Multi-view Depth Estimation using Epipolar Spatio-Temporal Networks

Xiaoxiao Long[1]   Lingjie Liu[2]   Wei Li[3]   Christian Theobalt[2]   Wenping Wang[1,4]

[1]The University of Hong Kong   [2]Max Planck Institute for Informatics
[3]Inceptio   [4]Texas A&M University

Abstract

We present a novel method for multi-view depth estimation from a single video, which is a critical task in various applications, such as perception, reconstruction and robot navigation. Although previous learning-based methods have demonstrated compelling results, most works estimate depth maps of individual video frames independently, without taking into consideration the strong geometric and temporal coherence among the frames. Moreover, current state-of-the-art (SOTA) models mostly adopt a fully 3D convolution network for cost regularization and therefore require high computational cost, thus limiting their deployment in real-world applications. Our method achieves temporally coherent depth estimation results by using a novel Epipolar Spatio-Temporal (EST) transformer to explicitly associate geometric and temporal correlation with multiple estimated depth maps. Furthermore, to reduce the computational cost, inspired by recent Mixture-of-Experts models, we design a compact hybrid network consisting of a 2D context-aware network and a 3D matching network which learn 2D context information and 3D disparity cues separately. Extensive experiments demonstrate that our method achieves higher accuracy in depth estimation and significant speedup than the SOTA methods.

1. Introduction

Multi-view depth estimation aims to recover 3D geometry of given images with known camera parameters, which is one of fundamental problems in computer vision. Many applications benefit from the recovered depth, such as autonomous driving [39], augmented reality [18], 3D modeling [13], and image-based rendering [10].

Recently, learning based depth estimation methods, whether designed for images [9, 23, 17, 42] or videos [22, 46, 37, 29, 27], have achieved great improvements against their traditional counterparts [45, 31, 1, 15]; however, these methods, especially for videos, have significant room for improvement in terms of temporal coherence and computational efficiency.
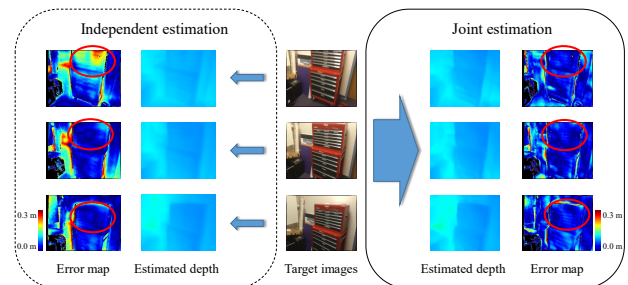

Figure 1. Independent estimation vs Joint estimation. Prior works estimate the depth maps of frames from a video independently despite their temporal coherence, thus the estimated depth maps may contain inconsistent values for the same region (red circles). Our model could estimate multiple depth maps jointly utilizing temporal coherence, enabling the estimated depth maps to be consistent.

Temporal coherence. Most of multi-view depth estimation methods [17, 42, 21, 26, 35] are designed for individual images so they are not suitable for temporally coherent videos. Directly extending the existing methods from images to video sequences causes flickering artifacts. i.e. inconsistent estimated depth maps across consecutive frames, because they do not take temporal coherence into consideration. As shown in Figure. 1, the independently estimated depth maps contain inconsistent values. Therefore, it is necessary to jointly estimate depth maps of a video sequence to produce temporally coherent results.

Existing works on video depth estimation can be divided into two categories: single-view methods [46, 6, 37, 29, 27] and multi-view methods [22], according to the input of the depth estimation network. Recurrent neural units are widely used in single-view video methods [46, 6, 37, 29] to encode temporal coherence implicitly. Although the inconsistency problem is alleviated by incorporating temporal coherence, these methods still suffer from the ambiguity of depth scale since depth estimation from a single image is an ill-posed problem.

Multi-view video methods are advantageous over single-view methods because epipolar geometry information provided by multiple images could be used to avoid depth scale

ambiguity problem. To the best of our knowledge, there has been only one multi-view method [22] for video depth estimation based on the epipolar geometry information. This method produces compelling results but it is restricted by its specific network design that it can only make use of one preceding estimation to improve current estimation. We improve upon this multi-view method by proposing a novel Epipolar Spatio-Temporal (EST) network that is capable of utilizing temporal coherence of multiple preceding estimations, thus producing more accurate depth maps with better temporal coherence.

Computational Efficiency. Top-performing multi-view depth estimation models [17, 21, 22] are slow with low computational efficiency, because they adopt a single fully 3D convolution networks for cost regularization to learn both local feature matching information and global context information, which have been shown to be critical for accurate depth estimation [2, 19]. While the local feature information is necessary for matching texture-rich regions, the global context information is crucial for scenes with textureless regions. The existing networks tend to use deeper and deeper networks to improve the ability of learning global context information, leading to increased computation cost. It should be pointed out that the global context information is essentially 2D, so learning it by deep 3D convolution layers will unnecessarily consume masses of computational resources.

Our insight is that for depth estimation the local feature information and global semantic information can be learned by two separate networks, as inspired by recent Mixture-of-Experts models [38, 47, 28]. Specifically, we propose a hybrid cost regularization network, consisting of two complementary expert sub-networks: a 2D *ContextNet* focusing on 2D global context information, and a shallow 3D *MatchNet* concentrating on 3D local matching information. By explicitly disentangles these two different types of information, our hybrid network consumes much less GPU computational resources and achieves faster running speed.

Our main contributions are summarized as follows:

- We proposed an Epipolar Spatio-Temporal (EST) transformer that propagates temporal coherence to perform joint depth estimation of multiple frames to make estimated depth maps more temporally coherent.

- We designed a hybrid network for cost regularization that consists of two expert networks to learn 3D local matching information and 2D global context information separately. This decoupling approach achieves faster speed and consumes less computational resources than using a single fully 3D network in prior works.

- Based on these two contributions we developed a new mutli-view method for generating temporally coherent

depth maps from videos. We conducted extensive experiments on several datasets to demonstrate that our method outperforms the SOTAs by a large margin in terms of accuracy and speed.

2. Related Work

Depth estimation from single image Learning based single-view depth estimation methods [9, 8, 3, 4, 36, 41, 5, 25, 24] have been extensively studied in recent years. Convolution neural network for depth estimation is first introduced by [9], which demonstrates the superior ability of CNN for depth regression. Later, [23] combines conditional random field (CRF) with CNN to further improve the quality of estimated depth. [11] proposed a seminal ordinal regression loss instead of metric $l1/l2$ loss, recasting depth regression as an ordinary regression problem. Some works [30, 43] introduce extra geometric constraints to improve depth estimation. However, these methods suffer from the scale ambiguity problem due to single-view depth estimation is an ill-posed problem.

Multi-view stereo depth estimation Recently, some learning based methods [42, 35, 17, 42] based plane-sweeping volume algorithm [12] have achieved promising improvements against their traditional counterparts [16, 14]. Some works [21, 26] introduce surface normal as extra constraint to further improve multi-view depth estimation. However, most of the methods are designed for individual images not suitable for videos, thus directly extending these methods from images to video sequence causes flickering artifacts. Moreover, these methods mostly adopt fully 3D convolution network for cost regularization, which is memory-consuming, expensive, and slow. Although one method [33] tries to avoid 3D convolution layers by replacing plane-sweeping construction by correspondence triangulation, singular value decomposition (SVD) operations used in its triangulation module lead to more time consumption than plane-sweeping methods.

Temporal coherence in video depth estimation To utilize temporal coherence to improve depth accuracy, some single image depth estimation methods [11, 37, 46, 29] exploit recurrent neural units to encode temporal correlation in latent space. However, these methods suffer from depth scale ambiguity problem and show poor generalization to frames with unseen camera motions. Another single depth estimation method [27] adopts an online training strategy, performing model re-training on testing videos with a geometric consistency loss function. However, the online training scheme costs more than 40 minutes for a video with 244 frames and might suffer from model degradation due to insufficient online training data. One multi-view method, neuralrgbd [22] exploits temporal coherence by accumulating depth probability over time. However, it can only make
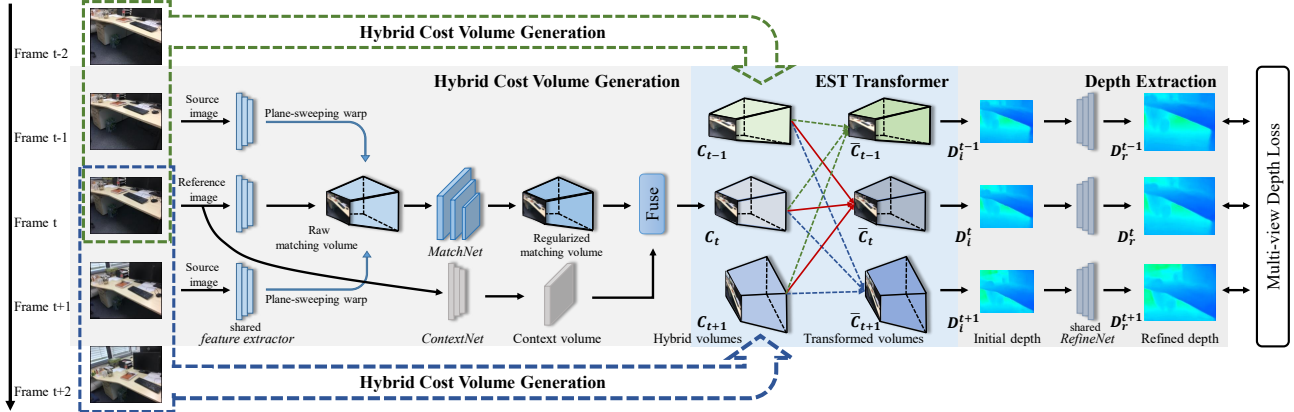
Figure 2. Overview of our method. In the training stage, our method takes a clip of video with five frames as input to estimate the depth maps of the middle three target frames. For each target image with its source images, we first extract learned feature maps by a shallow feature extractor. Subsequent plane-sweeping warping is applied to the feature maps to construct raw matching volume, and it is further regularized by *MatchNet* to obtain a regularized matching volume. In parallel, we utilize *ContextNet* to learn 2D context information from the target image, and concatenate context volume and matching volume together to get one hybrid volume. The Epipolar Spatio-Temporal (EST) transformer is applied on all obtained hybrid volumes to associate temporal coherence with them. Finally, we extract initial depth maps from the transformed volumes and further refine the initial depth maps by *RefineNet*.

use of one preceding estimation to improve the depth estimation of the current frame. In contrast, our method can take advantage of the temporal coherence of multiple preceding estimations for more accurate and temporally consistent depth maps.

3. Method

We propose an end-to-end pipeline for multi-view depth estimation as shown in Figure 2. In the training stage, our model takes a video clip with 5 images $\{I_{t-2}, I_{t-1}, I_t, I_{t+1}, I_{t+2}\}$ as input, and estimate the depth maps of three target images $\{I_{t-1}, I_t, I_{t+1}\}$ jointly with short-term temporal coherence. In the inference stage, our model could propagate long-term temporal coherence through the whole video by an Epipolar Spatio-Temporal Memory (ESTM) inference operation.

Briefly, our method has the following four parts, which will be discussed in detail in the subsequent sections:

i Hybrid cost volume generation (see Section 3.1). It obtains a regularized matching volume from each target image with its source images using *MatchNet*, and in parallel extracts context feature volume of the target image using *ContextNet*. Finally, it fuses the matching volume and context volume together to a hybrid cost volume.

ii Epipolar Spatio-Temporal Transformer (see Section 3.2). It is applied on the hybrid cost volumes of all the target images to enforce temporal coherence.

iii Depth extraction (see Section 3.3). In this part initial depth maps are first extracted from the transformed cost volumes and further refined by *RefineNet*.

iv Epipolar Spatio-Temporal Memory (ESTM) inference operation (see Section 3.4). In the inference stage, this operation propagates long-term temporal coherence through the whole video.

3.1. Hybrid cost volume generation

For computational efficiency, we utilize two expert sub-networks, *MatchNet* and *ContextNet*, to learn two types of cost volumes, 3D matching volume and 2D context volume, respectively.

3.1.1 Matching volume generation

To simplify the exposition, we take frame I_t as an example to construct its matching volume. Frame I_t and its neighboring frames I_{t-1}, I_{t+1} are taken as reference image and source images respectively, since they share the largest overlapping regions. Without loss of generality, here we only consider the reference image and one of its source images, denoted as I_r and I_s respectively.

Feature extraction We first pass I_r and I_s through a spatial pyramid pooling (SPP) *feature extractor* [2] to extract corresponding hierarchical feature maps F_r and F_s. For computational efficiency, the feature maps are down-sampled by four to the original image size with 32 channels.

Raw matching volume A raw matching volume for reference image I_r is constructed by backprojecting source feature map F_s into the coordinate system defined by I_r at a stack of fronto-parallel virtual planes. The virtual planes are uniformly sampled over a range of depths $z_1, z_2, \cdots, z_D$, which are chosen to span the ranges observed in the dataset. Following prior works [17, 26, 22, 21, 35], we set $D = 64$

in all experiments. The coordinate mapping from the source feature map F_s to the reference feature map F_r, at each depth z_m, is determined by a planar homography transformation:

$$u'_m \sim H_m u, u'_m \sim \mathbf{K}[\mathcal{R}_s|\mathbf{t}_s] \begin{bmatrix} (\mathbf{K}^{-1}u)z_m \\ 1 \end{bmatrix} \quad (1)$$

where the '$\sim$' denotes the projective equality, u is homogeneous coordinate of a pixel in the reference image, u'_m is the projected homogeneous coordinate of u on the paired source feature map. K denotes the intrinsic parameters of the camera, $\{\mathcal{R}_s, \mathbf{t}_s\}$ are the rotation and the translation of the source image I_s relative to the reference image I_r.

Based on the above mapping, we warp the source feature map into all the virtual planes to construct a feature volume with dimension $C \times D \times H \times W$. Finally, we concatenate the feature volume with F_r at each virtual plane to increase the dimension to $2C \times D \times H \times W$. By concatenation operation, the network could receive necessary information to perform feature matching between F_r and F_s without decimating the feature dimension [35, 26]. With N source images, we will obtain N raw matching volumes.

MatchNet The N raw matching volumes are first processed by three 3D convolution layers to decrease their dimension to $C \times D \times H \times W$. Then a view average pooling operation is performed on the volumes to aggregate information across different source images, yielding a single aggregated volume. Finally, the aggregated volume is further regularized by a series of 3D convolution layers. Note that our *MatchNet* is only responsible for learning local features for matching. We will use another network for learning global context information to complement *MatchNet* for more efficient depth estimation, as will be discussed in the next section.

3.1.2 Context volume generation

Prior methods [22, 17, 21] adopt a heavy 3D regularization networks to learn 2D context information together with 3D local matching clues in a mixed manner, and the network is made very deep for increased ability of learning the two types of information. We observe that the global context information is essentially 2D information, so it is unnecessary to use a 3D network to learn it. Hence, we decouple the global context information from the local matching information and use a 2D network, called *ContextNet*, to learn the former. This makes the network simpler and, consequently, more efficient to train and run.

Specifically, we use Resnet-50 as the *ContextNet*. The output of *ContextNet* is a learned feature volume with size $C' \times H \times W$, where C' is the number of feature channels and the same as the number of virtual planes D. To fuse 3D matching volume and the 2D context volume, we expand
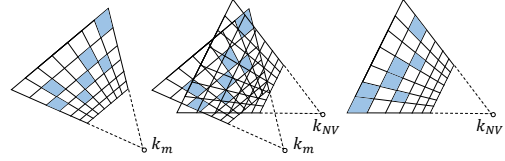

Figure 3. Epipolar warping. To apply EST transformer on the query volume $\mathbf{C}_q$, we need to resample the memory keys and memory values with respect to a novel view, the camera view of query volume. For better visualization, we only look at a planar slice of the key volumes. Each voxel of the volumes is encoded by a one-dimension feature, and here we only colorize some voxels for simplification.

the dimensions of context volume to $1 \times D \times H \times W$. Finally we concatenate the regularized matching volume and the expanded context volume together to get a hybrid cost volume, with size $(C + 1) \times D \times H \times W$. As shown in Figure 2, we repeatedly apply the hybrid cost volume generation operation for images I_{t-1}, I_t, I_{t+1} and obtain three corresponding hybrid cost volumes $\mathbf{C}_{t-1}, \mathbf{C}_t, \mathbf{C}_{t+1}$.

3.2. Epipolar Spatio-Temporal transformer

To associate temporal coherence with the three hybrid cost volumes $\mathbf{C}_{t-1}, \mathbf{C}_t, \mathbf{C}_{t+1}$, we propose a novel Epipolar Spatio-Temporal (EST) transformer.

Consistency constraint Our EST transformer is inspired by the photometric consistency assumption: a 3D point in world space will be projected into visible images I_{t-1}, I_t, I_{t+1}, and the image textures near their projections should bear high similarity. We formulate depth estimation as an occupancy estimation problem: if a pixel (u, v) of image I_t has depth value d then the voxel (u, v, d) in $\mathbf{C}_t$ is occupied, that is, the learned features of $\mathbf{C}_t$ encode the probability of occupancy for each voxel. The hybrid cost volumes are treated as multiple occupancy measurements for the same 3D world space in different viewpoints, namely, for a 3D point in world space, its corresponding voxels of the volumes $\mathbf{C}_{t-1}, \mathbf{C}_t, \mathbf{C}_{t+1}$ should keep similar embedding vectors.

Epipolar warping To associate temporal coherence with the hybrid volumes, we should first perform epipolar warping (see Figure 3) to convert the hybrid volumes into the same camera coordinate space. Assuming a 3D point (X_w, Y_w, Z_w) in the world coordinate is observed by target images I_t and I_{t+1}, that is, the voxel (u, v, d) of $\mathbf{C}_t$ and (u', v', d') of $\mathbf{C}_{t+1}$ are occupied. The coordinate mapping from (u, v, d) to (u', v', d') can be easily derived from Equation 1. Using this mapping, we warp $\mathbf{C}_{t-1}$ and $\mathbf{C}_{t+1}$ into the camera coordinate space of $\mathbf{C}_t$ and obtain two warped hybrid volumes $\mathbf{C}_{t-1}^{warp}, \mathbf{C}_{t+1}^{warp}$. After converted into the same camera coordinate, the two warped volumes and $\mathbf{C}_t$ should contain similar features in voxels for overlapped regions.
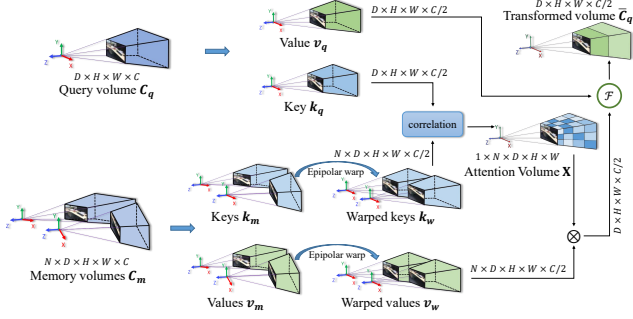
Figure 4. The structure of EST Transformer. Both query volume and memory volumes are encoded into pairs of **key** and **value** maps through shared convolution layers. The memory keys and memory values are epipolar warped into the camera space of the query volume. Similarities between query key and warped memory keys are computed to determine when-and-where to retrieve relevant memory values from. Finally we adaptively fuse the retrieved values with query value together for final output. Here $\otimes$ means matrix inner product, and $\mathcal{F}$ denotes fusion function.

EST transformer As depicted in Figure 4, we denote the hybrid volume to be transformed $\mathbf{C}_t$ as query volume and others ($\mathbf{C}_{t-1}$ and $\mathbf{C}_{t+1}$) as memory volumes. For computational efficiency, instead of applying EST transformer on the hybrid volumes, we first feed the query volume and memory volumes into two parallel and identical convolution layers to generate two new squeezed feature maps **key**s $k \in \mathbb{R}^{C/2 \times D \times H \times W}$ and **value**s $v \in \mathbb{R}^{C/2 \times D \times H \times W}$. The memory keys and memory values denoted as k_m and v_m, are first epipolar warped into the camera space of $\mathbf{C_t}$, obtaining warped keys and values denoted as k_w and v_w. Then we calculate the correlation between the query key k_q and the warped keys k_w, yielding a correlation volume, which measures the similarity of k_q and k_w. Finally we apply a softmax layer to get an attention volume $\mathbf{X} \in \mathbb{R}^{1 \times N \times D \times H \times W}$:

$$x_i = \frac{exp(k_q \cdot k_w^i)}{\sum_{i=1}^{N} exp(k_q \cdot k_w^i)} \quad (2)$$

where $x_i \in \mathbb{R}^{1 \times 1 \times D \times H \times W}$ measures the similarity of query to the warped key of ith memory volume, N is the number of memory volumes, and $\cdot$ means dot product.

The attention map $\mathbf{X}$ is used to retrieve relevant values from all warped memory values $v_w^i, i = 1, \cdots, N$. Finally we fuse the query value and retrieved values together to obtain the final output $\bar{\mathbf{C}}_q$:

$$\bar{\mathbf{C}}_q = f(v_q, \sum_{i=1}^{N} x_i v_w^i) \quad (3)$$

where v_w^i is the ith warped memory value, and $f(\cdot, \cdot)$ denotes a fusion function.

We explore two types of fusion function: concatenation fusion and adaptive fusion. For concatenation fusion, we just simply concatenate query value and the retrieved values together, which is straightforward and introduces no trainable parameters. However, due to occlusion or surface reflection, there might exist incorrect information in retrieved values but the concatenation operation integrates all information equally. To avoid propagating wrong information to current estimation, we propose an adaptive fusion operation to fuse the query value and the retrieved values:

$$f(v_q, y) = w \odot y + (1-w) \odot g(v_q, r \odot y), y = \sum_{i=1}^{N} x_i v_w^i \quad (4)$$

where $\odot$ means Hadamard product, $w, r \in \mathbb{R}^{D \times H \times W}$ are two learned weight volumes which measure the reliability of the retrieved values, $g(\cdot, \cdot)$ is a convolution layer. Unless otherwise specified, EST transformer refers to the adaptive transformer. As shown in Table 1, the adaptive fusion outperforms the simple concatenation fusion.

The EST transformer is applied to the other two hybrid cost volumes $\mathbf{C}_{t-1}$ and $\mathbf{C}_{t+1}$, yielding three transformed cost volumes $\bar{\mathbf{C}}_{t-1}, \bar{\mathbf{C}}_t$ and $\bar{\mathbf{C}}_{t+1}$. We first apply two convolution layers on the transformed volumes to produce intermediate volumes with size $D \times H \times W$, and then utilize the softmax operator over the depth dimension, so that we obtain probability volumes $\mathbf{P}$.

3.3. RefineNet and depth regression

We extract a depth map from probability volume $\mathbf{P}$ by *soft argmax* operation [19], which calculates the expected depth. We denote the depth map regressed from probability volume as initial depth map, since its size is downsized by four compared with original image size, whose fine grained features are lost and boundary edges are jagged. So we use a two-stage *RefineNet* to gradually upsample the initial depth maps and enhance fine-grained features, yielding a $1/2$ resolution depth map and a full-resolution depth map. Besides, we also extract a depth map from hybrid cost volumes before being applied EST transformer. Consequently, we obtain estimated depth maps of four stages, and we denote the four types of depth map from hybrid cost volumes, transformed cost volumes, and two-stage *RefineNet* as $\mathbf{D}_s, s = 0, 1, 2, 3$.

Unlike single-view depth loss used in prior works, we utilize multi-view depth loss to provide multiple supervision signals in different viewpoints:

$$loss = \frac{1}{N} \sum_{s=0}^{3} \sum_{i=1}^{N} \lambda^{s-3} \left\| \mathbf{D}_s^i - \hat{\mathbf{D}}_s^i \right\|_1 \quad (5)$$

where i is the the index of target image, N is the number of target images, $\hat{D}$ means ground truth depth map, and $\|\cdot\|_1$ means L1 norm. The weight λ is set to 0.8 in all experiments.
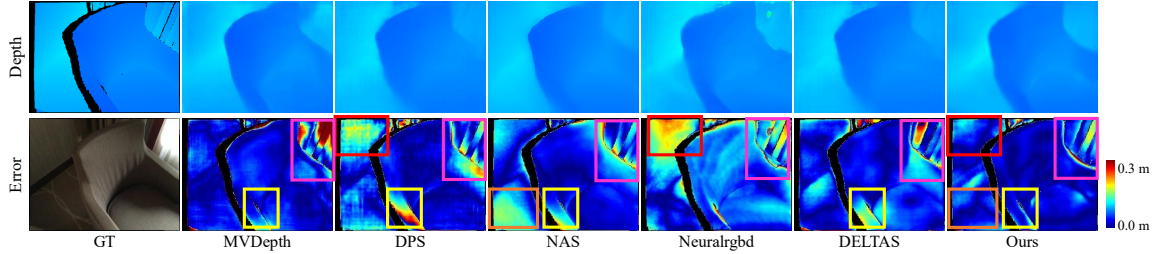
Figure 5. Depth comparison with the SOTAs. Our estimated depth map is more accurate and has less noises even in texture-less regions (the marked boxes). Black regions are invalid regions where gt depth values are missed. More comparisons are in the supplementary materials.
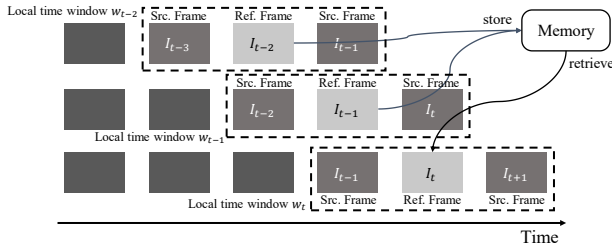

Figure 6. Epipolar Spatio-Temporal Memory inference operation.

3.4. Epipolar Spatio-Temporal Memory Inference

In the training stage, our model takes a short video sequence with 5 frames as input and jointly estimate the depth maps of three target images with short-term temporal coherence. To propagate long-term temporal coherence through the whole video, we propose an Epipolar Spatio-Temporal Memory (ESTM) inference operation. As depicted in Figure 6, we hold a sliding window containing one reference image and two source images to estimate the depth map of the current frame I_t. Using the EST transformer, we retrieve relevant values from a memory space storing the pairs of keys and values of N past frames, thus useful information at different space-time locations can be utilized for estimating the depth map of the current frame. When the sliding window moves on, the memory space will be also updated accordingly, by which operation the long-term temporal coherence is propagated through the whole video.

4. Datasets and Implementation details

Datasets We use ScanNet dataset [7] for training our end-to-end pipeline. The whole dataset consists of more than 1600 indoor scenes, which provides color images, ground truth depth maps and camera poses. Some MVS methods, DPS [17] and MVDepth [35] are trained on DeMoN [34] dataset, but DeMoN dataset mainly consists of two-view image pairs, which is not appropriate for our setting. For a fair comparison, we finetune DPS and MVDepth on ScanNet, and evaluate all methods on the official test split of ScanNet.

Furthermore, we test our method on 7scenes [32] and SUN3D [40] datasets for cross-dataset evaluation. 7scenes and SUN3D datasets also provide color images, gt depth maps and camera poses. Unlike prior works [34, 17, 21] sample two-view image pairs from videos, we directly adopt whole videos for evaluation.

Implementation In the training stage, we use a video clip with 5 frames as input. The frames are sampled from 30fps video with an interval of 10. Our model is implemented by Pytorch with Adam optimizer ($lr = 0.00004$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $weight\_decay = 0.00001$). We train our model for 7 epochs (115k iterations) with batch size 4 on four GeForce RTX 2080 Ti GPUs, and the learning rate is downscaled by a factor of 2 every two epochs.

5. Experiments

We compare our method with SOTA models in two aspects: depth accuracy and computational efficiency.

5.1. Evaluation metrics

To quantitatively evaluate the estimated depth, we use the standard metrics defined in [9]: i) inlier ratios ($\sigma < 1.25^i$ where $i \in \{1, 2, 3\}$); ii) Absolute error (Abs); iii) Absolute relative error (Abs Rel); iv) Square Relative error (Sq Rel); iv) Root Mean Square Error (RMSE); v) RMSE in log space (RMSE log).

To measure computational complexity of models, we adopt the following metrics: i) the number of trainable parameters; ii) the number of Multiply–Accumulate Operations (MACs); iii) inference memory usage; iv) inference time cost.

5.2. Depth evaluation

Depth accuracy Since MVDepth [35] and DPS [17] are trained on DeMon dataset [34] not on ScanNet [7] dataset. For fair comparisons, we fine-tune MVDepthNet and DPSNet on ScanNet. As shown in Table 1, our model significantly outperforms all other methods over both ScanNet and 7scenes datasets. Compared with Neuralrgbd [22], our model not only achieves better results in ScanNet dataset but also shows superior generalization ability on unseen

Table 1. Depth comparison over ScanNet [7] and 7scenes [32] datasets. Our method outperforms the other methods by a large margin. We report results in two depth ranges, since Neuralrgbd is trained in range $0 \sim 5m$. Complete tables are in the supplementary materials.

Range	Method	ScanNet					7scenes				
		Abs Rel	Abs	Sq Rel	RMSE	$\sigma < 1.25$	Abs Rel	Abs	Sq Rel	RMSE	$\sigma < 1.25$
10m	MVDepth [35]	0.1167	0.2301	0.0596	0.3236	84.53	0.2213	0.4055	0.2401	0.5154	67.33
	MVDepth-FT	0.1116	0.2087	0.0763	0.3143	88.04	0.1905	0.3304	0.1319	0.4260	71.93
	DPS [17]	0.1200	0.2104	0.0688	0.3139	86.40	0.1963	0.3471	0.1970	0.4625	72.51
	DPS-FT	0.0986	0.1998	0.0459	0.2840	88.80	0.1675	0.2970	0.1071	0.3905	76.03
	NAS [21]	0.0941	0.1928	0.0417	0.2703	90.09	0.1631	0.2885	0.1023	0.3791	77.12
	CNM [26]	0.1102	0.2129	0.0513	0.3032	86.88	0.1602	0.2751	0.0819	0.3602	76.81
	DELTAS [33]	0.0915	0.1710	0.0327	0.2390	91.47	0.1548	0.2671	0.0889	0.3541	79.66
	Ours-EST(concat)	0.0818	0.1536	0.0301	0.2234	92.99	**0.1458**	0.2554	0.0745	0.3436	79.82
	Ours-EST(adaptive)	**0.0812**	**0.1505**	**0.0298**	**0.2199**	**93.13**	<u>0.1465</u>	**0.2528**	**0.0729**	**0.3382**	**80.36**
5m	Neuralrgbd [22]	0.1013	0.1657	0.0502	0.2500	91.60	0.2334	0.4060	0.2163	0.5358	68.03
	Ours-EST(concat)	0.0811	0.1469	0.0279	0.2066	93.19	**0.1458**	0.2554	0.0745	0.3435	79.82
	Ours-EST(adaptive)	**0.0805**	**0.1438**	**0.0275**	**0.2029**	**93.33**	<u>0.1465</u>	**0.2528**	**0.0729**	**0.3382**	**80.36**

7scenes dataset. In Figure 5, compared with DPS [17], NAS [21], Neuralrgbd [22] and DELTAS [33], our estimated depth maps are more accurate and contain fewer noises. Our depth maps have smaller errors in texture-less regions, such as floor, wall, and sofa. Furthermore, we feed the estimated depth maps into TSDF fusion system for 3D reconstruction, as shown in Figure 7, our reconstructed model is more accurate and has fewer outliers.

Temporal coherence To measure the temporal consistency of the estimated depth maps, we adopt standard deviation of the mean absolute error of the estimated depth maps for evaluation. As shown in Table 2, the estimated depth maps of our methods are more accurate and temporally consistent than those of other methods (See supplementary materials for more details).

Table 2. Comparison of temporal coherence over ScanNet dataset with depth evaluation range $0 \sim 5m$.

Metric	DPS[17]	NAS [20]	Neuralrgbd [22]	DETALS [33]	Ours
Abs	0.1887	0.1823	0.1642	0.1650	**0.1432**
Std	0.2243	0.2177	0.1848	0.1886	**0.1673**

5.3. Analysis of computation complexity

To evaluate the computational efficiency of our model, we compare it with three plane-sweeping based methods, namely DPS, NAS, and neuralrgbd, plus the correspondences triangulation based method DELTAS. We run the models on one RTX 2080Ti GPU with the same setting: one reference image with two source images with a size of 320×256. We run our model using ESTM inference operation with 2 memory volumes.

Table 3 shows that DELTAS has the most trainable parameters but the lowest MACs, owing to the low computational cost of a fully 2D convolution network it uses. However, DELTAS consumes much more time than all the other plane-sweeping stereo methods because its correspondence triangulation module needs to perform time-consuming Singular Value Decomposition (SVD). Our method achieves
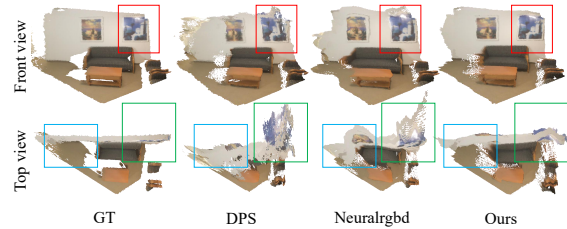

Figure 7. 3D reconstruction results by TSDF fuison method [44]. To evaluate temporal coherence, we fuse the estimated depth maps of 10 successive sampled video frames with the interval of 10. Compared to DPS [17] and Neuralrgbd [22], our method generates more faithful and temporally consistent depth maps.

Table 3. Memory and computation complexity analysis.

Model	Params(M)	MACs(G)	Memory(M)	Time(ms)
DPS [17]	**4.2**	442.7	**1595**	337
NAS [21]	18.0	527.7	1689	212
Neuralrgbd [22]	<u>5.3</u>	616.6	2027	195
DELTAS [33]	124.6	**98.6**	2395	495
Ours-ESTM(concat)	36.1	<u>165.6</u>	<u>1641</u>	**66**
Ours-ESTM(adaptive)	36.2	176.9	1799	<u>71</u>

significantly faster speed than the other methods. It takes about 40 milliseconds for our model to perform the adaptive ESTM operation, and 31 milliseconds to forward convolution layers, thus 71 milliseconds in total.

5.4. Ablation studies

In this section we evaluate the efficacy of our EST transformer and hybrid cost regularization network.

Epipolar Spatio-Temporal transformer We consider several variants of our method for ablation studies. We denote the depth estimated by the model without EST transformer as *independent depth*, the depth jointly estimated by the model with EST transformer as *joint depth*, and the depth sequentially estimated by the model using ESTM operation as *ESTM depth*. As shown in Table 4, when using the hybrid cost regularization network, both *joint depth* and *ESTM depth* are better than *independent depth*. However, when adopting the pure 3D regularization network (without *ContextNet*), *joint depth* is worse than *independent depth*,
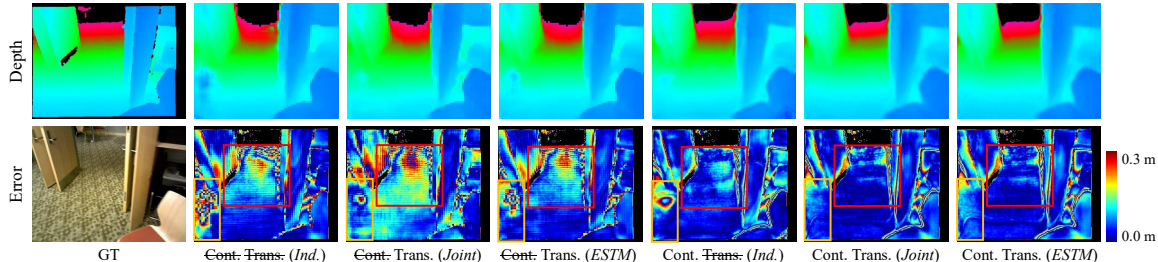
Figure 8. Effects of EST transformer and *ContextNet*. Here ~~Cont.~~ denotes models without *ContextNet*, ~~Trans.~~ denotes models without EST transformer, *Ind.* means that independently estimate depth maps by the model without EST transformer, *Joint* means that jointly estimate multiple depth maps at a time, and *ESTM* means that estimate depth maps sequentially by our ESTM inference operation.

Table 4. The usefulness of *ContextNet* and epipolar transformer. We test models with various settings on SUN3D [40].

Cont.	Trans.	Inference type	Abs	Sq Rel	RMSE	$\sigma < 1.25$
✗	✗	Independent	0.3333	0.0994	0.4897	80.89
✗	✓	Joint	0.3429	0.1291	0.4927	81.36
✗	✓	ESTM	0.3319	0.1073	0.4822	81.43
✓	✗	Independent	0.3220	0.0897	0.4657	82.82
✓	✓	Joint	**0.3133**	**0.0883**	**0.4556**	**83.52**
✓	✓	ESTM	0.3137	0.0884	0.4554	83.43

Table 5. The effect of different size of ESTM memory space. Experiments are done over 7scenes [32].

Memory size	Abs Rel	Abs	Sq Rel	RMSE	$\sigma < 1.25$
1	0.1530	0.2632	0.783	0.3494	79.07
2	0.1465	0.2528	0.0729	0.3382	80.36
3	**0.1460**	**0.2520**	**0.0727**	**0.3376**	**80.44**
4	0.1461	0.2521	0.0728	0.3377	80.44

Table 6. EST transformer vs RNN. We replace the EST transformer with Gated Recurrent Units in our model.

Model	ScanNet		SUN3D	
	Abs	$\sigma < 1.25$	Abs	$\sigma < 1.25$
Ours-RNN	0.1680	92.67	0.3401	82.33
Ours-ESTM	**0.1505**	**93.13**	**0.3137**	**83.52**

and the improvement of *ESTM depth* is trivial.

That is because if the estimated depth is not accurate enough, the joint estimation will propagate wrong information across the multiple depth maps. But ESTM suffers from it less because the errors can be alleviated gradually as more frames are processed sequentially. Overall, the combination of the hybrid regularization network and EST transformer can boost the best performance. Detailed discussions are included in the supplementary materials.

Furthermore, we compare our EST transformer with the recurrent neural network. We replace EST transformer by Gated Recurrent Units (GRU) in our model for comparison. As shown in Table 6, models with EST transformer outperforms the model with RNN.

Efficacy of ContextNet As shown in Table 4, our models with *ContextNet* significantly outperform those without *ContextNet* in all inference types: independently, jointly, and by ESTM operation. In Figure 8, the depth maps estimated by the models with *ContextNet* have smaller errors in pattern-rich regions (e.g. floor) and texture-less regions (e.g. pillar) than those without *ContextNet*. We show that our *ContextNet* is indispensable and provides complementary and important information for cost regularization.

Memory size of ESTM Using the ESTM inference operation, we can retrieve useful information from an arbitrary number of preceding estimations stored in *memory* space to enhance the estimation of the current frame. As shown in Table 5, the quality of estimated depth will be improved gradually with the increase of the number of stored preceding estimations, but the improvement becomes less significant with more than 2 stored estimations. Hence we set the optimal memory size to be 2, as the best trade-off between performance and computational efficiency.

6. Conclusions and limitations

In this paper, we present a novel multi-view depth estimation method for videos. There are two main contributions. First, we propose an Epipolar Spatio-Temporal transformer to associate temporal coherence with multiple estimated depth maps. Second, for better computational efficiency, we design a hybrid network for cost regularization, which disentangles the learning of 2D global context information from that of 3D local matching information. Experiments demonstrate that our method outperforms the state-of-the-art in terms of depth accuracy and speed.

There are several limitations we plan to tackle in the future. First, since our method takes video frames with known camera poses as input, the accuracy of camera poses will influence the performance of our EST transformer. A general framework that jointly estimates camera pose and depth maps may alleviate this problem. Second, our work takes only a short video clip into consideration, rather than estimate depth maps in a global optimization scheme for further potential improvement of temporal coherence.

Acknowledgments

References

[1] Michael Bleyer, Christoph Rhemann, and Carsten Rother. Patchmatch stereo-stereo matching with slanted support windows. In *Bmvc*, volume 11, pages 1–11, 2011. 1

[2] Jia-Ren Chang and Yong-Sheng Chen. Pyramid stereo matching network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5410–5418, 2018. 2, 3

[3] Richard Chen, Faisal Mahmood, Alan Yuille, and Nicholas J Durr. Rethinking monocular depth estimation with adversarial training. *arXiv preprint arXiv:1808.07528*, 2018. 2

[4] Weifeng Chen, Zhao Fu, Dawei Yang, and Jia Deng. Single-image depth perception in the wild. In *Advances in neural information processing systems*, pages 730–738, 2016. 2

[5] Xinjing Cheng, Peng Wang, and Ruigang Yang. Learning depth with convolutional spatial propagation network. *IEEE transactions on pattern analysis and machine intelligence*, 42(10):2361–2379, 2019. 2

[6] Arun CS Kumar, Suchendra M Bhandarkar, and Mukta Prasad. Depthnet: A recurrent neural network architecture for monocular depth prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 283–291, 2018. 1

[7] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5828–5839, 2017. 6, 7

[8] David Eigen and Rob Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *Proceedings of the IEEE international conference on computer vision*, pages 2650–2658, 2015. 2

[9] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *Advances in neural information processing systems*, pages 2366–2374, 2014. 1, 2, 6

[10] Christoph Fehn. Depth-image-based rendering (dibr), compression, and transmission for a new approach on 3d-tv. In *Stereoscopic Displays and Virtual Reality Systems XI*, volume 5291, pages 93–104. International Society for Optics and Photonics, 2004. 1

[11] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. Deep ordinal regression network for monocular depth estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2002–2011, 2018. 2

[12] David Gallup, Jan-Michael Frahm, Philippos Mordohai, Qingxiong Yang, and Marc Pollefeys. Real-time plane-sweeping stereo with multiple sweeping directions. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2007. 2

[13] Peter Henry, Michael Krainin, Evan Herbst, Xiaofeng Ren, and Dieter Fox. Rgb-d mapping: Using depth cameras for dense 3d modeling of indoor environments. In *Experimental robotics*, pages 477–491. Springer, 2014. 1

[14] Heiko Hirschmuller. Stereo processing by semiglobal matching and mutual information. *IEEE Transactions on pattern analysis and machine intelligence*, 30(2):328–341, 2007. 2

[15] Asmaa Hosni, Michael Bleyer, Margrit Gelautz, and Christoph Rhemann. Local stereo matching using geodesic support weights. In *2009 16th IEEE International Conference on Image Processing (ICIP)*, pages 2093–2096. IEEE, 2009. 1

[16] Asmaa Hosni, Christoph Rhemann, Michael Bleyer, Carsten Rother, and Margrit Gelautz. Fast cost-volume filtering for visual correspondence and beyond. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(2):504–511, 2012. 2

[17] Sunghoon Im, Hae-Gon Jeon, Stephen Lin, and In So Kweon. Dpsnet: end-to-end deep plane sweep stereo. *arXiv preprint arXiv:1905.00538*, 2019. 1, 2, 3, 4, 6, 7

[18] J Adam Jones, J Edward Swan, Gurjot Singh, Eric Kolstad, and Stephen R Ellis. The effects of virtual reality, augmented reality, and motion parallax on egocentric depth perception. In *Proceedings of the 5th symposium on Applied perception in graphics and visualization*, pages 9–14, 2008. 1

[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 *Proceedings of the IEEE International Conference on Computer Vision*, pages 66–75, 2017. 2, 5

[20] Uday Kusupati, Shuo Cheng, Rui Chen, and Hao Su. Normal assisted stereo depth estimation. *arXiv preprint arXiv:1911.10444*, 2019. 7

[21] Uday Kusupati, Shuo Cheng, Rui Chen, and Hao Su. Normal assisted stereo depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2189–2199, 2020. 1, 2, 3, 4, 6, 7

[22] Chao Liu, Jinwei Gu, Kihwan Kim, Srinivasa G Narasimhan, and Jan Kautz. Neural rgb (r) d sensing: Depth and uncertainty from a video camera. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10986–10995, 2019. 1, 2, 3, 4, 6, 7

[23] Fayao Liu, Chunhua Shen, and Guosheng Lin. Deep convolutional neural fields for depth estimation from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5162–5170, 2015. 1, 2

[24] Lina Liu, Yiyi Liao, Yue Wang, Andreas Geiger, and Yong Liu. Learning steering kernels for guided depth completion. *IEEE Transactions on Image Processing*, 30:2850–2861, 2021. 2

[25] Lina Liu, Xibin Song, Xiaoyang Lyu, Junwei Diao, Mengmeng Wang, Yong Liu, and Liangjun Zhang. Fcfr-net: Feature fusion based coarse-to-fine residual learning for monocular depth completion. *arXiv preprint arXiv:2012.08270*, 2020. 2

[26] Xiaoxiao Long, Lingjie Liu, Christian Theobalt, and Wenping Wang. Occlusion-aware depth estimation with adaptive normal constraints. *arXiv preprint arXiv:2004.00845*, 2020. 1, 2, 3, 4, 7

[27] Xuan Luo, Jia-Bin Huang, Richard Szeliski, Kevin Matzen, and Johannes Kopf. Consistent video depth estimation. *arXiv preprint arXiv:2004.15021*, 2020. 1, 2

[28] Jiaqi Ma, Zhe Zhao, Xinyang Yi, Jilin Chen, Lichan Hong, and Ed H Chi. Modeling task relationships in multi-task learning with multi-gate mixture-of-experts. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1930–1939, 2018. 2

[29] Vaishakh Patil, Wouter Van Gansbeke, Dengxin Dai, and Luc Van Gool. Don't forget the past: Recurrent depth estimation from monocular video. *arXiv preprint arXiv:2001.02613*, 2020. 1, 2

[30] Xiaojuan Qi, Renjie Liao, Zhengzhe Liu, Raquel Urtasun, and Jiaya Jia. Geonet: Geometric neural network for joint depth and surface normal estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 283–291, 2018. 2

[31] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4104–4113, 2016. 1

[32] Jamie Shotton, Ben Glocker, Christopher Zach, Shahram Izadi, Antonio Criminisi, and Andrew Fitzgibbon. Scene coordinate regression forests for camera relocalization in rgb-d images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2930–2937, 2013. 6, 7, 8

[33] Ayan Sinha, Zak Murez, James Bartolozzi, Vijay Badrinarayanan, and Andrew Rabinovich. Depth estimation by learning triangulation and densification of sparse points for multi-view stereo. *arXiv preprint arXiv:2003.08933*, 2020. 2, 7

[34] Benjamin Ummenhofer, Huizhong Zhou, Jonas Uhrig, Nikolaus Mayer, Eddy Ilg, Alexey Dosovitskiy, and Thomas Brox. Demon: Depth and motion network for learning monocular stereo. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5038–5047, 2017. 6

[35] Kaixuan Wang and Shaojie Shen. Mvdepthnet: real-time multiview depth estimation neural network. In *2018 International Conference on 3D Vision (3DV)*, pages 248–257. IEEE, 2018. 1, 2, 3, 4, 6, 7

[36] Peng Wang, Xiaohui Shen, Zhe Lin, Scott Cohen, Brian Price, and Alan L Yuille. Towards unified depth and semantic prediction from a single image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2800–2809, 2015. 2

[37] Rui Wang, Stephen M Pizer, and Jan-Michael Frahm. Recurrent neural network for (un-) supervised learning of monocular video visual odometry and depth. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5555–5564, 2019. 1, 2

[38] Xin Wang, Fisher Yu, Lisa Dunlap, Yi-An Ma, Ruth Wang, Azalia Mirhoseini, Trevor Darrell, and Joseph E Gonzalez. Deep mixture of experts via shallow embedding. In *Uncertainty in Artificial Intelligence*, pages 552–562. PMLR, 2020. 2

[39] Yan Wang, Wei-Lun Chao, Divyansh Garg, Bharath Hariharan, Mark Campbell, and Kilian Q Weinberger. Pseudo-lidar from visual depth estimation: Bridging the gap in 3d object detection for autonomous driving. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8445–8453, 2019. 1

[40] Jianxiong Xiao, Andrew Owens, and Antonio Torralba. Sun3d: A database of big spaces reconstructed using sfm and object labels. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1625–1632, 2013. 6, 8

[41] Yan Xu, Xinge Zhu, Jianping Shi, Guofeng Zhang, Hujun Bao, and Hongsheng Li. Depth completion from sparse lidar data with depth-normal constraints. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2811–2820, 2019. 2

[42] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. Mvsnet: Depth inference for unstructured multi-view stereo. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 767–783, 2018. 1, 2

[43] Wei Yin, Yifan Liu, Chunhua Shen, and Youliang Yan. Enforcing geometric constraints of virtual normal for depth prediction. In *The IEEE International Conference on Computer Vision (ICCV)*, 2019. 2

[44] Andy Zeng, Shuran Song, Matthias Nießner, Matthew Fisher, Jianxiong Xiao, and Thomas Funkhouser. 3dmatch: Learning local geometric descriptors from rgb-d reconstructions. In *CVPR*, 2017. 7

[45] Guofeng Zhang, Jiaya Jia, Tien-Tsin Wong, and Hujun Bao. Consistent depth maps recovery from a video sequence. *IEEE Transactions on pattern analysis and machine intelligence*, 31(6):974–988, 2009. 1

[46] Haokui Zhang, Chunhua Shen, Ying Li, Yuanzhouhan Cao, Yu Liu, and Youliang Yan. Exploiting temporal consistency for real-time video depth estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1725–1734, 2019. 1, 2

[47] Lianbo Zhang, Shaoli Huang, Wei Liu, and Dacheng Tao. Learning a mixture of granularity-specific experts for fine-grained categorization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8331–8340, 2019. 2