

# IterMVS: Iterative Probability Estimation for Efficient Multi-View Stereo

Fangjinhua Wang<sup>1</sup> Silvano Galliani<sup>2</sup> Christoph Vogel<sup>2</sup> Marc Pollefeys<sup>1,2</sup>

<sup>1</sup>Department of Computer Science, ETH Zurich

<sup>2</sup>Microsoft Mixed Reality & AI Zurich Lab

## Abstract

We present IterMVS, a new data-driven method for high-resolution multi-view stereo. We propose a novel GRU-based estimator that encodes pixel-wise probability distributions of depth in its hidden state. Ingesting multi-scale matching information, our model refines these distributions over multiple iterations and infers depth and confidence. To extract the depth maps, we combine traditional classification and regression in a novel manner. We verify the efficiency and effectiveness of our method on DTU, Tanks&Temples and ETH3D. While being the most efficient method in both memory and run-time, our model achieves competitive performance on DTU and better generalization ability on Tanks&Temples as well as ETH3D than most state-of-the-art methods. Code is available at <https://github.com/FangjinhuaWang/IterMVS>.

## 1. Introduction

Multi-view stereo (MVS) describes the technology to reconstruct dense 3D models of observed scenes from a set of calibrated images. MVS is a fundamental problem of geometric computer vision and a core technique for applications like augmented/virtual reality, autonomous driving and robotics. Albeit being studied extensively for decades, the conditions that occur in real-world application scenarios pose problems such as occlusion, illumination changes, low-textured areas and non-Lambertian surfaces [1, 18, 32, 49] that remain unsolved up to now.

Traditional methods [9, 10, 31, 41] suffer from hand-crafted modeling and matching metrics and struggle under those challenging conditions. In comparison, recent data-driven approaches [11, 39, 47, 48, 50] based on Convolutional Neural Networks (CNN) demonstrate a significantly improved performance on various MVS benchmarks [1, 18].

A popular representative, MVSNet [47], constructs a 3D cost volume that is regularized with a 3D CNN and regresses the depth map from the probability volume. While this methodology achieves impressive performance on benchmarks, it does not scale up well to high-resolution

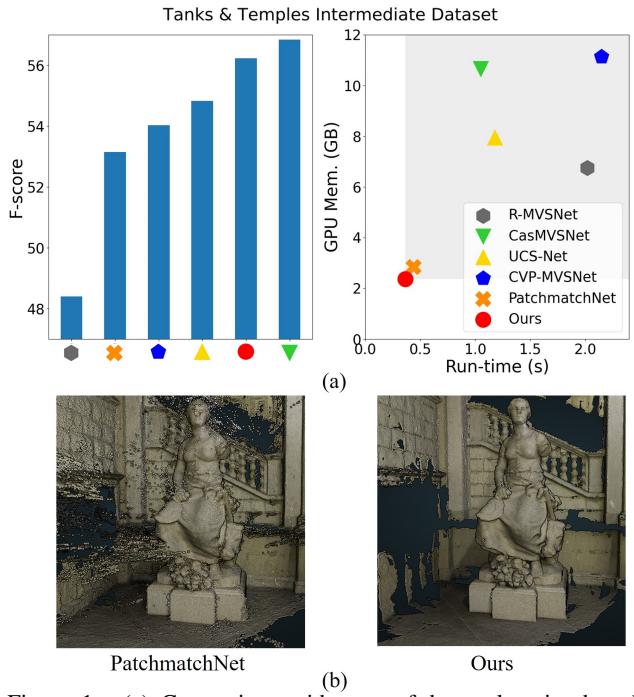


Figure 1. (a) Comparison with state-of-the-art learning-based MVS methods [6, 11, 39, 46, 48] on Tanks & Temples [18] (all methods are trained on DTU [1] only). **Left:** F-score ( $\uparrow$ ). **Right:** GPU memory and run-time (image size  $1920 \times 1024$ , 7 views). (b) Qualitative comparison with PatchmatchNet on ETH3D [32]. Our reconstruction contains significantly less noise.

images or large-scale scenes, because the 3D CNN is memory and run-time consuming. However, low run-time and power consumption are the key in most industrial applications and resource friendly methodologies become more important. Aiming to improve the efficiency, recent variants of MVSNet [47] are proposed, which can be mainly divided into two categories: recurrent methods [45, 48] and multi-stage methods [6, 11, 39, 46]. Several recurrent methods [45, 48] can relax the memory consumption by regularizing the cost volume sequentially with GRU [7] or convolutional LSTM [40], however, at the cost of an increase in run-time. In contrast, multi-stage methods [6, 11, 46] utilize cascade cost volumes and estimate the depth map from coarse to fine. While this approach can lead to high efficiency in both memory and run-time, a reduced search

range at finer stages implies a limitation in recovering from errors induced at coarse resolutions [34].

A current method that unites a competitive performance with highest efficiency in memory and run-time among all learning-based methods is PatchmatchNet [39]. Based on traditional PatchMatch [3, 10], PatchmatchNet combines learned adaptive propagation and evaluation modules with a cascaded structure. While sharing the common limitation of coarse-to-fine methods, the generalization ability of PatchmatchNet appears limited when compared to other multi-stage methods [6, 11, 46].

In this work, we propose IterMVS, a novel GRU-based iterative method aimed at further improving efficiency as well as performance for high-resolution MVS.

**Contributions:** **(i)** We propose a novel and lightweight GRU-based probability estimator that encodes the per-pixel probability distribution of depth in its hidden state. This compressed representation does not require to keep the probability volume in memory the whole time. In each iteration, multi-scale matching information is injected to update the pixel-wise depth distribution. Compared to coarse-to-fine methods, the GRU-based probability estimator always operates at the same resolution, utilizes a large search range and keeps track of the distribution over the full depth range. **(ii)** We propose a simple, yet effective depth estimation strategy that combines both classification and regression, which is robust to multi-modal distributions but also achieves sub-pixel precision. **(iii)** We verify the effectiveness of our method on various MVS datasets, *e.g.* DTU [1], Tanks & Temples [18] and ETH3D [32]. The results demonstrate that IterMVS achieves very competitive performance, while showing highest efficiency in both memory and run-time among all the learning-based methods, Fig. 1. Compared with PatchmatchNet [39], IterMVS is more efficient in both memory and run-time, achieves comparable performance on DTU [1] and demonstrates much better generalization ability on Tanks & Temples [18] and ETH3D [32].

## 2. Related Work

**Traditional MVS.** Based on the scene representations, traditional MVS methods can be divided into three main categories: volumetric, point cloud based and depth map based. Volumetric methods [19, 21, 33, 38] discretize 3D space into voxels and label each as inside or outside of the true surface. Operating in scene space usually comes at the price of large memory and run-time consumption, limiting applications to scenes of smaller scale. Point cloud based methods [9, 23] operate directly on 3D points and often employ propagation to gradually densify the reconstruction. By decoupling the problem into depth map estimation and fusion, depth map based methods [10, 31, 41, 43] are more con-

cise and flexible. Galliani *et al.* [10] propose Gipuma, a multi-view extension of Patchmatch stereo, which uses a red-black checkerboard pattern to parallelize propagation. In COLMAP, Schönberger *et al.* [31] jointly estimate pixel-wise view selection, depth map and surface normal. Although traditional depth map based methods can achieve impressive results, hand-crafted models and features limit the performance under challenging conditions.

**Data-driven MVS.** Recently, data-driven methods dominate the research for MVS. Several volumetric methods [14, 15] first compute a cost volume from multiple images and infer surface voxels after cost volume regularization with a 3D CNN. However, similar to traditional volumetric methods, they are restricted to small-scale reconstructions. More common are depth map based methods [5, 27, 42, 47] that often operate in similar fashion. MVSNet [47] can be seen as a blueprint. It computes an initial cost volume from the features that is regularized with a 3D CNN and regresses the depth map from the probability volume. The high memory consumption of 3D CNNs often limits these methods to down-sampled cost volumes and depth maps. Recently, several variants based on MVSNet [47] were proposed that aim to reduce memory and run-time consumption. The two main ideas involve recurrent [45, 48] and multi-stage methods [6, 11, 39, 44, 46, 51]. R-MVSNet [48] sequentially regularizes 2D slices of the cost volume with a GRU [7].  $D^2$ HC-RMVSNet [45] augments R-MVSNet with a complex convolutional LSTM [40]. The main drawback of these recurrent methods is the high run-time. In contrast, multi-stage methods [6, 11, 44, 46, 51] achieve efficiency in both memory and run-time. They operate on cascade cost volumes and estimate the depth map in a coarse-to-fine manner. First, a low resolution depth map is computed utilizing a large but coarse sampling interval. After upsampling, the estimation is refined at a higher sampling rate but at a smaller interval and search range. PatchmatchNet [39] further proposes an adaptive procedure based on PatchMatch [2, 3] that achieves superior efficiency among all learning-based methods. Despite their impressive performance, coarse-to-fine methods have difficulty to recover from errors introduced at coarse resolutions [34], where the sampling interval is large but the sampling frequency low. In contrast, we estimate the depth map at a relatively high resolution and generate hypotheses in a fixed large search range in each GRU iteration. Further, we let the hidden state of our GRU encode the probability distribution for the whole depth range.

**Iterative Update.** Recently, RAFT [34] proposes to estimate optical flow by iteratively updating a motion field through a GRU, which emulates first-order optimization. The idea is further adopted in stereo [26], scene flow [35] and SfM [12]. In our work, we let the GRU model a probability distribution per pixel from which we predict the depth

map. The hidden state is updated in each iteration to more accurately model the pixel-wise probability distribution.

### 3. Method

In this section, we introduce the detailed structure of Iter-MVS, illustrated in Fig. 2. It consists of a multi-scale feature extractor, an iterative GRU-based probability estimator that models a probability distribution of the depth at each pixel, and a spatial upsampling module.

#### 3.1. Multi-scale Feature Extractor

Given  $N$  input images of size  $W \times H$ , we use  $\mathbf{I}_0$  and  $\{\mathbf{I}_i\}_{i=1}^{N-1}$  to denote reference and source images respectively. Similar to [11, 39], we extract multi-scale features from the images with a Feature Pyramid Network (FPN) [25]. We attain features at 3 scale levels, and denote the feature of the  $i$ -th image at level  $l$  by  $\mathbf{F}_{i,l}$ . The features  $\mathbf{F}_{i,l}$  are stored at  $1/2^l$  resolution and possess  $C = 16, 32, 64$  channels at level  $l = 1, 2, 3$  respectively. Although we refrain from using an explicit coarse-to-fine structure, e.g. [11, 39], we can include multi-scale context information by using our multi-scale features for matching similarity computation in each iteration of the GRU. This improves performance as shown in Table 4.

#### 3.2. GRU-based Probability Estimator

Our core module, the GRU-based probability estimator, models the per-pixel probability distribution of depth with a hidden state of 32 dimensions. The GRU operates at  $1/4$  resolution, outputs a depth map  $\mathbf{D} \in \mathbb{R}^{W/4 \times H/4}$  and is unrolled for  $K$  iterations.

**Differentiable Warping.** Following most learning-based MVS methods [11, 39, 47, 48], we warp the source features into front-to-parallel planes w.r.t. the reference view at the given depth hypotheses. Specifically, for a pixel  $\mathbf{p}$  in the reference view and the  $j$ -th depth hypothesis  $d_j := d_j(\mathbf{p})$ , with known intrinsic  $\{\mathbf{K}_i\}_{i=0}^{N-1}$  and relative transformations  $\{[\mathbf{R}_{0,i} | \mathbf{t}_{0,i}]\}_{i=1}^{N-1}$  between reference view 0 and source view  $i$ , we can compute the corresponding pixel  $\mathbf{p}_{i,j} := \mathbf{p}_i(d_j)$  in the source view as:

$$\mathbf{p}_{i,j} = \mathbf{K}_i \cdot (\mathbf{R}_{0,i} \cdot (\mathbf{K}_0^{-1} \cdot \mathbf{p} \cdot d_j) + \mathbf{t}_{0,i}). \quad (1)$$

After de-homogenization, we obtain the warped feature  $\mathbf{F}_i(\mathbf{p}_{i,j})$  via bilinear interpolation.

**2-View Matching Similarity Computation.** Because the principle is same for all feature levels, we omit the subindices denoting the level. Given reference and warped  $i$ -th source feature,  $\mathbf{F}_0(\mathbf{p}), \mathbf{F}_i(\mathbf{p}_{i,j}) \in \mathbb{R}^C$ , we first use group-wise correlation [13, 39, 42] to reduce the dimension. By dividing the feature channels evenly into  $G = 8$  groups,

the  $g$ -th group similarity  $\mathbf{s}_i(\mathbf{p}, j)^g \in \mathbb{R}$  can be computed as:

$$\mathbf{s}_i(\mathbf{p}, j)^g = \frac{1}{C/G} \langle \mathbf{F}_0(\mathbf{p})^g, \mathbf{F}_i(\mathbf{p}_{i,j})^g \rangle, \quad (2)$$

where  $\langle \cdot, \cdot \rangle$  denotes the dot product. This results in the similarity  $\mathbf{s}_i \in \mathbb{R}^{W \times H \times D \times G}$ , where  $D$  is the number of depth hypotheses per pixel.

**Initialization.** To initialize the hidden state  $h \in \mathbb{R}^{W/4 \times H/4 \times 32}$  of the GRU, we only utilize the features on level 3 and then upsample the results to further reduce the computation. Per pixel at  $1/8$  resolution and within a pre-defined depth range  $[d_{min}, d_{max}]$ , we place  $D_1$  equidistant depth hypotheses in the *inverse* depth range. Sampling hypotheses uniformly in image space is more suitable for large-scale scenes [42, 48]. After differentiable warping, we can compute  $(N - 1)$  2-view matching similarities  $\mathbf{s}_i$  ( $i = 1, \dots, N - 1$ ) following Eq. 2.

For each source view, we further estimate a pixel-wise view weight [39, 44] that provides visibility information and increased robustness when integrating the information from all the source views. A lightweight 2D CNN is applied on the image space of  $\mathbf{s}_i$  to aggregate local information and reduce the feature channels from  $G$  to 1. Applying a *softmax* non-linearity along the depth dimension produces  $\mathbf{P}_i \in \mathbb{R}^{W/8 \times H/8 \times D_1}$ . For pixel  $\mathbf{p}$ , the view weight of source view  $i$  can then be computed as:

$$\mathbf{w}_i(\mathbf{p}) = \max \{ \mathbf{P}_i(\mathbf{p}, j) | j = 0, 1, \dots, D_1 - 1 \}. \quad (3)$$

Finally, the integrated matching similarity  $\mathbf{S}_{\text{initial}}(\mathbf{p}, j)$  for pixel  $\mathbf{p}$  and depth hypothesis  $d_j$  is given by:

$$\mathbf{S}_{\text{initial}}(\mathbf{p}, j) = \frac{\sum_{i=1}^{N-1} \mathbf{w}_i(\mathbf{p}) \cdot \mathbf{s}_i(\mathbf{p}, j)}{\sum_{i=1}^{N-1} \mathbf{w}_i(\mathbf{p})}. \quad (4)$$

To also consider spatial correlation of depth maps, we aggregate similarity information from neighboring pixels by applying a 2D U-Net [30] on  $\mathbf{S}_{\text{initial}} \in \mathbb{R}^{W/8 \times H/8 \times D_1 \times G}$ . This makes the matching more robust [11, 39, 47]. The final convolution layer outputs a 1-channel similarity  $\bar{\mathbf{S}}_{\text{initial}} \in \mathbb{R}^{W/8 \times H/8 \times D_1}$ . Then a 2D CNN,  $2 \times$  bilinear upsampling and the tanh nonlinearity are applied sequentially on  $\bar{\mathbf{S}}_{\text{initial}}$  to produce the initial hidden state  $h_0$ .

**Iterative Update.** For each pixel  $\mathbf{p}$  at iteration  $k$  and level  $l$ , we generate  $N_l$  new depth hypotheses sampled uniformly within an interval of size  $2R_l$  in the *normalized inverse* range  $[(\frac{1}{\mathbf{D}_{k-1}(\mathbf{p})} - \frac{1}{d_{max}})/(\frac{1}{d_{min}} - \frac{1}{d_{max}}) - R_l, (\frac{1}{\mathbf{D}_{k-1}(\mathbf{p})} - \frac{1}{d_{max}})/(\frac{1}{d_{min}} - \frac{1}{d_{max}}) + R_l]$ , centered at the previous depth estimate  $\mathbf{D}_{k-1} \in \mathbb{R}^{W/4 \times H/4}$ . Similar to [26], we further ensure that  $R_{l-1} < R_l$  to capitalize on the more high-frequent information of the higher resolution features.

We then compute matching similarities  $\bar{\mathbf{S}}_k$  from features at all 3 levels to include multi-scale information (see supplementary). We use the  $2 \times$  upsampled pixel-wise view weights to integrate matching similarities from source

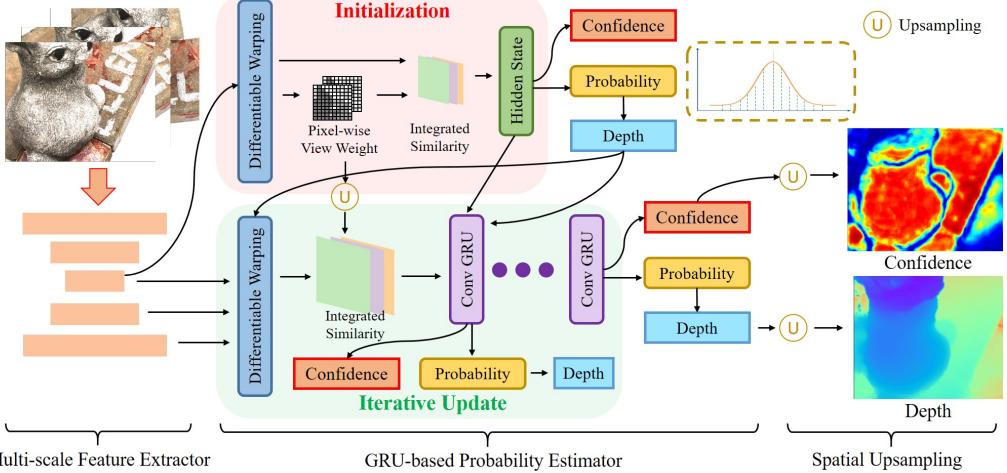


Figure 2. Detailed structure of IterMVS. With multi-scale features extracted from the feature extractor, the GRU-based probability estimator computes matching similarity for depth hypotheses on different scale levels. The GRU takes the matching information as input to iteratively update the hidden state, which encodes the per-pixel probability distributions of depth. The depth map and confidence are estimated from the hidden state. The spatial upsampling module takes the quarter-resolution depth and confidence output by the GRU to full resolution.

views and then pass them through a level-wise 2D U-Net to aggregate neighborhood information and reduce the number of channels to 1. The final matching similarities for 3 levels are concatenated into  $\bar{\mathbf{S}}_k \in \mathbb{R}^{W/4 \times H/4 \times (N_1+N_2+N_3)}$ . Concatenating  $\mathbf{D}_{k-1}$  and  $\bar{\mathbf{S}}_k$  to  $x_k$ , we use our convolutional GRU to update the hidden state, which further allows the network to propagate information from spatial neighbors:

$$\begin{aligned} z_k &= \sigma(\text{Conv}([h_{k-1}, x_k], W_z)), \\ r_k &= \sigma(\text{Conv}([h_{k-1}, x_k], W_r)), \\ \tilde{h}_k &= \tanh(\text{Conv}([r_k \odot h_{k-1}, x_k], W_h)), \\ h_k &= (1 - z_k) \odot h_{k-1} + z_k \odot \tilde{h}_k, \end{aligned} \quad (5)$$

where  $\sigma(\cdot)$  denotes the *sigmoid* nonlinearity,  $\odot$  denotes the Hadamard Product. With the multi-scale matching information injected in each iteration, the hidden state can encode the pixel-wise depth distribution more accurately (Table 5).

**Depth Prediction.** The depth map at iteration  $k$  is predicted from the hidden state  $h_k$  ( $k = 0, \dots, K$ ). For simplicity, we omit index  $k$  here. Recall that we let the hidden state encode the pixel-wise probability distribution of the depth. We extract those probabilities  $\mathbf{P} \in \mathbb{R}^{W/4 \times H/4 \times D_2}$ , for depths sampled uniformly at  $D_2$  locations over the *inverse* depth range, by applying 2D CNN on the hidden state, followed by a *softmax* nonlinearity along the depth dimension.

Usual strategies to predict a depth value from such sampled distributions are to take the *argmax* [45, 48] or the *soft argmax* [11, 16, 39, 47]. The former corresponds to measuring the Kullback-Leibler divergence between a one-hot encoding of the ground truth and  $\mathbf{P}$ , but cannot deliver solutions beyond the discretization level (e.g. ‘sub-pixel’ solutions). The latter corresponds to measuring the distance of the expectation of  $\mathbf{P}$  to the ground truth depth. While the expectation can take any continuous value, the measure

cannot handle multiple modes in  $\mathbf{P}$  and strictly prefers unimodal distributions. At this point we propose a new hybrid strategy that combines classification and regression. It is robust to multi-modal distributions but also achieves ‘sub-pixel’ precision that is not limited by the sampling resolution. Specifically, we find the index  $\mathbf{X}(\mathbf{p})$  with the highest probability for pixel  $\mathbf{p}$  from probability  $\mathbf{P}$ :

$$\mathbf{X}(\mathbf{p}) = \arg \max_j \mathbf{P}(\mathbf{p}, j). \quad (6)$$

With a radius  $r$ , we take the expectation in the *local inverse* range to compute the depth estimate  $\mathbf{D}(\mathbf{p})$ :

$$\mathbf{D}(\mathbf{p}) = \left( \frac{1}{\sum_{j=\mathbf{X}(\mathbf{p})-r}^{\mathbf{X}(\mathbf{p})+r} \mathbf{P}(\mathbf{p}, j)} \sum_{j=\mathbf{X}(\mathbf{p})-r}^{\mathbf{X}(\mathbf{p})+r} \frac{1}{d_j} \cdot \mathbf{P}(\mathbf{p}, j) \right)^{-1}, \quad (7)$$

where  $d_j$  is the  $j$ -th depth value.

**Confidence Estimation.** Because the hidden state of GRU models the pixel-wise depth probability distributions, we can estimate the uncertainty. We apply a 2D CNN followed by a *sigmoid* on the hidden state  $h$  to predict the confidence  $\mathbf{C} \in \mathbb{R}^{W/4 \times H/4}$ . The confidence is defined as the likelihood that the ground truth depth is located within a small range near the estimation [8, 29, 37, 47] and, thus, is used to filter outliers during the reconstruction [47] with a threshold  $\tau$ .

### 3.3. Spatial Upsampling

We upsample the depth map  $\mathbf{D}_K$  output by the GRU probability estimator at the final iteration from  $1/4$  to full resolution  $\mathbf{D}_{\text{upsample}}$ . The procedure is guided by image features: Given  $\mathbf{F}_{0,2}$ , the features of the reference, a 2D CNN predicts a mask  $\mathbf{M} \in \mathbb{R}^{W/4 \times H/4 \times (4 \times 4 \times 9)}$ , where the last dimension represents the weights for the 9 nearest neighbors at coarse resolution. The depth at full resolution is then computed as the normalized (using *softmax*) weighted com-

bination of those neighbors [34] (see supplementary). The confidence map is bilinearly upsampled to full resolution.

### 3.4. Loss Function

The loss function considers the depth and confidence estimated from the GRU, and the final upsampled depth. To facilitate convergence, we further utilize the coarse depth  $\mathbf{D}_{\text{initial}} \in \mathbb{R}^{W/8 \times H/8}$  predicted from the similarity  $\bar{\mathbf{S}}_{\text{initial}}$  in the Initialization phase. It is found as:

$$\mathbf{D}_{\text{initial}}(\mathbf{p}) = \left( \sum_{j=1}^{D_1} \frac{1}{d_j} \cdot \mathbf{P}_{\text{initial}}(\mathbf{p}, j) \right)^{-1}, \quad (8)$$

where the probability  $\mathbf{P}_{\text{initial}}$  is generated from  $\bar{\mathbf{S}}_{\text{initial}}$  by applying *softmax* along the depth dimension. Then we bilinearly upsample  $\mathbf{D}_{\text{initial}}$  to 1/4 resolution. The losses are based on  $\mathbf{D}'$  that is converted from a depth map  $\mathbf{D}$ :

$$\mathbf{D}'(\mathbf{p}) = \left( \frac{1}{\mathbf{D}(\mathbf{p})} - \frac{1}{d_{\max}} \right) / \left( \frac{1}{d_{\min}} - \frac{1}{d_{\max}} \right). \quad (9)$$

We can summarize our loss function as follows:

$$\begin{aligned} L_{\text{full}} = & \alpha^{K+1} L_{\text{initial}} + L_{\text{upsample}} \\ & + \sum_{k=0}^K \alpha^{K-k} (L_{\text{class},k} + L_{\text{regress},k} + L_{\text{conf},k}) \end{aligned} \quad (10)$$

and only consider  $N_{\text{valid}}$  pixels with valid ground truth depth. The loss function is composed of five component-level losses: classification loss  $L_{\text{class},k}$ , regression losses  $L_{\text{initial}}, L_{\text{regress},k}, L_{\text{upsample}}$  and confidence loss  $L_{\text{conf},k}$ ,

$$\begin{aligned} L_{\text{initial}} &= \beta \cdot \|\mathbf{D}'_{\text{initial}} - \mathbf{D}'_{\text{gt},2}\|_1, \\ L_{\text{class},k} &= \frac{1}{N_{\text{valid}}} \sum_{\mathbf{p}} \sum_{j=1}^{D_2} -\mathbf{Q}(\mathbf{p}, j) \log(\mathbf{P}_k(\mathbf{p}, j)), \\ L_{\text{regress},k} &= \beta \cdot \|\mathbb{I}\{|\mathbf{X}_{\text{gt},2} - \mathbf{X}_k| \leq r\} \cdot (\mathbf{D}'_k - \mathbf{D}'_{\text{gt},2})\|_1, \\ L_{\text{conf},k} &= -\frac{1}{N_{\text{valid}}} \sum_{\mathbf{p}} [\mathbf{C}_k^*(p) \log \mathbf{C}_k(p) \\ &+ (1 - \mathbf{C}_k^*(p)) \log(1 - \mathbf{C}_k(p))], \\ L_{\text{upsample}} &= \beta \cdot \|\mathbf{D}'_{\text{upsample}} - \mathbf{D}'_{\text{gt},0}\|_1, \end{aligned} \quad (11)$$

where  $\|\cdot\|_1$  is the  $l_1$  loss,  $\mathbf{D}'_{\text{gt},l}$  is computed from the ground truth depth at level  $l$  with Eq. 9,  $\mathbb{I}\{\cdot\}$  denotes the indicator function and  $\alpha = 0.8$  and  $\beta = D_2$  are weights. For the confidence loss  $L_{\text{conf},k}$ , we set  $\mathbf{C}_k^*$  to:

$$\mathbf{C}_k^*(\mathbf{p}) = \mathbb{I}\{|\mathbf{D}'_{\text{gt},2}(\mathbf{p}) - \mathbf{D}'_k(\mathbf{p})| \leq \gamma\}, \quad (12)$$

where  $\gamma$  is empirically set to 0.002. To train the classification, we define  $\mathbf{Q}$  as the one-hot encoding generated by the ground truth depth  $\mathbf{D}_{\text{gt},2}$ , peaking at the nearest discrete location and use cross-entropy loss [48] for  $L_{\text{class}}$ . The subsequent regression can only change the initial estimate of the

classification by  $r$  in any direction. Hence, the loss  $L_{\text{regress}}$  only considers pixels whose ground truth classification index,  $\mathbf{X}_{\text{gt},2}(\mathbf{p})$ , falls within a radius  $r$  of the estimated index  $\mathbf{X}_k(\mathbf{p})$ . For the 1<sup>st</sup> epoch of the training, we exclude  $L_{\text{regress}}$  and  $L_{\text{conf}}$  from the loss to warm up the classification.

## 4. Experiments

### 4.1. Datasets

The DTU dataset [1] is an indoor multi-view stereo dataset with 124 different scenes and 7 different lighting conditions. We use the training, testing and validation split introduced in [14]. BlendedMVS [49] is a large-scale dataset, which provides over 17k high-quality training samples of various scenes. Tanks & Temples [18] is a large-scale outdoor dataset consisting of complex environments that are captured under real-world conditions. The ETH3D benchmark [32] consists of calibrated high-resolution images of real-world scenes with strong viewpoint variations.

### 4.2. Implementation Details

Implemented with PyTorch, two models are trained on DTU [1] (**Ours**) and Large-Scale BlendedMVS [49] (**Ours-LS**) respectively. For DTU, we use an image resolution of  $640 \times 512$  and the number of input images to  $N = 5$ . Observing that a fixed depth range (e.g. MVSNet [47]) filters out some of the ground truth signals on DTU, we propose to estimate a depth range per-view from the sparse ground truth point clouds for both training and evaluation. This is comparable to the treatment of real-world scenes [18, 32] in MVS, where the per-view depth range is usually estimated from a sparse SfM reconstruction, e.g. [28, 31]. For BlendedMVS, we use an image resolution of  $768 \times 576$  and  $N = 5$  input images. Here, the depth range is provided by the dataset. To improve the robustness during training, we follow [39] to randomly select the source views and also scale the scene within the range of [0.8, 1.25].

We use  $D_1 = 32$  during initialization and perform  $K = 4$  update iterations. Here, we set  $R_1 = 2^{-7}, R_2 = 2^{-5}, R_3 = 2^{-3}$  and  $N_1 = 4, N_2 = 4, N_3 = 2$ . When predicting the depth, we use  $D_2 = 256$  and  $r = 4$ . We train for 16 epochs using Adam [17] ( $\beta_1 = 0.9, \beta_2 = 0.999$ ). The learning rate is initially set to 0.001, and halved after 4, 8 and 12 epochs. We set the batch size to 4 and 2 for the models trained on DTU and BlendedMVS. The models are trained on a single Nvidia RTX 2080Ti GPU. After depth estimation, we perform confidence filtering (threshold  $\tau$  is set to 0.3) and geometric consistency filtering to remove outliers and then reconstruct the point clouds (see supplementary).

### 4.3. Benchmark Performance

**Evaluation on DTU.** We use the DTU-only trained model for evaluation. Image size, number of views  $N$  and number

Methods	Acc.(mm) ↓	Comp.(mm) ↓	Overall(mm) ↓
Camp [4]	0.835	0.554	0.695
Furu [9]	0.613	0.941	0.777
Tola [36]	0.342	1.190	0.766
Gipuma [10]	<b>0.283</b>	0.873	0.578
MVSNet [47]	0.396	0.527	0.462
R-MVSNet [48]	0.383	0.452	0.417
CIDER [42]	0.417	0.437	0.427
P-MVSNet [27]	0.406	0.434	0.420
Point-MVSNet [5]	0.342	0.411	0.376
Fast-MVSNet [50]	0.336	0.403	0.370
CasMVSNet [11]	0.325	0.385	0.355
UCS-Net [6]	0.338	0.349	<b>0.344</b>
CVP-MVSNet [46]	0.296	0.406	0.351
Vis-MVSNet [51]	0.369	0.361	0.365
$D^2$ HC-RMVSNet [45]	0.395	0.378	0.386
PatchmatchNet [39]	0.427	<b>0.277</b>	0.352
Ours	0.373	0.354	0.363

Table 1. Quantitative results of different methods on DTU [1]. Methods are separated into two categories (from top to bottom): traditional and trained on DTU [1].

of iterations  $K$  are set to  $1600 \times 1152$ , 5 and 4. In our quantitative evaluation, we use the metrics provided by DTU [1] and calculate *accuracy*, *completeness* and *overall quality* as the mean of both metrics [47]. Shown in Table 1, our method achieves competitive performance in *overall quality*. Compared to the highly efficient PatchmatchNet [39], our method achieves higher *accuracy*.

**Memory and Run-time Comparison.** To demonstrate the high efficiency of our method, we compare to the learning-based multi-stage methods [6, 11, 39] that are tailored for efficiency in both memory and run-time. The experiments use the same environment as before and are summarized in Fig. 3. For our method, both memory consumption and run-time increase the slowest with the input size. For example, at a resolution of  $1600 \times 1152$ , memory consumption and run-time are reduced by 78.0% and 68.7% compared to CasMVSNet [11], by 71.6% and 73.5% compared to UCS-Net [6] and by 15.2% and 16.1% compared to PatchmatchNet [39], the most efficient learning-based MVS method so far. Combined with the quantitative analysis in Table 1, we conclude that our method demonstrates higher efficiency in both memory and run-time than most state-of-the-art learning-based methods, at a competitive performance.

**Evaluation on Tanks & Temples.** We set the input image size to  $1920 \times 1024$ , the number of views,  $N$ , to 7 and the number of iterations,  $K$ , to 4. The camera parameters and depth ranges are estimated with OpenMVG [28]. During evaluation, the GPU memory consumption and run-time for estimating each depth map are 2426 MB and 0.405 s respectively. Among the methods trained on DTU [1], as shown in Table 2, the *F-score* of our method on the intermediate dataset is similar to CasMVSNet [11] and much better than many other multi-stage methods [6, 39, 46]. On the complex advanced dataset, our method performs best in *Recall* and *F-score* among all the methods. Compared with Patch-

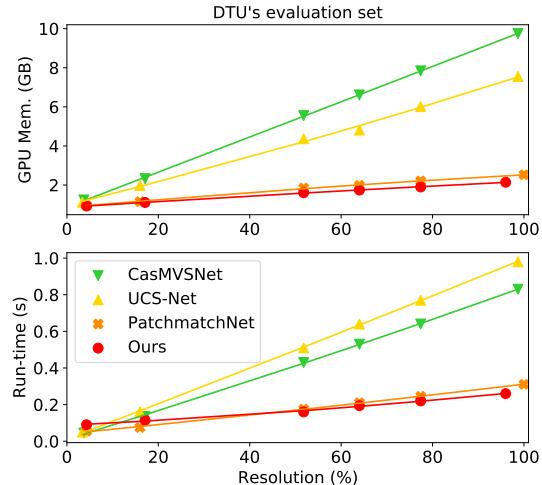


Figure 3. Relating GPU memory and run-time to the image resolution on DTU's evaluation set [1]. The original image resolution is  $1600 \times 1200$  (100%).

matchNet [39], our method performs much better on all the metrics of both datasets. Training on the large-scale BlendedMVS [49] dataset appears to improve the generalization ability of our method on both datasets compared with the model trained on DTU only. On the advanced dataset, our method performs better in *F-score* than Vis-MVSNet [51], which is a state-of-the-art multi-stage method. Overall, operating at a notably low memory and run-time, our method demonstrates very competitive generalization performance compared to most learning-based methods.

**Evaluation on ETH3D Benchmark.** We set the input image size to  $1920 \times 1280$ , the number of views  $N$  to 7 and the number of iterations  $K$  to 4. The camera parameters and depth ranges are estimated with COLMAP [31]. During evaluation, the GPU memory consumption and run-time for estimating each depth map are 2898 MB and 0.410 s respectively. The results are summarized in Table 3. When trained on DTU [1], our method performs much better in *accuracy* than PVSNet [44] and PatchmatchNet [39]. On the training dataset, our method achieves comparable performance in *F<sub>1</sub>-score* as PVSNet [44] and COLMAP [31]. On the test dataset, the performance of our method in *F<sub>1</sub>-score* is better than PatchmatchNet [39], PVSNet [44] and COLMAP [31], which is competitive to ACMH [41]. When trained on BlendedMVS [49], the performance of our method improves significantly in all metrics on both datasets compared with the model solely trained on DTU. Our method performs best in *F<sub>1</sub>-score* on both datasets among all competitors. This analysis further demonstrates the effectiveness, efficiency and generalization capabilities of our method.

#### 4.4. Ablation Study

We conduct an ablation study of our model trained on DTU [1] to analyze the effectiveness of its components. Re-

Methods	Intermediate Dataset			Advanced Dataset		
	Precision $\uparrow$	Recall $\uparrow$	F-score $\uparrow$	Precision $\uparrow$	Recall $\uparrow$	F-score $\uparrow$
COLMAP [31]	43.16	44.48	42.14	<b>33.65</b>	23.96	27.24
MVSNet [47]	40.23	49.70	43.48	-	-	-
R-MVSNet [48]	43.74	57.60	48.40	31.47	22.05	24.91
CIDER [42]	42.79	55.21	46.76	26.64	21.27	23.12
P-MVSNet [27]	49.93	63.82	55.62	-	-	-
Point-MVSNet [5]	41.27	60.13	48.27	-	-	-
Fast-MVSNet [50]	39.98	59.89	47.39	-	-	-
CasMVSNet [11]	47.62	74.01	56.84	29.68	35.24	31.12
UCS-Net [6]	46.66	70.34	54.83	-	-	-
CVP-MVSNet [46]	51.41	60.19	54.03	-	-	-
$D^2$ HC-RMVSNet [45]	49.88	74.08	59.20	-	-	-
PatchmatchNet [39]	43.64	69.37	53.15	27.27	41.66	32.31
Ours	46.82	73.50	56.22	28.04	42.60	33.24
PatchMatch-RL [22]	45.91	62.30	51.81	30.57	36.73	31.78
Vis-MVSNet [51]	<b>54.44</b>	70.48	<b>60.03</b>	30.16	41.42	33.78
Ours-LS	47.53	<b>74.69</b>	56.94	28.70	<b>44.19</b>	<b>34.17</b>

Table 2. Results of different methods on Tanks & Temples [18]. Methods are separated into three categories (from top to bottom): traditional, trained on DTU [1] and trained on BlendedMVS [49].

Methods	Training Dataset			Test Dataset		
	Accuracy $\uparrow$	Completeness $\uparrow$	$F_1$ -score $\uparrow$	Accuracy $\uparrow$	Completeness $\uparrow$	$F_1$ -score $\uparrow$
Gipuma [10]	84.44	34.91	36.38	86.47	24.91	45.18
PMVS [9]	90.23	32.08	46.06	90.08	31.84	44.16
COLMAP [31]	<b>91.85</b>	55.13	67.66	<b>91.97</b>	62.98	73.01
ACMH [41]	88.94	61.59	70.71	89.34	68.62	75.89
PVSNet [44]	67.84	69.66	67.48	66.41	80.05	72.08
PatchmatchNet [39]	64.81	65.43	64.21	69.71	<b>77.46</b>	73.12
Ours	73.62	61.87	66.36	76.91	72.65	74.29
PatchMatch-RL [22]	76.05	62.22	67.78	74.48	72.06	72.38
Ours-LS	79.79	<b>66.08</b>	<b>71.69</b>	84.73	76.49	<b>80.06</b>

Table 3. Results of different methods on ETH3D [32] (evaluation threshold is 2cm). Methods are separated into three categories (from top to bottom): traditional, trained on DTU [1] and trained on BlendedMVS [49].

sults of the first six experiments are summarized in Table 4.

**Depth Prediction.** For the depth estimated from GRU, instead of using both classification and regression, we also try both *argmax* (classification only) and *soft argmax* (regression only), *i.e.* taking the expectation on all the  $D_2$  depth samples. Our hybrid strategy performs better on both DTU [1] and ETH3D [32]. Apparently, the gap on the unseen ETH3D data widens when (also) utilizing the classification loss, whereas classification only leads to inferior accuracy on DTU. We conjecture that *soft argmax* (regression) only training could be more prone to overfitting, implied by the tendency to force single peaked distributions.

**Confidence.** We compare our idea of explicitly learning the confidence from the hidden state to a commonly used strategy that defines the confidence as the sum of 4 probability samples near the estimate [47] from the probability volume  $\mathbf{P}$ . Following MVSNet [47], we use a threshold of  $\tau = 0.8$  for 3D reconstruction. Our strategy performs slightly better on DTU [1] and ETH3D [32]. A visualization in Fig. 4 shows that our learned confidence can deliver more accurate predictions at object boundaries. When taking confidence as a sum of probabilities, the un-matchable uniformly colored background pixel near the boundary receives a confidence as high as foreground pixels, due to an

over-smoothing effect induced by averaging probabilities.

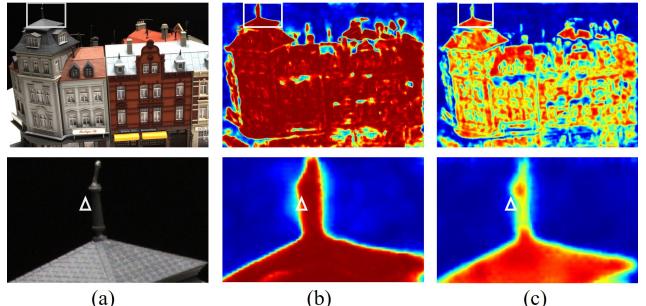


Figure 4. Visualization of confidence. (a) Reference image (chosen background pixel near the object boundary is highlighted). (b) Confidence as sum of probabilities. (c) Our learned confidence. Note that our method correctly outputs a lower confidence for the background pixel compared to the foreground pixels.

**Scale of Feature.** Recall that we utilize multi-scale features for similarity computation throughout the Iterative Update. Here, we also try to only use level 2 features (we still employ level 3 features in the Initialization phase to reduce computation). Our analysis shows that our model benefits from using multi-scale context information on both DTU [1] and ETH3D [32].

**Depth Upsampling.** We compare learned upsampling to

Experiments	Methods	DTU dataset			ETH3D Training
		Acc.(mm) ↓	Comp.(mm) ↓	Overall(mm) ↓	$F_1$ -score ↑
Depth Prediction	Classification+Regression	0.373	0.354	0.363	66.36
	Classification Only	0.448	0.408	0.428	63.34
	Regression Only	0.400	0.415	0.408	50.19
Confidence	Learned	0.373	0.354	0.363	66.36
	Sum of Probabilities	0.374	0.355	0.364	66.04
Scale of Feature	Multi-scale	0.373	0.354	0.363	66.36
	Single-scale	0.381	0.359	0.370	64.02
Depth Upsampling	Learned	0.373	0.354	0.363	66.36
	Bilinear	0.409	0.366	0.387	63.79
Inverse Depth Loss	Yes	0.373	0.354	0.363	66.36
	No	0.379	0.349	0.364	64.51
Pixel-wise View Weight	Yes	0.373	0.354	0.363	66.36
	No	0.371	0.368	0.369	60.02

Table 4. Ablation study of the first six experiments on DTU [1] and ETH3D [32]. Settings used in our method are underlined.

$K$	Acc.(mm) ↓	Comp.(mm) ↓	Overall(mm) ↓	Time (s)
1	0.389	0.403	0.396	<b>0.190</b>
2	0.383	0.381	0.382	0.220
3	0.377	0.363	0.370	0.250
4	0.373	0.354	0.363	0.260
8	0.366	0.339	0.353	0.375
16	<b>0.363</b>	<b>0.333</b>	<b>0.348</b>	0.610

Table 5. Ablation study on the number of GRU iterations  $K$  on DTU [1].

simple bilinear upsampling. Clearly, the learned approach delivers better results on both DTU [1] and ETH3D [32].

**Inverse Depth Loss.** By default, we compute the losses  $L_{\text{initial}}$ ,  $L_{\text{regress}}$ ,  $L_{\text{upsample}}$  w.r.t. the *inverse* depth. We also try computing the losses without converting to the *inverse* depth range. While the result on DTU [1] is similar, the performance on the large-scale ETH3D [32] dataset, is explicitly improved when using inverse depth.

**Pixel-wise View Weight.** By default, we estimate the pixel-wise view weight to robustly integrate the matching information from the source views. Here, we compare to the simpler approach of taking the average over the views. The large performance difference on ETH3D [32] can be explained by strong viewpoint variations within the dataset, such that our robust integration of the matching information across the views becomes very impactful.

**Number of Iterations.** Table 5 relates the performance in *accuracy*, *completeness* and *overall quality* to the number of iterations on the GRU. Running more iterations allows to inject more matching information into the hidden state, leading to more accurate depth probability distributions per pixel. Although we limit our model to  $K = 4$  iterations during training, the performance keeps improving beyond that in all metrics. Comparing the result with Table 1, we can even observe that our method achieves state-of-the-art performance at  $K = 16$  on DTU. Despite the increased runtime, the inference is still faster than most multi-stage methods [6, 11, 46]. We conclude by noting that the flexibility to run our model for an arbitrary number of iterations enables the user to trade-off time efficiency for performance in de-

$N$	Acc.(mm) ↓	Comp.(mm) ↓	Overall(mm) ↓
2	0.376	0.556	0.466
3	<b>0.370</b>	0.387	0.379
4	0.372	0.360	0.366
5	0.373	0.354	0.363
6	0.373	<b>0.352</b>	<b>0.362</b>
7	0.371	0.356	0.364

Table 6. Ablation study on the number of views  $N$  on DTU [1].

pendence of the application.

**Number of Views.** We vary the number of views  $N$  and summarize the results in Table 6. Multi-view information can help to alleviate problems such as occlusions and the reconstruction quality improves at a higher value, saturating at around 6 views.

## 4.5. Limitations

While our network structure allows to trade-off speed and accuracy by adjusting the number of iterations during inference, we need to fix the number of samples that our probability distributions consist of. This number,  $D_2$ , is afterwards determined by the network structure and cannot be adjusted for different scenes. Likewise predetermined is the range of the data term samples placed around the current solution that are ingested into the CNN. In our model those samples cover  $2R_3 = 1/4^{\text{th}}$  of the total *inverse* depth range.

## 5. Conclusion

We present IterMVS, a novel learning-based MVS method combining highest efficiency and competitive reconstruction quality. We propose to explicitly encode a pixel-wise probability distribution of depth in the hidden state of a GRU-based estimator. In each iteration, we inject multi-scale matching information and extract the – in the *inverse* depth range – uniformly sampled depth distribution to estimate depth map and confidence. Extensive experiments on DTU, Tanks & Temples and ETH3D show highest efficiency in both memory and run-time, and a better generalization ability than many state-of-the-art learning-based methods.

## Appendix

### 1. Matching Similarities with Multi-scale Features

When performing the iterative updates, we compute the matching similarities from features on all levels to include multi-scale information. For a pixel  $\mathbf{p}$  with coordinates  $(x, y)$  in the depth-map  $\mathbf{D} \in \mathbb{R}^{W/4 \times H/4}$ , we first find its corresponding position  $\mathbf{p}_l$  for level  $l$  ( $l = 1, 2, 3$ ) in the reference feature map at the coordinates  $(x/2^{l-2}, y/2^{l-2})$ . Then, the reference feature of  $\mathbf{p}_l$ ,  $\mathbf{F}_{0,l}(\mathbf{p}_l)$ , is found via bilinear interpolation. Afterwards, with  $N_l$  new depth hypotheses, known camera parameters (for each level  $l$ ) and source features  $\{\mathbf{F}_{i,l}\}_{i=1}^{N-1}$ , we warp (using differentiable warping)  $\mathbf{p}_l$  into the respective source view and compute the matching similarities between reference and each source view. Finally, we use the  $2 \times$  upsampled pixel-wise view weights to compute the integrated matching similarities and pass them through a level-wise 2D U-Net to aggregate the neighborhood information.

### 2. Depth Upsampling

Following RAFT [34], we upsample the depth map from  $1/4$  to full resolution. Specifically, the depth of each pixel in the high resolution depth map is a convex combination of its 9 neighbors at the coarse resolution. The weights are learned from the reference feature map. Fig. 5 illustrates the upsampling process.

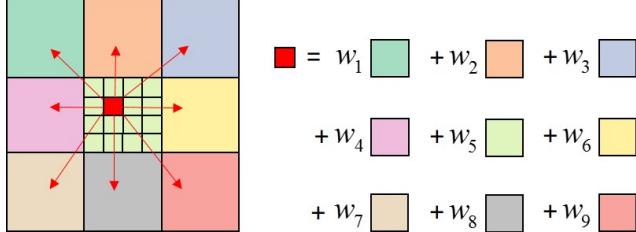


Figure 5. Illustration of depth upsampling. In the high resolution depth map, the depth of each pixel is the weighted sum of its 9 coarse resolution neighbors.

### 3. Point Cloud Reconstruction

Before fusing the depth maps, we filter out unreliable depth estimates, following MVSNet [47]. There are two filtering steps: geometric consistency filtering and confidence filtering.

**Geometric Consistency Filtering.** Following MVSNet [47], we apply a geometric constraint to measure the consistency of depth estimates among multiple views. For each pixel  $\mathbf{p}$  in the reference view, we project it, using its

depth  $d_0$ , to a pixel  $\mathbf{p}_i$  in the  $i$ -th source view. After looking up its depth  $d_i$  in the source view, we reproject  $\mathbf{p}_i$  into the reference view, and look up the depth  $d_{\text{reproj}}$  at this location,  $\mathbf{P}_{\text{reproj}}$ . We consider pixel  $\mathbf{p}$  and its depth  $d_0$  as consistent to the  $i$ -th source view, if the distances, in image space and depth, between the original estimate and its reprojection satisfy:

$$|\mathbf{p}_{\text{reproj}} - \mathbf{p}| < \delta, |d_{\text{reproj}} - d_0|/d_0 < \varepsilon, \quad (13)$$

where  $\delta = 1$  and  $\varepsilon = 0.01$  are two thresholds. Finally, we accept the estimations as reliable, if they are consistent in at least  $N_{\text{geo}}$  source views.

**Confidence Filtering.** Since our learned confidence indicates how close the estimation is to the ground truth depth, we use it to filter out estimations with high uncertainty. Specifically, we use a confidence threshold  $\tau = 0.3$  throughout the experiments to filter out all the pixels with confidence lower than it.

### 4. Visualization of Probability

Our GRU-based probability estimator encodes the per-pixel probability distribution of depth with the hidden state. A 2D CNN is applied on the hidden state to estimate the probability of  $D_2$  samples that are uniformly distributed in the *inverse* depth range for each pixel. We visualize this probability distribution for various scenes in Fig. 6. For pixels with distinct features, the probability has a single dominant peak and the estimation is precise. For some challenging situations, where distributions are non-peaky or multimodal, *e.g.* in textureless areas, our hybrid depth estimation strategy can still robustly produce estimations as accurate as possible. We also visualize the update process of probability distribution in Fig. 7. We observe that the probability distribution becomes more focused, several local maxima get suppressed and the estimation becomes more precise with more GRU iterations. In each iteration, multi-scale matching information is injected into the hidden state. This allows the hidden state to more accurately model the per-pixel probability distribution of depth with each iteration.

### 5. Visualization of Pixel-wise View Weight

Several examples for our estimated pixel-wise view weights are depicted in Fig. 8. Comparing the view weights with the visible areas in the reference validates that the all visible areas receive higher weights, while occluded and invisible parts have very low weights. Interestingly, in the first two images, pixels on the windows have low weights. Here, especially the upper row of windows mirror the surrounding buildings and cannot provide reasonable matching information. The other images have low weights in visible regions at areas with strong perspective and specular reflec-

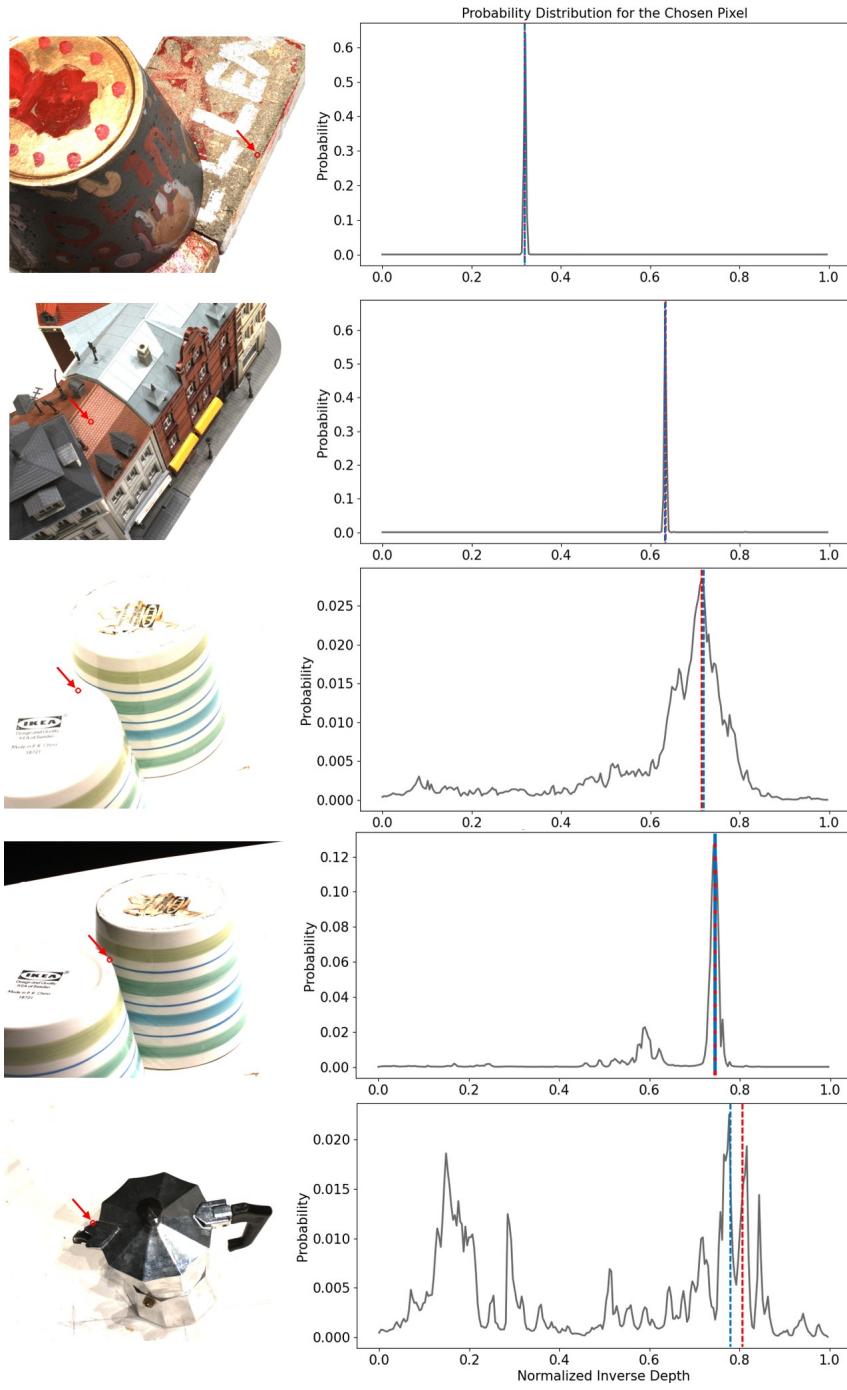


Figure 6. Visualization of probability. Left: Reference images (the chosen pixels are highlighted). Right: Probability distribution of depth for the chosen pixels. Red line denotes ground truth depth and blue line denotes our estimation.

tions as well as occlusions, while fronto-parallel and textured regions achieve higher weights. We conclude that our pixel-wise view weight is capable to determine co-visible areas between the reference and source images.

## 6. Visualization of Point Clouds

We visualize the reconstructed point clouds from DTU’s evaluation set [1], Tanks & Temples dataset [18] and ETH3D benchmark [32] in Fig. 9, 10 and 11.

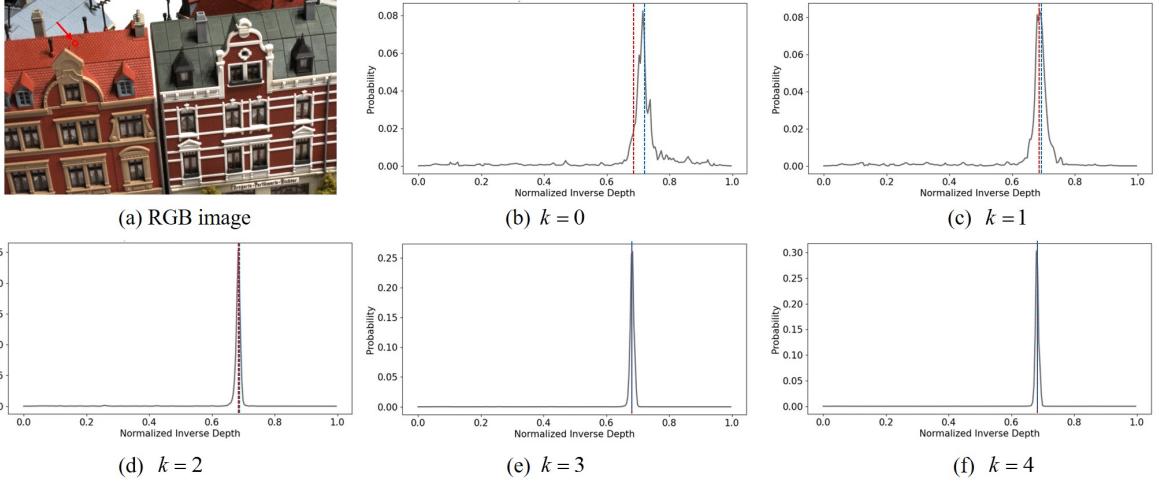


Figure 7. (a) Reference image (the chosen pixel is highlighted). (b)-(f) Probability distribution of depth for the chosen pixel in the  $k$ -th GRU iteration ( $k = 0$  represents the probability distribution from initial hidden state  $h_0$ ). **Red line** denotes ground truth depth and **blue line** denotes our estimation. Note our estimation becomes more accurate while several local maxima get suppressed in the distribution that converges to a single, more pronounced peak with more iterations.

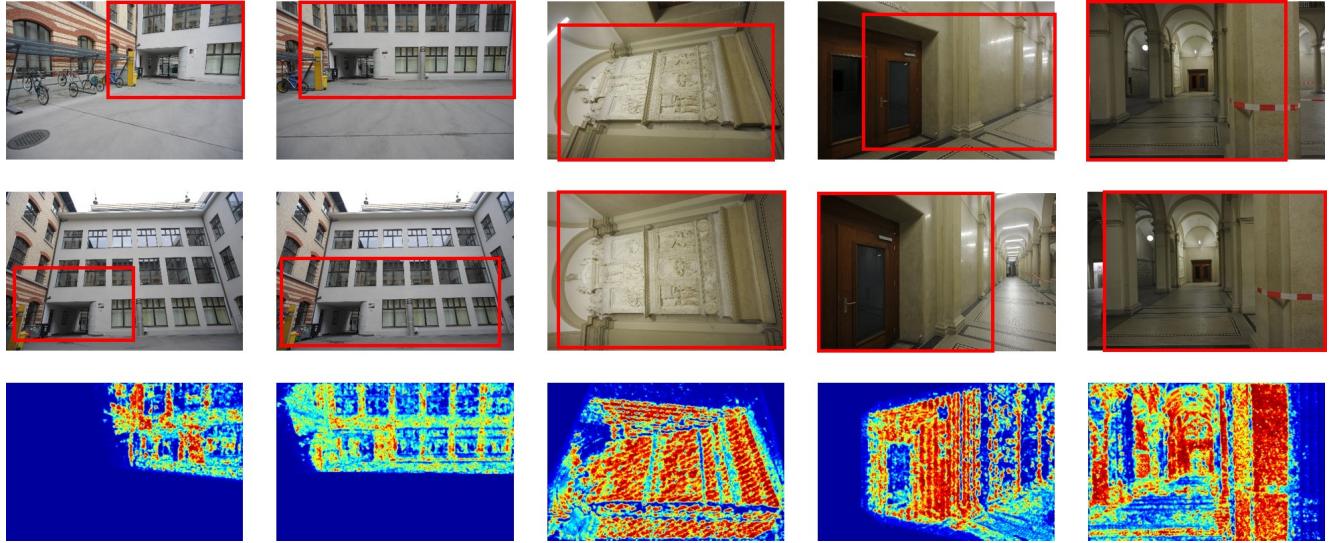


Figure 8. Visualization of our learned pixel-wise view weight on ETH3D [32]. Top row: reference images. Middle row: source images. Bottom row: pixel-wise view weight. Areas marked with boxes in reference images and source images are co-visible.

## 7. Future Work

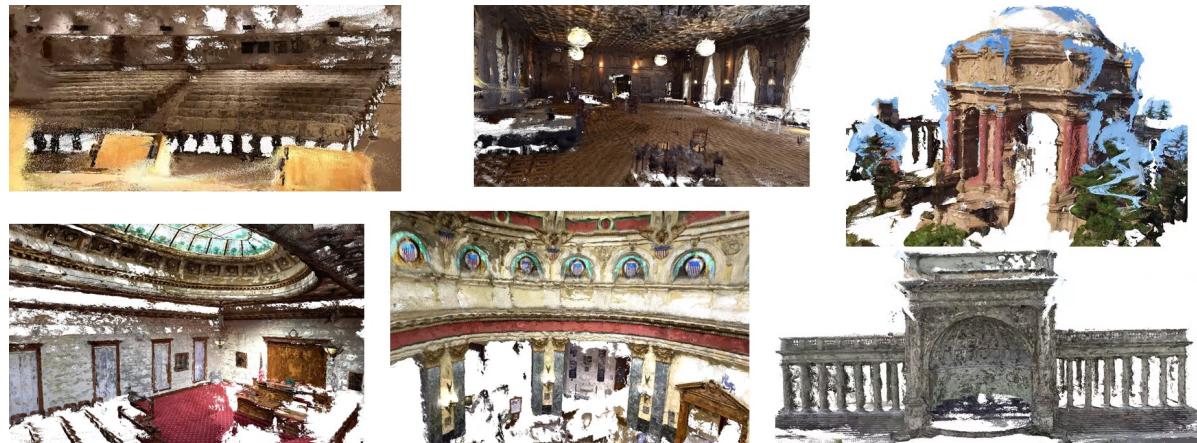
Currently, the learned confidence is only used to filter out unreliable estimates before depth fusion. However, we believe it will be a promising direction to further exploit the confidence in each GRU iteration. For example, one can refine the depth of unconfident areas with the information propagated from those confident areas [20, 24]. Another idea would be to focus more effort on unconfident areas only, while leaving the confident areas unchanged, which further should improve efficiency.



Figure 9. Reconstruction results on DTU's evaluation set [1].



(a) Intermediate dataset



(b) Advanced dataset

Figure 10. Reconstruction results on Tanks & Temples dataset [18].



(a) Training dataset



(b) Test dataset

Figure 11. Reconstruction results on ETH3D Benchmark [32].

## References

- [1] Henrik Aanæs, Rasmus Ramsbøl Jensen, George Vogiatzis, Engin Tola, and Anders Bjørholm Dahl. Large-scale data for multiple-view stereopsis. *IJCV*, 2016. [1](#), [2](#), [5](#), [6](#), [7](#), [8](#), [10](#), [12](#)
- [2] Connnelly Barnes, Eli Shechtman, Adam Finkelstein, and Dan B Goldman. Patchmatch: A randomized correspondence algorithm for structural image editing. *TOG*, 28(3):24, 2009. [2](#)
- [3] Michael Bleyer, Christoph Rhemann, and Carsten Rother. Patchmatch stereo - stereo matching with slanted support windows. In *BMVC*, 2011. [2](#)
- [4] Neill D. F. Campbell, George Vogiatzis, Carlos Hernández, and Roberto Cipolla. Using multiple hypotheses to improve depth-maps for multi-view stereo. In *ECCV*, 2008. [6](#)
- [5] Rui Chen, Songfang Han, Jing Xu, and Hao Su. Point-based multi-view stereo network. In *ICCV*, 2019. [2](#), [6](#), [7](#)
- [6] Shuo Cheng, Zexiang Xu, Shilin Zhu, Zhuwen Li, Li Erran Li, Ravi Ramamoorthi, and Hao Su. Deep stereo using adaptive thin volume representation with uncertainty awareness. In *CVPR*, 2020. [1](#), [2](#), [6](#), [7](#), [8](#)
- [7] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014. [1](#), [2](#)
- [8] Zehua Fu, Mohsen Ardabiliyan, and Guillaume Stern. Stereo matching confidence learning based on multi-modal convolutional neural networks. In *International Workshop on Representations, Analysis and Recognition of Shape and Motion From Imaging Data*, pages 69–81. Springer, 2017. [4](#)
- [9] Yasutaka Furukawa and Jean Ponce. Accurate, dense, and robust multiview stereopsis. *PAMI*, 2010. [1](#), [2](#), [6](#), [7](#)
- [10] Silvano Galliani, Katrin Lasinger, and Konrad Schindler. Massively parallel multiview stereopsis by surface normal diffusion. In *ICCV*, 2015. [1](#), [2](#), [6](#), [7](#)
- [11] 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*, 2020. [1](#), [2](#), [3](#), [4](#), [6](#), [7](#), [8](#)
- [12] Xiaodong Gu, Weihao Yuan, Zuozhuo Dai, Chengzhou Tang, Siyu Zhu, and Ping Tan. Dro: Deep recurrent optimizer for structure-from-motion. *arXiv preprint arXiv:2103.13201*, 2021. [2](#)
- [13] Xiaoyang Guo, Kai Yang, Wukui Yang, Xiaogang Wang, and Hongsheng Li. Group-wise correlation stereo network. In *CVPR*, 2019. [3](#)
- [14] Mengqi Ji, Juergen Gall, Haitian Zheng, Yebin Liu, and Lu Fang. Surfacenet: An end-to-end 3D neural network for multiview stereopsis. In *ICCV*, 2017. [2](#), [5](#)
- [15] Abhishek Kar, Christian Häne, and Jitendra Malik. Learning a multi-view stereo machine. In *NIPS*, 2017. [2](#)
- [16] Alex Kendall, Hayk Martirosyan, Saumitra Dasgupta, Peter Henry, Ryan Kennedy, Abraham Bachrach, and Adam Bry. End-to-end learning of geometry and context for deep stereo regression. In *ICCV*, 2017. [4](#)
- [17] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. [5](#)
- [18] Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and temples: Benchmarking large-scale scene reconstruction. *TOG*, 2017. [1](#), [2](#), [5](#), [7](#), [10](#), [13](#)
- [19] Ilya Kostrikov, Esther Horbert, and Bastian Leibe. Probabilistic labeling cost for high-accuracy multi-view reconstruction. In *CVPR*, pages 1534–1541, 2014. [2](#)
- [20] Andreas Kuhn, Christian Sormann, Mattia Rossi, Oliver Erdler, and Friedrich Fraundorfer. Deepc-mvs: Deep confidence prediction for multi-view stereo reconstruction. In *3DV*, pages 404–413. IEEE, 2020. [11](#)
- [21] Kiriakos N Kutulakos and Steven M Seitz. A theory of shape by space carving. *IJCV*, 38(3):199–218, 2000. [2](#)
- [22] Jae Yong Lee, Joseph DeGol, Chuhang Zou, and Derek Hoiem. Patchmatch-rl: Deep mvs with pixelwise depth, normal, and visibility. In *CVPR*, pages 6158–6167, 2021. [7](#)
- [23] Maxime Lhuillier and Long Quan. A quasi-dense approach to surface reconstruction from uncalibrated images. *PAMI*, 27(3):418–433, 2005. [2](#)
- [24] Zhaoxin Li, Wangmeng Zuo, Zhaoqi Wang, and Lei Zhang. Confidence-based large-scale dense multi-view stereo. *TIP*, 29:7176–7191, 2020. [11](#)
- [25] Tsung-Yi Lin, Piotr Dollár, Ross B. Girshick, Kaiming He, Bharath Hariharan, and Serge J. Belongie. Feature pyramid networks for object detection. *CVPR*, 2017. [3](#)
- [26] Lahav Lipson, Zachary Teed, and Jia Deng. Raft-stereo: Multilevel recurrent field transforms for stereo matching. *arXiv preprint arXiv:2109.07547*, 2021. [2](#), [3](#)
- [27] Keyang Luo, Tao Guan, Lili Ju, Haipeng Huang, and Yawei Luo. P-MVSNet: Learning patch-wise matching confidence aggregation for multi-view stereo. In *ICCV*, 2019. [2](#), [6](#), [7](#)
- [28] Pierre Moulon, Pascal Monasse, Romuald Perrot, and Renaud Marlet. Openmvg: Open multiple view geometry. In *International Workshop on Reproducible Research in Pattern Recognition*, 2016. [5](#), [6](#)
- [29] Matteo Poggi and Stefano Mattoccia. Learning from scratch a confidence measure. In *BMVC*, 2016. [4](#)
- [30] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. [3](#)
- [31] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-Motion Revisited. In *CVPR*, 2016. [1](#), [2](#), [5](#), [6](#), [7](#)
- [32] Thomas Schöps, Johannes L. Schönberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *CVPR*, 2017. [1](#), [2](#), [5](#), [7](#), [8](#), [10](#), [11](#), [14](#)
- [33] Steven M Seitz and Charles R Dyer. Photorealistic scene reconstruction by voxel coloring. *IJCV*, 35(2):151–173, 1999. [2](#)
- [34] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *ECCV*, pages 402–419. Springer, 2020. [2](#), [5](#), [9](#)
- [35] Zachary Teed and Jia Deng. Raft-3d: Scene flow using rigid-motion embeddings. In *CVPR*, pages 8375–8384, 2021. [2](#)

- [36] Engin Tola, Christoph Strecha, and Pascal Fua. Efficient large-scale multi-view stereo for ultra high-resolution image sets. *Machine Vision and Applications*, 2012. 6
- [37] Fabio Tosi, Matteo Poggi, Antonio Benincasa, and Stefano Mattoccia. Beyond local reasoning for stereo confidence estimation with deep learning. In *ECCV*, pages 319–334, 2018. 4
- [38] Ali Osman Ulusoy, Michael J. Black, and Andreas Geiger. Semantic multi-view stereo: Jointly estimating objects and voxels. In *CVPR*, 2017. 2
- [39] Fangjinhua Wang, Silvano Galliani, Christoph Vogel, Pablo Speciale, and Marc Pollefeys. Patchmatchnet: Learned multi-view patchmatch stereo. In *CVPR*, pages 14194–14203, June 2021. 1, 2, 3, 4, 5, 6, 7
- [40] SHI Xingjian, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. In *NIPS*, pages 802–810, 2015. 1, 2
- [41] Qingshan Xu and Wenbing Tao. Multi-scale geometric consistency guided multi-view stereo. In *CVPR*, 2019. 1, 2, 6, 7
- [42] Qingshan Xu and Wenbing Tao. Learning inverse depth regression for multi-view stereo with correlation cost volume. In *AAAI*, 2020. 2, 3, 6, 7
- [43] Qingshan Xu and Wenbing Tao. Planar prior assisted patch-match multi-view stereo. In *AAAI*, pages 12516–12523, 2020. 2
- [44] Qingshan Xu and Wenbing Tao. PVSNet: Pixelwise visibility-aware multi-view stereo network. *ArXiv*, 2020. 2, 3, 6, 7
- [45] Jianfeng Yan, Zizhuang Wei, Hongwei Yi, Mingyu Ding, Runze Zhang, Yisong Chen, Guoping Wang, and Yu-Wing Tai. Dense hybrid recurrent multi-view stereo net with dynamic consistency checking. In *ECCV*, pages 674–689. Springer, 2020. 1, 2, 4, 6, 7
- [46] Jiayu Yang, Wei Mao, Jose M. Alvarez, and Miaomiao Liu. Cost volume pyramid based depth inference for multi-view stereo. In *CVPR*, 2020. 1, 2, 6, 7, 8
- [47] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. MVSNet: Depth inference for unstructured multi-view stereo. In *ECCV*, 2018. 1, 2, 3, 4, 5, 6, 7, 9
- [48] Yao Yao, Zixin Luo, Shiwei Li, Tianwei Shen, Tian Fang, and Long Quan. Recurrent MVSNet for high-resolution multi-view stereo depth inference. In *CVPR*, 2019. 1, 2, 3, 4, 5, 6, 7
- [49] Yao Yao, Zixin Luo, Shiwei Li, Jingyang Zhang, Yufan Ren, Lei Zhou, Tian Fang, and Long Quan. Blendedmvs: A large-scale dataset for generalized multi-view stereo networks. In *CVPR*, pages 1790–1799, 2020. 1, 5, 6, 7
- [50] Zehao Yu and Shenghua Gao. Fast-MVSNet: Sparse-to-dense multi-view stereo with learned propagation and gauss-newton refinement. In *CVPR*, 2020. 1, 6, 7
- [51] Jingyang Zhang, Yao Yao, Shiwei Li, Zixin Luo, and Tian Fang. Visibility-aware multi-view stereo network, 2020. 2, 6, 7