

Alma Mater Studiorum Università di Bologna  
Archivio istituzionale della ricerca

TemporalStereo: Efficient Spatial-Temporal Stereo Matching Network

This is the final peer-reviewed author's accepted manuscript (postprint) of the following publication:

*Published Version:*

Zhang, Y., Poggi, M., Mattoccia, S. (2023). TemporalStereo: Efficient Spatial-Temporal Stereo Matching Network [10.1109/iros55552.2023.10341598].

*Availability:*

This version is available at: <https://hdl.handle.net/11585/961717> since: 2024-02-26

*Published:*

DOI: <http://doi.org/10.1109/iros55552.2023.10341598>

*Terms of use:*

Some rights reserved. The terms and conditions for the reuse of this version of the manuscript are specified in the publishing policy. For all terms of use and more information see the publisher's website.

This item was downloaded from IRIS Università di Bologna (<https://cris.unibo.it/>).  
When citing, please refer to the published version.

(Article begins on next page)

# TemporalStereo: Efficient Spatial-Temporal Stereo Matching Network

Youmin Zhang   Matteo Poggi   Stefano Mattoccia  
CVLAB, Department of Computer Science and Engineering (DISI)  
University of Bologna, Italy

<https://youmi-zym.github.io/projects/TemporalStereo/>  
{youmin.zhang2, m.poggi, stefano.mattoccia}@unibo.it

## Abstract

We present *TemporalStereo*, a coarse-to-fine stereo matching network that is highly efficient, and able to effectively exploit the past geometry and context information to boost matching accuracy. Our network leverages sparse cost volume and proves to be effective when a single stereo pair is given. However, its peculiar ability to use spatio-temporal information across stereo sequences allows *TemporalStereo* to alleviate problems such as occlusions and reflective regions while enjoying high efficiency also in this latter case. Notably, our model – trained once with stereo videos – can run in both single-pair and temporal modes seamlessly. Experiments show that our network relying on camera motion is robust even to dynamic objects when running on videos. We validate *TemporalStereo* through extensive experiments on synthetic (*SceneFlow*, *TartanAir*) and real (*KITTI 2012*, *KITTI 2015*) datasets. Our model achieves state-of-the-art performance on any of these datasets.

## 1. Introduction

Stereo reconstruction is a fundamental problem in computer vision. It aims at recovering the 3D geometry of the sensed scene by computing the disparity between corresponding pixels in a rectified pair of images. In robotic [36], virtual/augmented reality [57], autonomous driving [39], and more, depth information is crucial for scene understanding, with accurate yet fast solutions being particularly attractive.

Recently, deep learning approaches have shown tremendous improvements, especially in the supervised setting [7, 22, 52]. Common to most architectures is the cost volume, which encodes the feature similarity and plays a key role in matching pixels. In particular, features from the left image are concatenated (in 3D networks) or correlated (in 2D networks) and processed through convolutions, which are costly and prevent real-time execution. To address this is-

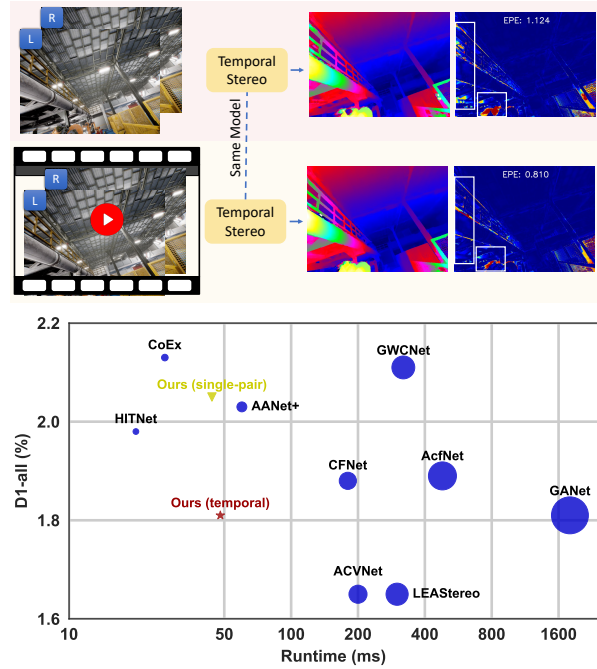


Figure 1. **Overview of TemporalStereo.** Our model can seamlessly process a stereo pair, as well as switch to temporal mode when multiple pairs are available, to increase accuracy. This allows *TemporalStereo* for achieving a favorable accuracy/speed trade-off with respect to existing networks. GMacs are measured with a resolution of 960×572.

sue, more efficient stereo models have been proposed, taking advantage of efficient cost aggregation [3, 53], coarse-to-fine [20, 50] and cost volume pruning [8, 27, 38] strategies to preserve computation and accuracy. Nonetheless, the sparse and constant disparity candidates used for matching cost computation can easily miss the ground truth disparity, thus deteriorating the accuracy of prediction. The case becomes even worse when dealing with challenging regions. For example, although occlusion is an ill-posed problem for stereo matching based on a single stereo pair, popular super-

vised stereo methods *have not ever* taken advantage of the intrinsic correlation among frames when stereo videos are available. Even if commercial cameras can easily acquire high FPS videos (e.g., 30 or more), in which likely past context is strictly related to the current one, state-of-the-art stereo methods generally process each frame independently. We argue that previous disparities can be relevant cues to improve estimates, especially in complex regions where the current pair alone is not sufficiently meaningful, such as near object boundaries and occlusions.

Inspired by these observations, we propose *TemporalStereo*, a novel lightweight deep stereo model (as depicted in Fig. 1) being not only suited for single pair fast inference, but also able to leverage the rich spatial-temporal cues provided by stereo videos, when available. In *single-pair* mode (i.e., when only the current pair is available), our network follows a coarse-to-fine scheme [13, 20, 50] to enable efficient inference. In order to increase the capacity of the constructed sparse cost volume and also the quality of sparse disparity candidates at every single stage, we enhance the cascade-based architecture in several aspects: 1) firstly, we enrich the cost of each queried disparity candidate with context from non-local neighborhoods during cost computation; 2) as the number of disparity candidates could be decreased to very few (e.g., 5), in addition to exploiting pure 3D convolutions for cost aggregation, we also perform multiple statistics (e.g., mean and maximum statistics) over a large window to grasp the global cost distribution of each pixel at one stage; 3) differently from previous strategies [20, 27, 38, 50], for which the candidates are constant throughout each stage, we rely on the aggregated cost volume itself to *shift* the candidates towards a better location and thus improve the quality of predicted disparity map; 4) additionally, when multiple pairs are available, our network can easily switch to *temporal* mode, in which past disparities, costs and cached features could also be aligned to the current reference frame and used to boost current estimates with a negligible runtime increase (4ms as shown in Tab. 6). Experimental results on synthetic (SceneFlow, TartanAir) and real (KITTI 2012, KITTI 2015) benchmarks support the following claims:

- Our method is accurate yet fast and achieves state-of-the-art results when a single stereo pair is available. In particular, the improved coarse-to-fine design based on very few (e.g., 5) candidates allows TemporalStereo to unlock fast execution and high performance.
- When multiple temporally adjacent stereo pairs are available, TemporalStereo is the first supervised stereo network that can effortlessly cache the past context and use it to improve ongoing predictions, e.g., in occluded and reflective regions. The model trained in temporal mode is effective also in single-pair mode, allowing

the deployment of the same model in both settings.

- The proposed temporal cues can be widely applied to boost the matching accuracy of current efficient stereo methods, e.g., on TartanAir [49] dataset, the improvements of CoEX [3] and StereoNet [20] are 14.6% and 26.1% respectively. Nonetheless, our TemporalStereo is still the best one for utilizing temporal information.

## 2. Related Work

**Single Pair Stereo Matching.** Traditional methods [16] employ handcrafted schemes to find local correspondences [35, 56] and approximate global optimization by exploiting spatial context [15]. Recent deep learning strategies [33] adopt CNNs and can be broadly divided into two categories according to how they build the cost volume. On the one hand, 2D networks follow DispNetC [28] employing a correlation-based cost volume and applying 2D convolutions to regress a disparity map. Stacked refinement sub-networks [23] and multi-task learning [40, 55] have been proposed to improve the accuracy. GCNet [19] represents a milestone for 3D networks. Following works improve its results thanks to spatial pyramid pooling [6], group-wise correlations [14], forcing unimodal [11, 60] or bimodal [48] cost distributions. 3D methods usually outperform 2D architecture by a large margin on popular benchmarks [12, 28, 29], paying more in terms of computational requirement. In this work, we leverage 3D sparse cost volumes, and the peculiar proposed architecture proves to be fast and accurate at disparity estimation.

**Efficient Stereo Matching with Deep-Learning.** A popular strategy to decrease the runtime consists in performing computation at a single yet small resolution (e.g., 1/4) and obtaining the final disparity through upsampling. CoEx [3] adopts this strategy. Coarse-to-fine [2, 54] design further improves efficiency, since the overall computation is split into many stages. To further reduce the disparity search range for each pixel, StereoNet [20] and AnyNet [50] add a constant offset to disparity maps produced in the previous stage to avoid checking all the disparity candidates. However, as reported in [27], the constant offset strategy is not robust against large errors in initial disparity maps. DeepPruner [8] prunes the search space using a differentiable variant of PatchMatch [4], obtaining sparse cost volumes. In contrast, methods such as [27, 38] employ uncertainty to evaluate the outcome of previous stages and then build the current cost volume accordingly. However, since the cost volume itself expresses the goodness of candidates (the better the candidates, the more representative the cost volume), it could be used to *check* and eventually *correct* the candidates themselves. Nonetheless, previous methods in literature cannot exploit this cue since their candidates

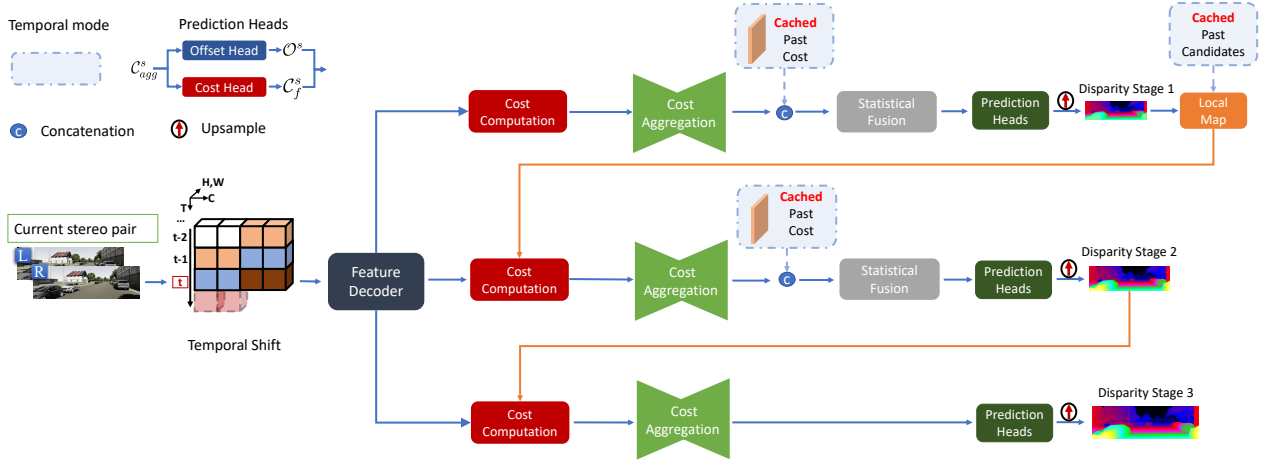


Figure 2. **TemporalStereo Architecture.** In single-pair mode, the model predicts the disparity map in a coarse-to-fine manner. If past pairs are available, the same model switches to temporal mode and employs features, costs, and candidates cached from past pairs to improve the current prediction.

come from the earlier level and are constant through the current stage. Conversely, this paper proposes inferring an offset for each disparity candidate from the current aggregated cost volume, which helps to ameliorate the disparity. Moreover, to improve network efficiency on high-resolution images, we perform the coarse-to-fine predictions only at downsampled feature maps rather than at full image resolution like HITNet [43].

**Multiple Pairs Stereo Matching.** When two temporally adjacent stereo pairs are available, optical flow [28, 45] or 3D scene flow [47] is often taken into account to link the images with the reconstructed 2D/3D motion field. Then, the stereo problem is tackled by leveraging a multi-task model [1, 21, 51] or by casting it as a special case of optical flow [25]. However, these methods are not able to capitalize longer stereo sequences. Conversely, OpenStereoNet [61] adopts recurrent units to capture temporal dynamics and correlations available in videos for unsupervised disparity estimation. We argue that past information could be beneficial to understand the current scene – especially in difficult areas – for *free* since past context and predictions can be cached easily, although none of the above methods make use of them explicitly in videos by taking into account geometry. On the contrary, TemporalStereo fully exploits this potential.

### 3. Method

We first describe TemporalStereo in single-pair mode, then we fully unlock the potential of our architecture enabling temporal mode. Fig. 2 depicts our model.

#### 3.1. Single Pair Mode

Given a single stereo pair, a backbone extracts multi-scale features at  $1/4$ ,  $1/8$ , and  $1/16$  of the original resolution. Then, three stages predict disparity maps starting from these features. In particular, each stage performs feature decoding, cost volume computation, and aggregation.

**Feature decoding** In each stage  $s \in \{1, 2, 3\}$ , a decoder processes the current features, together with those from the previous stage  $F_l^{s-1}$ ,  $F_r^{s-1}$  if  $s > 1$ . In particular,  $F_l^{s-1}$ ,  $F_r^{s-1}$  are bilinearly upsampled by a factor 2 and concatenated with left and right features from the backbone. Then, the resulting feature maps are processed by two 2D convolutions with kernels of 3, obtaining  $F_l^s$ ,  $F_r^s$ .

**Cost volume computation.** A 4D cost volume  $\mathcal{C}^s \in \mathbb{R}^{Ch^s \times H^s \times W^s \times |D^s|}$  is constructed by concatenating  $F_l^s$  with the corresponding  $F_r^s$  for each disparity  $d$  in the set  $D^s = \{d_1, d_2, \dots, d_n\}$ , with  $Ch^s$  respectively the number of channels of feature  $F_l^s$  and disparity candidates used in the stage:

$$\mathcal{C}_{concat}^s(\cdot, u, v, d) = \oplus \{F_l^s(\cdot, u, v), F_r^s(\cdot, u - d, v)\}, \quad (1)$$

where  $u, v$  are horizontal and vertical pixel coordinates, respectively, while  $\oplus$  stands for concatenation on channel dimension. However, by doing so,  $\mathcal{C}^s$  only contains local information, which is a major limitation in the case of a sparse set of candidates. To alleviate it, we enrich  $\mathcal{C}^s$  with multi-level costs [45] encoding a larger neighborhood. Specifically, we build a three-level cost volume from feature maps  $F_l^s$ ,  $F_r^s$  and by downsampling them with factors 2, 4. At each level, we perform group-wise correlations [14], then costs are bilinearly upsampled to  $\mathcal{C}^s$  resolution and stacked



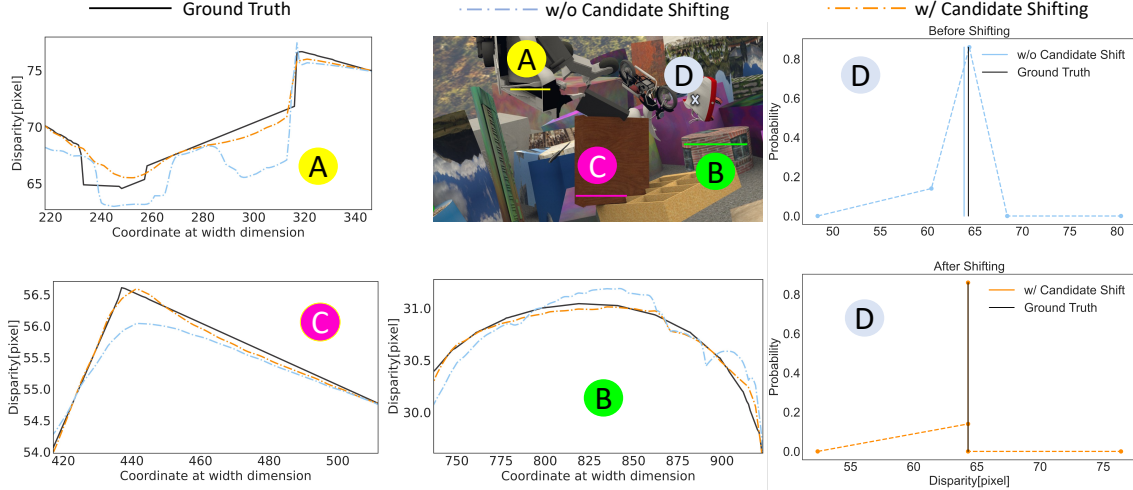


Figure 3. **Adaptive Shifting in action.** We show the ground truth disparities and the predictions with and without adaptive shifting for three horizontal regions (i.e., A, B, C) and one pixel (D) in the image. For A, B, C, our strategy better estimates the actual disparity distributions. For D, the matching distribution of 5 disparity candidates switches from a flattened unimodal to one-hot distribution centered at ground truth by adaptively shifting candidates.

together, obtaining  $\mathcal{C}_{gwc}^s$ . The final multi-level cost is:

$$\mathcal{C}^s(\cdot, u, v, d) = \oplus \{ \mathcal{C}_{concat}^s(\cdot, u, v, d), \mathcal{C}_{gwc}^s(\cdot, u, v, d) \}. \quad (2)$$

**Cost aggregation.** We now describe how we compute the aggregated cost volume  $\mathcal{C}_{agg}^s$ . In each stage, preliminary spatial cost aggregation is performed by one 3D convolution with kernel size  $3 \times 3 \times 3$  followed by an hourglass network and another 3D convolution with kernel size  $3 \times 3 \times 3$ . Afterward, a Statistical Fusion module further improves the aggregation with mean and maximum statistics [59], especially along the disparity dimension (which contains as few as 5 candidates): 4 parallel layers (identity, convolutional with kernel size  $5 \times 1 \times 1$ , average pooling and max pooling both with kernel size  $5 \times 5 \times 5$ , respectively) extract global statistics from the preliminary aggregated volume, then their outcomes are stacked together and processed by one 3D convolution. In temporal mode, Statistical Fusion also helps to merge past costs with current ones, as we will discuss later. We employ Statistical Fusion only in stages  $s = 1, 2$  to lighten the computation, while in  $s = 3$  past costs are not used. Finally, a cost prediction head with two 3D convolutions with kernel size  $3 \times 3 \times 3$  is in charge of predicting the final cost volume  $\mathcal{C}_f^s$  from  $\mathcal{C}_{agg}^s$ . From  $\mathcal{C}_f^s$  and  $\mathcal{D}^s$  we could obtain the disparity  $\hat{d}^s$  of the current stage by means of soft-argmin operator [19], but  $\mathcal{C}_{agg}^s$  can help us to improve the candidates – as we will describe thereafter. For this reason, before computing  $\hat{d}^s$ , we have to introduce our adaptive candidate shifting strategy.

**Adaptive candidate shifting.** Previous works [20, 27] consider the candidates  $\mathcal{D}^s$  as constants throughout the current stage  $s$ . However, the aggregated cost volume itself

contains cues about the candidates used to build it. Since it has been constructed from  $\mathcal{D}^s$ , it represents how well the candidates can explain the current scene. Moreover, in the cost aggregation step, each cost has been enriched with all the costs in its neighborhood, collecting matching results of nearby pixels. We propose to leverage a dedicated prediction head to infer, from  $\mathcal{C}_{agg}^s$  as shown in Fig. 2, an offset value for each candidate, which represents how much we have to *shift* the candidate towards a better location. Thus, the candidates are pixel-wisely different and able to adaptively adjust according to the learned context. This strategy gives the network the ability to adapt  $\mathcal{D}^s$  and improve  $\hat{d}^s$ , as depicted in Fig. 3. Specifically, an independent head – with the same structure of the cost prediction head, i.e., two 3D convolutions, – is in charge of inferring the offset volume  $\mathcal{O}^s \in \mathbb{R}^{H^s \times W^s \times |\mathcal{D}^s|}$  from  $\mathcal{C}_{agg}^s$  as input. Although offset learning [26] is not new and has been used in [11] to convert the dense disparity candidates, which always tightly cover the ground truth disparity, into continuous ones so that the predicted disparity value will not be restricted to integers, in contrast, our disparity candidates are few and sparse, which could easily miss the ground truth. By enriching the context of each sparse candidate, the proposed shifting strategy is able to *correct* the constant and flawed candidates for better disparity estimation.

**Disparity prediction and candidate sampling.** Given  $\mathcal{D}^s$ ,  $\mathcal{O}^s$  and  $\mathcal{C}_f^s$ , soft-argmin operator [19] could be applied to predict the final disparity map. However, this strategy results sub-optimal in the case of multi-modal distributions [11, 48, 60]. To overcome this limitation, we predict the disparity following the strategy proposed in [3]: we select top- $K$  values of  $-\mathcal{C}_f^s$  for each pixel across the disparity

dimension, and we normalize them with the softmax operator  $\sigma$ . The predicted disparity map  $\hat{d}^s$  results:

$$\hat{d}^s = \sum_{d \in \{d_1^{top}, \dots, d_K^{top}\}} (d + o_d) \times \sigma(-c_d), \quad (3)$$

where  $d_k^{top} = \operatorname{argmax}_d^k(-c_d)$ ,  $\operatorname{argmax}^k(\cdot)$  is the  $k$ -th maximal value for  $k \in \{1, 2, \dots, K\}$ ,  $c_d$  the cost in  $\mathcal{C}_f^s$  of candidate  $d$  for each pixel, and  $o_d$  is the offset value in  $\mathcal{O}^s$  for the candidate  $d$ . Convex upsampling [45] is used to double the resolution of  $\hat{d}^1$  and  $\hat{d}^2$ , obtaining  $\hat{d}_\uparrow^1, \hat{d}_\uparrow^2$ . To bring  $\hat{d}_\uparrow^3$  to full resolution, following [3], we bilinearly upsample  $\hat{d}_\uparrow^3$  to full resolution and for each pixel in the upsampled resolution, a weighted average of a  $3 \times 3$  superpixel surrounding it is calculated to get the final prediction  $\hat{d}_\uparrow^3$ . Finally, we obtain the set of candidates  $\mathcal{D}^{s+1}$  for the next stage applying an inverse transform sampling. We sample according to a normal distribution  $\mathcal{N}(\hat{d}_\uparrow^s, 1)$  in the range  $[\hat{d}_\uparrow^s - \beta, \hat{d}_\uparrow^s + \beta]$ , with  $\beta$  a constant value. This strategy allows to include more candidates near the predicted disparity. The first stage is initialized with  $\mathcal{D}^1$  uniformly sampled across the full range to provide an overview of current scene geometry.

**Loss function.** Following [6], we minimise the smooth  $L_1$  loss, i.e., Huber loss at full resolution between the ground truth disparity map  $d^{gt}$  and the predictions at the three stages  $\{\hat{d}_\uparrow^1, \hat{d}_\uparrow^2, \hat{d}_\uparrow^3\}$ .  $\hat{d}_\uparrow^1, \hat{d}_\uparrow^2$  and  $\hat{d}_\uparrow^3$  are bilinearly upsampled to full resolution and a factor  $\lambda^s$  weights each predicted map. The loss results:

$$\begin{aligned} \mathcal{L}_{huber} = & \frac{1}{|\mathcal{N}|} \lambda^0 \cdot \operatorname{smooth}_{L_1}(d^{gt} - \hat{d}_\uparrow^0) \\ & + \sum_{s=1}^3 \frac{1}{|\mathcal{N}|} \lambda^s \cdot \operatorname{smooth}_{L_1}(d^{gt} - \hat{d}_\uparrow^s), \end{aligned} \quad (4)$$

where  $\mathcal{N}$  is the number of pixel with valid ground truth.

Furthermore, to supervise the learning of offsets in three stages  $\{O^1, O^2, O^3\}$ , we adopt the Warsserstein loss [11] for training. Specifically, for each stage  $s$ , we downsample the ground truth disparity map  $d^{gt}$  to the resolution of corresponding offset for supervision, obtaining  $d_\downarrow^{gt,s}$ . As the offset value can range in a quite large value space, we improve the Warsserstein loss as:

$$\mathcal{L}_{war} = \sum_{s=1}^3 \frac{1}{|\mathcal{N}|} \lambda^s \cdot \sum_{d \in \mathcal{D}^s} (|d + o_d - d_\downarrow^{gt,s}| \times (\sigma(-c_d) + \alpha)) \quad (5)$$

where  $c_d$  the cost value in  $\mathcal{C}_f^s$  of candidate  $d$  for each pixel, and  $o_d$  is the offset value in  $\mathcal{O}^s$  for the candidate  $d$ . We set  $\alpha = 0.25$ , i.e., even for the candidate with extremely low matching probability  $\sigma(-c_d)$ , the network still learns an offset to enforce it moving towards the ground truth. With a weight factor  $\lambda_{final}$ , our final loss becomes:

$$\mathcal{L} = \mathcal{L}_{huber} + \lambda_{final} \cdot \mathcal{L}_{war}. \quad (6)$$

### 3.2. Temporal Mode

So far, we have presented the model suited for single-pair mode. Now, we illustrate how the model works in temporal mode. Differently from multi-view stereo models [17, 44], our network processes stereo videos one stereo pair at a time. This behaviour, which helps to save computation, is possible thanks to the sparse cost volume. In fact, we can easily add disparity candidates from the past inferences to the current set  $\mathcal{D}^s$ , thus increasing the search range with other plausible solutions. The past candidates, however, have to be aligned with the current frame to be meaningful. Optical flow/3D scene flow could be used to tackle the issue, but predicting flow fields is expensive in terms of memory footprint and time. Instead, in this work, we suppose that the camera is calibrated and the pose is given – since pose can be provided by external IMU sensors or estimated from the stereo pairs – to build the rigid motion field needed to align the candidates. Nonetheless, rigid flow formulation only holds for static objects in the scene, but since our network always relies on the current stereo pair, TemporalStereo is robust even in case of wrong flows (e.g., due to bad poses) or moving objects. In the remainder, we detail how we do use temporal information to improve stereo-matching results.

**Local map.** As highlighted before, occlusions represent a major issue in stereo matching. However, we argue that currently occluded regions might have been visible in past frames, e.g., due to camera motion. For this reason, the issue can be alleviated by adequately exploiting past estimates and at a minimal cost since they can be easily cached. Furthermore, we cache costs according to a keyframe strategy to bind the complexity in the case of long video sequences and ensure enough motion parallax. Following [41], a new incoming frame is promoted to keyframe if its relative translation and rotation are greater than  $\mathbf{t}_{max}$  and  $\mathbf{R}_{max}$  respectively. Then, a memory bank collects disparity  $\hat{d}_\uparrow^3$  of the last  $N_{key}$  keyframes. Since each cached map represents the scene geometry in the past, to use it in the current computation we need to update the disparity values and their coordinates in the current image. Inspired by SLAM literature [30], we leverage a Local Map strategy to this aim. In detail, given the camera model  $\pi : \mathbb{R}^3 \rightarrow \mathbb{R}^2$ , a 3D point  $P(X, Y, Z)$  can be projected to a 2D pixel  $p(u, v)$ :

$$\pi(P) = \left( f_x \frac{X}{Z} + c_x, f_y \frac{Y}{Z} + c_y \right), \quad (7)$$

where  $(f_x, f_y, c_x, c_y)$  are the known camera intrinsics. Similarly, a pixel  $p$  can be back-projected to a 3D point  $P$ :

$$\pi^{-1}(p, d) = \frac{b \cdot f_x}{d} \left( \frac{u - c_x}{f_x}, \frac{v - c_y}{f_y}, 1 \right)^\top, \quad (8)$$

where  $b$  is the baseline between the left and the right cameras and  $d$  the disparity. With the relative extrinsic  $\mathbf{T}_{j \rightarrow t} \in$

$\text{SE}(3)$  from keyframe  $j$  to the current frame  $t$ , we can update every disparity value  $d_j$  of keyframe  $j$  as:

$$d_j^{proj} = \frac{b \cdot f_x}{Z_j}, \text{ with } P_j(X_j, Y_j, Z_j) = T_{j \rightarrow t}^{-1} \cdot \pi^{-1}(p, d_j), \quad (9)$$

where  $d_j^{proj}$  is the updated disparity value. At this point, we obtain the coordinates of  $d_j^{proj}$  in  $t$  through forward warping. To preserve end-to-end requirement, we use the differentiable Softmax Splatting  $\vec{\sigma}$  [31]:

$$\hat{d}_j^{proj} = \vec{\sigma}(d_j^{proj}, \pi(P_j) - p) \quad (10)$$

Finally, the Local Map is defined as follows:

$$\mathcal{D}^2 = \oplus \left\{ \mathcal{D}^2, \hat{d}_1^{proj}, \hat{d}_2^{proj}, \dots, \hat{d}_{N_{key}}^{proj} \right\} \quad (11)$$

It is worth noticing that, in single-pair mode, the Local Map only contains candidates from the current pair. Moreover, it boosts candidate selection only for  $s = 2$ , since this stage represents the best trade-off between accuracy and speed. In fact, in  $s = 1$  candidates are looked over the full search range at the lowest resolution, while in  $s = 3$  a larger cost volume involves a much more expensive computation.

**Temporal shift and past costs.** Past semantic and matching scores are crucial for temporal stereo processing. Although Local Map effectively proposes previous depth cues about the scene, it lacks in providing the context behind these guesses. To address the problem, we introduce Temporal Shift and Past Costs modules. Temporal Shift enriches current feature maps with those computed in the past. Specifically, we adopt the TSM [24] strategy to facilitate the feature exchange among neighboring frames because it does not introduce additional computation or parameters: past features are *shifted* along the time dimension and then merged with current ones as shown in Fig. 2. Doing so, TSM provides *spatio-temporal* capabilities to our backbone not originally designed for temporal modeling. In practice, every feature map from the backbone is cached and used by Temporal Shift in the next prediction. Notably, as TSM does not rely on pose, we can still perform stereo matching in temporal mode and benefit from past information when pose is not available as shown in Tab. 2.

Similarly, Past Costs module adds past matching scores to current cost volumes. Given  $\mathcal{C}_f^3$  volume computed at time  $t - 1$ , first we update the value of its candidates to  $t$  according to Eq. (9). Then, cost values and candidates are forward warped to  $t$  using Eq. (10). Finally, the warped cost volume is downsampled by a factor of 2 and 4 and concatenated with current  $\mathcal{C}_{agg}^2$  and  $\mathcal{C}_{agg}^1$  respectively, to be further aggregated by the Statistical Fusion module. To preserve computation, we only warp top- $K$  candidates and their costs to  $t$ . In single-pair mode, both Temporal Shift and Past Costs are the identity function. This strategy allows us to run the

	Multi-Level Cost	Statistical Fusion	Adaptive Shifting	EPE			3PE			Runtime (ms)
				ALL	OCC	NOC	ALL	OCC	NOC	
(A)	✗	✗	✗	0.600	1.949	0.369	2.85	11.62	1.36	<b>40.62</b>
(B)	✓	✗	✗	0.587	1.924	0.360	2.79	11.44	1.33	43.41
(C)	✓	✓	✗	0.581	1.908	0.356	2.78	11.39	<b>1.31</b>	44.54
(D)	✗	✗	✓	0.564	1.923	0.334	2.87	11.78	1.36	41.58
(E)	✓	✓	✓	<b>0.532</b>	<b>1.830</b>	<b>0.315</b>	<b>2.75</b>	<b>11.37</b>	<b>1.31</b>	45.42

	Candidates Number	Disparity Range	EPE	Runtime (ms)	Model Variants	Depth-wise 3DCNN	EPE	Runtime (ms)
(F)	3	[-4, 4]	0.565	<b>41.61</b>	(K) Baseline	✗	0.589	63.91
(G)	9	[-4, 4]	<b>0.531</b>	55.84	(L) Baseline	✓	0.600	<b>40.62</b>
(H)	5	[-4, 4]	0.532	45.42	(M) Full	✗	0.535	75.42
(I)	5	[-2, 2]	0.543	45.42	(N) Full	✓	<b>0.532</b>	45.42
(J)	5	[-8, 8]	0.568	45.42				

Table 1. **Single-pair mode ablation.** We assess on SceneFlow the key components of the proposed architecture.

	Temporal Shift	Local Map	Past Costs	EPE			3PE			Runtime (ms)
				ALL	OCC	NOC	ALL	OCC	NOC	
(A)	✗	✗	✗	0.647	1.899	0.420	3.96	14.21	2.08	<b>36.85</b>
(B)	✓	✗	✗	0.643	1.842	0.435	4.00	14.55	2.15	<b>36.85</b>
(C)	✓	✓	✗	0.624	1.799	<b>0.413</b>	3.81	13.60	<b>2.02</b>	38.22
(D)	✓	✓	✓	<b>0.610</b>	<b>1.637</b>	<b>0.413</b>	<b>3.73</b>	<b>12.60</b>	2.03	40.13

Table 2. **Temporal mode ablation.** We evaluate the temporal components on TartanAir, with  $W_{tr} = W_{test} = 4$ .

model trained in temporal mode also in single-pair mode with a limited drop in accuracy, e.g., at bootstrap.

## 4. Experiments

This section describes the experimental setups used to evaluate on popular datasets, including: SceneFlow [28], TartanAir [49], KITTI 2012 [12] and KITTI 2015 [29]. As standard practice in this field [58, 60], we compute the end-point-error (EPE) and the percentage of points with a disparity error  $> 3$  pixels (3PE,  $> 5$  for 5PE) as error metrics in non-occluded (NOC), occluded (OCC) and both (ALL) the regions. Moreover, following the literature [29], we measure the D1 error on KITTI instead of the 3PE in background (BG), foreground (FG), and both (ALL) areas. We do so by computing the percentage of points with error  $> 3$  pixels and  $> 5\%$  than the ground truth. Runtime reported in Tab 1, 2, 6 is measured in the corresponding image resolution on each dataset, on a single NVIDIA 3090 GPU. The [supplementary material](#) outlines more details about implementation, dataset description, and training setups.

### 4.1. Ablation Study

**Single-pair mode.** We leverage SceneFlow [28] to assess the impact of main components of TemporalStereo in single-pair mode. Tab. 1 reports this ablation study, witnessing how each module helps to improve the results consistently. In particular, the multi-level cost computation with group-wise convolutions (B) is effective in enriching the cost volume of each stage, and the baseline model (A) also benefits from the further aggregation of Statistical Fusion (C). Nonetheless, the adaptive shifting strategy (E) provides without any doubt the main boost in performance. To be noticed, without the enriched context of each sparse candidate,

the adaptive shifting strategy shows limitations on disparity correcting and its performance gain (D) decreases a lot. Furthermore, results shown in (F, G, H) illustrates our TemporalStereo is able to predict accurate disparity with as few as 5 candidates. With the ability to shift the candidates towards a better solution, our model can search in a quite large space (e.g.,  $\beta = 2, 4, 8$  as reported in H, I, J) to get the best result when  $\beta = 4$ . 3D convolutions are the common operations to aggregate the cost in recent methods [3, 20, 60]. Our baseline model (A) benefits from the 3D convolutions (K) when compared to depth-wise [34] 3D convolutions (L), but the runtime increases a lot. In contrast, the full model (E) gives much higher improvement (N) with negligible runtime increase (4.8ms), and the time-consuming 3D convolutions are not necessary (M) for accuracy improvement.

**Temporal mode.** Before presenting the results achieved in temporal mode, we illustrate the protocol adopted at training and test time on TartanAir dataset [49]. *Given a temporal window containing  $W$  frames*, initial  $f = \{1, 2, \dots, W - 1\}$  frames are processed by the network sequentially without computing error metrics and blocking the gradients. Specifically, each outcome is cached and used in the next frame prediction. When  $f = W$ , we compute errors (and backpropagation at training time). Tab. 2 reports the ablation conducted on TartanAir [49], with  $W = 4$  both at train and test time, aimed at evaluating the importance of each temporal module. Specifically, we train the model in the single-pair mode for 20 epochs (i.e., the model learns how to solve stereo matching task), then we enable temporal components for 20 more epochs (i.e., the model now focuses on how to use past information). We can notice how the baseline (A), i.e., the model trained in single-pair mode for 40 epochs, could be improved using Temporal Shift to fuse past features with current ones (B). However, the benefit due to Temporal Shift is much lower than the gain provided by Local Map, which largely improves the performance (C). Finally, including cached past costs as well (D) provides an additional boost in accuracy.

**Impact of temporal window.** Tab. 3 reports the results achieved by TemporalStereo using different values of  $W$ . In particular, we could have two different values for  $W$ , that are  $W_{tr}$  and  $W_{test}$  for train and test respectively. In addition to the baseline  $W_{tr} = 1$  and the temporal  $W_{tr} = 4$  models, we also train a  $W_{tr} = 8$  model following the same configuration as for  $W_{tr} = 4$ . When compared with existing models such as PSMNet [6] and CoEx [3], our baseline outperforms them by a large margin in EPE metric. As for temporal mode, in general, the more frames joining the training or testing phases, the better result we can get in all regions. Notably,  $W_{test} = 4$  and  $W_{test} = 8$  are always beneficial in OCC, witnessing that TemporalStereo can effectively exploit more frames to deal with such difficult areas. Moreover, when  $W_{test} > 1$ , the results consistently

Method	$W_{test}$	EPE			3PE			
		ALL	OCC	NOC	ALL	OCC	NOC	
$W_{tr} = 1$	PSMNet [6]	1	0.866	2.654	0.558	4.80	18.63	2.48
	StereoNet [20]	1	0.888	2.647	0.578	5.15	19.34	2.68
	CoEx [3]	1	0.714	2.074	0.463	<b>3.83</b>	14.57	<b>1.93</b>
	Ours (single-pair)	1	<b>0.647</b>	<b>1.899</b>	<b>0.420</b>	3.96	<b>14.21</b>	2.08
$W_{tr} = 4$	Ours (single-pair)	1	0.665	1.731	0.459	3.95	13.34	2.16
	Ours (temporal)	4	0.610	1.637	0.413	3.73	12.60	2.03
	Ours (temporal)	8	<b>0.607</b>	<b>1.634</b>	<b>0.409</b>	<b>3.71</b>	<b>12.59</b>	<b>2.01</b>
$W_{tr} = 8$	Ours (single-pair)	1	0.673	1.912	0.450	4.00	14.03	2.17
	Ours (temporal)	4	0.609	1.625	0.412	3.74	12.58	2.03
	StereoNet [20] (temporal)	8	0.656	1.928	0.428	3.97	14.45	2.06
	CoEx [3] (temporal)	8	0.610	1.637	0.413	3.73	12.60	2.03
	Ours (temporal)	8	<b>0.601</b>	<b>1.615</b>	<b>0.405</b>	<b>3.71</b>	<b>12.54</b>	<b>2.02</b>

Table 3. **Impact of different frames in temporal mode.** Models are tested with  $W_{test} = 1, 4$  and 8.

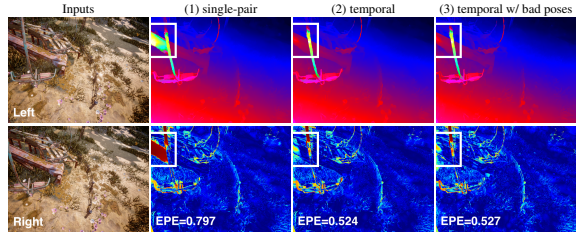


Figure 4. **Benefits of temporal mode.** Compared to single-pair mode (1), temporal mode (2) is more accurate at occlusions, even with noisy poses (3) – colder colors encode lower error.

overcome baseline with single stereo pair. It indicates our network can benefit from past information with only a few frames (e.g., 4, 8) after the network startup. Finally, we can notice how temporal models tested with  $W_{test} = 1$  obtain results closer to the baseline. This outcome implies that, in practical video applications, TemporalStereo can be trained once in temporal mode and used in single-pair way for the first inference (yet providing a good initial estimate) and in the temporal mode for all the others. Fig. 4 visualises the benefits of temporal model ( $W_{tr} = 8$ ,  $W_{test} = 8$ ) to alleviate occlusion errors. The temporal mode (2) largely outperforms the single-pair one (1). Furthermore, we highlight how TemporalStereo is also robust against inaccurate poses: although in (3) the camera pose is set to an identity matrix and only Temporal Shift module could help, it still outperforms (1). Eventually, the proposed temporal cues can be plugged into recent efficient methods (e.g., CoEx [3] and StereoNet [20]) for further improvement. Although CoEX still decreases from 0.714 to 0.610, with near 14% improvement. Nonetheless, our TemporalStereo utilizes past cues more effectively.

**Pose Analysis.** To assess the impact of pose input, we evaluate our model with setting  $W_{tr} = 8$  and  $W_{test} = 8$  on 4 scenes (2 indoor and 2 outdoor scenes) of TartanAir dataset [49] with hard motion patterns. The overall results of several pose inputs are listed in Tab 4. Specifically, the noise of rotation follows  $\mathcal{N}(0, \sigma_R)$  in degree( $^\circ$ ) and noise



Pose Type	Outdoor				Indoor			
	Amusement		SoulCity		Carwelding		Hospital	
	EPE	3PE	EPE	3PE	EPE	3PE	EPE	3PE
Single-pair	0.562	2.62	0.529	2.60	0.508	2.84	0.904	3.74
Identity	0.569	2.58	0.534	2.66	0.538	2.96	0.927	4.21
DROID-SLAM [46]	<b>0.545</b>	<b>2.51</b>	0.509	<b>2.52</b>	0.479	<b>2.63</b>	<b>0.848</b>	<b>3.30</b>
GT	<b>0.545</b>	<b>2.51</b>	<b>0.508</b>	<b>2.52</b>	<b>0.478</b>	<b>2.63</b>	<b>0.848</b>	<b>3.30</b>
$\sigma_R$ ( $^\circ$ )								
$\sigma_t$ (m)								
GT+Noise								
10	0.50	0.716	3.52	0.791	4.86	0.608	3.34	1.152
10	0.05	0.713	3.46	0.776	4.55	0.611	3.37	1.096
1	0.50	0.577	2.58	0.592	2.83	0.531	2.91	0.934
1	0.05	<b>0.546</b>	<b>2.54</b>	0.518	2.64	0.506	<b>2.72</b>	0.880
1	0.01	<b>0.546</b>	<b>2.54</b>	<b>0.517</b>	<b>2.62</b>	<b>0.505</b>	2.74	<b>0.873</b>
								<b>3.66</b>

Table 4. **Impact of different pose input in temporal mode.** “Single-pair” denotes our model is in single-pair mode (no pose needed); “Identity” means identity matrix; “DROID-SLAM” means pose estimated by DROID-SLAM (ORB-SLAM3 fails on these 4 scenes); and “GT+Noise” is the ground truth pose with manually added Gaussian noise.

Pretrain	SF		SF+Pseudo		TartanAir+Pseudo			
	Method	CoEx	Ours	CoEx	Ours	CoEx	Ours	Ours†
D1-BG	1.79	2.17	1.73	1.89	1.74	1.89	1.71	<b>1.61</b>
D1-FG	3.82	2.96	3.60	2.85	3.49	3.03	<b>2.78</b>	<b>2.78</b>
D1-ALL	2.13	2.30	2.04	2.05	2.03	2.07	1.89	<b>1.81</b>

Table 5. **Impact of pretraining.** We study the importance of pre-training by evaluating the D1 metric on KITTI 2015 test dataset. Results by both CoEx and Ours in single-pair and temporal mode (†) are reported accordingly. ‘SF’ and ‘TartanAir’ denotes SceneFlow and TartanAir datasets respectively. ‘+Pseudo’ means further training on KITTI raw sequences with our generated pseudo label.

of translation follows  $\mathcal{N}(0, \sigma_t)$  in meter(m). As reported, 1) poses from ground truth (GT) or estimated by DROID-SLAM [46] yield almost the same EPE and 3PE metrics. It means, when the ground truth pose is not available, an actual SLAM system like DROID-SLAM could be an alternative scheme. 2) Pose with small rotation error (e.g.,  $\sigma_R \leq 1^\circ$ ) and translation error (e.g.,  $\sigma_t \leq 0.05m$ ) gets very close results to the one with ground truth pose, and it always surpasses the results by single-pair mode. Even when the pose error comes to the maximum level of the dataset (i.e.,  $\sigma_R = 10^\circ, \sigma_t = 0.5m$ ), our model does not crash down and provides impressive predictions, proving the robustness of our model to inaccurate pose. 3) For outdoor scenes, since the view change between frames is much smaller than indoor scenes, the accuracy drops negligibly even when identity transformation is applied.

## 4.2. Evaluations on KITTI Benchmarks

To conclude, we run TemporalStereo on KITTI 2012 and KITTI 2015 test data and submit to the online leaderboard.

**Pretraining.** As KITTI is very challenging due the lack of a big training set, pretraining on SceneFlow dataset [28] is a common training schedule for deep learning-based stereo methods [6, 7]. Recent methods [38, 54] also introduce extra data, e.g., HR-VS [54], to augment KITTI during training. Following the knowledge distillation strat-

Method	SceneFlow [28]	KITTI 2012 [12]				KITTI 2015 [29]		
		Reflective		All		D1		
		EPE	3PE	SPE	5PE	BG	FG	ALL
slow	PSMNet [6]	1.09	10.18	5.64	1.89	1.15	1.86	4.62
	GwcNet-gc [14]	0.77	9.28	5.22	1.70	1.03	1.74	3.93
	GANet-Deep [58]	0.78	7.92	4.41	1.60	1.02	1.48	3.46
	ActNet [60]	0.87	8.52	5.28	1.54	1.01	1.51	3.80
	LEAStereo [7]	0.78	<b>6.50</b>	<b>3.18</b>	1.45	0.88	1.40	<b>2.91</b>
	ACVNet [52]	<b>0.46</b>	7.03	4.14	<b>1.13</b>	<b>0.71</b>	<b>1.37</b>	3.07
	CFNet [38]	-	7.29	3.81	1.58	0.94	1.54	3.56
	DWARF† [1]	-	-	-	-	-	3.20	3.94
	DTF-SENSE† [37]	-	-	-	-	-	2.08	3.13
	SENSE† [18]	-	-	-	-	-	2.07	3.01
fast	StereoNet [20]	1.10	-	-	6.02	-	4.30	7.45
	DeepPruner-Fast [8]	0.97	-	-	-	-	2.32	3.91
	AAANet+ [53]	0.72	9.10	5.12	2.04	1.30	1.65	3.96
	CoEx [3]	0.69	8.63	4.49	1.93	1.13	1.79	3.82
	HITNet [43]	<b>0.53</b>	7.54	4.01	1.89	1.29	1.74	3.20
	Ours (single-pair)	<b>0.53</b>	6.99	3.52	1.94	1.08	1.89	2.85
	Ours (temporal)	-	<b>6.14</b>	<b>3.08</b>	<b>1.61</b>	<b>0.88</b>	<b>1.61</b>	<b>2.78</b>
							<b>1.81</b>	

Table 6. **Comparison with state-of-the-art methods – slow and fast.** We report results of state-of-the-art methods on SceneFlow and KITTI. *fast* denotes models allowing real-time inference. ‡ means 3D scene flow-based methods.

egy proposed in AAANet+ [53], we augment the KITTI dataset by leveraging the prediction results from pretrained LEAStereo [7] to generate pseudo labels on KITTI raw sequences [10]. As a result, we get 61 stereo video sequences (containing 42K pairs) with pseudo labels for pre-training, and poses are calculated from the GPS/OXTS data on KITTI. We study the influence of different pretraining strategies on the D1 metric on KITTI 2015 test dataset. As shown in Tab. 5, compared with CoEx [3], our model in single-pair mode performs better in D1-FG and worse in D1-ALL, D1-BG metrics when pretrained on SceneFlow dataset only. After further training on pseudo labels, we get the almost same result as CoEx, which demonstrates our model requires more data to achieve better performance. Replacing the SceneFlow dataset with TartanAir, both the result of CoEx and ours in single-pair mode do not improve and achieve almost the same accuracy, i.e., D1-ALL 2.03% for CoEx and 2.07% for ours. However, by leveraging temporal information, the temporal mode can boost the accuracy of both TemporalStereo and CoEx further by a large margin, i.e. reducing D1-ALL to 1.81% and 1.89% respectively, supporting the major impact of our temporal paradigm over the pseudo labels training. We also point out that, despite the poses of KITTI raw sequences and KITTI 2015 are estimated by GPS/OXTS data and ORB-SLAM3 [5] respectively, TemporalStereo demonstrates its robustness to the pose error again.

Tab. 6 collects results achieved by a variety of deep stereo models, both on Scene Flow and the KITTI online benchmarks. For KITTI, we report the error rates achieved by TemporalStereo, both in single-pair and temporal mode ( $W_{tr}, W_{test} = 11$ ), finetuned from models pretrained on KITTI raw sequences [10]. In single-pair mode, our TemporalStereo achieves results already on par with state-of-the-art on all datasets. When switching to the temporal mode,



our network surpasses all fast stereo architectures by a large margin. In particular, our result on D1-FG is even better than the one achieved by slow models [7, 58] running in hundreds of milliseconds. It is worth mentioning that TemporalStereo in temporal mode, benefiting from past semantic and geometric information, achieves better results compared to single-pair mode on D1-FG metric (the D1 error on foreground areas, i.e., moving cars), which proves the robustness of our model to **dynamic objects** as well. The same advantage is also evident on 3PE and 5PE in **reflective regions** on KITTI 2012. Finally, we also compare with **3D scene flow** based state-of-the-art methods [1, 18, 37], which project with dense 3D motion field and output disparity. In contrast, simply using camera pose, our TemporalStereo is the obvious winner in both accuracy and efficiency on KITTI 2015.

## 5. Conclusions and Limitations

We presented TemporalStereo, a novel network devoted to fast stereo matching. The enhanced coarse-to-fine design and sparse cost volumes allow for fast inference and high performance. Moreover, TemporalStereo can exploit past information to ameliorate predictions, especially in occluded regions. The same model, trained once, can handle either single or multiple stereo pairs effectively. Considering the requirement of camera poses as its main limitation, empowering our system with pose estimation will be our future research direction.

**Acknowledgment.** We sincerely thank the scholarship supported by China Scholarship Council (CSC).

## References

- [1] Filippo Aleotti, Matteo Poggi, Fabio Tosi, and Stefano Mattochia. Learning end-to-end scene flow by distilling single tasks knowledge. In *AAAI*, 2020. 3, 8, 9
- [2] Abhishek Badki, Alejandro Troccoli, Kihwan Kim, Jan Kautz, Pradeep Sen, and Orazio Gallo. Bi3d: Stereo depth estimation via binary classifications. In *CVPR*, pages 1600–1608, 2020. 2
- [3] 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 *IROS*, 2021. 1, 2, 4, 5, 7, 8, 12, 13, 14, 15
- [4] Connelly Barnes, Eli Shechtman, Adam Finkelstein, and Dan B Goldman. Patchmatch: A randomized correspondence algorithm for structural image editing. *ACM TOG*, 28(3):24, 2009. 2
- [5] Carlos Campos, Richard Elvira, Juan J. Gómez, José M. M. Montiel, and Juan D. Tardós. ORB-SLAM3: An accurate open-source library for visual, visual-inertial and multi-map SLAM, 2020. 8, 12, 14
- [6] Jia-Ren Chang and Yong-Sheng Chen. Pyramid stereo matching network. In *CVPR*, pages 5410–5418, 2018. 2, 5, 7, 8
- [7] Xuelian Cheng, Yiran Zhong, Mehrtash Harandi, Yuchao Dai, Xiaojun Chang, Tom Drummond, Hongdong Li, and Zongyuan Ge. Hierarchical neural architecture search for deep stereo matching. *NeurIPS*, 2020. 1, 8, 9, 13, 14
- [8] Shivam Duggal, Shenlong Wang, Wei-Chiu Ma, Rui Hu, and Raquel Urtasun. Deeppruner: Learning efficient stereo matching via differentiable patchmatch. In *ICCV*, 2019. 1, 2, 8
- [9] Geoffrey Egnal and Richard P Wildes. Detecting binocular half-occlusions: Empirical comparisons of five approaches. *IEEE TPAMI*, 24(8):1127–1133, 2002. 13
- [10] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. *NeurIPS*, 27, 2014. 8, 12, 13, 14
- [11] Divyansh Garg, Yan Wang, Bharath Hariharan, Mark Campbell, Kilian Q Weinberger, and Wei-Lun Chao. Wasserstein distances for stereo disparity estimation. *NeurIPS*, 2020. 2, 4, 5
- [12] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, pages 3354–3361. IEEE, 2012. 2, 6, 8, 13, 14
- [13] Xiaodong Gu, Zhiwen Fan, Siyu Zhu, Zuozhuo Dai, Feitong Tan, and Ping Tan. Cascade cost volume for high-resolution multi-view stereo and stereo matching. In *CVPR*, June 2020. 2
- [14] Xiaoyang Guo, Kai Yang, Wukui Yang, Xiaogang Wang, and Hongsheng Li. Group-wise correlation stereo network. In *CVPR*, pages 3273–3282, 2019. 2, 3, 8, 12, 15
- [15] Heiko Hirschmüller. Stereo processing by semiglobal matching and mutual information. *IEEE TPAMI*, 30(2):328–341, 2007. 2
- [16] Asmaa Hosni, Christoph Rhemann, Michael Bleyer, Carsten Rother, and Margrit Gelautz. Fast cost-volume filtering for visual correspondence and beyond. *IEEE TPAMI*, 35(2):504–511, 2012. 2
- [17] Po-Han Huang, Kevin Matzen, Johannes Kopf, Narendra Ahuja, and Jia-Bin Huang. Deepmvs: Learning multi-view stereopsis. In *CVPR*, pages 2821–2830, 2018. 5
- [18] Huaizu Jiang, Deqing Sun, Varun Jampani, Zhaoyang Lv, Erik Learned-Miller, and Jan Kautz. Sense: A shared encoder network for scene-flow estimation. In *ICCV*, October 2019. 8, 9
- [19] 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 *ICCV*, pages 66–75, 2017. 2, 4
- [20] Sameh Khamis, Sean Fanello, Christoph Rhemann, Adarsh Kowdle, Julien Valentin, and Shahram Izadi. Stereonet: Guided hierarchical refinement for real-time edge-aware depth prediction. In *ECCV*, pages 573–590, 2018. 1, 2, 4, 7, 8
- [21] Hsueh-Ying Lai, Yi-Hsuan Tsai, and Wei-Chen Chiu. Bridging stereo matching and optical flow via spatiotemporal correspondence. In *CVPR*, June 2019. 3
- [22] 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 *CVPR*, pages 16263–16272, 2022. 1
- [23] Zhengfa Liang, Yiliu Feng, Yulan Guo, Hengzhu Liu, Wei Chen, Linbo Qiao, Li Zhou, and Jianfeng Zhang. Learning for disparity estimation through feature constancy. In *CVPR*, pages 2811–2820, 2018. 2
- [24] Ji Lin, Chuang Gan, and Song Han. Tsm: Temporal shift module for efficient video understanding. In *ICCV*, 2019. 6, 14, 15
- [25] Pengpeng Liu, Irwin King, Michael R Lyu, and Jia Xu. Flow2stereo: Effective self-supervised learning of optical flow and stereo matching. In *CVPR*, pages 6648–6657, 2020. 3
- [26] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *ECCV*, pages 21–37. Springer, 2016. 4
- [27] 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 *ICCV*, pages 6311–6319, 2021. 1, 2, 4
- [28] 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 *CVPR*, pages 4040–4048, 2016. 2, 3, 6, 8, 12, 14
- [29] Moritz Menze and Andreas Geiger. Object scene flow for autonomous vehicles. In *CVPR*, pages 3061–3070, 2015. 2, 6, 8, 13, 14
- [30] Raúl Mur-Artal and Juan D. Tardós. ORB-SLAM2: an open-source SLAM system for monocular, stereo and RGB-D cameras. *IEEE Transactions on Robotics*, 33(5):1255–1262, 2017. 5
- [31] Simon Niklaus and Feng Liu. Softmax splatting for video frame interpolation. In *CVPR*, 2020. 6
- [32] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *NeurIPS*, pages 8024–8035. Curran Associates, Inc., 2019. 12
- [33] Matteo Poggi, Fabio Tosi, Konstantinos Batsos, Philippos Mordohai, and Stefano Mattoccia. On the synergies between machine learning and binocular stereo for depth estimation from images: a survey. *IEEE TPAMI*, 2021. 2
- [34] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *CVPR*, pages 4510–4520, 2018. 7, 12
- [35] Daniel Scharstein and Richard Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *IJCV*, 47(1-3):7–42, 2002. 2
- [36] Korbinian Schmid, Teodor Tomic, Felix Ruess, Heiko Hirschmüller, and Michael Suppa. Stereo vision based indoor/outdoor navigation for flying robots. In *IROS*, pages 3955–3962. IEEE, 2013. 1
- [37] René Schuster, Christian Unger, and Didier Stricker. A deep temporal fusion framework for scene flow using a learnable motion model and occlusions. In *WACV*, 2021. 8, 9
- [38] Zhelun Shen, Yuchao Dai, and Zhibo Rao. Cfnet: Cascade and fused cost volume for robust stereo matching. In *CVPR*, pages 13906–13915, June 2021. 1, 2, 8
- [39] Sayanan Sivaraman and Mohan M Trivedi. A review of recent developments in vision-based vehicle detection. In *IV*, pages 310–315. IEEE, 2013. 1
- [40] Xiao Song, Xu Zhao, Hanwen Hu, and Liangji Fang. Edgestereo: A context integrated residual pyramid network for stereo matching. In *ACCV*, pages 20–35. Springer, 2018. 2
- [41] Jiaming Sun, Yiming Xie, Linghao Chen, Xiaowei Zhou, and Hujun Bao. Neuralrecon: Real-time coherent 3d reconstruction from monocular video. In *CVPR*, pages 15598–15607, 2021. 5
- [42] Mingxing Tan and Quoc Le. Efficientnetv2: Smaller models and faster training. In *ICML*, pages 10096–10106. PMLR, 2021. 12, 14
- [43] 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 *CVPR*, pages 14362–14372, June 2021. 3, 8
- [44] Zachary Teed and Jia Deng. Deepv2d: Video to depth with differentiable structure from motion. *ICLR*, 2020. 5
- [45] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *ECCV*, pages 402–419. Springer, 2020. 3, 5
- [46] Zachary Teed and Jia Deng. DROID-SLAM: Deep Visual SLAM for Monocular, Stereo, and RGB-D Cameras. *NeurIPS*, 2021. 8
- [47] Zachary Teed and Jia Deng. Raft-3d: Scene flow using rigid-motion embeddings. In *CVPR*, 2021. 3
- [48] Fabio Tosi, Yiyi Liao, Carolin Schmitt, and Andreas Geiger. Smd-nets: Stereo mixture density networks. In *CVPR*, pages 8942–8952, 2021. 2, 4
- [49] Wenshan Wang, Delong Zhu, Xiangwei Wang, Yaoyu Hu, Yuheng Qiu, Chen Wang, Yafei Hu, Ashish Kapoor, and Sebastian Scherer. Tartanair: A dataset to push the limits of visual slam. In *IROS*, 2020. 2, 6, 7, 12, 14
- [50] Yan Wang, Zihang Lai, Gao Huang, Brian H Wang, Laurens Van Der Maaten, Mark Campbell, and Kilian Q Weinberger. Anytime stereo image depth estimation on mobile devices. In *ICRA*, pages 5893–5900. IEEE, 2019. 1, 2
- [51] Yang Wang, Peng Wang, Zhenheng Yang, Chenxu Luo, Yi Yang, and Wei Xu. Unos: Unified unsupervised optical-flow and stereo-depth estimation by watching videos. In *CVPR*, pages 8071–8081, 2019. 3
- [52] Gangwei Xu, Junda Cheng, Peng Guo, and Xin Yang. Attention concatenation volume for accurate and efficient stereo matching. In *CVPR*, pages 12981–12990, 2022. 1, 8

- [53] Haofei Xu and Juyong Zhang. Aanet: Adaptive aggregation network for efficient stereo matching. In *CVPR*, pages 1959–1968, 2020. 1, 8
- [54] Gengshan Yang, Joshua Manela, Michael Happold, and Deva Ramanan. Hierarchical deep stereo matching on high-resolution images. In *CVPR*, June 2019. 2, 8, 12
- [55] Guorun Yang, Hengshuang Zhao, Jianping Shi, Zhidong Deng, and Jiaya Jia. Segstereo: Exploiting semantic information for disparity estimation. In *ECCV*, pages 636–651, 2018. 2
- [56] Qingxiong Yang. A non-local cost aggregation method for stereo matching. In *CVPR*, pages 1402–1409. IEEE, 2012. 2
- [57] Nadia Zenati and Noureddine Zerhouni. Dense stereo matching with application to augmented reality. In *NeurIPS*, pages 1503–1506. IEEE, 2007. 1
- [58] Feihu Zhang, Victor Prisacariu, Ruigang Yang, and Philip HS Torr. Ga-net: Guided aggregation net for end-to-end stereo matching. In *CVPR*, pages 185–194, 2019. 6, 8, 9
- [59] Feihu Zhang, Oliver J. Woodford, Victor Adrian Prisacariu, and Philip H.S. Torr. Separable flow: Learning motion cost volumes for optical flow estimation. In *ICCV*, pages 10807–10817, 2021. 4
- [60] Youmin Zhang, Yimin Chen, Xiao Bai, Suihanjin Yu, Kun Yu, Zhiwei Li, and Kuiyuan Yang. Adaptive unimodal cost volume filtering for deep stereo matching. In *AAAI*, volume 34, pages 12926–12934, 2020. 2, 4, 6, 7, 8
- [61] Yiran Zhong, Hongdong Li, and Yuchao Dai. Open-world stereo video matching with deep rnn. In *ECCV*, pages 101–116, 2018. 3

# – Supplementary Material –

## TemporalStereo: Efficient Spatial-Temporal Stereo Matching Network

Youmin Zhang   Matteo Poggi   Stefano Mattoccia  
CVLAB, Department of Computer Science and Engineering (DISI)  
University of Bologna, Italy

{youmin.zhang2, m.poggi, stefano.mattoccia}@unibo.it

### A. Video Demo

For a better understanding of the effect of TemporalStereo both in single-pair and temporal mode when dealing with a continuous video, and also as a comparison with state-of-the-art method [3], we conduct inference on another KITTI raw sequence ‘2011\_09\_28\_drive\_0037’ (a campus scene, its poses are computed by ORB-SLAM3 [5] algorithm.) which doesn’t appear in any our training process. The results are displayed in the appended **video demo**. As shown in the video, even without access to the ground truth pose, our TemporalStereo in temporal mode is robust to dynamic objects (*e.g.*, pedestrians). Benefiting from past context, for both static and dynamic regions, it provides more impressive predictions compared to CoEx [3] and our model in single-pair mode, *e.g.*, accurate estimations on occluded regions and fine-grained details on pillars of the building and street signs.

### B. Implementation Details

We implement our network in PyTorch [32], using RM-SProp as optimizer to train all models in an end-to-end fashion. We train the models from scratch for 40, 40 epochs on SceneFlow and TartanAir respectively. Given the pre-trained model from SceneFlow, TartanAir or KITTI Raw Sequences [10] as we discussed in Section 4.2 and Tab.5 of the main manuscript, we finetune our model for 16, 16 epochs on KITTI 2012 and KITTI 2015 separately. Scale weights are  $\lambda_0 = 1.0$ ,  $\lambda_1 = 0.5$ ,  $\lambda_2 = 0.7$ ,  $\lambda_3 = 2.0$ ,  $\lambda_{final} = 2.0$ , while we set  $K = 2$  for top- $K$  selection, number of candidates  $n = 12$  in stage 1 while  $n = 5$  in stages  $s = 2, 3$ ,  $\beta = 4$  for sampling and  $N_{key} = 3$  in Local Map. As for keyframe selection strategy, we set  $|\mathbf{t}_{max}| = 0.1\text{m}$  and  $|\mathbf{R}_{max}| = 15^\circ$  respectively. We leverage EfficientNetV2-S [42] as backbone feature extractor, while the hourglass network proposed in [14] for cost aggregation. For all 3D convolutions with kernel size  $k \times k \times k$  where  $k > 1$ , we implement it in depthwise [34] manner which consists two convolutions with kernels of size

$k \times 1 \times 1$  and  $1 \times k \times k$  respectively to save computation. For all datasets, we perform asymmetric chromatic augmentation as described in [54] to mitigate the effect of varying lighting and exposure. Furthermore, we also add random patching [54] on the right image to help the network deal with occluded areas. The maximum disparity value is  $D_{max} = 192$ . The smooth  $L_1$  loss, *i.e.*, Huber loss used in Eq.(4) of the main manuscript is defined as:

$$smooth_{L_1}(x) = \begin{cases} 0.5x^2, & \text{if } |x| < 1, \\ |x| - 0.5, & \text{otherwise.} \end{cases} \quad (12)$$

Furthermore, **our code** will be publicly available after the paper is accepted.

### C. Dataset Description and Qualitative Results

#### C.1. SceneFlow

It is a large synthetic dataset [28] including 35,454 training and 4,370 test images with a resolution of  $540 \times 960$ . We only use the Flyingthings part in “finalpass” format for training and testing. Specifically, we randomly crop the image of  $512 \times 960$  and set the batch size as 4. Since it does not provide ground truth camera poses, we use this dataset for single-pair mode only. The model is trained for 40 epochs with the initial learning rate of 0.001 decaying by a factor of 0.1 at epochs 30.

We validate the effectiveness of our model in single-pair mode on SceneFlow dataset and qualitative results are shown in Fig. A. We can notice that our network achieves extremely low error in both cases.

#### C.2. TartanAir

It’s a challenging synthetic dataset [49] with moving objects and various light and weather conditions. To adapt it for stereo matching, we manually split the dataset with hard motion (which has 6DoF motion, the max translation and rotation are 0.5 meters and  $10^\circ$  respectively.) for training and testing. Specifically, the scenes consisted the testing



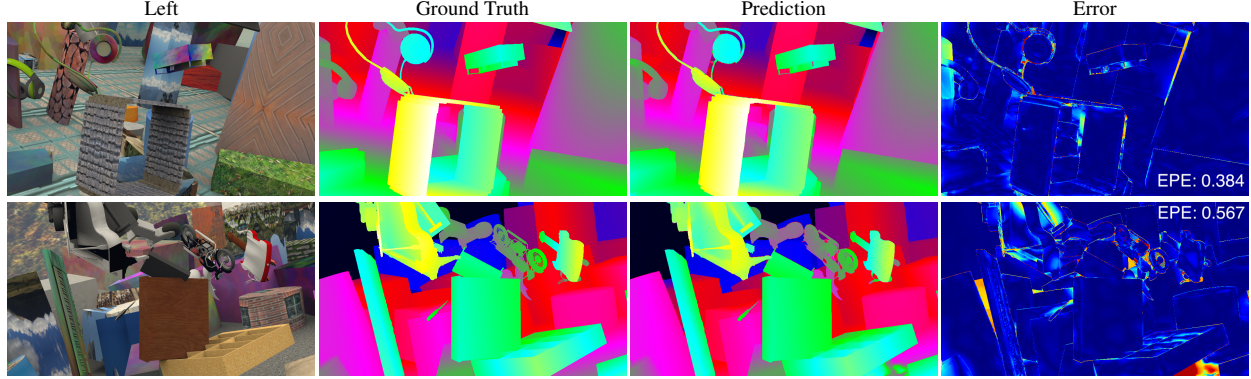


Figure A. **Qualitative results on SceneFlow.** From left to right, the reference/left image, the ground truth disparity map, the prediction in single-pair mode and its error (darker the color, lower the error).

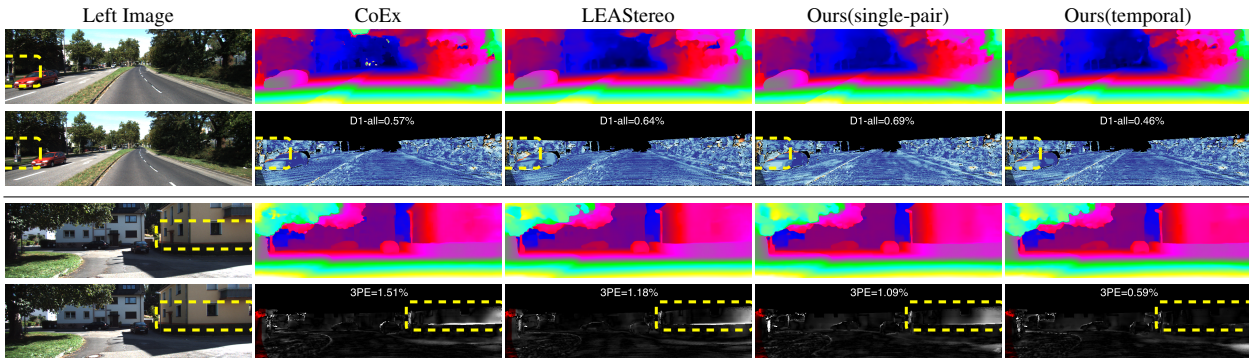


Figure B. **Results on KITTI 2015 and 2012 testing set.** Compared to existing methods [3, 7], TemporalStereo exploits time to improve accuracy at occlusions (top) or texturless regions (bottom). For the error maps on second and forth row, colder and darker color means lower the error respectively.

Scene	Part
abandonedfactory	P002
amusement	P007
carwelding	P003
endofworld	P006
gascola	P001
hospital	P042
office	P006
office2	P004
oldtown	P006
seasonforest	P002
seasonforest_winter	P015
soucity	P008

Table A. **TartanAir Testing Dataset.** We list Scene-Part pairs used in our testing set.

dataset are listed in Tab. A. The other parts of the corresponding scene are used for training. Overall, we collect nearly 66K stereo pairs for training and 5K for testing. The model is trained for 40 epochs with the initial learning rate of 0.001, reduced to 0.0001 at epoch 30. As for the first 20 epochs, we train the model in the single-pair mode to

let the network learn to perform stereo matching. Then, we set the model in the temporal mode for the left 20 epochs. For all experiments, we keep the image at full resolution (i.e.,  $480 \times 640$ ) with batch size 16, and use the same training schedule for all experiments on the TartanAir dataset. As the TartanAir dataset targets at visual SLAM and thus ground truth poses are provided. In temporal experiments, we leverage ground truth poses for warping.

### C.3. KITTI Raw Sequences

As we discussed in the main paper, we augment the KITTI [12, 29] datasets with KITTI Raw Sequences [10]. Specifically, the KITTI Raw Sequences are composed of several outdoor scenes captured with car-mounted cameras and depth sensors. Following [10], we get 61 stereo video sequences (containing 42K pairs) with pseudo labels for pretraining, and poses are calculated from the GPS/OXTS devices on KITTI. As for pseudo label generation, we leverage the prediction results from LEAStereo [7] which has been finetuned on KITTI 2015 [29] training dataset. We also perform left-right consistency [9] to filter out outliers



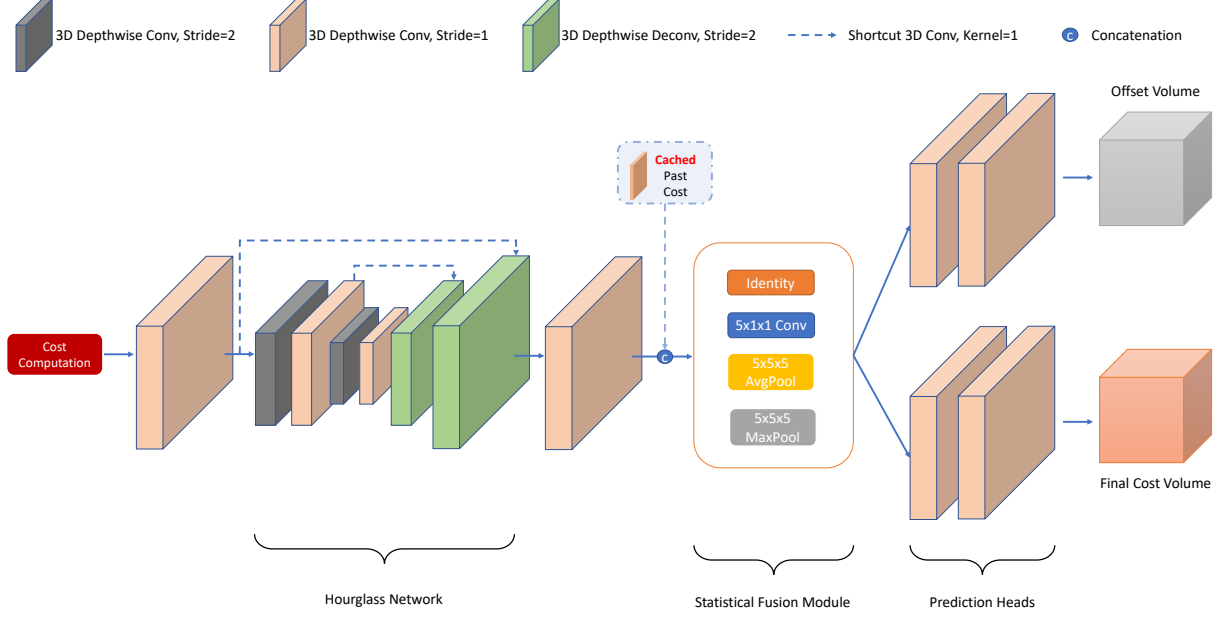


Figure C. The structure of our proposed network at stage 2.

in the generated pseudo labels. Given the model trained on SceneFlow or TartanAir dataset, we further pretrain the network for 10 epochs with an initial learning rate of 0.001 and decrease it to 0.0001 at epochs 8. Besides, we set the batch size as 4 and randomly crop the image into  $320 \times 1024$ .

#### C.4. KITTI 2012&2015

They are both real-world datasets collected by a driving car and depicting urban scenes. KITTI 2015 [29] contains 200 training and 200 testing stereo pairs while KITTI 2012 [12] has 194 and 195 pairs for training and test, respectively. Both the splits provide sparse ground truth depth labels. For temporal mode evaluation, we leverage the multi-view split, which contains, for each sample, also the 10 previous pairs. We estimate camera poses with an off-the-shelf SLAM algorithm [5]. We randomly crop the image into resolution  $320 \times 1024$ , use batch size as 4, and train our models on KITTI 2012 [12] and KITTI 2015 [29] with the same schedule. More specifically, for both single-pair and temporal mode, we use the pre-trained model from KITTI Raw Sequences [10] and finetune it for 16 epochs with the initial learning rate of 0.0001 and reduce it by a factor of 0.1 at epochs 12.

Fig. B and appended **video demo** show qualitative comparisons between existing networks [3, 7] and Temporal-Stereo, highlighting the benefits yielded by temporal mode.

#### D. Model Architecture Details

Tab. B presents the details of the TemporalStereo which is used in experiments to produce state-of-the-art accuracy

on Scene Flow dataset [28], TartanAir [49] and KITTI benchmarks [12, 29]. The feature extraction backbone is based on EfficientNetV2-S [42] and further equipped with Residual Temporal Shift Module (TSM) [24] to encode temporal info. For a better understanding of our architecture, we give an illustration of Stage 2 in Fig C.

Name	Layer setting	Output dimension
<b>Feature Extraction</b>		
input		$H \times W \times 3$
conv_stem	EfficientNetV2-S.Conv3 $\times$ 3	$\frac{1}{2}H \times \frac{1}{2}W \times 24$
block0	EfficientNetV2-S.Stage1	$\frac{1}{2}H \times \frac{1}{2}W \times 24$
block1	EfficientNetV2-S.Stage2	$\frac{1}{4}H \times \frac{1}{4}W \times 48$
block2	EfficientNetV2-S.Stage3	$\frac{1}{8}H \times \frac{1}{8}W \times 64$
block3	EfficientNetV2-Stage4,5 with Residual TSM [24] at each Inverted Residual Block	$\frac{1}{16}H \times \frac{1}{16}W \times 160$
block4	EfficientNetV2-S.Stage6,7 with Residual TSM [24] at each Inverted Residual Block	$\frac{1}{32}H \times \frac{1}{32}W \times 272$
conv_out4	$[3 \times 3, 320]$	$\frac{1}{32}H \times \frac{1}{32}W \times 320$
conv_out3	concat [conv_out4 $\uparrow$ , block3] $3 \times 3, 256$ $3 \times 3, 256$	$\frac{1}{16}H \times \frac{1}{16}W \times 256$
conv_out2	concat [conv_out3 $\uparrow$ , block2] $3 \times 3, 128$ $3 \times 3, 128$	$\frac{1}{8}H \times \frac{1}{8}W \times 128$
conv_out1	concat [conv_out2 $\uparrow$ , block1] $3 \times 3, 64$ $3 \times 3, 64$	$\frac{1}{4}H \times \frac{1}{4}W \times 64$
<b>Stage 1</b>		
cost_volume	concat and multiscale groupwise left and shifted right conv_out3	$\frac{1}{16}D \times \frac{1}{16}H \times \frac{1}{16}W \times 704$
aggregation_1	$1 \times 3 \times 3, 32, \quad 3 \times 1 \times 1, 32$ Hourglass Network [14] $1 \times 3 \times 3, 32, \quad 3 \times 1 \times 1, 32$	$\frac{1}{16}D \times \frac{1}{16}H \times \frac{1}{16}W \times 32$
fusion_1	concat PastCost / Identity Statistical Fusion Module	$(\frac{1}{16}D + \text{top-}K) \times \frac{1}{16}H \times \frac{1}{16}W \times 32$
output_1	Prediction Heads Convex Upsample	$\frac{1}{8}H \times \frac{1}{8}W$
<b>Stage 2</b>		
cost_volume	concat and multiscale groupwise left and shifted right conv_out2	$(n + N_{key}) \times \frac{1}{8}H \times \frac{1}{8}W \times 294$
aggregation_2	$1 \times 3 \times 3, 16, \quad 3 \times 1 \times 1, 16$ Hourglass Network [14] $1 \times 3 \times 3, 16, \quad 3 \times 1 \times 1, 16$	$(n + N_{key}) \times \frac{1}{8}H \times \frac{1}{8}W \times 16$
fusion_2	concat PastCost / Identity Statistical Fusion Module	$(n + N_{key} + \text{top-}K) \times \frac{1}{8}H \times \frac{1}{8}W \times 16$
output_2	Prediction Heads Convex Upsample	$\frac{1}{4}H \times \frac{1}{4}W$
<b>Stage 3</b>		
cost_volume	concat and multiscale groupwise left and shifted right conv_out1	$n \times \frac{1}{4}H \times \frac{1}{4}W \times 156$
aggregation_3	$1 \times 3 \times 3, 8, \quad 3 \times 1 \times 1, 8$ Hourglass Network [14] $1 \times 3 \times 3, 8, \quad 3 \times 1 \times 1, 8$	$n \times \frac{1}{4}H \times \frac{1}{4}W \times 8$
output_3	Prediction Heads Upsample Module from [3]	$H \times W$

Table B. Parameters of the network architecture of TemporalStereo.  $\uparrow$  means bilinearly upsampling with a factor of 2.  $n$  is the number of disparity candidates when applying inverse transform sampling,  $N_{key}$  is the number of last keyframes in the memory bank, while top- $K$  is the number of selected values in aggregated cost volume.