Guided Depth Estimation Substantial research [18, 23, 24, 30, 35, 48] demonstrates that utilizing surface normal contributes to depth estimation. Most methods [18, 23, 24, 48] design loss functions to leverage normal for constraining the depth map. Geo-Net [30] utilizes constraints between the normal and depth during inference, but this constraint is limited to local areas. HITNet [35] up-samples the disparity map through predicted disparity gradients to constrain the local surfaces. In this work, we iteratively refine results in the disparity and disparity gradient spaces, propagating the surface constraint globally.

3 Method

Our method processes stereo video sequences in an online manner and outputs the temporally consistent disparity maps. The pipeline of TC-Stereo is depicted in Fig. 2. In the following subsections, we will provide a detailed introduction to three key components of our TC-Stereo: temporal disparity completion (Sec. 3.1), temporal state fusion (Sec. 3.2), and iterative dual-space refinement (Sec. 3.3). Finally, we will present the loss functions involved (Sec. 3.4).

3.1 Temporal Disparity Completion

The aim of temporal disparity completion is to provide a well-initialized disparity map by leveraging the result of the previous frame. However, for the initial frame, temporal information is unattainable. As a result, there is no disparity point to use as the hint for the completion module, causing it to degrade into a monocular module and affecting the reliability of the initialization. To address this, we propose leveraging a semi-dense disparity map derived from the cost volume to compensate for the lack of temporal information in the initial frame.

Semi-dense disparity map from the cost volume We define the cost volume $C \in [0, 1]^{H \times W \times D}$ as the cosine similarity between the left and right features for each disparity hypothesis $d \in \mathbb{D} = \{0, 1, \dots, D - 1\}$. This is expressed as:

$$C(v, u, d) = \frac{\langle F_l(v, u), F_r(v, u - d) \rangle}{\|F_l(v, u)\|_2 \times \|F_r(v, u - d)\|_2}, \quad (1)$$

where $F_l(v, u)$ and $F_r(v, u - d)$ represent the feature vectors at pixel coordinates (v, u) in the left image and $(v, u - d)$ in the right image, respectively. Here, $\langle \cdot, \cdot \rangle$ denotes the inner product and $\|\cdot\|_2$ is Euclidean norm of a feature vector. We first compute the disparity from the cost volume by the winner-take-all strategy. Then, a threshold is applied to filter out outliers and obtain a semi-dense disparity map. This process can be summarized by the following formula:

$$\begin{aligned} d_1 &= \arg \max_{d \in \mathbb{D}} C(d), \\ d_2 &= \arg \max_{d \in \mathbb{D} \setminus \{d_1 - 1, d_1, d_1 + 1\}} C(d), \\ d_s &= \begin{cases} d_1, & \text{if } C(d_1) - C(d_2) > \theta \\ 0, & \text{otherwise} \end{cases}, \end{aligned} \quad (2)$$

where d_1 is the disparity with the highest similarity, and d_2 is the second-highest disparity, excluding the immediate neighbors of d_1 . The semi-dense disparity map d_s retains d_1 only if the $C(d_1)$ exceeds the $C(d_2)$ by a threshold θ ; otherwise, it is set to 0, indicating an invalid point. This process filters out disparities with high uncertainty, yielding a more reliable semi-dense disparity map for the completion. Moreover, to effectively learn the semi-dense disparity map from the cost volume, we apply a contrastive loss [35] to the cost volume. We discuss the loss in detail in Section 3.4.

Semi-dense disparity map from the previous timestamp For subsequent frames, we project the previous disparity map into the current image coordinates based on the intrinsic parameters and poses. This process can be represented as:

$$\begin{aligned} p^{t-1 \rightarrow t} &= \Pi_c(T^{t-1 \rightarrow t} \cdot \Pi_c^{-1}(p^{t-1}, z^{t-1})), \\ z^{t-1 \rightarrow t} &= (T^{t-1 \rightarrow t} \cdot \Pi_c^{-1}(p^{t-1}, z^{t-1}))_z, \\ d^{t-1 \rightarrow t} &= \frac{bf}{z^{t-1 \rightarrow t}}, \\ d_s^t &= \text{warp}(d^{t-1 \rightarrow t}, p^{t-1 \rightarrow t}). \end{aligned} \quad (3)$$

Here, p^{t-1} and z^{t-1} denote the coordinates of pixels and depths at time $t - 1$, respectively. Π_c represents the image projection process, which maps 3D points onto the 2D image. $T^{t-1 \rightarrow t}$ is the relative transformation from time $t - 1$ to t . b is the baseline of stereo camera, and f is the pixel-presented focal length. We forward warp $d^{t-1 \rightarrow t}$ by the pixel mapping $p^{t-1 \rightarrow t}$ to get the semi-dense disparity map d_s^t at time t . The flexible sources of the semi-dense disparity map enable our model to perform inference in both single-frame and multi-frame modes.

Disparity completion Our temporal disparity completion module utilizes contexts from the feature encoder, the semi-dense disparity, and a binary mask that indicates the sparsity as inputs. It employs a lightweight encoder-decoder network to regress a dense disparity map and outputs state features of the map for temporal state fusion. The detailed architecture of the completion module is provided in the supplementary materials.

3.2 Temporal State Fusion

Different from XR-Stereo [8], which directly uses the warped previous hidden states to initialize the refinement module, we argue that the previous states may not be the best choice for initializing the current states. On the one hand, the past states only encode the state information along the previous line of sight, which may fail in the current viewpoint. On the other hand, for non-overlapping regions between two frames, there are no past hidden states available. Therefore, we leverage a lightweight GRU-like module to fuse the current state features c^t from the TDC module with the past hidden states h_N^{t-1} (N denotes the number of refinement iterations) to provide initial hidden states h_0^t for the current refinement module:

$$\begin{aligned} z^t &= \sigma(W_z \cdot [c^t, h_{N-1}^{t-1}]), \\ r^t &= \sigma(W_r \cdot [c^t, h_{N-1}^{t-1}]), \\ q^t &= \tanh(W_q \cdot [r^t \odot c^t, h_{N-1}^{t-1}]), \\ h_0^t &= z^t \odot c^t + (1 - z^t) \odot q^t. \end{aligned} \quad (4)$$

Here, W_z , W_r , and W_q are the parameters of the network, $\sigma(\cdot)$ denotes the sigmoid function, and \odot is the element-wise dot. This simple module plays a crucial role in enhancing stability and accuracy, which we will demonstrate in the experimental section.

3.3 Dual-space Refinement

We observed that for some challenging regions for the disparity space, such as reflective areas, the gradient values can be easily estimated in the disparity gradient space. Based on this observation, we design a dual-space refinement module to refine the results in both the disparity space and the gradient space. As shown in Fig. 3, the pipeline of the dual-space module consists of three steps: disparity space refinement, gradient space refinement, and gradient-guided disparity propagation.

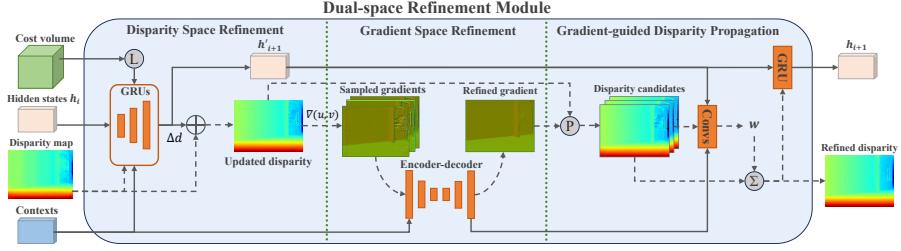


Fig. 3: Architecture of the dual-space refinement module. The disparity map corresponds to a scene consisting of a wall, a floor, and a transparent glass door. The encircled \mathbf{L} denotes the lookup operation to the cost volume, \mathbf{P} represents the gradient-guided disparity propagation, and Σ means weighted summation.

Disparity space refinement Follow RAFT-Stereo [22], we leverage the multi-level GRUs to refine the disparity map. It takes the hidden state h_i (i denotes the iteration), contextual features, the looked-up costs, and the disparity map as inputs. It outputs an intermediate hidden state h'_{i+1} and a step size Δd to update the disparity map.

Gradient space refinement We convert the updated disparity map to its disparity gradient space. For a point (u_0, v_0, d_0) in the disparity space, we sample two neighbor points (u_1, v_1, d_1) and (u_2, v_2, d_2) , which yields two vectors, $\mathbf{x}_1 = (\Delta u_1, \Delta v_1, \Delta d_1) = (u_1 - u_0, v_1 - v_0, d_1 - d_0)$ and $\mathbf{x}_2 = (\Delta u_2, \Delta v_2, \Delta d_2) = (u_2 - u_0, v_2 - v_0, d_2 - d_0)$. We assume that \mathbf{x}_1 and \mathbf{x}_2 are non-collinear, and the rotation from \mathbf{x}_1 to \mathbf{x}_2 in the u - v coordinate is clockwise. The formula of the disparity gradient can be represented as:

$$\begin{pmatrix} \frac{\partial d}{\partial u} \\ \frac{\partial d}{\partial v} \end{pmatrix} = \begin{pmatrix} \frac{\Delta v_1 \Delta d_2 - \Delta v_2 \Delta d_1}{\Delta u_2 \Delta v_1 - \Delta u_1 \Delta v_2} \\ \frac{\Delta u_2 \Delta d_1 - \Delta u_1 \Delta d_2}{\Delta u_2 \Delta v_1 - \Delta u_1 \Delta v_2} \end{pmatrix}. \quad (5)$$

By sampling different points within the neighborhood, we can obtain a series of gradient maps. The sampled gradients and contextual features are input into an encoder-decoder network to regress a refined gradient map.

Gradient-guided disparity propagation We utilize the refined gradients to guide the propagation of refined disparity in the disparity space. Specifically, for each pixel (u, v) with disparity d , we propagate the disparity to its neighborhood (u_n, v_n) based on the local planar hypothesis:

$$\hat{d}_n = d + (u_n - u) \frac{\partial d}{\partial u} + (v_n - v) \frac{\partial d}{\partial v}, \quad (6)$$

where \hat{d}_n is a propagated disparity candidate. We concatenated the h'_{i+1} , contexts, and the propagated disparity candidates and fed them into several convolutional layers with a softmax function to regress a set of weights, denoted as w . The refined disparity is regressed through the weighted summation of the disparity candidates. Finally, a lightweight GRU network with 1×1 convolutional

layers updates the intermediate hidden state using the refined disparity. The updated hidden state h_{i+1} and the disparity are then used as inputs for the next iteration. By iteratively refining the results in both disparity and gradient space, the surface constraints are propagated globally, improving results in ill-posed areas.

3.4 Loss functions

Our loss function consists of three components: the cost volume loss \mathcal{L}_{cv} , the disparity loss \mathcal{L}_{disp} , and the disparity gradient loss \mathcal{L}_{grad} . For the \mathcal{L}_{cv} , we adopt the contrastive loss proposed by HITNet [35] to supervise the cost volume:

$$\begin{aligned}\mathcal{L}_{cv} &= 1 - \psi(d^{gt}) + \max(\eta + \psi(d^{nm}) - \psi_{detach}(d^{gt}), 0), \\ d^{nm} &= \arg \max_{d \in \mathbb{D} \setminus [d^{gt} - 1.5, d^{gt} + 1.5]} C(d), \\ \psi(d) &= (d - \lfloor d \rfloor)C(\lfloor d \rfloor + 1) + (\lfloor d \rfloor + 1 - d)C(\lfloor d \rfloor).\end{aligned}\quad (7)$$

$\psi(d)$ is the cost (strictly speaking, similarity) for a sub-pixel disparity d . It is obtained through linear interpolation between $C(\lfloor d \rfloor + 1)$ and $C(\lfloor d \rfloor)$, where $\lfloor \cdot \rfloor$ is the floor function. The aim of \mathcal{L}_{cv} is to maximize $\psi(d^{gt})$, while penalizing $\psi(d^{nm})$ to ensure it remains at least a threshold η lower than $\psi(d^{gt})$.

As for the disparity loss \mathcal{L}_{disp} , it comprises three subparts: **disparity completion loss** \mathcal{L}_{dc} , **disparity space refinement loss** \mathcal{L}_{dsr} , and **gradient-guided disparity propagation loss** \mathcal{L}_{gdp} , which can be expressed as:

$$\mathcal{L}_{disp} = \lambda_{dc}\mathcal{L}_{dc} + \sum_{i=1}^N \gamma^{N-i}(\mathcal{L}_{dsr}^i + \lambda_{gdp}\mathcal{L}_{gdp}^i). \quad (8)$$

N represents the number of iterations, γ denotes the decay coefficient, and λ_{dc} and λ_{gdp} represent the balancing scalars. All these loss functions utilize the L1 loss between the d^{gt} and the disparity outputs at the corresponding stage.

The disparity gradient losses \mathcal{L}_{grad} can be formulated as:

$$\begin{aligned}\mathcal{L}_{grad} &= \|g^{gsr} - g^{gt}\|_1 + \|g^{gdp} - g^{gt}\|_1, \\ g^{gt} &= \nabla_{u,v}d^{gt}, \\ g^{gdp} &= \nabla_{u,v}d^{gdp}.\end{aligned}\quad (9)$$

The g^{gsr} represents the refined gradient map in the gradient space refinement. The g^{gt} and g^{gdp} are derived by taking the gradients of d^{gt} and d^{gdp} with respect to u and v . The final loss is the combination of \mathcal{L}_{cv} , \mathcal{L}_{disp} and \mathcal{L}_{grad} :

$$\mathcal{L} = \mathcal{L}_{cv} + \mathcal{L}_{disp} + \mathcal{L}_{grad}. \quad (10)$$

4 Experiments

4.1 Datasets

TartanAir [43] is a synthetic dataset for visual SLAM, covering various challenging indoor and outdoor scenarios. It encompasses over a thousand stereo

videos, totaling approximately 306K stereo pairs, making it well-suited for training our temporal model. In this study, we employ this dataset for pre-training and conduct ablation experiments on it.

Sceneflow [27] is a synthetic dataset containing three sub-datasets covering indoor and outdoor scenes. This dataset includes short stereo video sequences (10 frames), with large motions between frames. We utilize the dataset for temporal training and compare our method with others.

KITTI [28] is a real-world dataset for autonomous driving. It provides a leaderboard to evaluate the methods on a test set consisting of 200 images. Our work utilizes video sequences provided by KITTI for inference.

ETH3D SLAM [31] is a real-world dataset for SLAM in indoor environments. It provides stereo video sequences, as well as depth and pose ground truth. We evaluated the generalization performance of our method on this dataset.

4.2 Implementation Details

Our method is based on the implementation of RAFT-Stereo [22]. For both training and testing, we set the number of refinement iterations N to 5. During the training, we utilize a slice of stereo sequences as input and output the corresponding disparities frame by frame. We compute the loss for each frame’s output and accumulate the gradients across all frames before updating the parameters. In our experiment, we set the sequence length to 2 or 4 frames for training. During the inference process, our method takes a video stream of arbitrary length as input and outputs disparity predictions in an online manner. For the hyperparameters, we set $\theta = 0.3$, $\eta = 0.5$, $\gamma = 0.9$, and λ_{dc} and λ_{gdp} to 0.1 and 1.2, respectively. We use the AdamW optimizer and a one-cycle learning rate schedule where the maximum learning rate is set to 0.0002. The batch size is set to 8 for all the experiments. For the TartanAir dataset, we use all data (including Easy and Hard) for the ablation study, and the detailed training-testing split can be found in the supplementary material. We train for 100k steps on this dataset. To make a fair comparison with TemporalStereo [50], we retrain our model with a sequence length of 2 and 200k steps according to their train-test split. For the SceneFlow dataset, we train with the provided sequence data for 200k steps, with a sequence length of 2. For the KITTI dataset, due to the lack of temporally annotated data, we follow the setting of TemporalStereo [50] and train our model using pseudo-label data exported by the SOTA method LEAStereo [7] on the KITTI raw data. We train for 50k steps with a sequence length of 4 and a max learning rate of 0.0001, based on the model pre-trained on TartanAir. All pose data are either provided by the dataset or derived from SLAM algorithms. All experiments are conducted on two NVIDIA A40 GPUs.

4.3 Ablation Study and Analysis

Ablation Study We demonstrate the effectiveness of our designs through ablation experiments conducted on the TartanAir dataset. As shown in Table 1, we investigate the impacts of sequence length for training, temporal information

Setting	Seq length	N	Past disp & state	State fusion	TDC module	Dual-space refinement	ALL				OCC					
							>1px (%)	>3px (%)	EPE (px)	$ \Delta d $ (px)	$Relu(\Delta e)$ (px)	>1px (%)	>3px (%)	EPE (px)	$ \Delta d $ (px)	$Relu(\Delta e)$ (px)
(A)	2	5					8.04	2.98	0.53	0.35	0.16	34.91	16.60	1.63	1.15	0.47
(B)	2	32					6.02	2.26	0.41	0.28	0.13	25.89	12.39	1.28	0.93	0.39
(C)	2	5	✓				8.77	3.27	0.60	0.29	0.12	35.04	16.01	1.67	0.94	0.34
(D)	2	5	✓	✓			6.98	2.55	0.45	0.25	0.10	29.89	13.33	1.32	0.86	0.31
(E)	2	5	✓	✓	✓		6.10	2.33	0.42	0.23	0.10	25.97	12.40	1.21	0.78	0.28
(F)	2	5	✓	✓	✓	✓	6.28	2.42	0.42	0.21	0.09	25.89	11.86	1.18	0.74	0.28
(G)	2	5	✓	✓	✓	✓	5.98	2.19	0.40	0.21	0.09	24.67	11.28	1.14	0.71	0.27
(H)	4	5	✓	✓	✓	✓	5.84	2.05	0.39	0.20	0.08	23.85	10.50	1.06	0.68	0.26
(I)	4	5	✓	✓	✓	✓	6.50	2.29	0.43	0.28	0.13	26.80	12.48	1.26	0.90	0.38

Table 1: Ablation study on the TartanAir dataset. ALL and OCC respectively represent metrics for all regions and occluded regions.

(i.e. Past disp & state), temporal state fusion, temporal disparity completion (TDC) module, and dual-space refinement module on the accuracy in all regions and occluded regions. Initial settings (A) and (B) serve as baselines, representing RAFT-Stereo [22] with 5 and 32 iterations respectively. For a fair comparison, we set the sequence length during training to 2 for them, even though they don't use temporal information. Similar to XR-Stereo [8], setting (C) utilizes the past disparity and hidden states as initialization, but direct iterations on this initialization reduce performance, as analyzed in Section 3.2. As shown by the results of setting (D), the performance is improved in both all and occluded regions compared with (C), demonstrating the necessity of temporal state fusion. The accuracy of (E) shows further improvement, particularly on the 1-px error rate, which benefits from the well-initialized disparity provided by the TDC module. This enables the refinement module to be devoted to improving the fine-grained matching. (F) demonstrates that the dual-space refinement effectively improves the results in occluded regions. By combining (E) and (F), (G) achieves an even more precise stereo matching capability, exemplifying the synergy between the modules. Furthermore, we extend the training sequence length from 2 to 4 and get a better disparity output, suggesting that the performance of our model scales with a more extensive temporal context. Setting (I) corresponds to the model of (H) in the single-frame mode. Although there is a slight decrease in accuracy compared to setting (H), the results are still superior to (A). This demonstrates the versatility of our method in the single-frame and multi-frame mode, while also emphasizing the importance of temporal information. To sum up, our TC-Stereo (G) and (H), with only 5 iterations, significantly improve the results of RAFT-Stereo (A) and even surpass the performance of RAFT-Stereo with 32 iterations (B) on all metrics, greatly enhancing efficiency.

Temporal Consistency We design two types of metrics to evaluate the temporal consistency. For the first one, We convert the predicted disparity map d^{t+1} to the image coordinate at time t through poses and ground truth optical flow, denoting as \tilde{d}^{t+1} . We use the absolute difference between \tilde{d}^{t+1} and d^t , denoted as $|\Delta d| = |\tilde{d}^{t+1} - d^t|$, to evaluate temporal consistency. However, evaluating temporal absolute differences alone is one-sided. If a disparity d^t is incorrect, the model should be allowed to make correct predictions at the next time step $t + 1$, even if this results in temporary inconsistencies. To evaluate

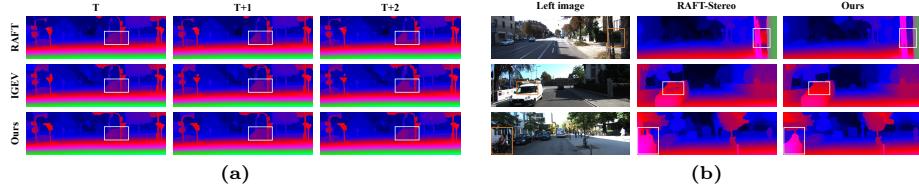


Fig. 4: Visualizations on KITTI 2015. (a) Comparison of temporal disparity sequences from RAFT-Stereo [22], IGEV [45] and our method. (b) Comparison of disparities in ill-posed regions between RAFT-Stereo [22] and our method.

temporal consistency comprehensively, we design a more lenient metric to allow the temporal change towards d^{gt} . Concretely, we calculate the error maps $e = |d - d^{gt}|$ for timestamp t and $t + 1$, and then compute the change in error, $\Delta e = \tilde{e}^{t+1} - e^t$. We use $Relu(\Delta e)$ to evaluate the extent of error divergence over time (*i.e.*, error increased from the previous frame for the same 3D point). Table 1 proves that our final models, (G) and (H), achieve the best temporal consistency and convergence compared to other settings. Settings (D), (E), and (F) respectively demonstrate the effectiveness of the state fusion, TDC module, and Dual-space refinement module in improving temporal consistency. Notably, (G) is comparable in accuracy to the baseline with 32 iterations (B), yet (G) significantly outperforms (B) in temporal metrics, especially in occlusion areas, further demonstrating the effectiveness of our TC-Stereo in mitigating temporal inconsistencies. Additionally, although (I) surpasses (D) in accuracy metrics, it is inferior to (D) in temporal metrics, further revealing the crucial role of temporal information in improving temporal consistency. We also provide the results of the 3-pixel error of consistency metrics in the supplementary materials.

We present visualizations of disparity sequences to illustrate the improvement in temporal consistency achieved by our method. As shown in Fig. 4a, our method produces more consistent disparity maps in the highlighted white box areas compared to other SOTA methods. More visualizations of temporal consistency are given in the appendix and our project website.

Local Searching Range and Fast Convergence Fig. 5 illustrates the disparity maps and update step sizes of our method and RAFT-Stereo, step by step. It can be seen that our method iterates based on a well-initialized disparity, providing a local disparity searching range. This leads to smaller update step sizes and faster convergence to the ground truth.

Improvement on Ill-posed Areas Fig. 4b shows the results of occlusion, transparent areas, and out-of-view areas, respectively. Thanks to the tempo-

Method	RAFT-Stereo [22]	LEAStereo [7]	DLNR [52]	DynamicStereo [15]	TemporalStereo [50]	CODD [20]	Ours
EPE (px)	0.65	0.78	0.47	0.71	0.53	0.60	0.50
>3px (%)	3.13	-	-	3.38	2.75	2.80	2.39
>1px (%)	6.67	7.82	5.38	7.57	6.38	-	4.93

Table 2: Benchmark results on Sceneflow dataset.

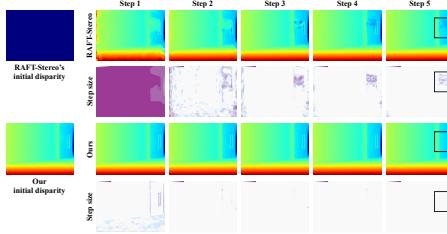


Fig. 5: Visualizations of the disparity map **Table 3:** Benchmark results of KITTI2015. and the update step size at each iteration.

Method	Nov			All			Time (s)
	D1-bg (%)	D1-fg (%)	D1-all (%)	D1-bg (%)	D1-fg (%)	D1-all (%)	
RAFT-Stereo [22]	1.45	2.94	1.69	1.58	3.05	1.82	0.38
CREStereo [19]	1.33	2.6	1.54	1.45	2.86	1.69	0.41
LEAStereo [7]	1.29	2.65	1.51	1.4	2.91	1.65	0.3
DLNR [52]	1.45	2.39	1.61	1.6	2.59	1.76	0.3
IGEV-Stereo [45]	1.27	2.62	1.49	1.38	2.67	1.59	0.18
PCW-Net [33]	1.26	2.93	1.53	1.37	3.16	1.67	0.44
HITNet [35]	1.54	2.72	1.74	1.74	3.2	1.98	0.02
TemporalStereo [50]	1.52	2.58	1.70	1.61	2.78	1.81	0.04
TC-Stereo(ours)	1.21	2.24	1.38	1.29	2.33	1.46	0.09

Table 4: Left: Results of Tartanair dataset based on the train-test splits of TemporalStereo [50]. * represents using the weights provided by the author. **Right:** Zero-shot results on ETH3D SLAM dataset.

ral information and the design of dual-space refinement, our method achieved smoother disparity output and clearer boundaries in these ill-posed regions.

Robustness Analysis Eq. 3 is based on the static scene assumption, resulting in unreliable initial disparities in dynamic areas. However, our method remains robust in ordinary dynamic scenes since we only use the temporal information for initialization. Disparities in dynamic regions will be corrected by iterative refinements. The moving car in Fig. 4a is a good example of this robustness. Additionally, we discuss the robustness of our method on incorrect poses and large motions in the supplementary materials.

Zero-shot Generalization We evaluated the zero-shot generalization performance on the ETH3D SLAM dataset, as shown in Table 4 (right). All models are only trained on SceneFlow. The metrics exclude regions with EPE>50 due to significant noise of depth ground truth around object boundaries. Our method surpasses other methods in generalization with fewer iterations. We also provide the qualitative results of our method on ETH3D SLAM, as shown in Fig. 6.

4.4 Benchmark Results

Sceneflow We conducted evaluations on the SceneFlow dataset, as shown in Table 2. Overall, our method outperforms the other methods except DLNR on the EPE metric, which may be attributed to their use of a more powerful backbone and more iterations. Our method exhibits a clear advantage over others in the >1px and >3px error rate metrics.

TartanAir We compared our method with TemporalStereo [50] and RAFT-Stereo [22] on the TartanAir dataset using the train-test split of TemporalStereo. As shown in Table 4 (left), our model surpassed RAFT-Stereo with only 5 iter-

Method	>1px			>3px			EPE
	(%)	(%)	(px)	(%)	(%)	(px)	
TemporalStereo* [50]	39.78	18.46	1.60	11.18	3.81	0.66	
RAFT-Stereo [22] (N=32)	34.20	16.67	1.45	8.25	3.33	0.53	
Ours (N=5)	34.68	16.36	1.36	8.51	3.31	0.52	

Method	>1px			>3px			EPE
	(%)	(%)	(px)	(%)	(%)	(px)	
IGEV-Stereo [45] (N = 32)	37.72	24.82	6.41				
RAFT-Stereo [22] (N = 32)	38.19	24.95	6.36				
Ours (N = 5)	37.00	23.29	6.27				

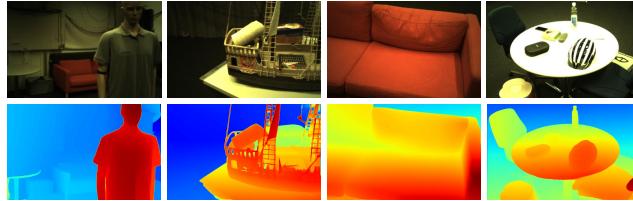


Fig. 6: Zero-shot qualitative results of our method on ETH3D SLAM dataset.

ations. Furthermore, it demonstrates significant advantages over the other two methods in terms of the EPE metric in occluded regions. This indicates that our design can enhance performance in ill-posed areas.

KITTI 2015 We present a comprehensive comparison of our method with SOTA methods on the KITTI 2015 dataset. As shown in Table 3, our proposed method, TC-Stereo, outperforms the other methods in most key metrics. Specifically, it achieves the lowest error rates in both non-occluded (Noc) and all regions, with a D1-all error rate of 1.46%, which is an 8.18% improvement over the best-performing IGEV-Stereo [45]. Among all these methods, our method exhibits the smallest difference between the Noc and All metrics, demonstrating our superior performance in occluded areas. Additionally, TC-Stereo showcases exceptional efficiency, with a processing time of only 0.09 seconds per frame, which is significantly faster than other methods with high performance. This performance shows that TC-Stereo excels in both accuracy and speed, making it well-suited for online applications.

5 Limitation

Despite the robustness of our model for ordinary dynamic regions and pose noise, it remains sensitive to large motions of dynamic objects, large errors in camera pose, and little overlap between adjacent frames, as we discuss in the supplementary materials. In these challenging cases, our method degrades to searching for the ground truth in a global disparity range like RAFT-Stereo.

6 Conclusion

In this paper, we proposed a temporally consistent stereo matching method that exploits temporal information through our temporal disparity completion module and temporal state fusion module, providing a well-initialized disparity map and hidden states. We further proposed a dual-space refinement module that iteratively refines the temporally initialized results in both disparity and disparity gradient spaces, improving estimations in ill-posed regions. Experiments demonstrated that our method effectively mitigates temporal inconsistency in video stereo matching and achieves SOTA results on challenging datasets.

Acknowledgements

This work was supported by the Natural Science Foundation of Shenzhen under Grant No. JCYJ20230807142703006, Natural Science Foundation of China (NSFC) under Grants No. 62172041 and No. 62176021, and Key Research Platforms and Projects of the Guangdong Provincial Department of Education under Grant No.2023ZDZX1034.

References

1. Bartolomei, L., Poggi, M., Conti, A., Tosi, F., Mattoccia, S.: Revisiting depth completion from a stereo matching perspective for cross-domain generalization. arXiv preprint arXiv:2312.09254 (2023)
2. Bartolomei, L., Poggi, M., Tosi, F., Conti, A., Mattoccia, S.: Active stereo without pattern projector. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 18470–18482 (2023)
3. Campos, C., Elvira, R., Gomez, J.J., Montiel, J.M.M., Tardós, J.D.: ORB-SLAM3: An accurate open-source library for visual, visual-inertial and multi-map SLAM. *IEEE Transactions on Robotics* **37**(6), 1874–1890 (2021)
4. Chang, J.R., Chen, Y.S.: Pyramid stereo matching network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5410–5418 (2018)
5. Cheng, X., Wang, P., Guan, C., Yang, R.: Cspn++: Learning context and resource aware convolutional spatial propagation networks for depth completion. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 10615–10622 (2020)
6. Cheng, X., Wang, P., Yang, R.: Depth estimation via affinity learned with convolutional spatial propagation network. In: Proceedings of the European conference on computer vision (ECCV). pp. 103–119 (2018)
7. Cheng, X., Zhong, Y., Harandi, M., Dai, Y., Chang, X., Li, H., Drummond, T., Ge, Z.: Hierarchical neural architecture search for deep stereo matching. *Advances in Neural Information Processing Systems (NeurIPS)* **33**, 22158–22169 (2020)
8. Cheng, Z., Yang, J., Li, H.: Stereo matching in time: 100+ fps video stereo matching for extended reality. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 8719–8728. IEEE (2024)
9. Choi, K., Jeong, S., Kim, Y., Sohn, K.: Stereo-augmented depth completion from a single rgb-lidar image. In: 2021 IEEE International Conference on Robotics and Automation (ICRA). pp. 13641–13647. IEEE (2021)
10. Deng, K., Liu, A., Zhu, J.Y., Ramanan, D.: Depth-supervised nerf: Fewer views and faster training for free. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12882–12891 (2022)
11. Duggal, S., Wang, S., Ma, W.C., Hu, R., Urtasun, R.: Deeppruner: Learning efficient stereo matching via differentiable patchmatch. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 4384–4393 (2019)
12. Gu, X., Fan, Z., Zhu, S., Dai, Z., Tan, F., Tan, P.: Cascade cost volume for high-resolution multi-view stereo and stereo matching. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2495–2504 (2020)

13. Guo, X., Yang, K., Yang, W., Wang, X., Li, H.: Group-wise correlation stereo network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3273–3282 (2019)
14. Kalman, R.E.: A new approach to linear filtering and prediction problems (1960)
15. Karaev, N., Rocco, I., Graham, B., Neverova, N., Vedaldi, A., Rupprecht, C.: Dynamicstereo: Consistent dynamic depth from stereo videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13229–13239 (2023)
16. Kendall, A., Martirosyan, H., Dasgupta, S., Henry, P., Kennedy, R., Bachrach, A., Bry, A.: End-to-end learning of geometry and context for deep stereo regression. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 66–75 (2017)
17. Khamis, S., Fanello, S., Rhemann, C., Kowdle, A., Valentin, J., Izadi, S.: Stereonet: Guided hierarchical refinement for real-time edge-aware depth prediction. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 573–590 (2018)
18. Kusupati, U., Cheng, S., Chen, R., Su, H.: Normal assisted stereo depth estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2189–2199 (2020)
19. Li, J., Wang, P., Xiong, P., Cai, T., Yan, Z., Yang, L., Liu, J., Fan, H., Liu, S.: Practical stereo matching via cascaded recurrent network with adaptive correlation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16263–16272 (2022)
20. Li, Z., Ye, W., Wang, D., Creighton, F.X., Taylor, R.H., Venkatesh, G., Unberath, M.: Temporally consistent online depth estimation in dynamic scenes. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 3018–3027 (2023)
21. Liang, Z., Feng, Y., Guo, Y., Liu, H., Qiao, L., Chen, W., Zhou, L., Zhang, J.: Learning deep correspondence through prior and posterior feature constancy. arXiv preprint arXiv:1712.01039 (2017)
22. Lipson, L., Teed, Z., Deng, J.: Raft-stereo: Multilevel recurrent field transforms for stereo matching. In: 2021 International Conference on 3D Vision (3DV). pp. 218–227. IEEE (2021)
23. Long, X., Lin, C., Liu, L., Li, W., Theobalt, C., Yang, R., Wang, W.: Adaptive surface normal constraint for depth estimation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 12849–12858 (2021)
24. Long, X., Liu, L., Theobalt, C., Wang, W.: Occlusion-aware depth estimation with adaptive normal constraints. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16. pp. 640–657. Springer (2020)
25. Ma, F., Karaman, S.: Sparse-to-dense: Depth prediction from sparse depth samples and a single image. In: 2018 IEEE international conference on robotics and automation (ICRA). pp. 4796–4803. IEEE (2018)
26. Mao, Y., Liu, Z., Li, W., Dai, Y., Wang, Q., Kim, Y.T., Lee, H.S.: Uasnet: Uncertainty adaptive sampling network for deep stereo matching. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6311–6319 (2021)
27. Mayer, N., Ilg, E., Hausser, P., Fischer, P., Cremers, D., Dosovitskiy, A., Brox, T.: A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4040–4048 (2016)

28. Menze, M., Geiger, A.: Object scene flow for autonomous vehicles. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3061–3070 (2015)
29. Mur-Artal, R., Tardós, J.D.: ORB-SLAM2: an open-source SLAM system for monocular, stereo and RGB-D cameras. *IEEE Transactions on Robotics* **33**(5), 1255–1262 (2017). <https://doi.org/10.1109/TRO.2017.2705103>
30. Qi, X., Liao, R., Liu, Z., Urtasun, R., Jia, J.: Geonet: Geometric neural network for joint depth and surface normal estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 283–291 (2018)
31. Schops, T., Sattler, T., Pollefeys, M.: Bad slam: Bundle adjusted direct rgb-d slam. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 134–144 (2019)
32. Shen, Z., Dai, Y., Rao, Z.: Cfnet: Cascade and fused cost volume for robust stereo matching. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 13906–13915 (2021)
33. Shen, Z., Dai, Y., Song, X., Rao, Z., Zhou, D., Zhang, L.: Pcw-net: Pyramid combination and warping cost volume for stereo matching. In: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXII. pp. 280–297. Springer (2022)
34. Tang, J., Tian, F.P., Feng, W., Li, J., Tan, P.: Learning guided convolutional network for depth completion. *IEEE Transactions on Image Processing* **30**, 1116–1129 (2020)
35. Tankovich, V., Hane, C., Zhang, Y., Kowdle, A., Fanello, S., Bouaziz, S.: Hitnet: Hierarchical iterative tile refinement network for real-time stereo matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14362–14372 (2021)
36. Teed, Z., Deng, J.: Raft: Recurrent all-pairs field transforms for optical flow. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16. pp. 402–419. Springer (2020)
37. Tonioni, A., Tosi, F., Poggi, M., Mattoccia, S., Stefano, L.D.: Real-time self-adaptive deep stereo. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 195–204 (2019)
38. Triggs, B., McLauchlan, P.F., Hartley, R.I., Fitzgibbon, A.W.: Bundle adjustment—a modern synthesis. In: Vision Algorithms: Theory and Practice: International Workshop on Vision Algorithms Corfu, Greece, September 21–22, 1999 Proceedings. pp. 298–372. Springer (2000)
39. Usenko, V., Engel, J., Stückler, J., Cremers, D.: Direct visual-inertial odometry with stereo cameras. In: 2016 IEEE International Conference on Robotics and Automation (ICRA). pp. 1885–1892. IEEE (2016)
40. Wang, J., Scharstein, D., Bapst, A., Blackburn-Matzen, K., Yu, M., Lehman, J., Alsisan, S., Wang, Y., Tsai, S., Frahm, J.M., et al.: A practical stereo depth system for smart glasses. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21498–21507 (2023)
41. Wang, J., Wang, P., Long, X., Theobalt, C., Komura, T., Liu, L., Wang, W.: Neuris: Neural reconstruction of indoor scenes using normal priors. In: European Conference on Computer Vision. pp. 139–155. Springer (2022)
42. Wang, T.H., Hu, H.N., Lin, C.H., Tsai, Y.H., Chiu, W.C., Sun, M.: 3d lidar and stereo fusion using stereo matching network with conditional cost volume normalization. In: 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 5895–5902. IEEE (2019)

43. Wang, W., Zhu, D., Wang, X., Hu, Y., Qiu, Y., Wang, C., Hu, Y., Kapoor, A., Scherer, S.: Tartanair: A dataset to push the limits of visual slam. In: 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 4909–4916. IEEE (2020)
44. Xu, G., Cheng, J., Guo, P., Yang, X.: Acvnet: Attention concatenation volume for accurate and efficient stereo matching. arXiv preprint arXiv:2203.02146 (2022)
45. Xu, G., Wang, X., Ding, X., Yang, X.: Iterative geometry encoding volume for stereo matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21919–21928 (2023)
46. Xu, H., Zhang, J.: Aanet: Adaptive aggregation network for efficient stereo matching. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1959–1968 (2020)
47. Yang, G., Song, X., Huang, C., Deng, Z., Shi, J., Zhou, B.: Drivingstereo: A large-scale dataset for stereo matching in autonomous driving scenarios. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
48. Yin, W., Liu, Y., Shen, C., Yan, Y.: Enforcing geometric constraints of virtual normal for depth prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5684–5693 (2019)
49. Zeng, J., Yao, C., Yu, L., Wu, Y., Jia, Y.: Parameterized cost volume for stereo matching. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 18347–18357 (2023)
50. Zhang, Y., Poggi, M., Mattoccia, S.: Temporalstereo: Efficient spatial-temporal stereo matching network. In: 2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 9528–9535. IEEE (2023)
51. Zhang, Y., Tosi, F., Mattoccia, S., Poggi, M.: Go-slam: Global optimization for consistent 3d instant reconstruction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3727–3737 (2023)
52. Zhao, H., Zhou, H., Zhang, Y., Chen, J., Yang, Y., Zhao, Y.: High-frequency stereo matching network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1327–1336 (2023)