One at a Time: Progressive Multi-Step Volumetric Probability Learning for Reliable 3D Scene Perception

Bohan Li[1, 2], **Yasheng Sun**[3], **Jingxin Dong**[2], **Zheng Zhu**[4], **Jinming Liu**[1, 2], **Xin Jin**[2*], **Wenjun Zeng**[1, 2]
[1]Shanghai Jiao Tong University, Shanghai, China
[2]Ningbo Institute of Digital Twin, Eastern Institute of Technology, Ningbo, China
[3]Tokyo Institute of Technology, Tokyo, Japan
[4]PhiGent Robotics, Beijing, China
{bohan_li, jmliu206}@sjtu.edu.cn, sun.y.aj@m.titech.ac.jp,
jingxin_dong@outlook.com, zhengzhu@ieee.org, {jinxin, wenjunzengvp}@eias.ac.cn

Abstract

Numerous studies have investigated the pivotal role of reliable 3D volume representation in scene perception tasks, such as multi-view stereo (MVS) and semantic scene completion (SSC). They typically construct 3D probability volumes directly with geometric correspondence, attempting to fully address the scene perception tasks in a single forward pass. However, such a single-step solution makes it hard to learn accurate and convincing volumetric probability, especially in challenging regions like unexpected occlusions and complicated light reflections. Therefore, this paper proposes to decompose the complicated 3D volume representation learning into a sequence of generative steps to facilitate fine and reliable scene perception. Considering the recent advances achieved by strong generative diffusion models, we introduce a multi-step learning framework, dubbed as VPD, dedicated to progressively refining the Volumetric Probability in a Diffusion process. Specifically, we first build a coarse probability volume from input images with the off-the-shelf scene perception baselines, which is then conditioned as the basic geometry prior before being fed into a 3D diffusion UNet, to progressively achieve accurate probability distribution modeling. To handle the corner cases in challenging areas, a Confidence-Aware Contextual Collaboration (CACC) module is developed to correct the uncertain regions for reliable volumetric learning based on multi-scale contextual contents. Moreover, an Online Filtering (OF) strategy is designed to maintain representation consistency for stable diffusion sampling. Extensive experiments are conducted on scene perception tasks including multi-view stereo (MVS) and semantic scene completion (SSC), to validate the efficacy of our method in learning reliable volumetric representations. Notably, for the SSC task, our work stands out as the first to surpass LiDAR-based methods on the SemanticKITTI dataset.

Introduction

Obtaining a dependable 3D representation is of critical importance in the realm of computer vision, particularly for tasks involving 3D scene perception, such as multi-view stereo (MVS) (Yao et al. 2018; Chen et al. 2020; Zhang et al. 2020) and semantic scene completion (SSC) (Li et al. 2023b; Miao et al. 2023; Li et al. 2023a). The existing 2D-based approaches implicitly learned 3D features by harnessing contextual information (Mayer et al. 2016; Wang et al. 2020, 2021), which often struggle with precise geometric modeling due to the inherent ambiguity of 2D representations. On the other hand, some researchers have sought to enforce geometric constraints by utilizing 3D probability volumes to model correspondences across various depth hypothesis planes, which attracts growing attention (Yin, Darrell, and Yu 2019; Gu et al. 2020; Ding et al. 2022).

Nevertheless, many complex real-world scenarios, characterized by incomplete observations and intricate reflection conditions, pose substantial challenges when striving for precise geometric modeling. Existing 3D probability volume-based approaches have made strides by devising sophisticated architectures (Gu et al. 2020; Chen et al. 2020; Ding et al. 2022) and refining loss functions (Peng et al. 2022; Wang et al. 2022b) to acquire reliable probability volumes. However, these methods generally resolve the problem with a single-step approximation solution, imposing a substantially heavy burden on the learning process. To mitigate these learning challenges, another line of research has introduced GRU-based architectures (Yao et al. 2019; Wang et al. 2022a; Xu et al. 2023) to facilitate the acquisition of a dependable 3D volumetric representation through iterative refinement. Nevertheless, these approaches typically rely on 2D convolutional GRU mechanisms, which are susceptible to cumulative errors (Li et al. 2018; Mao and Sejdić 2022). This, in turn, motivates us to explore the potential of iterative refinement in the context of 3D volumetric probability.

Based on the above analysis, we propose a **Volumetric Probability Diffusion (VPD)** framework, which progressively models the volumetric probability and thus achieves reliable geometry estimation in the MVS and SSC tasks. As depicted in Fig. 1, the core idea is to *devise a multi-step learning scheme that models the probability volumes and progressively refine them*. Inspired by the powerful probability distribution modeling capabilities exhibited by generative diffusion models (Saharia et al. 2022; Müller et al. 2023), we propose a progressive optimization paradigm based on the diffusion process for reliable probability volume modeling. To leverage the geometry prior extracted

*Corresponding author.

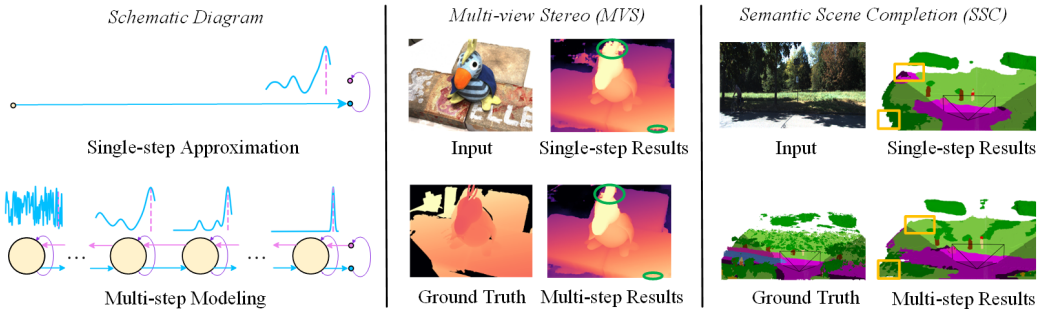

Figure 1: Comparison between single-step approximation and multi-step modeling for 3D scene perception tasks including multi-view stereo (MVS) and semantic scene completion (SSC). We demonstrate the qualitative results of these two methods. The multi-step modeling yields significantly more accurate and reliable results.

from the input images with pre-trained models, our VPD is conditioned with the extracted coarse volumes and contextual features to guide the diffusion progress. Specifically, the coarse volumes are employed as basic geometry prior, which is concatenated with the noisy input volume of the diffusion framework as prior volume condition. Despite the effectiveness of the prior volume condition in high-confidence regions, the low-confidence mismatch issue in challenging regions (e.g. non-Lambertian surfaces, thin structures and reflections) still exists, which impairs the learning of probability distribution approximation to the target volumes. Therefore, we further introduce a **Confidence-Aware Contextual Collaboration (CACC)** module to correct the uncertain regions of the predicted 3D volumes with rich contextual information. In detail, CACC first prunes the 3D volumes using confidence-aware filtering. Next, the fine-grained features and geometric details are retrieved from multi-scale contextual contents to complement the information in the low-confidence regions of the volumes. Moreover, to avoid perturbations in the diffusion sampling process, we introduce an **Online Filtering (OF)** strategy to maintain the consistency of the representations for a stable diffusion. In summary, the main contributions of this paper are listed as:

- We pinpoint the limitation of sing-step-based strategies, and correspondingly propose a novel Volumetric Probability Diffusion (VPD) framework, which fully exploits the strong generative ability of diffusion models for fine and reliable volumetric representation.
- We propose a Confidence-Aware Contextual Collaboration (CACC) module to enhance the reliability of volumetric learning in VPD. Additionally, we develop an Online Filtering (OF) strategy to maintain representation consistency during the reverse sampling process.
- Extensive experiments validate the effectiveness of our approach. We achieve state-of-the-art results on various scene perception tasks, including 1) MVS: DTU (Aanæs et al. 2016), BlendedMVS (Yao et al. 2020) and ScanNet (Dai et al. 2017); 2) SSC: SemanticKITTI (Behley et al. 2019). Notably, to the best of our knowledge, VPD is the first camera-based method that surpasses LiDAR-based methods on the SemanticKITTI.

Related Works

Learning-based 3D Scene Perception

With the development of learning-based methods, the quality of 3D representation for scene perception has been steadily improved (Okae et al. 2021; Xie et al. 2023). Recently, stereo matching has been explored in semantic scene completion (SSC) (Li et al. 2023b,a). In StereoScene (Li et al. 2023a), a stereo volume constructor is proposed to generate a geometric cost volume to enhance the perception of 3D scenarios. For multi-view stereo (MVS), a set of images are employed to construct 3D cost volumes with epipolar constrain (Gu et al. 2020; Ding et al. 2022; Peng et al. 2022). CasMVSNet (Gu et al. 2020) employs cascade cost volumes with different scales to form a coarse-to-fine depth estimation framework. TransMVSNet (Ding et al. 2022) leverages global context information with a feature matching transformer to exploit long-range aggregation across input images. Different from all the previous methods that try to approximate the ground truth in a single step, we propose to formulate depth estimation as progressive distribution modeling, which decomposes the issue into multiple steps to further improve performance in challenging scenarios.

Denoising Diffusion Models

Denoising diffusion models (DDMs) are a novel class of generative models derived from nonequilibrium thermodynamics (Sohl-Dickstein et al. 2015) and have achieved astounding results in the field of computer vision (Luo and Hu 2021; Rombach et al. 2022; Ramesh et al. 2021). DiffRF (Müller et al. 2023) adopts a set of posed images as additional conditions for radiance field synthesis with a rendering loss to resolve ambiguities. Different from the one-to-many mapping in the generation process of DiffRF, we employ the diffusion process as a one-to-one mapping, leveraging geometry prior for accurate and reliable 3D scene perception. DiffuStereo (Shao et al. 2022) leverages an iterative diffusion model to obtain highly accurate depth maps for automatic high-quality human reconstruction from sparse-view inputs as conditions. However, DiffuStereo directly refines the depth maps generated by the off-the-shelf algorithms, without fully exploring the geometric constraints in

the matching process. In contrast, we propose volumetric probability diffusion (VPD) to make full use of the correspondence distribution across different depth hypothesis planes, which is more advisable because the diffusion process excels at modeling distributions.

Methodology

In this work, we formulate the 3D perception in MVS and SSC tasks as multi-step conditional volumetric probability learning, and propose Volumetric Probability Diffusion (VPD). As shown in Figure 2, given input images, we first construct diffusion conditions with coarse probabilistic volumes and multi-scale contextual features extracted from off-the-shelf scene perception baselines. Next, we progressively estimate a refined volume over multiple steps by diffusing a noisy volume with the constructed conditions. The refined volumes are finally fed into the task-specific head to generate depth maps in MVS or occupancy grids in SSC. Please refer to the **Supplementary Material** for the details on the task-specific head. In detail, the proposed VPD mainly consists of the following components:

I. A **Volumetric Diffusion** model that is implemented with a 3D UNet (Ronneberger, Fischer, and Brox 2015). In the forward process, the target volumes are constructed from ground truth depth maps. In the reverse process, an Online Filtering (**OF**) strategy is further developed to maintain the unique peak distribution in the estimated volumes.

II. The diffusion conditions including the basic prior volume condition and the contextual feature condition constructed with the Confidence-Aware Contextual Collaboration (**CACC**) module.

Volumetric Diffusion

The standard generative diffusion models aim to form one-to-many mappings with a forward and reverse process. In our scenario, we employ a volumetric diffusion model to learn the parametric approximation to the target volume based on the guidance of conditions.

In the forward process, we construct target unimodel volume $\boldsymbol{y}_0$ from ground truth depth map d^{gt}, and progressively corrupt the target volume to $\boldsymbol{y}_T \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ in $\boldsymbol{T}$ time steps. In the reverse process, the 3D diffusion UNet estimates a refined volume $\tilde{\boldsymbol{y}}_0$ to approximate the target volume $\boldsymbol{y}_0$ from noisy input volume $\boldsymbol{y}_T$ and we consider conditions $\boldsymbol{x}$ to guide the estimation.

Volumetric Gaussian Forward Process. Given a ground truth depth map d^{gt}, we first construct the target volume $\boldsymbol{y}_0$ following the unimodel projection (Ding et al. 2022; Peng et al. 2022) along depth dimension D as diffusion input:

$$\boldsymbol{y}_0 = Project^{Uni}\left\{d^{gt}, Dim = D\right\}, \tag{1}$$

We gradually add noise on $\boldsymbol{y}_0$ to generate the noisy volume $\boldsymbol{y}_T$ over $\boldsymbol{T}$ steps following a discrete-time Markov chain. Given distribution of $\boldsymbol{y}_0$, the forward process can be characterized as:

$$q\left(\boldsymbol{y}_t \mid \boldsymbol{y}_0\right) = \mathcal{N}\left(\boldsymbol{y}_t \mid \sqrt{\bar{\alpha}_t}\boldsymbol{y}_0, \left(1-\bar{\alpha}_t\right)\boldsymbol{I}\right), \tag{2}$$

where $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$ and α_t is the pre-defined coefficient. $\mathcal{N}$ and $\boldsymbol{I}$ denote the normal distribution and the identical matrix, respectively.

Iterative Conditional Reverse Process. The conditional reverse sampling process is dedicated to iteratively denoise $\boldsymbol{y}_T$ with conditions to recover $\boldsymbol{y}_0$. Each step of the reverse process can be defined as conditional distribution transition (Saharia et al. 2022), which is formulated as:

$$p_\theta\left(\boldsymbol{y}_{0:T} \mid \boldsymbol{x}\right) = p\left(\boldsymbol{y}_T\right) \prod_{t=1}^{T} p_\theta\left(\boldsymbol{y}_{t-1} \mid \boldsymbol{y}_t, \boldsymbol{x}\right), \tag{3}$$

where p_θ represents the reverse function, $\boldsymbol{x}$ denotes the two conditions of the diffusion model.

Online Filtering. Since VPD is dedicated to learning for approximating the target volume $\boldsymbol{y}_0$ with unimodal distribution, we propose to filter the predicted $\boldsymbol{y}_t$ online at each iteration to suppress the perturbation caused by the generated multi-model representation before sending it to the next reverse sampling step. Our implementation is formed as:

$$\boldsymbol{y}'_t = Project^{Uni}\left\{WTA^D\left(\boldsymbol{y}_t\right), Dim = D\right\}, \tag{4}$$

where $Project^{Uni}$ denotes unimodel projection same as GT volume construction in Section . WTA^D represents *Winner-Takes-All* operation (Cheng et al. 2020), which maintains the unique peak along the depth dimension.

Condition Construction

In this section, we introduce the condition construction in VPD. As shown in Figure 2, we extract coarse probabilistic volumes and multi-scale contextual features from input images with the off-the-shelf MVS or SSC baselines (correspond to the Feature Net and the Volume Net in Figure 2). Next, we employ them as the prior volume condition and the contextual feature condition to constrain the learning of distribution transition in the diffusion process, respectively.

Coarse Volume Probabilization. We employ Coarse Volume Probabilization (CVP) to construct the coarse probabilistic volume, which is concatenated with the input noisy volume as the basic prior volume condition of the diffusion model. Given a coarse cost volume $\mathbf{V}_{cost}$ from baseline networks, we employ $softmax$ along the depth dimension for each pixel (h, w) in space to implement the volume probabilization, which is formally written as:

$$\mathbf{V}_{prob}^{h,w,m} = Softmax(\mathbf{V}_{cost}^{h,w,m}) = \frac{\exp(d_m^{h,w})}{\sum_{n=1}^{D_{max}} \exp(d_n^{h,w})}, \tag{5}$$

where $d_m^{h,w}$ represents cost value of m^{th} depth hypothesis plane ($1 < m < D_{max}$). D_{max} denotes the number of depth hypothesis planes. For multi-view stereo (MVS), we adopt regularized cost volumes (Gu et al. 2020; Long et al. 2021; Ding et al. 2022; Peng et al. 2022) as the volume $\mathbf{V}_{cost}$, while the geometric cost volumes (Li et al. 2023a) are employed for semantic scene completion (SSC).

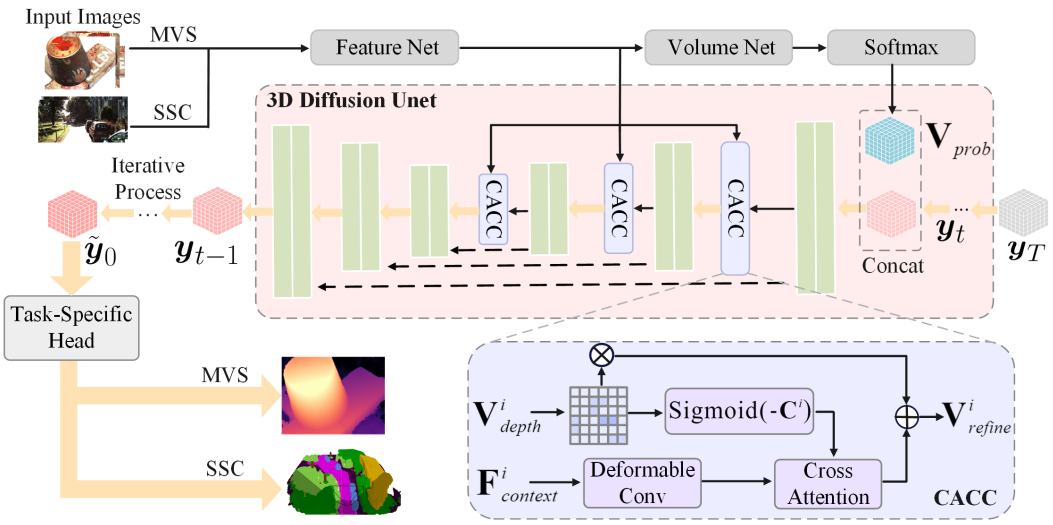

Figure 2: Our volumetric probability diffusion (VPD). Given input images, We first extract multi-scale contextual features $\mathbf{F}^i_{context}$ and coarse probabilistic volumes $\mathbf{V}_{prob}$ with off-the-shelf scene perception baselines. Then, $\mathbf{V}_{prob}$ concatenated with the random noisy volume $\boldsymbol{y}_t$ as input is fed into the 3D diffusion UNet for refinement, while $\mathbf{F}^i_{context}$ are employed as conditions in CACC to continuously refine the depth volume $\mathbf{V}^i_{depth}$ in the 3D UNet. Following an iterative process, we progressively estimate a refined volume $\tilde{\boldsymbol{y}}_0$ over multiple steps with diffusion. The estimated volumes are finally fed to the task-specific head to generate depth maps for MVS or occupancy grids for SSC.

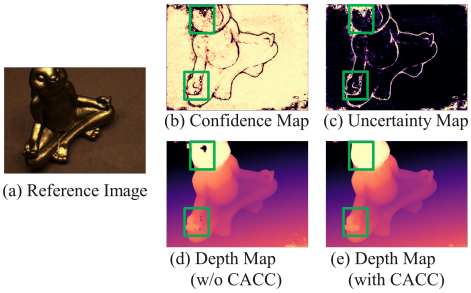

Figure 3: Visualization results in the confidence-aware contextual collaboration (CACC) module. The confidence map and the uncertainty map illustrate the regions with poor estimation, which are effectively refined with CACC.

Confidence-Aware Contextual Collaboration. Although the CVP provides basic geometry prior, it is still hard to achieve compelling results, especially in challenging regions like occlusions, reflections, textureless regions, etc. Thus, we propose a Confidence-Aware Contextual Collaboration (CACC) module to further apply continuous refinement with the contextual feature condition on the estimated volumes.

The overall structure of CACC is shown in Figure 2. Given a depth volume $\mathbf{V}^i_{depth} \in \mathbb{R}^{C\times D\times H\times W}$ in i^{th} downsample block of the 3D UNet (i.e., diffusion model) and i^{th} scale contextual features $\mathbf{F}^i_{context} \in \mathbb{R}^{C'\times H\times W}$ from feature extraction networks, our goal is to retrieval reliable multi-scale contextual features from $\mathbf{F}^i_{context}$, and refine $\mathbf{V}^i_{depth}$ according to the confidence information along the spatial dimension. It is worth noting that we directly obtain multi-scale contextual features from the off-the-shelf baseline networks for computational efficiency.

Specifically, we form a confidence map $\mathbf{C}^i \in \mathbb{R}^{C\times H\times W}$ by checking the highest probability value among all depth hypothesis planes across the depth dimension. Next, we reverse the values in $\mathbf{C}^i$ to obtain query Q^i for cross attention that measures the matching uncertainty in $\mathbf{V}^i_{depth}$:

$$Q^i = Sigmoid(-\mathbf{C}^i) = Sigmoid\left\{-\left(WTA^D\left(\mathbf{V}^i\right)\right)\right\}, \tag{6}$$

where WTA^D denotes *Winner-Takes-All* along depth dimension. To generate key K^i and value V^i, we apply deformable convolution on the corresponding contextual features $\mathbf{F}^i_{context}$ for efficient geometric transformation modeling and receptive fields adaption. For each point $\mathbf{p}$ on the contextual features $\mathbf{F}^i_{context}$, the process is formulated as:

$$K^i(\mathbf{p}), V^i(\mathbf{p}) = S\left\{\sum_{c=0}^{C'-1} W\cdot \mathbf{F}^i_{context}\left((\mathbf{p})+\Delta_{(\mathbf{p})}, \mathbf{c}\right)\right\}, \tag{7}$$

where W and $\Delta_{(\mathbf{p})}$ denotes the deformable weight and learnable offset, respectively. S represents splitting the input into halves along the feature channel. To reduce computation cost, we adopt linear attention (Shen et al. 2021) as:

$$\mathbf{F}^i_{conf} = Atten(Q^i, K^i, V^i) = \phi_q(Q^i)\left(\phi_k(K^i)^T V^i\right), \tag{8}$$

where $\mathbf{F}^i_{conf}$ represents confidence-aware context. ϕ_q and ϕ_k are $softmax$ operations along each row and column of the input matrix, respectively. In this way, the relevant information of contextual features is retrieved according to the matching uncertainty of the depth volume $\mathbf{V}^i_{depth}$.

Subsequently, we implement element-wise multiplication between the depth volume $\mathbf{V}^i_{depth}$ and confidence map $\mathbf{C}^i$ to obtain a filtered volume. To match in dimension, $\mathbf{F}^i_{conf}$ is projected into 3D contextual volume $\mathbf{V}^i_{context}$ before adding to the filtered volume following lift operation (Philion and Fidler 2020). Finally, the refined volume is constructed by element-wise summation between the filtered volume and the contextual volume:

$$\mathbf{V}^i_{refine} = \mathbf{V}^i_{depth} \odot \mathbf{C}^i + \mathbf{V}^i_{context}, \tag{9}$$

where $\odot$ denotes element-wise multiplication. Note that CACC is applied on each downsample block of the 3D UNet with different dimension sizes. Through the refinement operation of CACC on the depth volume, volumetric distribution in high-confidence regions is retained, while that in low-confidence regions is optimized with multi-scale contexts.

In Figure 3, we visualize the confidence map $\mathbf{C}^i$, uncertainty map $Sigmoid(-\mathbf{C}^i)$, estimated depth map without CACC and estimated depth map with CACC. It can be seen that the model without CACC struggles to achieve compelling results in challenging regions (e.g. object boundaries, low-texture regions). The confidence map and the uncertainty map illustrate the regions with poor estimation, which are effectively refined by retrieving information from the contextual features with CACC.

Training Objective

In this work, we adopt an end-to-end joint training pipeline for the whole framework, and our training objective is to optimize the volumetric diffusion model for target volume approximation. Different loss functions are applied to achieve the object according to the representations of coarse probabilistic volumes (in Section), which is consistent with baseline networks (Gu et al. 2020; Peng et al. 2022; Ding et al. 2022; Long et al. 2021; Li et al. 2023a).

Experiments

We evaluate the proposed Volumetric Probability Diffusion (VPD) on the 3D scene perception tasks of multi-view stereo (MVS) and semantic scene completion (SSC).

Multi-view Stereo (MVS)

Datasets. DTU (Aanæs et al. 2016) is a large-scale indoor dataset, which consists of 124 different scenes with 7 different illumination conditions. We split the dataset into training, validation, and test set following the setting of MVSNet (Yao et al. 2018). BlendedMVS (Yao et al. 2020) dataset is a synthetic dataset that consists of 106 training scans and 7 validation scans. ScanNet (Dai et al. 2017) is an RGB-D video dataset that consists of more than 1600 scans, annotated with depth maps and camera poses.

Implementation Details. Our model is implemented on the Pytorch platform with 4 NVIDIA A100 GPUs. We train our model for 16 epochs on the DTU dataset and 7 epochs on the ScanNet dataset. For the BlendedMVS dataset, we implement tests using the model trained on the DTU dataset to evaluate the generalization ability. The initial learning rate is set to 2.5×10^{-5}, which decays following the same strategy of baseline networks. The batch size is set to 12 and Adam is adopted as optimizer. The diffusion forward step $\boldsymbol{T}$ is set to 1000 and we adopt 4 iterations in the reverse process.

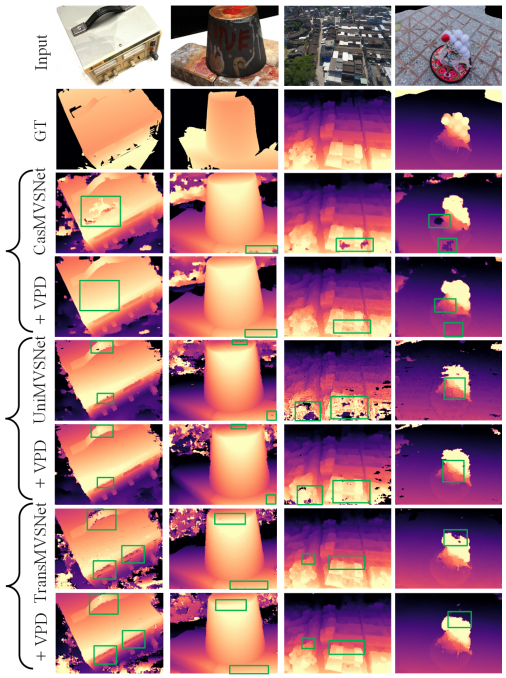

Figure 4: Qualitative results for MVS on DTU test set (left two columns) and BlendedMVS validation set (right two columns). Our approach consistently generates more complete predictions in low-texture regions, as well as more accurate and fine-grained results in thin-structure regions.

Performance. For quantitative evaluation, we conduct standard metrics (Eigen, Puhrsch, and Fergus 2014; Long et al. 2021), including absolute relative error (Abs Rel), absolute error (Abs), square relative error (Sq Rel), root mean square error in linear scale (RMSE), threshold distance error (Th) and inlier ratios (δ). As reported in Table 1, our method shows significant improvements compared to TransMVSNet, reducing Abs by 31.46%.

We evaluate the zero-shot generalization ability of our method from DTU to BlendedMVS validation set without fine-tuning. As shown in Table 2, our method has a notable performance gain compared to baseline networks, which demonstrates that our approach generalizes well across different datasets without post-processing. Table 3 shows quantitative results on the ScanNet test set. ESTD (Long et al. 2021) with VPD outperforms other methods in terms of accuracy, which indicates that our method has strong modeling capability for temporal cost volumes.

Moreover, We visualize qualitative results on the DTU test set and BlendedMVS validation set in Figure 4. Our approach significantly enhances outcomes from baseline models, generating more complete depth maps with heightened accuracy, particularly in challenging areas like object boundaries and repetitive patterns.

Method	Abs Rel↓	Abs↓	Sq Rel↓	Th8↓	Th20↓	$\delta_1 < 1.25$↑	$\delta_2 < 1.25^2$↑
MVSNet	0.0139	11.5502	2.0383	0.1378	0.0932	0.9845	0.9966
CasMVSNet	0.0097	7.4381	1.6300	0.0872	0.0570	0.9887	0.9976
UniMVSNet	0.0095	7.2756	1.3163	0.0837	0.0547	0.9858	0.9934
TransMVSNet	0.0094	7.2096	1.2712	0.0842	0.0541	0.9905	0.9982
TransMVSNet+VPD	**0.0067**	**4.9416**	**0.9918**	**0.0510**	**0.0333**	**0.9918**	**0.9984**

Table 1: Quantitative results on DTU test set for MVS. The best performers are marked bold.

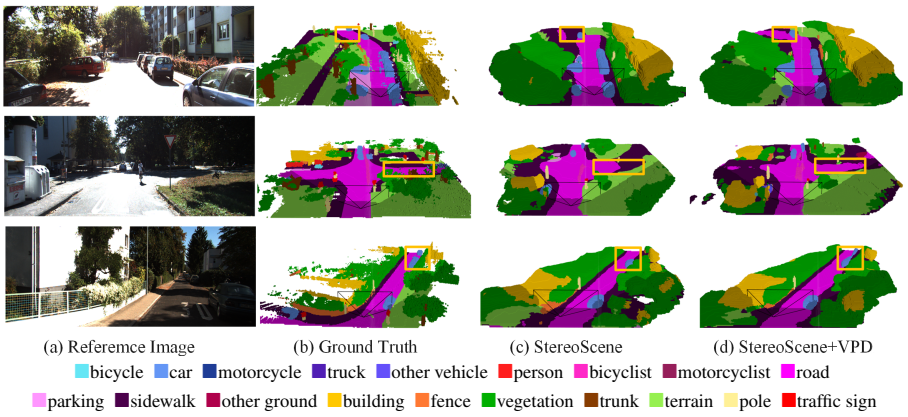

Figure 5: Qualitative results for SSC on SemanticKITTI validation set. The shadow areas denote unseen scenery out of the camera's field of view. Our proposed VPD improves the performance of the baseline in challenging regions.

Method	Abs Rel↓	Abs↓	$\delta_1 < 1.25$↑
MVSNet	0.0915	2.6554	0.9135
CasMVSNet	0.0665	1.7102	0.9349
UniMVSNet	0.0825	1.8744	0.9320
TransMVSNet	0.0657	1.9216	0.9402
CasMVSNet+VPD	0.0404	1.4122	0.9604
UniMVSNet+VPD	0.0496	1.3128	**0.9425**
TransMVSNet+VPD	**0.0376**	**1.2267**	0.9596

Table 2: Zero-shot generalization from DTU to BlendedMVS validation set for MVS.

Methods	Abs Rel↓	Abs↓	Sq Rel↓	RMSE↓	$\delta_1 < 1.25$↑
MVDepth	0.1167	0.2301	0.0596	0.3236	0.8453
DPSNet	0.1200	0.2104	0.0688	0.3139	0.8640
DELTAS	0.0915	0.1710	0.0327	0.2390	0.9147
NRGBD	0.1013	0.1657	0.0502	0.2500	0.9160
PairNet	0.0895	0.1709	0.0615	0.2734	0.9172
ESTD	0.0812	0.1505	0.0298	0.2199	0.9313
ESTD+VPD	**0.0753**	**0.1497**	**0.0237**	**0.2149**	**0.9483**

Table 3: Quantitative results on ScanNet test set for MVS.

Semantic Scene Completion (SSC)

Datasets. SemanticKITTI (Behley et al. 2019) is a popular semantic scene completion dataset, which contains 22 outdoor driving scenes. SemanticKITTI holds LiDAR annotations that are voxelized as 256×256×32 grid of 0.2m voxels. The target voxel grids are labeled with 21 classes (1 free, 1 unknown, and 19 semantics). Following (Li et al. 2023a), we only adopt RGB images of the dataset as inputs.

Implementation Details. We extend StereoScene (Li et al. 2023a) with our proposed VPD for SSC evaluation. Specifically, the geometric cost volume in StereoScene is leveraged as the coarse volume. The model is trained for 30 epochs with a learning rate of 2.5×10^{-5}. AdamW is adopted as a training optimizer following (Li et al. 2023a).

Performance. For quantitative evaluation, we adopt mIoU (mean Intersection over Union) to account for the SSC task. We compare our method with other state-of-the-art SSC networks: (1) camera-based methods including MonoScene (Cao and de Charette 2022), VoxFormer (Li et al. 2023b) and StereoScene (Li et al. 2023a), (2) LiDAR-based method of SSCNet (Song et al. 2017). As shown in Table 4, our method surpasses StereoScene by 6.18% in terms of mIoU, which demonstrates the application of VPD effectively improves the accuracy of depth estimation and thereby enhances the performance of semantic scene completion. It's worth noting that our method even surpasses the LiDAR-based method of SSCNet in terms of mIoU. Figure 5 visualizes the qualitative results, our method produces more accurate and complete results in large-scale driving scenarios compared with StereoScene.

Method	road (15.30%)	sidewalk (11.13%)	parking (1.12%)	other-grnd (0.56%)	building (14.1%)	car (3.92%)	truck (0.16%)	bicycle (0.03%)	motorcycle (0.03%)	other-veh. (0.20%)	vegetation (39.3%)	trunk (0.51%)	terrain (9.17%)	person (0.07%)	bicyclist (0.07%)	motorcyclist. (0.05%)	fence (3.90%)	pole (0.29%)	traf.-sign (0.08%)	mIoU
MonoScene	54.70	27.10	24.80	5.70	14.40	18.80	3.30	0.50	0.70	4.40	14.90	2.40	19.50	1.00	1.40	0.40	11.10	3.30	2.10	11.08
VoxFormer-S	53.90	25.30	21.10	5.60	19.80	20.80	3.50	1.00	0.70	3.70	22.40	7.50	21.30	1.40	**2.60**	0.20	11.10	5.10	4.90	12.20
VoxFormer-T	54.10	26.90	25.10	7.30	23.50	21.70	3.60	1.90	1.60	4.10	24.40	8.10	24.20	1.60	1.10	0.00	13.10	6.60	5.70	13.41
SSCNet	51.15	30.76	27.12	6.44	**34.53**	24.26	1.18	0.54	0.78	4.43	**35.25**	1.18	29.01	0.25	0.25	**0.78**	**19.87**	**13.10**	6.73	16.14
StereoScene	**61.90**	31.20	**30.70**	10.70	24.20	22.80	2.80	3.40	2.40	6.10	23.80	8.40	27.00	**2.90**	2.20	0.50	16.50	7.00	7.20	15.36
StereoScene+VPD	61.76	**32.41**	20.39	**11.11**	24.43	**32.12**	**7.33**	**3.74**	**2.53**	**9.26**	25.87	**8.89**	**37.68**	2.02	0.99	0.00	10.09	11.72	**7.53**	**16.31**

Table 4: Quantitative results on SemanticKITTI test set against the state-of-the-art SSC methods (higher is better). Our method even surpasses temporal stereo-based VoxFormer-T and LiDAR-based SSCNet in terms of mIoU.

Model Settings			Evaluation Metrics	
CVP	CACC	OF	Abs Rel↓	Abs↓
			0.0097	7.4381
✓			0.0083	6.3111
✓	✓		0.0078	5.9829
✓	✓	✓	**0.0075**	**5.7275**

Table 5: Ablation study for different model settings. The basic CasMVSNet is employed as the baseline setting.

Ablation Study

We conduct extensive ablation studies on DTU test set with different model settings. We extend CasMVSNet with VPD of different settings for the evaluation.

Effect of Model Settings. For the experiment in the first row, we adopt basic CasMVSNet as the baseline setting. The reverse iterations are set to 4 unless otherwise stated. As shown in Table 5, the VPD framework brings significant improvement with the basic prior volume condition of CVP. The CACC and OF obviously enhance the depth estimation performance with reliable volumetric learning, reducing Abs by 5.20% and 4.27%, respectively.

Efficiency Analyse. We report the running time and memory consumption of several schemes that are built upon different baselines equipped with our proposed VPD on the NVIDIA A100 GPU, which are detailed in Table 6. It's worth noting that our method could effectively achieve compelling performance gains with acceptable time consumption. As shown in the table, the proposed VPD delivers satisfactory performance improvements with relatively slight increments in time consumption, while the memory consumption of our proposed method remains constant regardless of the number of reverse steps. In addition, the performance gain of more than 4 reverse steps with more running time is not obvious, thus we adopt 4 steps as the default setting to balance efficiency and effectiveness.

Method	Abs↓	Time(s)↓	Mem(G)↓
MVSNet	11.55	0.75	12.74
CasMVSNet	7.44	**0.41**	**6.32**
CasMVSNet+VPD (4 Steps)	5.73	0.69	8.11
TransMVSNet	7.21	0.87	8.27
TransMVSNet+VPD (1 Step)	6.27	0.96	10.48
TransMVSNet+VPD (2 Steps)	5.40	1.04	10.48
TransMVSNet+VPD (4 Steps)	4.94	1.24	10.48
TransMVSNet+VPD (6 Steps)	4.79	1.39	10.48
TransMVSNet+VPD (8 Steps)	**4.72**	1.52	10.48

Table 6: Running time and memory consumption of the proposed method. We adopt 4 reverse steps as the default setting to balance efficiency and effectiveness.

Conclusion

In this work, we propose a novel framework of Volumetric Probability Diffusion (VPD) for 3D scene perception tasks including MVS and SSC. Different from previous single-step approximation solutions, we employ multi-step generative diffusion to progressively model volumetric probability for more reliable estimation. Specifically, we introduce a Confidence-Aware Contextual Collaboration (CACC) module to correct the uncertain regions for reliable target volume approximation. In the sampling process, we develop an Online Filtering (OF) strategy to maintain consistency in estimated volume representations. Our method achieves state-of-the-art performance on multiple MVS/SSC benchmarks and even surpasses the LiDAR-based method with only camera-based inputs.

Acknowledgements

This paper is supported in part by NSFC under Grant 62302246 and ZJNSFC under Grant LQ23F010008.

References

Aanæs, H.; Jensen, R.; Vogiatzis, G.; Tola, E.; and Dahl, A. 2016. Large-Scale Data for Multiple-View Stereopsis. *International Journal of Computer Vision*.

Behley, J.; Garbade, M.; Milioto, A.; Quenzel, J.; Behnke, S.; Stachniss, C.; and Gall, J. 2019. Semantickitti: A dataset for semantic scene understanding of lidar sequences. In *ICCV*.

Cao, A.-Q.; and de Charette, R. 2022. Monoscene: Monocular 3d semantic scene completion. In *CVPR*.

Chen, P.-H.; Yang, H.-C.; Chen, K.-W.; and Chen, Y.-S. 2020. Mvsnet++: Learning depth-based attention pyramid features for multi-view stereo. *IEEE Transactions on Image Processing*, 29.

Cheng, X.; Zhong, Y.; Harandi, M.; Dai, Y.; Chang, X.; Li, H.; Drummond, T.; and Ge, Z. 2020. Hierarchical neural architecture search for deep stereo matching. *NeurIPS*, 33.

Dai, A.; Chang, A. X.; Savva, M.; Halber, M.; Funkhouser, T.; and Nießner, M. 2017. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *CVPR*.

Ding, Y.; Yuan, W.; Zhu, Q.; Zhang, H.; Liu, X.; Wang, Y.; and Liu, X. 2022. Transmvsnet: Global context-aware multi-view stereo network with transformers. In *CVPR*.

Eigen, D.; Puhrsch, C.; and Fergus, R. 2014. Depth map prediction from a single image using a multi-scale deep network. *NeurIPS*, 27.

Gu, X.; Fan, Z.; Zhu, S.; Dai, Z.; Tan, F.; and Tan, P. 2020. Cascade cost volume for high-resolution multi-view stereo and stereo matching. In *CVPR*.

Li, B.; Sun, Y.; Jin, X.; Zeng, W.; Zhu, Z.; Wang, X.; Zhang, Y.; Okae, J.; Xiao, H.; and Du, D. 2023a. StereoScene: BEV-Assisted Stereo Matching Empowers 3D Semantic Scene Completion. *arXiv preprint arXiv:2303.13959*.

Li, Y.; Yu, R.; Shahabi, C.; and Liu, Y. 2018. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. In *ICLR*.

Li, Y.; Yu, Z.; Choy, C.; Xiao, C.; Alvarez, J. M.; Fidler, S.; Feng, C.; and Anandkumar, A. 2023b. VoxFormer: Sparse Voxel Transformer for Camera-based 3D Semantic Scene Completion. *CVPR*.

Long, X.; Liu, L.; Li, W.; Theobalt, C.; and Wang, W. 2021. Multi-view depth estimation using epipolar spatio-temporal networks. In *CVPR*, 8258–8267.

Luo, S.; and Hu, W. 2021. Diffusion probabilistic models for 3d point cloud generation. In *CVPR*, 2837–2845.

Mao, S.; and Sejdić, E. 2022. A review of recurrent neural network-based methods in computational physiology. *IEEE Transactions on Neural Networks and Learning Systems*.

Mayer, N.; Ilg, E.; Hausser, P.; Fischer, P.; Cremers, D.; Dosovitskiy, A.; and Brox, T. 2016. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *CVPR*.

Miao, R.; Liu, W.; Chen, M.; Gong, Z.; Xu, W.; Hu, C.; and Zhou, S. 2023. OccDepth: A Depth-Aware Method for 3D Semantic Scene Completion. *arXiv preprint arXiv:2302.13540*.

Müller, N.; Siddiqui, Y.; Porzi, L.; Bulò, S. R.; Kontschieder, P.; and Nießner, M. 2023. DiffRF: Rendering-Guided 3D Radiance Field Diffusion. In *CVPR*.

Okae, J.; Li, B.; Du, J.; and Hu, Y. 2021. Robust Scale-Aware Stereo Matching Network. *IEEE Transactions on Artificial Intelligence*.

Peng, R.; Wang, R.; Wang, Z.; Lai, Y.; and Wang, R. 2022. Rethinking depth estimation for multi-view stereo: A unified representation. In *CVPR*.

Philion, J.; and Fidler, S. 2020. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In *ECCV*.

Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Radford, A.; Chen, M.; and Sutskever, I. 2021. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, 8821–8831. PMLR.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *CVPR*, 10684–10695.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*.

Saharia, C.; Ho, J.; Chan, W.; Salimans, T.; Fleet, D. J.; and Norouzi, M. 2022. Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Shao, R.; Zheng, Z.; Zhang, H.; Sun, J.; and Liu, Y. 2022. Diffustereo: High quality human reconstruction via diffusion-based stereo using sparse cameras. In *ECCV 2022*.

Shen, Z.; Zhang, M.; Zhao, H.; Yi, S.; and Li, H. 2021. Efficient attention: Attention with linear complexities. In *WACV*.

Sohl-Dickstein, J.; Weiss, E.; Maheswaranathan, N.; and Ganguli, S. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*. PMLR.

Song, S.; Yu, F.; Zeng, A.; Chang, A. X.; Savva, M.; and Funkhouser, T. 2017. Semantic scene completion from a single depth image. In *CVPR*.

Wang, F.; Galliani, S.; Vogel, C.; and Pollefeys, M. 2022a. IterMVS: iterative probability estimation for efficient multi-view stereo. In *CVPR*.

Wang, F.; Galliani, S.; Vogel, C.; Speciale, P.; and Pollefeys, M. 2021. PatchmatchNet: Learned Multi-View Patchmatch Stereo. In *CVPR*.

Wang, Q.; Shi, S.; Zheng, S.; Zhao, K.; and Chu, X. 2020. Fadnet: A fast and accurate network for disparity estimation. In *ICRA*.

Wang, X.; Zhu, Z.; Huang, G.; Qin, F.; Ye, Y.; He, Y.; Chi, X.; and Wang, X. 2022b. MVSTER: epipolar transformer for efficient multi-view stereo. In *ECCV*.

Xie, B.; Li, B.; Zhang, Z.; Dong, J.; Jin, X.; Yang, J.; and Zeng, W. 2023. NaviNeRF: NeRF-based 3D Representation Disentanglement by Latent Semantic Navigation. In *ICCV*.

Xu, G.; Wang, X.; Ding, X.; and Yang, X. 2023. Iterative Geometry Encoding Volume for Stereo Matching. In *CVPR*.

Yao, Y.; Luo, Z.; Li, S.; Fang, T.; and Quan, L. 2018. Mvsnet: Depth inference for unstructured multi-view stereo. In *ECCV*.

Yao, Y.; Luo, Z.; Li, S.; Shen, T.; Fang, T.; and Quan, L. 2019. Recurrent mvsnet for high-resolution multi-view stereo depth inference. In *CVPR*.

Yao, Y.; Luo, Z.; Li, S.; Zhang, J.; Ren, Y.; Zhou, L.; Fang, T.; and Quan, L. 2020. Blendedmvs: A large-scale dataset for generalized multi-view stereo networks. In *CVPR*.

Yin, Z.; Darrell, T.; and Yu, F. 2019. Hierarchical discrete distribution decomposition for match density estimation. In *CVPR*, 6044–6053.

Zhang, Y.; Chen, Y.; Bai, X.; Yu, S.; Yu, K.; Li, Z.; and Yang, K. 2020. Adaptive unimodal cost volume filtering for deep stereo matching. In *AAAI*.