IS-MVSNet: Importance Sampling-based MVSNet

Likang Wang[1], Yue Gong[2], Xinjun Ma[2], Qirui Wang[2], Kaixuan Zhou[3,4], and Lei Chen[1]

[1] Department of Computer Science and Engineering
Hong Kong University of Science and Technology
{lwangcg, leichen}@ust.hk
[2] Distributed and Parallel Software Lab, Huawei Technologies
{gongyue1, maxinjun1, wangqirui1}@huawei.com
[3] Riemann Lab, Huawei Technologies
[4] Fundamental Software Innovation Lab, Huawei Technologies
zhoukaixuan2@huawei.com

Abstract. This paper presents a novel coarse-to-fine multi-view stereo (MVS) algorithm called importance-sampling-based MVSNet (IS-MVSNet) to address a crucial problem of limited depth resolution adopted by current learning-based MVS methods. We proposed an importance-sampling module for sampling candidate depth, effectively achieving higher depth resolution and yielding better point-cloud results while introducing no additional cost. Furthermore, we proposed an unsupervised error distribution estimation method for adjusting the density variation of the importance-sampling module. Notably, the proposed sampling module does not require any additional training and works reasonably well with the pre-trained weights of the baseline model. Our proposed method leads to up to 20× promotion on the most refined depth resolution, thus significantly benefiting most scenarios and excellently superior on fine details. As a result, IS-MVSNet **outperforms all the published papers on TNT's intermediate benchmark** with an F-score of 62.82%. Code is available at github.com/NoOneUST/IS-MVSNet.

Keywords: 3D reconstruction; multi-view stereo; importance sampling

1 Introduction

Multi-view stereo (MVS) is one of the most fundamental computer vision challenges. MVS aims to reconstruct the 3D structure of scenes from multiple 2D image slots taken at different angles and positions. Most existing MVS algorithms formulate the reconstruction task as a problem of maximizing the geometric consistency among views. Inspired by the great successes of deep learning in visual perception [7, 11, 17], MVSNet [27] introduced convolutional neural networks (CNNs) for better reconstruction quality. While these learning-based methods are proven effective, they encountered difficulties handling large-scale scenes due to the heavy computational overhead. For example, the maximum resolution of

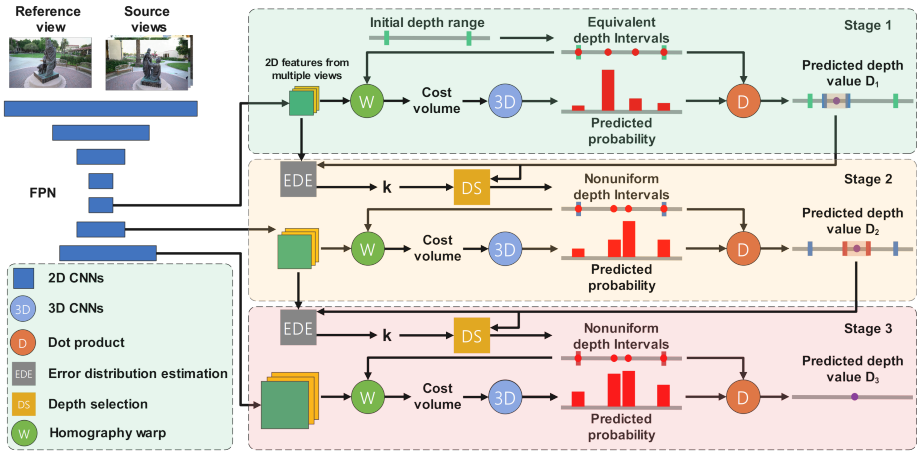


Fig. 1: Model structure of IS-MVSNet. At stage $s = 1$, we uniformly sample depth candidates. At each stage $s > 1$, we first estimate the former error distribution on the whole dataset, then we sample depth candidates according to the former prediction and error distribution. Next, we establish a cost volume at the sampled depths. Afterward, we adopt a 3D CNN to estimate the probability of each sampled depth to be true. The depth prediction (purple point) is calculated as the dot product of the sampled depths and the corresponding probability estimations.

MVSNet is limited to $1152 \times 864 \times 192$ (width, height, depth) given a GPU with 11GB graphics memory [3]. Follow-up works [6,29] partially eased the resolution constraints via predicting the depth maps in a coarse-to-fine manner. Specifically, these papers start from a low-resolution depth prediction, then gradually enlarge the resolution while shrinking the depth range and reducing the candidate depth number. In the end, we can obtain a depth prediction with a higher resolution. The basic assumption behind these coarse-to-fine algorithms is that the coarse prediction is a reliable ground truth estimation.

Even with the coarse-to-fine strategy, the depth resolution is still a crucial factor preventing us from high accuracy and efficiency at the same time [4]. We argue that the existing coarse-to-fine algorithms [6, 26, 29] did not fully use the reliable former prediction assumption because these methods equally treat each candidate within the depth range. In this paper, instead, we put more effort into the most promising candidates. Then, the new problem is to distinguish which depths are the most trustworthy. Although the coarse prediction is assumed to be close to the actual depth, it is not 100% accurate. Thus, towards a more precise localization of the ground truth, it becomes crucial to estimate the error distribution of the coarse prediction.

Based on such considerations, we propose a novel MVS algorithm named importance-sampling-based MVSNet (IS-MVSNet), introducing an effective can-

didate depth sampling strategy to conduct a more precise depth prediction via significantly enlarging the depth resolution around the ground truth in a cost-free manner.

Inspired by the importance sampling theory [12], we sample the candidate depths following the estimated error distribution of the last stage instead of blindly treating the error distribution uniform as all the existing coarse-to-fine approaches. If we can estimate the error distribution effectively, the new depth prediction will undoubtedly be more accurate. Towards a general estimation strategy of the error distribution, we fit it with a simple but effective uni-modal probability density function. For better local consistency, IS-MVSNet adopts a geometric interval sequence to place hypothesized depths. In this way, we significantly increase the depth resolution at the ground truth. In most cases, our method samples more densely around the previous stage's prediction while giving less attention to the furthest points. Our sampling strategy outperforms the uniform-sampling-based solution adopted by the existing models in most cases while introducing no additional computational overhead.

Besides the importance-sampling-based strategy, we propose an unsupervised algorithm to estimate the error distribution based on photometric consistency. In this way, our model becomes adaptive to new datasets. With these properties, IS-MVSNet generalizes very well on various unseen scenarios.

Finally, our extensive experiments on the most popular MVS datasets, including Tanks & Temples [10] (TNT), ETH3D [19], and DTU [1], demonstrate IS-MVSNet's superiority over current SOTAs. With an F-score of 62.82%, IS-MVSNet surpasses all the published MVS algorithms on TNT's intermediate benchmark by a clear margin.

2 Related Work

Multi-view stereo has been studied for decades as a fundamental computer vision task. Before the prosperity of deep learning, there had been various hand-crafted methods [5, 18, 21]. Despite the traditional solutions' success, learning-based algorithms provide better semantic insights and are more robust in illumination and parallax.

As the first learning-based MVS algorithm, SurfaceNet [8] suffers significant GPU memory overhead and applicable restricted scenarios due to the divide-and-conquer framework and the adopted 3D CNNs. Most modern models inherit the main framework proposed by MVSNet [27] to ease the constraints above. MVSNet separates the depth map prediction from the point cloud fusion and establishes a differentiable end-to-end depth prediction network containing four sequential sub-steps: representation learning for input views, geometric-consistency-based scoring at hypothesized depths, scoring refinement, and depth regression. Although MVSNet demonstrates a practical and universal pipeline, its performance is intensively restricted by the image resolution and depth sampling. Specifically, MVSNet can only achieve an F-score of 43.48% on a GPU with 11GB of graphics memory.

For the sake of high-quality, large-scale 3D reconstruction, the follow-up papers generally ease MVSNet's resolution barrier in two ways: RNN-based [24,25] and CNN-based [4,6,26]. The RNN-based methods utilize GRU or LSTM instead of CNN to regularize the cost volume. While the RNN-based models trade temporal overheads for spatial advantages, the CNN-based models generally inherit the coarse-to-fine framework proposed in [16].

CasMVSNet [6] first sparsely samples hypothesized depths from a wide range and generates a rough depth estimation, then repeatedly shrinks the depth range and refines the depth prediction. CVP-MVSNet [26] tunes the depth range based on the image resolution. UCSNet [4] decides the depth range according to the former stage's confidence. However, its depth resolution may be lower than other fixed-depth-range methods because of its unstable depth range determining strategy [14].

In this paper, we inherit both the learning-based framework and the hierarchical pipeline. However, we promote the depth resolution around the ground truth via non-uniformly sampling of the candidate depths. Specifically, inspired by the importance sampling theory, we first unsupervisedly estimate the previous stage's error distribution, then based on which, we sample the current stage's candidate depths. Compared to CasMVSNet, we retain both the depth range and the depth number while sampling at different locations. Compared to CVP-MVSNet and UCSNet, we aim to find better candidates within an arbitrary depth range while not caring about the depth range or the depth number itself. There are also non-MVS methods considering candidate selection. For example, AdaBins [2] assumes the model can learn how to tune the bin density with an additional sub-network. UASNet [14] and NeRF [15] first infer an initial prediction, then sample based on it. We argue these methods generalize worse on unseen datasets because a) UASNet, UCSNet, and NeRF estimate distributions for every pixel individually, which is difficult. We estimate the distribution of the whole dataset, which is statistically more stable and accurate; b) we can unsupervisedly adjust the sampling strategy on unseen datasets, but UASNet and UCSNet cannot, thus are easy to overfit.

3 Methodology

In this section, we first present the main structure of IS-MVSNet and then provide a comprehensive introduction to the model's details. Following CasMVSNet [6] and VisMVSNet [29], IS-MVSNet inherits a coarse-to-fine network structure. The overall framework of our model is shown in Fig. 1. Firstly, IS-MVSNet adopts a feature pyramid network [13] (FPN) to extract the hierarchical representations for both the reference and source images. The FPN allows IS-MVSNet to capture both the global contexts and the local pixel-wise details. Then, a group of hypothesized depths is sampled for further evaluation.

For the coarsest stage $s = 1$, we uniformly sample hypothesized depths within the pre-defined depth range. For stages $s > 1$, we propose an importance-sampling-based hypothesized depth selection strategy, which is formally de-

scribed in Sec. 3.1. This strategy provides IS-MVSNet with much higher sampling effectiveness without sacrificing efficiency. In Sec. 3.2, we propose an unsupervised method to estimate appropriate hyper-parameters for importance sampling.

After this, we project the source feature maps to the reference view at the chosen hypothesized depths and calculate the inter-view matching cost at each hypothesized depth to form a cost volume. Next, we adopt a 3D CNN to regularize the cost volume and predict the probability of each hypothesized depth as the ground truth. Finally, the current stage's depth prediction is calculated as the inner product of the depth samples and the corresponding probability predictions.

3.1 Importance-sampling based hypothesized depth selection

As a coarse-to-fine algorithm, IS-MVSNet gradually refines the depth prediction. Given stage $s > 1$, although the former prediction D_p^{s-1} is generally close to the actual depth d_{gt}, there is still a gap between them. Suppose we can estimate each pixel's depth prediction error and further sample hypothesized depth around the ground truth with greater resolution. In that case, the model's capability of capturing fine details can be immensely enhanced.

Although it is difficult and impractical to estimate the error for each pixel, we propose to estimate the error distribution for the whole dataset and adjust the hypothesized depth sampling accordingly. While all the existing MVS algorithms did not consider the error estimation and blindly treated the prediction error as a uniform random variable. In IS-MVSNet, we propose a method to find out n_s promising candidate depth values $\{d_i^s\}_{i=1}^{n_s}$ for each pixel at stage $s > 1$, based on both the former stage's depth prediction D_p^{s-1} and the probability density function (PDF) $f_e^{s-1}(\delta)$ of the depth prediction error $\delta \sim \Delta_p^{s-1} = d_{gt} - D_p^{s-1}$, where d_{gt} denotes the pixel's real depth, estimated on all the pixels within the dataset. Then, we sample at $\{d_i^s\}_{i=1}^{n_s}$ to generate a more precise depth prediction $D_p^s = \sum_{i=1}^{n_s} d_i^s \cdot p(d_i^s)$, where $p(d_i^s)$ denotes the probability that the candidate depth $d_i^s \in \{d_j^s\}_{j=1}^{n_s}$ is the nearest neighbor of d_{gt}.

In this way, we can locate the most promising candidate depths more precisely and then allocate more attention to them. The result is that depth precision gets promoted due to the finest depth resolution increment around the ground truth.

Error formulation The first problem is how to formulate the error distribution. We argue it is reasonable to approximate the error PDF as a uni-modal function for three reasons. Firstly, since many factors influence the prediction error, the central limit theorem [9] suggests that the error tends to follow a zero-mean uni-modal distribution. Secondly, the coarse prediction is generated via uniform sampling, which leads to unbiased estimation [12]. Thirdly, our experiments on various real datasets verify that the error indeed follows a uni-modal distribution with a mean close to zero. Notably, we do not require the former stage to give a uni-modal probability prediction for a given pixel's hypothesized depths. Instead,

we prefer the distance from the actual depth to the depth prediction calculated from all the hypothesized depths to follow a uni-modal distribution.

Suppose most pixels' depths are correctly estimated in the former stage, it is clear that our method outperforms uniform sampling. In Fig. 4d, our experiments on real datasets show that sampling following zero-mean Gaussian distribution indeed significantly surpasses uniform distribution. Moreover, even in extreme cases where most pixels' depths are wrongly estimated in the former stage, sampling following Gaussian distribution benefits the majority of pixels by providing a higher sampling density at these pixels' actual depths. Even if we do not estimate the mean and sample following a zero-mean Gaussian distribution, our method still benefits more pixels than uniform sampling. Our sampling method is better or comparable to uniform sampling even in the regions containing the most wrong former predictions, e.g., repetitive and textureless regions, small and thin objects distant from backgrounds.

Discrete interval Compared to sampling from a continuous PDF, discretized intervals have two advantages. First, given a limited depth number, e.g., 8, discretized intervals lead to a sampling density more stable and closer to the actual error distribution than i.i.d. sampling. Second, discretized intervals benefit the convolutions because the neighboring pixels have similar sampled depths, and the spatial correlation is crucial for convolution.

We further propose to sample the candidate depths following a pre-defined interval sequence unevenly based on such considerations. Precisely, the error PDF should control the depth interval: in positions with larger PDF, the interval should be smaller; otherwise, it should be larger. Let μ_e^{s-1} denote the mean error at stage $s-1$, then the depth interval close to $D_p^{s-1} + \mu_e^{s-1}$ should be smaller, otherwise larger. We adopt a simple and typical geometric sequence to fit the interval pattern to satisfy the requirement. Note that other sequences with similar trends are also acceptable as if they have similar properties with a Gaussian distribution $N(\mu_e^{s-1}, \sigma_e^{s-1})$, i.e., both have only one single mode at μ_e^{s-1} and the sequence has a parameter with similar effect as σ_e^{s-1} of $N(\mu_e^{s-1}, \sigma_e^{s-1})$. In addition, it is unnecessary to strictly force the interval sequence to converge to $N(\mu_e^{s-1}, \sigma_e^{s-1})$ when the number of intervals $\to \infty$. For example, the arithmetic sequence also works well. In this way, we sample the depths following the error distribution while preserving the local consistency. Our detailed importance sampling algorithm is described below.

Detailed algorithm We use discretized intervals to put depth hypotheses in the depth range rather than directly sampling depths from a continuous PDF. In the first stage, we divide the whole depth range R_1 into $n_1 - 1$ equivalent intervals of size ${}^{R_1}/_{n_1-1}$ because there is no prior unbiased depth estimation given at stage $s = 1$. In the following stages $s \in \{2, 3, \ldots\}$, we adopt a trivial geometric progression for generating the depth hypothesis with promoted sampling density in the central area. The discretized intervals are parameterized by k_s, a hyperparameter determining the shape of the intervals. As illustrated in Fig. 2, the minimum interval is reduced to $\frac{1}{k_s}$ and the change velocity of interval lengths is

c_s, which is controlled by k_s. A larger k_s means to sample more densely near the rectified former prediction $D_p^{s-1} + \mu_e^{s-1}$. When $k_s > 1$, the central hypothesized depths have intervals reduced to $1/k_s$, while the fringing depths have intervals enlarged. In other words, the central interval r_s/k_s gets smaller than the uniform sampling interval r_s by a factor of $1/k_s$. When $k_s = 1$, our importance sampling down-grades to uniform sampling. When $0 < k_s < 1$, our method can deal with the case that most former predictions are wrong.

To be specific, the depth intervals form a symmetric geometric progressions $T = [\frac{1}{k_s} r_s c_s^{\frac{n_s}{2}-1}, \cdots, \frac{1}{k_s} r_s c_s^2, \frac{1}{k_s} r_s c_s, \frac{1}{k_s} r_s, \frac{1}{k_s} r_s c_s, \frac{1}{k_s} r_s c_s^2, \cdots, \frac{1}{k_s} r_s c_s^{\frac{n_s}{2}-1}]$, where $t_j = \frac{1}{k_s} r_s c_s^{|\frac{n_s}{2}-j|}$ is the jth interval, $r_s = \frac{R_s}{n_s - 1}$ is the depth interval in uniform sampling, and c_s is the common ratio between adjacent intervals. Since we want to keep IS-MVSNet's depth range and the number of hypothesized depths the same as those of the baseline model, i.e., $\sum_{i=1}^{n_s-1} t_i = R_s$, c_s is uniquely controlled by k_s, r_s, and n_s according to Eq. (1). In practice, c_s is numerically calculated as the root of Eq. (1).

$$c_s^{\frac{n_s}{2}} - 1 = \frac{1}{2}(k_s n_s - k_s + 1)(c_s - 1) \tag{1}$$

The depth candidates are defined uniquely for each pixel. To be specific, first, each pixel has its own set of discrete depth candidates defined by the interval sequence; second, the intervals between depth candidates and the depth range R (i.e., the sum of intervals) are consistent among all the pixels in terms of sizes; third, the center position of the depth range R along the depth axis is set to each pixel's previous depth estimate D_p^{s-1}. As a result, each pixel has a unique set of depth candidates whose intervals are but the same among pixels; fourth, if the mean error μ_e^{s-1} is estimated, the position of the range is further "rectified" to $D_p^{s-1} + \mu_e^{s-1}$.

3.2 Unsupervised error distribution estimation

In IS-MVSNet, we introduce two new hyper-parameters k_s and μ_s, to adjust the sampling function $g^s(x)$'s shape in stages $s > 1$. In practice, the depth estimation error concentratedly distributes around zero. Thus, in default, we treat the mean error $\mu_s = 0$ and only estimate k_s. However, the k_s estimation solution proposed in this section is also applicable to μ_s. If we want to estimate both k_s and μ_s, we first fix k_s and estimate μ_s, then fix μ_s and estimate k_s.

As analyzed in Sec. 3.1, the optimal k_s can be uniquely determined by minimizing the sampling function $g^s(x)$'s difference to the actual error distribution with the true depths known. However, we do not know the actual depth in real scenarios, and the scale, illumination, and camera intrinsics are distinct in different datasets. Thus, it is necessary to estimate a k_s for each dataset. We treat the matching costs as cues for the actual depths and demonstrate that estimating the error distribution is equivalent to minimizing the matching costs, which is always obtainable. This section proposes a general unsupervised hyper-parameter k_s selection strategy, making the importance-based sampling module hyper-parameter-free in all scenarios.

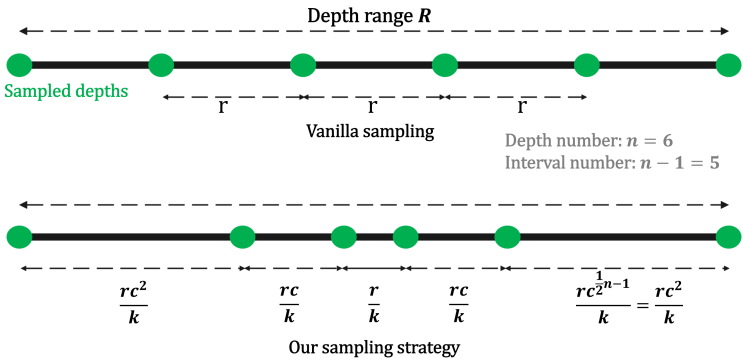


Fig. 2: The illustration of depth selection given a depth number of 6. In our sampling strategy, the depth range is retained the same. The minimum depth interval is reduced to $\frac{1}{k_s}$ and the interval lengths are increasing at the ratio of c_s, which is uniquely controlled by k_s following Eq. (1). A greater k_s leads to a smaller minimum interval, a greater c_s, and a higher changing velocity of the interval lengths.

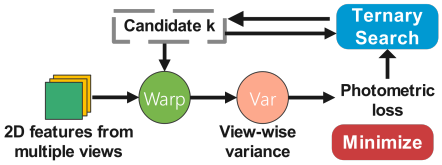


Fig. 3: The illustration of the error distribution estimation module. We evaluate k with the photometric loss and apply a ternary search to find the optimal k following Algorithm 1 and Algorithm 2.

Algorithm 1: Unsupervised k selection

Data: k_l: the minimum hypothesis k,
k_r: the maximum hypothesis k
Result: k_e^*: the estimated optimal k

while *within the allowed iterations* **do**
  $k_{mid} \leftarrow \frac{k_l+k_r}{2}$, $k_{midmid} \leftarrow \frac{k_{mid}+k_r}{2}$;
  Measure the photometric cost $C_l, C_r, C_{mid}, C_{midmid}$ with Algorithm 2;
  if $C_{mid} < C_{midmid}$ **then**
    $k_r \leftarrow k_{midmid}$;
  else
    $k_l \leftarrow k_{mid}$;
  end
end
$k_e^* \leftarrow \arg\min_k(C_l, C_r, C_{mid}, C_{midmid})$ % the minimum C must be in the four;

Recall that in MVS, the input 2D images and the camera parameters are always obtainable, and photometric consistency exists among different views. Given a 3D point P with the depth d_r and projection P_r in the reference view, then, P's projection in the vth source view P_v's coordinate can be computed as $P_v = H_v(d_r)P_r$, where $H_v(d_r) = K_v(R_v - \frac{1}{d_r}t_v{a_r}^T){K_v}^{-1}$ is a homography matrix, a_r denotes the reference view's principal axis, and K_v, R_v, t_v denote the vth camera's intrinsics, relative rotations, and translations, separately.

Algorithm 2: Photometric cost calculation

Data: k: the sampling hyper-parameter,
F_i: representation for the ith view,
V: number of views,
q: number of selected reference views,
D_p^{s-1}: the former stage's depth prediction
Result: C_k: the photometric cost

for *each scene in the dataset* **do**
  Randomly select q views as the reference views ;
  for *each reference view* i **do**
    for *each source view* j **do**
      Sample hypothesized depths according to k and the former stage's depth prediction D_p^{s-1} following Sec. 3.1;
      Infer each hypothesized depth's probability with the trained model;
      Calculate the depth prediction D_p^s for each pixel as the dot product of the sampled depths and the predicted probabilities;
      Map each pixel in F_j to view i with the homography matrix H_j at D_p^s: $F_j \leftarrow H_j F_j$;
    end
    Calculate the inter-view variance for all pixels:
    $var \leftarrow \frac{\sum {F_j}^2}{V} - (\frac{\sum F_j}{V})^2$, then append var to $VarSet$
  end
end
$C_k \leftarrow \overline{Varset}$

Suppose the depth estimation D_p^s is correct, then $P_v^s = H_v(D_p^s)P_r^s$ should represent the same 3D point as P_r^s, saying that P_r^s's feature $F_r^s = F_v^s$. Since multiple views are given, we use the variance $Var[F_v^s]$ to measure their similarity. Thus, the best depth estimation $D_p^* = \arg\min_{d_e} Var[F_v^s]$.

As mentioned in Sec. 3.1, k determines the estimated error distribution's PDF. Specifically, a larger k refers to an error distribution with a smaller variance. When $k = 1$, the importance sampling performs the same as the uniform sampling; only one candidate has the chance to be sampled when $k = \infty$. Clearly, $k = \{1, \infty\}$ both lead to a non-minimum difference between the estimation and the actual PDF. Thus, as shown in Fig. 4a, when k increases starting from 1, the model's performance first gets promoted, then gradually decreases. We use a

uni-modal function to approximate the performance-k curve. Based on such consideration, we proposed a ternary-search-based unsupervised hyper-parameter k selection algorithm as described in Algorithm 1, Algorithm 2 and Fig. 3. Since the ternary search reduces the search range by a constant ratio in each iteration, it converges very fast. Generally, 3 to 5 iterations are enough to find a satisfying k. Our experiments in Fig. 4c show that randomly picking two reference views from each scan is enough for k's determination.

4 Experiment

Our experiments adopt the most popular MVS datasets: Tanks & Temples (TNT), ETH3D, DTU, and BlendedMVS [28]. We summarize their properties in Tab. 1. We compare IS-MVSNet to the SOTA learning-based algorithms, e.g., VisMVSNet [29], CasMVSNet, CVP-MVSNet, UCS-MVSNet, Patchmatch-Net [20], and traditional algorithms, e.g., COLMAP [18], ACMM [21], ACMP [22]. On Tanks & Temples and ETH3D datasets, the metric is F-score, while on the DTU dataset, the metric is the overall distance. Our model only requires the number of stages $S > 1$. In our experiments, we set $S = 3$ following most existing models [6, 29] for two reasons. First, in this way, we can conduct more fair comparisons to the mainstream models. Second, $S = 3$ provides satisfying precision while maintaining high efficiency. Note that for a fair comparison, we use the same hypothesized depth number (even) as Vis-MVSNet. Thus, D_p^{s-1} is not sampled to make the sampling symmetric. It is reasonable to adopt any progression with increasing intervals. Here we choose the geometric progression T to approximate the error distribution instead of other sequences for the following two reasons. The first reason is that it is both trivial and easy to implement. The second reason is that compared with the arithmetic sequence, the interval in the geometric progression increases much faster and thus can mimic the Gaussian-like noise more accurately.

Table 1: Adopted datasets

Dataset	Indoor scenes	Outdoor scenes	High resolution
TNT [10]	✔	✔	✔
BlendedMVS [28]	✔	✔	✔
ETH3D [19]	✔	✔	✔
DTU [1]	✔	✖	✖

Training setting Following VisMVSNet, we train IS-MVSNet on two datasets: BlendedMVS and DTU. When the model is evaluated on the DTU's testing set, we train IS-MVSNet on DTU's training set; otherwise, we train it on the BlendedMVS's training set. While training, the image resolution is fixed at $640\times$

512, the number of source views is three, and the total number of stages is three. We sample 32 hypothesized depths with equivalent intervals in the first stage. While in the second and third stages, the depth intervals are determined following our novel sampling strategy described in Sec. 3.1, and the numbers of hypothesized depths are 16 and 8, respectively. We use an Adam optimizer to train the model for ten epochs. The batch size is four, and the initial learning rate is 10^{-3}. The learning rate is decayed by a factor of 0.5 in epochs 6, 8, and 9, respectively.

Evaluation setting We evaluate IS-MVSNet on three datasets: Tanks & Temples, ETH3D, and DTU, without fine-tuning. When synthesizing the point clouds, we adopt the dynamic consistency checking approach [25].

Tanks & Temples Dataset When predicting the depth maps, the number of source views is seven, the minimum consistency among views is four, the input image size is 1920×1056, and the estimated $k^* = 10$. Thus, the finest depth resolution is $10\times$ promoted. As shown in Tab. 2, IS-MVSNet surpasses all published methods on the intermediate benchmark. Note that IS-MVSNet is superior to the SOTA method Vis-MVSNet in nearly all scenes. At the same time, IS-MVSNet also achieves a higher F-score than nearly all published learning-based methods on the advanced dataset. Although ACMP [22] achieves a higher F-score on the advanced dataset, it does not hurt our importance sampling's superiority. This is because ACMP's advantage comes from its use of planar information, which does not conflict with our sampling strategy. Without the planar information, ACMP degrades to ACMM [21], which performs worse than our method on both datasets. The point cloud generated by our algorithm is of high reconstruction quality and precise details.

Table 2: F-score (higher is better) results on the Tanks & Temples [10].

	Intermediate set %									Advanced set %						
Method	**Mean**	Fam.	Franc.	Horse	Light.	M60	Pan.	Play.	Train	**Mean**	Audi.	Ballr.	Courtr.	Museum	Palace	Temple
COLMAP [18]	42.14	50.41	22.25	25.63	56.43	44.83	46.97	48.53	42.04	27.24	16.02	25.23	34.70	41.51	18.05	27.94
CVP-MVSNet [26]	54.03	76.50	47.74	36.34	55.12	57.28	54.28	57.43	47.54	-	-	-	-	-	-	-
CasMVSNet [6]	56.84	76.37	58.45	46.26	55.81	56.11	54.06	58.18	49.51	31.12	19.81	38.46	29.10	43.87	27.36	28.11
UCSNet [4]	54.83	76.09	53.16	43.03	54.00	55.60	51.49	57.38	47.89	-	-	-	-	-	-	-
Vis-MVSNet [29]	60.03	77.40	60.23	47.07	**63.44**	62.21	57.28	60.54	52.07	-	-	-	-	-	-	-
ACMM [21]	57.27	69.24	51.45	46.97	63.20	55.07	57.64	60.08	54.48	34.02	23.41	32.91	41.17	48.13	23.87	34.60
ACMP [22]	58.41	70.31	54.06	**54.11**	61.65	54.16	57.60	58.12	**57.25**	**37.44**	**30.12**	34.68	**44.58**	**50.64**	27.20	37.43
PatchmatchNet [20]	53.15	66.99	52.64	43.24	54.87	52.87	49.54	54.21	50.81	32.31	23.69	37.73	30.04	41.80	28.31	32.29
Ours	**62.82**	**79.92**	**62.05**	52.54	62.68	**63.65**	**62.57**	**62.94**	56.21	34.87	20.54	**39.88**	33.07	47.73	**30.12**	**37.91**

ETH3D Dataset The number of source views is seven, the minimum consistency among views is four, the input image size is 3072×2048, and the estimated $k^* = 6$. Thus, the finest depth resolution is $6\times$ promoted. As shown in Tab. 3, IS-MVSNet offers significant advantages over the learning-based SOTAs PVS-Net [23] and PatchmatchNet [20]. Although ACMP shows a higher F-score on the

training set, our method performs better on the testing set, which is obviously more important than the training set.

Table 3: F-score ↑ on the ETH3D high-res set at evaluation threshold 2cm.

Method	Training set (%)	Testing set (%)
Gipuma [5]	36.48	45.18
COLMAP [18]	67.66	73.01
PVSNet [23]	67.48	72.08
PatchmatchNet [20]	64.21	73.12
ACMP [22]	**79.79**	81.51
Ours	73.33	**83.15**

Table 4: Overall distance (mm), accuracy distance (mm), and completeness distance (mm) on the DTU testing set. All three metrics are preferred to be smaller.

Method	Overall distance	Accuracy distance	Completeness distance
COLMAP [18]	0.532	0.400	0.664
MVSNet [27]	0.462	0.396	0.527
VisMVSNet [29]	0.365	0.369	0.361
UCSNet [4]	**0.344**	0.338	**0.349**
CasMVSNet [6]	0.355	**0.325**	0.385
Ours	0.355	0.351	0.359

DTU Dataset The number of source views is four, the minimum consistency among views is five, the input image size is 1152×864, and the estimated $k^* = 20$. Thus, the finest depth resolution is $20\times$ promoted. The $20\times$ promotion is compared to the finest stage of Vis-MVSNet (a typical uniform-sampling-based method). For example, Vis-MVSNet's finest depth resolution on DTU is 2.65 mm, while our method is 0.13 mm. It is hard to quantitatively compare to UCSNet and UASNet because their resolutions rely on network predictions. Still, we generalize better and are more stable, as stated in Sec. 2.

Following Vis-MVSNet, we predict depth maps of half sizes, while other mentioned methods are in full sizes. Since objects in DTU are pretty small, the depth maps require higher plane resolution. In consequence, our improvements on TNT are more significant than on DTU. As shown in Tab. 4, our method outperforms Vis-MVSNet, where all the improvements come from our sampling strategy. Although UCSNet shows a better overall distance, its advantage relies on the depth range determination strategy, which does not conflict with our depth-range-agnostic sampling algorithm.

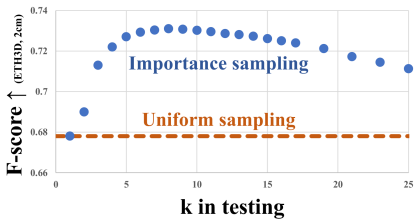
5 Ablation study

5.1 Error distribution

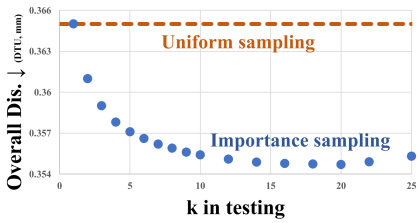
Without adopting the importance sampling strategy, we measure the coarse stage's error distributions on BlendedMVS and DTU and find out the distribution is indeed uni-modal. We calculate the per-pixel error as the predicted depth's difference to the ground truth: $\delta = d_{gt} - d_p$. Moreover, the error distribution concentrates around 0. Thus, it is reasonable to trust the last stage's prediction and sample more densely around it.

5.2 The effectiveness of k

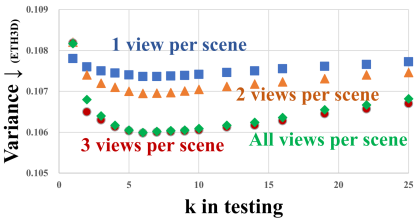
To analyze k's impact on the model's performance, we test IS-MVSNet with different k. The performance to k curve is shown in Fig. 4a and Fig. 4b. When $k = 1$, the importance sampling is equivalent to the uniform sampling. As k increases, the performance first gets better, then gradually decreases. It can be observed that an extensive range of k brings about significant improvement.



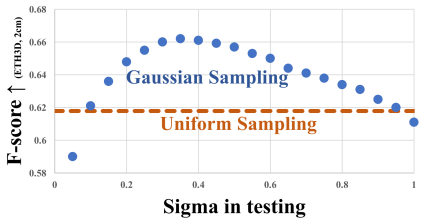
(a) F-score ↑ on ETH3D



(b) Overall distance ↓ (mm) on DTU



(c) Variance ↓ estimation with varying k



(d) F-score ↑ with varying σ (Gaussian)

Fig. 4: **Ablation studies.** (a) and (b) demonstrate that importance sampling is superior to uniform sampling under a wide range of k. (c) reveals that two to three samples per scene are enough to estimate k accurately. (d) shows that importance sampling does not rely on the probability approximation function because continuous Gaussian sampling also outperforms uniform sampling.

Reliability of coarse prediction On the ETH3D training set, the F-score at threshold 50 cm is as high as 0.96, 0.97, and 0.98 at stages $1 \rightarrow 3$, respectively; on DTU, the overall distance is as low as 0.73, 0.41, and 0.35 mm at stages $1 \rightarrow 3$, respectively. Regarding the 2 cm F-score measured on the ETH3D training set, uniform sampling achieves 0.14, 0.32, and 0.44 at stages $1 \rightarrow 3$, respectively, while our sampling strategy achieves 0.14, 0.41, and 0.58 at stages $1 \rightarrow 3$, respectively. These facts suggest that the coarse prediction on real datasets is reliable, and our method works well in reality.

5.3 The necessity of interval sequence sampling

Although i.i.d. sampling is straightforward, it suffers from the local consistency problem mentioned in Sec. 3.1. Empirically, converting the i.i.d uniform sampling to identical interval sampling leads to 6.4% F-score improvement on the ETH3D high-res training set. Moreover, as shown in Fig. 4d, i.i.d. sampling following Gaussian distribution with a proper variance σ significantly outperforms uniform distribution. These facts suggest that sampling following an increasing interval sequence is beneficial and necessary.

5.4 Unsupervised k selection

To validate the effectiveness of our unsupervised k selection algorithm, we fix the weights and use different k to generate the point clouds. It can be observed in Fig. 4a and Fig. 4c that the variance curve matches very well with the F-score curve. The minimum variance on the whole ETH3D dataset occurs when $k = 6$, exactly where the highest F-score is achieved according to Fig. 4a. Moreover, in Fig. 4c, we show that randomly evaluating two reference images from each scene is enough for variance estimation. Thus, our hyper-parameter selection algorithm is lightweight.

6 Conclusion

This paper presents an effective importance-sampling-based multi-view stereo network and the corresponding hyper-parameter estimation algorithm. Both theoretical analysis and extensive experiments strongly prove our method's superiority. Although, like other coarse-to-fine models, our model is limited to corner cases, in which the coarse prediction is too far from the ground truth. Our depth sampling and hyper-parameter estimation techniques could benefit most coarse-to-fine solutions.

Acknowledgments We gratefully acknowledge the support from MindSpore.

References

1. Aanæs, H., Jensen, R.R., Vogiatzis, G., Tola, E., Dahl, A.B.: Large-scale data for multiple-view stereopsis. International Journal of Computer Vision **120**(2), 153–168 (2016)
2. Bhat, S.F., Alhashim, I., Wonka, P.: Adabins: Depth estimation using adaptive bins. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4009–4018 (2021)
3. Chen, R., Han, S., Xu, J., Su, H.: Point-based multi-view stereo network. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1538–1547 (2019)
4. Cheng, S., Xu, Z., Zhu, S., Li, Z., Li, L.E., Ramamoorthi, R., Su, H.: Deep stereo using adaptive thin volume representation with uncertainty awareness. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2524–2534 (2020)
5. Galliani, S., Lasinger, K., Schindler, K.: Massively parallel multiview stereopsis by surface normal diffusion. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 873–881 (2015)
6. Gu, X., Fan, Z., Zhu, S., Dai, Z., Tan, F., Tan, P.: Cascade cost volume for high-resolution multi-view stereo and stereo matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2495–2504 (2020)
7. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
8. Ji, M., Gall, J., Zheng, H., Liu, Y., Fang, L.: Surfacenet: An end-to-end 3d neural network for multiview stereopsis. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2307–2315 (2017)
9. Kallenberg, O., Kallenberg, O.: Foundations of modern probability, vol. 2. Springer (1997)
10. Knapitsch, A., Park, J., Zhou, Q.Y., Koltun, V.: Tanks and temples: Benchmarking large-scale scene reconstruction. ACM Transactions on Graphics (ToG) **36**(4), 1–13 (2017)
11. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems **25**, 1097–1105 (2012)
12. Li, D.: Statistical computing. Higher Education Press (2017)
13. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2117–2125 (2017)
14. Mao, Y., Liu, Z., Li, W., Dai, Y., Wang, Q., Kim, Y.T., Lee, H.S.: Uasnet: Uncertainty adaptive sampling network for deep stereo matching. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6311–6319 (2021)
15. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. In: European conference on computer vision. pp. 405–421. Springer (2020)
16. Okutomi, M., Kanade, T.: A multiple-baseline stereo. IEEE Trans. Pattern Anal. Mach. Intell. **15**(4), 353–363 (1993)
17. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: towards real-time object detection with region proposal networks. IEEE transactions on pattern analysis and machine intelligence **39**(6), 1137–1149 (2016)

18. Schönberger, J.L., Zheng, E., Frahm, J., Pollefeys, M.: Pixelwise view selection for unstructured multi-view stereo. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III. Lecture Notes in Computer Science, vol. 9907, pp. 501–518. Springer (2016)
19. Schöps, T., Schönberger, J.L., Galliani, S., Sattler, T., Schindler, K., Pollefeys, M., Geiger, A.: A multi-view stereo benchmark with high-resolution images and multi-camera videos. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
20. Wang, F., Galliani, S., Vogel, C., Speciale, P., Pollefeys, M.: Patchmatchnet: Learned multi-view patchmatch stereo. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021. pp. 14194–14203. Computer Vision Foundation / IEEE (2021)
21. Xu, Q., Tao, W.: Multi-scale geometric consistency guided multi-view stereo. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5483–5492 (2019)
22. Xu, Q., Tao, W.: Planar prior assisted patchmatch multi-view stereo. In: The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020. pp. 12516–12523. AAAI Press (2020)
23. Xu, Q., Tao, W.: Pvsnet: Pixelwise visibility-aware multi-view stereo network. arXiv preprint arXiv:2007.07714 (2020)
24. Yan, J., Wei, Z., Yi, H., Ding, M., Zhang, R., Chen, Y., Wang, G., Tai, Y.W.: Dense hybrid recurrent multi-view stereo net with dynamic consistency checking. In: European Conference on Computer Vision. pp. 674–689. Springer (2020)
25. Yan, J., Wei, Z., Yi, H., Ding, M., Zhang, R., Chen, Y., Wang, G., Tai, Y.W.: Dense hybrid recurrent multi-view stereo net with dynamic consistency checking. In: European Conference on Computer Vision. pp. 674–689. Springer (2020)
26. Yang, J., Mao, W., Alvarez, J.M., Liu, M.: Cost volume pyramid based depth inference for multi-view stereo. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4877–4886 (2020)
27. Yao, Y., Luo, Z., Li, S., Fang, T., Quan, L.: Mvsnet: Depth inference for unstructured multi-view stereo. In: Proceedings of the European Conference on Computer Vision (ECCV) (September 2018)
28. Yao, Y., Luo, Z., Li, S., Zhang, J., Ren, Y., Zhou, L., Fang, T., Quan, L.: Blendedmvs: A large-scale dataset for generalized multi-view stereo networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1790–1799 (2020)
29. Zhang, J., Yao, Y., Li, S., Luo, Z., Fang, T.: Visibility-aware multi-view stereo network. In: 31st British Machine Vision Conference 2020, BMVC 2020, Virtual Event, UK, September 7-10, 2020. BMVA Press (2020)