CVRecon: Rethinking 3D Geometric Feature Learning For Neural Reconstruction

Ziyue Feng[1], Liang Yang[2], Pengsheng Guo[3], and Bing Li[1]

[1]Clemson University [2]City University of New York [3]Carnegie Mellon University

Abstract

Recent advances in neural reconstruction using posed image sequences have made remarkable progress. However, due to the lack of depth information, existing volumetric-based techniques simply duplicate 2D image features of the object surface along the entire camera ray. We contend this duplication introduces noise in empty and occluded spaces, posing challenges for producing high-quality 3D geometry. Drawing inspiration from traditional multi-view stereo methods, we propose an end-to-end 3D neural reconstruction framework CVRecon, designed to exploit the rich geometric embedding in the cost volumes to facilitate 3D geometric feature learning. Furthermore, we present Ray-contextual Compensated Cost Volume ($RCCV$), a novel 3D geometric feature representation that encodes view-dependent information with improved integrity and robustness. Through comprehensive experiments, we demonstrate that our approach significantly improves the reconstruction quality in various metrics and recovers clear fine details of the 3D geometries. Our extensive ablation studies provide insights into the development of effective 3D geometric feature learning schemes. Project page: `https://cvrecon.ziyue.cool`

1. Introduction

Monocular 3D reconstruction is a fundamental task in computer vision with wide-ranging applications, including augmented reality [31, 24], virtual reality, robotics [25, 7], and autonomous driving[38, 10]. In recent years, learning-based methods [1, 21, 24, 29, 31] have shown promising results for this task. These methods use a sequence of posed images to predict a Truncated Signed Distance Field (TSDF) volume as the 3D reconstruction. They can be broadly categorized into two groups: volumetric-based and depth-based.

Existing volumetric-based methods [1, 21, 29, 31] suffer from a critical limitation where the image feature is in the 2D while the reconstruction target is in 3D. As shown in Fig. 1, previous works simply duplicate the 2D features along the entire camera ray to the 3D space. When 2D image features representing object surfaces are placed in empty or occluded spaces without differentiation, it can complicate feature fusion and TSDF prediction in later stages, introduce noise and limit the model's ability to predict fine geometries. Therefore, it is more logical to directly build 3D feature representations that encode the geometry clue instead of simply filling it with 2D feature copies.

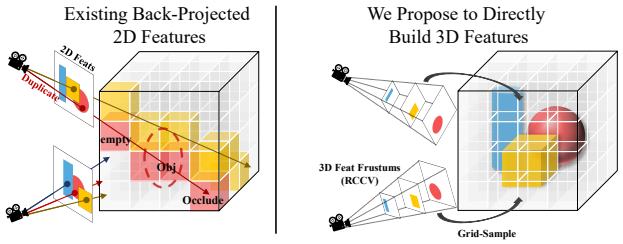

Figure 1. **Novel 3D geometric feature learning paradigm.** Existing volumetric-based neural reconstruction methods simply back-project and duplicate the 2D image features along the entire camera ray, including empty spaces in front of the object and occluded areas behind the object, which will introduce noise and degrade the performance. We propose Ray-contextual Compensated Cost Volume ($RCCV$) as a novel 3D geometric feature representation. Geometries could be inferred from a fusion of the $RCCV$s.

The cost volume, widely used in the depth prediction task, is a great choice for 3D geometric feature representation. It is composed of multi-view feature matching confidences across a set of pre-defined depth planes. A higher feature matching confidence at a given position indicates a greater likelihood of that position being an object surface. Compared to the simply duplicated 2D image features, cost volume explicitly encodes the depth probability distribution along the ray. However, existing depth-based reconstruction methods [3, 16, 18, 24, 33, 40] predict 2D depth maps from the 3D cost volume and then reconstruct the 3D structure

through the TSDF Fusion [22]. The 3D-2D-3D pipeline is not optimal because it doesn't take into account global context and frame consistency, which can result in the loss of structural details or the introduction of floating artifacts during the 3D-2D transformation. In contrast, if the 3D feature representation of each keyframe is retained and fused for the final reconstruction, the model will be able to estimate the structure holistically and unleash a greater potential.

In this paper, we propose to directly construct a novel Ray-contextual Compensated Cost Volume ($RCCV$) as a 3D geometric feature representation, which is a multi-view cost volume with 2 contributions: (1) For a keyframe image, we argue that the feature matching cost at a single location is not sufficient for inferring the geometry. As shown in Fig 4, the confidence distribution along the entire camera ray is needed for reference: the highest matching confidence position within a camera ray is more likely to be the surface. Therefore, we devised a distribution feature for individual camera rays, which is subsequently integrated into each voxel along the said ray. (2) We observe the cost volume fails to encode any useful information in the non-overlapping regions and is excessively noisy in areas with low-contrasting textures as shown in Fig 3. Therefore, besides the feature matching confidence, the original 2D image feature is also fused to the cost volume and we name it "Contextual Compensation". With these two contributions, our proposed $RCCV$ encodes comprehensive 3D geometric information. The reconstruction result could be inferred from a fusion of our $RCCV$s from multiple keyframes. Extensive experiments have demonstrated, as a generic representation, our $RCCV$ is agnostic to the downstream fusion and prediction models, and provides a more effective 3D geometric feature learning scheme.

To summarize, Our contributions are as follows:

- We identify fundamental limitations of the existing feature learning scheme in the neural reconstruction field and accordingly propose to leverage the multi-view cost volume as a direct 3D geometric feature representation.
- We observe the widely-used standard cost volume lacks the distribution reference information along the camera ray and propose the Ray Compensation mechanism to solve this problem.
- To improve the robustness of the cost volume in the non-overlapping and low-texture areas, we propose a novel Contextual Compensation module.
- Our extensive experiments show the effectiveness of our proposed $RCCV$, and its agnostic nature with downstream fusion and prediction models.

2. Related Works

In this section, we first review relevant volumetric-based neural reconstruction methods and analyze their limitations. Then we briefly introduce the cost volume in the depth prediction field and survey its applications in the 3D reconstruction task.

Volumetric-based 3D Reconstructions. In recent years 3D computer vision research [27, 6, 5, 4, 42] have shown remarkable progress, especially volumetric-based 3D reconstruction [1, 21, 29, 31, 28]. These methods usually extract 2D image features and back-project them into the 3D feature volume, 3D dense or sparse convolutions are later applied to predict the TSDF volume from it. Finally, the 3D mesh can easily be obtained by marching cubes [19]. Atlas [21] is the pioneering work that achieved promising results with simple back-projection and 3D convolution. NeuralRecon [31] is later proposed to introduce a fragmenting strategy and RNN-based global fusion to handle large-scale environments. Moreover, VoRTX [29], TransformerFusion [1], and concurrent work SDF-Former [43] explore the Transformer [32] mechanism for view selection, fusion, and aggregation, are complementary to our contribution.

Despite the progress made by volumetric-based methods, they suffer from fundamental limitations due to their feature representation. The duplications of image feature does not represent 3D geometry, inherently still a 2D feature. It also introduces noise by placing surface features in the front empty space and behind occluded areas, adding burdens to downstream models to estimate the 3D geometry, degrading the performance. In contrast, our novel $RCCV$ directly and explicitly encodes the 3D geometric information via cost volume representation (higher matching confidence indicates higher probability as object surface).

Depth-based 3D Reconstructions. The traditional 3D reconstruction methods typically involve predicting a dense depth map for each keyframe, which is then fused into a 3D structure. COLMAP [25] provides an impressive depth prediction baseline but suffers from low-texture regions. In the era of deep learning, neural network-based depth prediction models [3, 15, 16, 18, 33, 40] usually build a cost volume to aid in the depth prediction. SimpleRecon [24] proposed to improve the robustness and accuracy by adding camera metadata to the cost volume. TSDF Fusion [22], RoutedFusion [36], and NeuralFusion [37] are proposed to fuse the predicted depth maps into a global TSDF representation.

However, depth-based methods have some essential limitations. First, the depth maps of different frames are predicted separately, leading to inevitable inconsistencies between the frames. Second, the depth representation only encodes the structure of the predicted 2D manifold, any imperfection in the depth prediction can cause to loss of actual surface geometry information from the cost volume. In contrast, our proposed method fuses all the geometric infor-

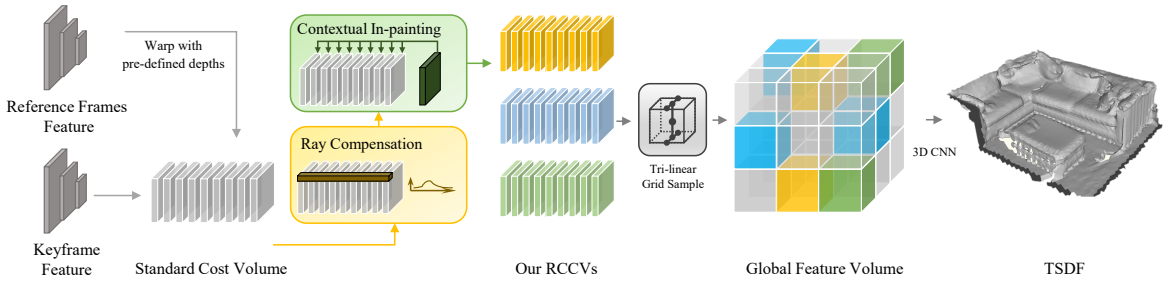

Figure 2. **CVRecon Architecture:** We first build standard cost volumes for each keyframe with reference frames. Novel $RCCV$s are then generated with our proposed Ray Compensation and Contextual Compensation. We tri-linear grid-sample and fuse the $RCCV$s as a global feature volume and the TSDF reconstruction is later inferred with 3D Convolutions.

mation of our novel $RCCV$s and predicts the 3D geometry holistically, ensuring consistency, accuracy, and completeness.

Cost Volume in Depth Predictions. The cost volume is a successful geometric learning paradigm that has found wide application in flow estimation [17, 30, 39], stereo matching [2, 12, 20, 40, 34], and depth prediction fields [11, 13, 24, 35, 41]. In the depth scenario, multi-view RGB pixels or CNN features are projected onto pre-defined depth planes in the view-frustum, and pixel or feature matching costs are used to encode the depth probability distribution. The higher matching confidence regions in a camera ray are generally indicative of the object's surface. Existing methods flatten the 3D cost volume as a 2D feature map such that the cost distribution within each camera ray will be used to predict the depth value for the corresponding pixel. Conversely, we retain the 3D structure and introduce the Ray Compensation and Contextual Compensation to enable the costs in each 3D location to represent the geometry independently and robustly.

3. Methodology

Given a sequence of monocular images $\{I^i\}_{i=1}^M \in R^{3\times H\times W}$ with its 6-DOF poses $\{P^i\}_{i=1}^M \in SE(3)$ and intrinsics $\{K^i\}_{i=1}^M$, the goal of the 3D neural reconstruction is to predict a global TSDF volume S. Before the training, we use the TSDF Fusion [22] to generate the ground-truth TSDF S_{gt} from the ground truth depths. During the testing time, we do not have access to the ground truth depth.

3.1. Method Overview

The framework of our method is shown in Fig. 2. To improve the efficiency, redundant image frames are removed and only N keyframes $\{I^{i,0}\}_{i=1}^N$ are kept based on the pose distance. Each keyframe $I^{i,0}$ is associated with a set of reference frames $\{I^{i,k}\}_{k=1}^K$ with proper camera pose difference and view overlapping. We first build a standard cost volume CV for each of the keyframes and then enhance them with our proposed Ray Compensation and Contextual Compensation. The generated $RCCV$s are integrated into a global feature volume through the process of grid sampling. Subsequently, a 3D CNN is employed to transform the volumetric representation into a TSDF volume.

Our key insights are as follows: (1) As shown in Fig. 1, we directly build $RCCV$ as a 3D geometric feature representation of the input image. Compare to the existing back-projection mechanism, our approach avoids introducing noise and improves the reconstruction quality. (2) We avoid the use of a 2D depth map as the intermediate representation, which suffers from inconsistency and would lose the information on the actual surface when the depth prediction is imperfect. Instead, we propose an end-to-end 3D framework CVRecon, which preserves all the geometric information to ensure an accurate holistic reconstruction. (3) We observe that the standard cost volume lacks the global context. As shown in Fig 4, the cost distribution within a camera ray is not normalized and has multiple peaks. Predicting the geometry from a single cost value needs the ray distribution for reference. (4) The cost volume in the non-overlapping and texture-less areas does not carry much useful information as shown in Fig 3. We accordingly propose Ray Compensation and Contextual Compensation to improve the integrity and robustness of the standard cost volume.

In the following sections, we first introduce the standard cost volume construction, followed by a detailed explanation of our proposed Ray Compensation and Contextual Compensation. Subsequently, the paper delves into the explication of the fusion mechanism, TSDF prediction, and loss function designs.

3.2. Ray-contextual Compensated Cost Volume

Standard Cost Volume Construction. Consider an image keyframe $I^{t,0}$ with a set of reference frames $\{I^{t,k}\}_{k=1}^K$, cost volume is to encode the feature matching confidence at

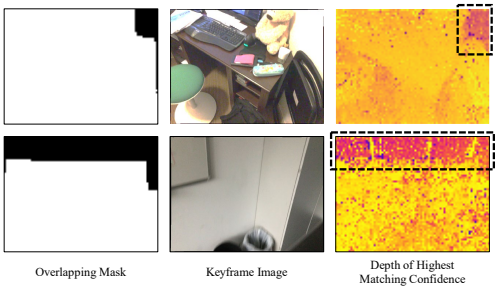

Figure 3. **Standard Cost Volumes are Noisy in Non-overlapping and Low-texture Areas.** The black masks and dashed rectangular markers indicate there is no reference frame overlap with the keyframe. We propose Contextual Compensation to improve the robustness in the non-overlapping and low-texture areas.

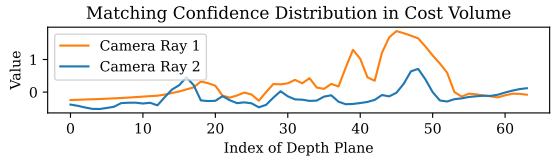

Figure 4. **Camera Ray Distributions in Cost Volume.** We visualize a channel of matching confidence distributions of two sample camera rays in the cost volume. These distributions have multiple peaks and the values are not normalized in magnitude. The model tends to decode local maxima as the object surfaces without the ray distribution as a reference, causing floating artifacts as shown in Fig 7.

different 3D locations. The feature maps $\{F^{t,k}\}_{k=0}^{K}$ of all images are first extracted by a 2D CNN model called Matching Encoder θ_{ME} and then projected to a set of pre-defined depth hypothesis planes $\{D_i\}_{i=1}^{|D|}$. $|D|$ is the number of the depth hypothesis. For the keyframe $I^{t,0}$, the ith plane CV_i^t in the standard cost volume CV^t is the concatenation of the dot products of the projected reference frame features $\{\hat{F}_i^{t,k}\}_{k=1}^{K}$ and the keyframe feature $F^{t,0}$. $\hat{F}_i$ indicates the feature F is perspective projected with depth D_i.

$$F^{t,k} = \theta_{ME}(I^{t,k}), \quad (1)$$

$$\hat{F}_i^{t,k} = \pi_0(\pi_k^{-1}(F^{t,k}, D_i)), \quad (2)$$

$$CV_i^t = \left\langle F^{t,0} \cdot \hat{F}_i^{t,k} \right\rangle_{k=1}^{K}, \quad (3)$$

$$CV^t = \left\langle CV_i^t \right\rangle_{i=1}^{|D|}, \quad (4)$$

where π is the perspective projection function based on the corresponding intrinsic and pose, and $\langle\rangle$ is the concatenation operation.

Camera Ray Compensation. We observe that directly using the above standard cost volume as 3D geometric feature does not shown significant performance improvement because it lacks camera ray information. Specifically, for each keyframe $I^{t,0}$, we build a standard cost volume

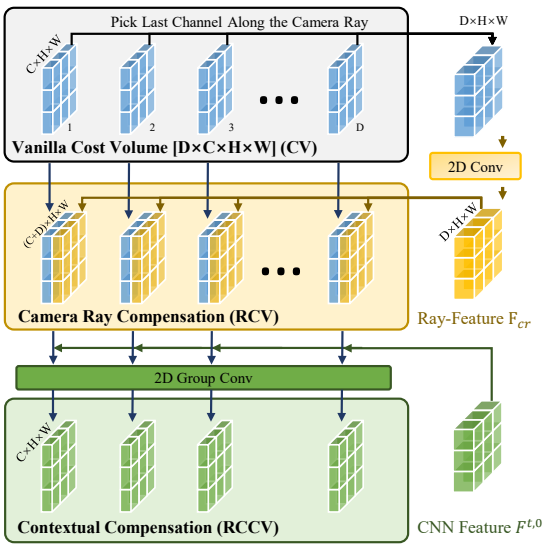

Figure 5. **Ray Compensation and Contextual Compensation.** Decoding the geometry from the cost value requires the overall distribution information within its camera ray for reference. We propose Ray Compensation to concatenate the ray information to all depth planes of the cost volume. The matching cost is too noisy in the non-overlapping and low-texture areas, the Contextual Compensation is proposed to fuse 2D CNN features to improve the robustness.

$CV^t \in R^{|D|\times C\times H\times W}$ where C is the channel number and H, W are widths and heights in the image plane. The camera ray $R^{|D|\times C}$ of each pixel $[h, w]$ encodes its C channels depth probability distribution as shown in Fig 4. This standard cost volume is view-dependent, whether a cost value represents an object surface is conditioned on all front cost values in its camera ray. Conventional depth prediction methods reshape this standard cost volume as a 2D feature map $R^{(|D|\cdot C)\times H\times W}$ and predict each pixel's depth with its entire camera ray information. However, for 3D reconstruction we expect a view-independent 3D feature representation. If we directly employ the standard cost volume as a 3D feature representation, the single feature R^C of a 3D location $[d, h, w]$ solely is not sufficient to decode its geometry without the overall camera ray distribution. Based on this observation, as shown in Fig 5, we propose a Camera Ray Compensation module that builds a camera ray feature F_{cr}^t and concatenates it into this standard cost volume CV^t to form our Ray-compensated Cost Volume RCV^t.

$$F_{cr}^t = \text{Conv2d}(\langle CV_i^t[-1]\rangle_{i=1}^{|D|}), \quad (5)$$

$$RCV_i^t = \langle CV_i^t, F_{cr}^t\rangle, \quad , \quad (6)$$

$$RCV^t = \langle RCV_i^t\rangle_{i=1}^{|D|} \quad (7)$$

where $[-1]$ is the tensor indexing operation that picks the

Eval	Method	3D Feature Source	The lower the better			The higher the better		
			Accuracy	Completeness	Chamfer	Precision	Recall	F1 Score @ 5cm
Atlas [21]	DeepVideoMVS [9]	2D Depth	0.079	0.133	0.106	0.521	0.454	0.474
	SimpleRecon [24]	2D Depth	0.065	0.078	0.071	0.641	0.581	0.608
	NeuralRecon [31]	Proj 2D feats	0.054	0.128	0.091	0.684	0.479	0.562
	Atlas [21]	Proj 2D feats	0.068	0.098	0.083	0.640	0.539	0.583
	TransformerFusion [1]	Proj 2D feats	0.078	0.099	0.088	0.648	0.547	0.591
	VoRTX [29]	Proj 2D feats	0.054	0.090	0.072	0.708	0.588	0.641
	Ours	**3D RCCV**	**0.045**	**0.077**	**0.061**	**0.753**	**0.639**	**0.690**
TransformerFusion [1]	COLMAP [25]	2D Depth	0.102	0.118	0.110	0.509	0.474	0.489
	DPSNet [16]	2D Depth	0.119	0.075	0.097	0.474	0.519	0.492
	DELTAS [26]	2D Depth	0.119	0.074	0.097	0.478	0.533	0.501
	DeepVideoMVS [9]	2D Depth	0.106	0.069	0.087	0.541	0.592	0.563
	3DVNet [23]	2D Depth	0.077	0.067	0.072	0.655	0.596	0.621
	SimpleRecon [24]	2D Depth	0.055	**0.060**	0.058	0.686	0.658	0.671
	NeuralRecon [31]	Proj 2D feats	0.051	0.091	0.071	0.630	0.612	0.619
	Atlas [21]	Proj 2D feats	0.072	0.076	0.074	0.675	0.605	0.636
	TransformerFusion [1]	Proj 2D feats	0.055	0.083	0.069	0.728	0.600	0.655
	VoRTX [29]	Proj 2D feats	0.043	0.072	0.057	0.767	0.651	0.703
	Our CVRecon	**3D RCCV**	**0.038**	0.067	**0.053**	**0.794**	**0.685**	**0.735**

Table 1. **3D Mesh Evaluation on ScanNet2.** The upper part follows the evaluation metric from Atlas [21] while the lower part follows TransformerFusion [1]. Methods in each category are sorted by the F1 Score. The **best** and second-best scores are marked respectively. Existing volumetric methods simply duplicate 2D image features to the 3D space, depth-based methods predict 2D depth maps and use TSDF-Fusion to retrieve the 3D geometry. In contrast, our end-to-end 3D framework CVRecon with novel 3D geometric feature representation $RCCV$ is a more natural and logical design.

last element of the channel dimension.

Before the ray compensation, the channel number of standard cost volume is reduced by a small MLP to $C = 7$. Each of the C channels is an aggregation of all information from all frames. We find that using all channels to construct the ray context feature only increases 0.002 of the F1-score but consumes 7 times memory. We find that only picking one channel performs well enough and which channel to pick does not matter.

Contextual Compensation. As shown in Fig 3, we still observe some fundamental limitations of our Ray-compensated Cost Volume RCV: (1)When the camera is moving backward or rotating, some areas in the new image frame may not overlap with the previous frames; (2)Some objects like floors and walls have very low contrast in textures. In these areas, the feature matching cost is noisy and could not provide reliable geometric information. To solve these problems, as shown in Fig 5, we propose a Contextual Compensation module to fuse the 2D CNN feature $F^{t,0}$ of the keyframe t into the RCV^t to form the Ray-contextual Compensated Cost Volume $RCCV^t$. In the non-overlapping and low-texture areas, instead of fully noise in standard cost volume, our $RCCV$ will degrade and similar to duplicated 2D CNN features in the existing works, improving the robustness. We observe that using separate convolution kernel weights for each depth plane generates significantly better results and we hypothesize it is related to the different spatial scales across depth planes. We efficiently implement the Contextual Compensation by concatenating the cost volume feature and 2D feature, followed by a group convolution operation with the $|D|$ as the group number.

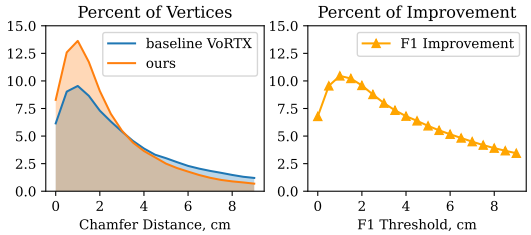

Figure 6. **Analyze of 3D Mesh Improvements.** The left figure is the chamfer distance distribution between the predicted and ground-truth meshes, which is generally lower in our predictions. The right figure shows our improvement percentage of the F1 score with different evaluation thresholds, the widely used 5cm F1 score could not fully capture our advantage.

$$RCCV_i^t = \text{Conv2d}_i(\langle RCV_i^t, F^{t,0}\rangle), \quad (8)$$

$$RCCV = \langle RCCV_i\rangle_{i=1}^{|D|} \quad (9)$$

Fusion of the RCCVs. After obtaining the $RCCV$ for each image keyframe, we generate a global feature volume by grid sampling with tri-linear interpolation. Given the downstream operation-agnostic nature of our proposed RCCV feature, we have found that it can be seamlessly integrated with various inter-frame feature fusion techniques, such as the multi-head self-attention module [29] or naive averaging operation [21, 31].

TSDF Prediction. We employ 3D dense [21] or

	Method	Depth-Supervised	The lower the better					The higher the better		
			Abs Diff	Abs Rel	Sq Rel	RMSE	logRMSE	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Depth-based	COLMAP [25]	Yes	0.264	0.137	0.138	0.502	-	83.4	-	-
	GPMVS [14]	Yes	0.239	0.130	0.339	0.472	0.108	90.6	96.7	98.0
	MVDepthNet [33]	Yes	0.191	0.098	0.061	0.293	0.116	89.6	97.7	99.4
	DPSNet [16]	Yes	0.158	0.087	0.035	0.232	0.110	92.5	98.4	99.5
	DELTAS [26]	Yes	0.149	0.078	0.027	0.221	0.107	93.7	-	-
	SimpleRecon [24]	Yes	0.088	0.043	0.012	0.146	0.067	98.0	-	-
Volumetric	Atlas [21]	No	0.124	0.065	0.043	0.251	-	93.6	97.1	98.6
	NeuralRecon [31]	No	0.106	0.065	0.031	0.195	-	94.8	96.1	97.5
	VoRTX [29]	No	0.096	0.061	0.038	0.205	-	94.3	97.3	98.7
	Our CVRecon	No	**0.078**	**0.047**	**0.028**	**0.181**	**0.094**	**96.3**	**98.2**	**99.1**

Table 2. **2D Depth Evaluation on ScanNet2.** The upper parts are depth-based methods, which predict depths and are directly supervised by the ground-truth depths. The lower parts are volumetric methods supervised by the ground-truth TSDF, we render the depth maps from its mesh predictions. All methods are ranked by the absolute depth difference within their category, best scores are in **bold**.

sparse [29, 31] convolution modules for geometry prediction. The predictions at the coarse and medium levels are occupancy grids to sparsify the feature grid while at the fine level, we directly predict the TSDF volume.

Loss Functions. Following NeuralRecon [31], we apply binary cross-entropy (BCE) loss function to the coarse and medium level occupancy predictions and $l1$ loss function to the fine level TSDF prediction. The TSDF ground truth is in $4\ cm$ resolution. Following Atlas [21], we mark all unobserved columns of the ground truth TSDF volume as unoccupied.

3.3. Implementation Details

For the cost volume, we set the depth range from 0.25 to 5.0 meters and divide it into 64 depth planes evenly in the log space. We construct the cost volume at $1/8(80 \times 60)$ of the input image resolution (640×480). Following SimpleRecon [24], we concatenate metadata (rays, angles, pose, etc..) into the cost volume and use an MLP to reduce the channel dimension to 7. In Ray Compensation, the channel number of our ray feature F_{cr} is the same as $|D|$ which is 64. After the Contextual Compensation, our group convolution reduces the channel number back to 7. The final $RCCV$ is a tensor of $R^{64\times7\times60\times80}$ which consumes around $4MB$ of memory in the FP16 precision and $15ms$ computation time.

Our proposed $RCCV$ is agnostic to the downstream fusion and prediction models which consume most of the computation power. When applying our $RCCV$ to the VoRTX [29] or Atlas [21] downstream models, the training takes around $40-45$ hours on two Nvidia A100 GPUs, showing minor computational overhead.

4. Experiments

To evaluate the effectiveness of our new 3D geometric feature representation $RCCV$, we apply it to a volumetric 3D reconstruction baseline VoRTX [29] to replace the simple back-projection mechanism and name it CVRecon. We evaluate its performance on the challenging ScanNet2 [8] dataset, which comprises 1513 RGB-D scans from 707 indoor spaces. Following the official split, our training, validation, and testing sets consist of 1201, 312, and 100 scans, respectively.

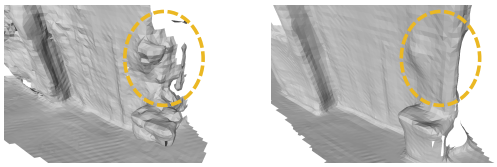

Figure 7. **Floating Artifacts without Ray-compensation.** Without the ray distribution as a reference, the model tends to decode local maxima of the matching confidence in the cost volume as object surfaces, causing floating artifacts.

In the following sections, we evaluate our reconstruction quality with 3D mesh metrics and 2D depth metrics. Then we conduct a comprehensive ablation study to dig into each part of our contributions to analyze their qualitative and quantitative impact. We also adopt our $RCCV$ to different baselines to verify it is agnostic to downstream modules.

4.1. 3D Reconstruction Evaluations

The quantitative 3D mesh evaluation results on the ScanNet2 [8] dataset are shown in Table 1. We evaluate our method with the standard protocol proposed by Atlas [21] and a more reasonable protocol proposed by TransformerFusion [1] which excluded the unobserved areas. The ground-truth meshes are incomplete at the unseen areas in the input monocular sequences while many learning-based methods are able to hallucinate those structures. The TransformerFusion [1] evaluation protocol uses an occlusion mask to prevent penalizing the more complete reconstruction.

Our CVRecon with the novel $RCCV$ outperforms a wide range of state-of-the-art reconstruction methods by a large margin. The depth-based methods [9, 16, 23, 24,

Back-Project 2D Image Feats	Grid Sample Cost Volume	Camera Ray Compensation	Contextual Inpainting			3D Mesh Metrics			
			Concat	Uni−Conv	Group−Conv	Acc ↓	Comp ↓	Chamfer ↓	F1-Score ↑
Evaluating 3D Feature Construction									
✓						0.043	0.072	0.057	0.703
	✓					0.041	0.071	0.056	0.711
Evaluating Camera Ray Compensation									
	✓					0.041	0.071	0.056	0.711
	✓	✓				0.040	0.068	0.054	0.728
Evaluating Contextual Inpainting									
	✓	✓				0.040	0.068	0.054	0.728
	✓	✓	✓			0.039	0.067	0.053	0.730
	✓	✓	✓	✓		0.038	0.067	0.053	0.731
	✓	✓	✓		✓	**0.038**	**0.067**	**0.053**	**0.735**

Table 3. **Ablation Study:** Evaluating the effectiveness of our proposed novel 3D geometric feature learning scheme, including the cost volume, Camera Ray Compensation, and Contextual In-painting on the ScanNet2 [8] dataset with the evaluation protocol from the TransformerFusion [1].

25, 26] use predicted 2D depth maps as the intermediate representation and retrieve 3D structure by the TSDF Fusion [22], these methods suffer from (1) inconsistent depth prediction and (2) potentially lost information of the actual surface whenever the depth prediction is imperfect. Volumetric-based methods simply back-project 2D image features to all voxels along the camera ray, introducing noise to empty and occluded spaces. In contrast, our novel 3D feature representation $RCCV$ and the end-to-end 3D framework CVRecon solve these problems and significantly improve the reconstruction quality.

Qualitative visualizations are shown in Fig 8. Compared to existing volumetric methods [21, 29], our CVRecon generates much more clear and more complete fine details. Our proposed novel 3D geometric feature representation $RCCV$ significantly unlocks the potential of downstream models for better reconstruction quality. Compared to the state-of-the-art depth-based model SimpleRecon [24], one of the major advantages of our method is to holistically generate more consistent and clean geometries. More analysis could be found in the supplementary materials.

Moreover, we would like to point out that the default 5 cm threshold for F1-score is excessively large and fails to adequately capture our superior performance in modeling fine details of the geometry. As shown in Fig 6, our method reduces a significant portion of mesh vertex chamfer distances to under 2 cm, smaller thresholds demonstrate much higher improvements.

4.2. 2D Depth Evaluations

We evaluate the 2D depth metrics on the ScanNet2 [8] dataset and compare them with existing state-of-the-art methods in Table 2. Depth-based reconstruction methods are supervised by ground-truth depths. Volumetric methods like ours directly predict the 3D geometry and are supervised by the ground-truth TSDF volume. We render pseudo-depths from predicted meshes to perform this evaluation for volumetric methods.

Method	The lower the better			The higher the better		
	Accuracy	Completeness	Chamfer	Precision	Recall	F1 Score
Atlas [21] + Back-proj	0.076	0.085	0.081	0.644	0.559	0.596
Atlas [21] + RCCV	**0.070**	**0.073**	**0.071**	**0.683**	**0.620**	**0.648**
Improvement	**7.9%**	**14.1%**	**12.3%**	**6.1%**	**10.9%**	**8.7%**

Table 4. **Atlas with Our RCCV.** To verify our proposed novel RCCV is a general 3D geometric feature representation and not limited to a specific framework, we additionally apply it to Atlas [21] to replace its simple back-projected feature. The results meet our expectations.

Our CVRecon significantly outperforms all other volumetric methods on all metrics by a large margin. We even outperform all the depth-based methods on the absolute difference metric. We attribute this performance boost to the rich geometric encoding in our $RCCV$. Another interesting observation is that depth-based methods generally perform better on relative metrics but worse on absolute metrics because they focused too much on the near objects.

4.3. Ablation Study

In this section, we conduct extensive ablation studies on the ScanNet2 [8] dataset to comprehensively evaluate the effectiveness of our proposed novel 3D geometric feature learning scheme. As shown in Table 3, we perform 3 groups of experiments to evaluate the standard cost volume, Ray Compensation, and Contextual Compensation.

Cost Volume as 3D Feature Representation. In the first group of our experiment, we compare the simple back-projection of the 2D image feature and the grid sampling of the standard cost volume. The result confirms that 3D reconstruction quality could be improved by not polluting the feature volume of the empty and occluded areas, as well as the rich 3D geometric encoding in the standard cost volume.

Ray Compensation. As analyzed in section 3.2, a single value in the standard cost volume is not sufficient to decode the geometry without its camera ray distribution as a reference. Our experiment in the second group verifies this assumption by showing a noticeable performance boost with

the Ray Compensation module. As shown in Fig 7, our ray compensation module solves the floating artifacts problem and generates more clear geometry.

Contextual Compensation. In the last group, we explore the necessity of our proposed contextual compensation mechanism. 3 sets of experiments are designed as follows: (1)We simply concatenate the keyframe 2D image feature to the Ray-compensated Cost Volume. All 3D mesh metrics are improved by this concatenation, confirming the effectiveness of the contextual feature. (2)A 2D convolutional layer is employed to better fuse the contextual feature with the cost volume, the improvements are relatively minor. (3)We use a group convolution layer to limit the perceptive field of the depth planes in the cost volume to be within itself. A more significant boost of the F1-score is observed, which shows the convolution layer is unable to preserve the structural order of the depth planes and manual regularization is necessary.

Downstream Module Agnostic. Our proposed $RCCV$ is a general 3D geometric feature representation and is agnostic to the downstream fusion and prediction modules. The above experiments employ the VoRTX [29] for the downstream prediction, which is similar to other models like the NeuralRecon [31] and TransformerFusion [1]. To further verify the generalization ability of our $RCCV$, we apply it to the Atlas [21] framework to replace its simple back-projected feature. As shown in Table 4, our method improved all evaluation metrics by a large margin.

5. Additional Discussions

5.1. Information Lost of Depth-based Methods

The cost volume is widely used in depth-based reconstruction methods. Compared to their existing 3D-2D-3D pipelines, our end-to-end 3D volumetric reconstruction from the cost volumes have several fundamental advantages. In addition to the qualitative and quantitative evaluations in the main paper, here we analyze a typical case in the ScanNet2 [8] dataset testing split.

In the supplementary material, we visualize the meshes, point clouds, a sample keyframe, and the matching confidence distribution of a sample pixel from the state-of-the-art depth-based method SimpleRecon [24]. The ground truth depth is 3.3 meters for the sample pixel, which is correctly reflected by its overall matching confidence distribution. However, SimpleRecon mistakenly predicts a depth of 2.92 meters due to a glitch in the cost distribution. Since the cost distribution information is discarded after the depth prediction, the downstream TSDF Fusion [22] is unable to filter out this outlier and generates a floating surface artifact. In contrast, our end-to-end 3D volumetric framework preserves the cost volume information of all the keyframes and holistically reconstructs clear geometry.

5.2. Use of Reference Frames

The construction of our keyframe cost volumes requires reference image frames. While most of these reference images come from the keyframe pool, our method may utilize slightly more image frames than some volumetric baselines, depending on the chosen frame selection strategy. To determine if our improved performance is due to this additional information, we conducted two experiments. (1) As mentioned in Sec 4.3 of the main paper, we apply our $RCCV$ to the Atlas [21] baseline. Both the baseline and our modified Atlas were using all available image frames, ensuring a fair comparison. (2) We evaluated our baseline VoRTX [29] using the same frames as our method and found that the F1-Score only improved from 0.703 to 0.705, indicating a negligible difference in performance that does not affect our conclusions.

5.3. Computation Efficiency

Constructing cost volumes requires additional computation time and memory compared to existing volumetric baselines. In our experiment, we find reducing the channel number of the cost volume from 7 to 1 and the number of depth planes from 64 to 32 does not noticeably affect reconstruction quality but will greatly reduce the computation overhead. The $RCCV$ of $R^{32\times1\times60\times80}$ only consumes $300KB$ of the memory and $5ms$ of the GPU time.

5.4. RCCV fusion strategy:

We adopted the VoRTX [29] (attention-based) and Atlas [21] (averaging) way of fusion strategy and both of them work great with our RCCV. We will clarify this in the future version.

5.5. Limitations

One major limitation of volumetric reconstruction methods like ours is the update of results is slower than depth-based methods, which could be alleviated by a proper fragmenting strategy.

6. Conclusion

In this paper, we identify fundamental limitations of the existing neural reconstruction methods and present an end-to-end 3D reconstruction framework named CVRecon with novel cost-volume-based 3D geometric feature representation $RCCV$. We significantly outperformed existing state-of-the-art methods and provide valuable insights into the development of 3D geometric feature learning schemes.

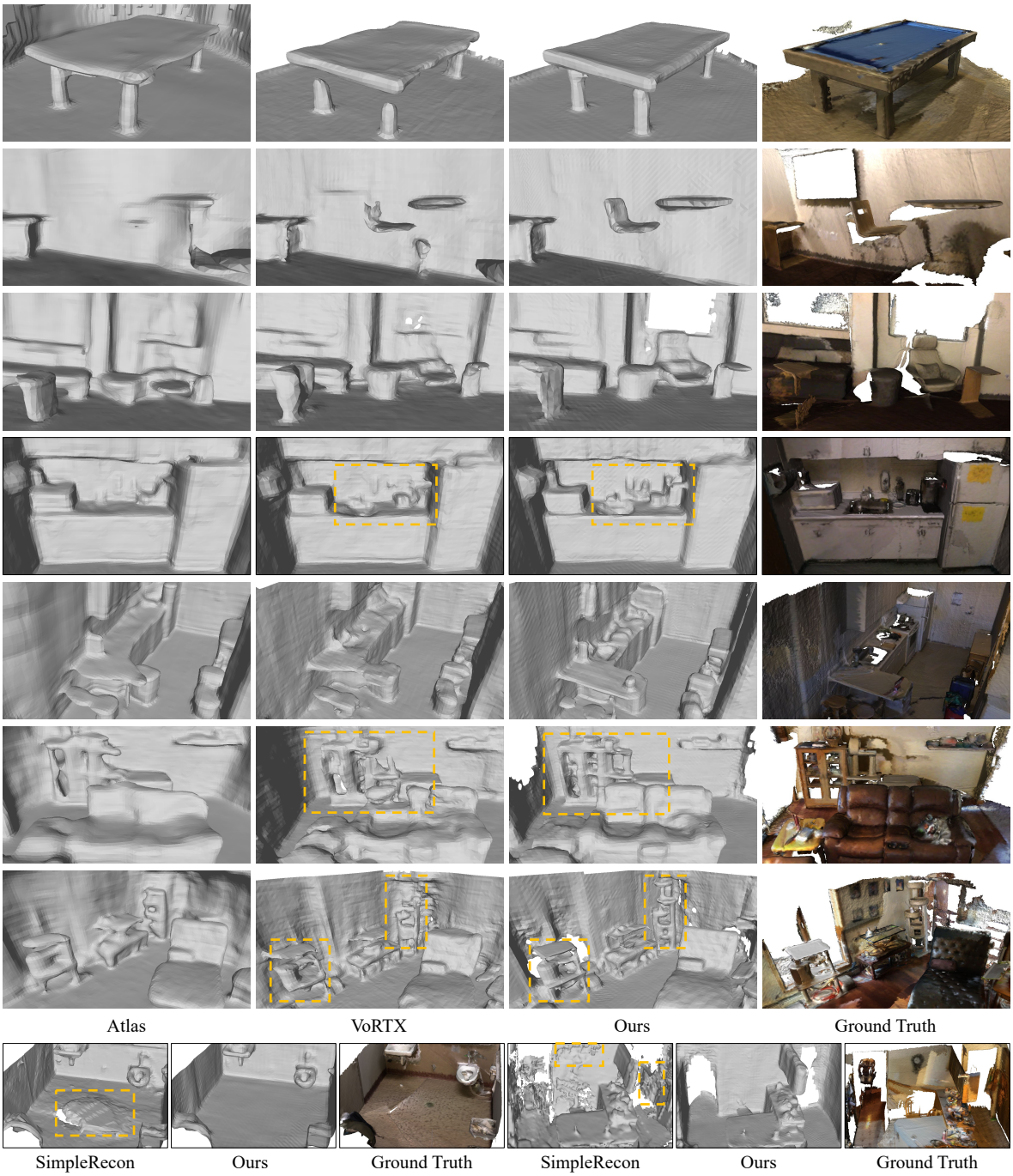

Figure 8. **Qualitative Comparison:** In the above part we compare our method with Atlas [21] and the current state-of-the-art VoRTX [29] on the ScanNet2 [8] dataset test set. Note that the only difference with the VoRTX [29] is we use our $RCCV$ as the 3D geometric feature representation, which leads to significantly clear geometry details. We compare our method with the state-of-the-art depth-based method SimpleRecon [24] in the last row. The lack of translation parallax in narrow spaces like the left sample and the faraway texture-less walls in the right sample will lead to severe inconsistency and degrade the performance of the depth-based methods. In contrast, our holistic prediction generates a much clear and smooth reconstruction. More samples and analysis are in the supplementary materials.

References

[1] Aljaz Bozic, Pablo Palafox, Justus Thies, Angela Dai, and Matthias Nießner. Transformerfusion: Monocular rgb scene reconstruction using transformers. *Advances in Neural Information Processing Systems*, 34:1403–1414, 2021.

[2] Jia-Ren Chang and Yong-Sheng Chen. Pyramid stereo matching network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5410–5418, 2018.

[3] Rui Chen, Songfang Han, Jing Xu, and Hao Su. Point-based multi-view stereo network. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1538–1547, 2019.

[4] Zhimin Chen, Longlong Jing, Yang Liang, YingLi Tian, and Bing Li. Multimodal semi-supervised learning for 3d objects. *arXiv preprint arXiv:2110.11601*, 2021.

[5] Zhimin Chen, Longlong Jing, Liang Yang, Yingwei Li, and Bing Li. Class-level confidence based 3d semi-supervised learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 633–642, 2023.

[6] Zhimin Chen and Bing Li. Bridging the domain gap: Self-supervised 3d scene understanding with foundation models. *arXiv preprint arXiv:2305.08776*, 2023.

[7] Zheng Chen and Lantao Liu. Navigable space construction from sparse noisy point clouds. *IEEE Robotics and Automation Letters*, 6(3):4720–4727, 2021.

[8] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2017.

[9] Arda Duzceker, Silvano Galliani, Christoph Vogel, Pablo Speciale, Mihai Dusmanu, and Marc Pollefeys. Deepvideomvs: Multi-view stereo on video with recurrent spatio-temporal fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15324–15333, 2021.

[10] Ziyue Feng, Longlong Jing, Peng Yin, Yingli Tian, and Bing Li. Advancing self-supervised monocular depth learning with sparse lidar. In *Conference on Robot Learning*, pages 685–694. PMLR, 2022.

[11] Ziyue Feng, Liang Yang, Longlong Jing, Haiyan Wang, YingLi Tian, and Bing Li. Disentangling object motion and occlusion for unsupervised multi-frame monocular depth. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXII*, pages 228–244. Springer, 2022.

[12] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2495–2504, 2020.

[13] Vitor Guizilini, Rareș Ambruș, Dian Chen, Sergey Zakharov, and Adrien Gaidon. Multi-frame self-supervised depth with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 160–170, 2022.

[14] Yuxin Hou, Juho Kannala, and Arno Solin. Multi-view stereo by temporal nonparametric fusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2651–2660, 2019.

[15] Junjie Hu, Mete Ozay, Yan Zhang, and Takayuki Okatani. Revisiting single image depth estimation: Toward higher resolution maps with accurate object boundaries. In *2019 IEEE winter conference on applications of computer vision (WACV)*, pages 1043–1051. IEEE, 2019.

[16] Sunghoon Im, Hae-Gon Jeon, Stephen Lin, and In So Kweon. Dpsnet: End-to-end deep plane sweep stereo. *arXiv preprint arXiv:1905.00538*, 2019.

[17] Haisong Liu, Tao Lu, Yihui Xu, Jia Liu, Wenjie Li, and Lijun Chen. Camliflow: Bidirectional camera-lidar fusion for joint optical flow and scene flow estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5791–5801, June 2022.

[18] Xiaoxiao Long, Lingjie Liu, Wei Li, Christian Theobalt, and Wenping Wang. Multi-view depth estimation using epipolar spatio-temporal networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8258–8267, 2021.

[19] William E Lorensen and Harvey E Cline. Marching cubes: A high resolution 3d surface construction algorithm. *ACM siggraph computer graphics*, 21(4):163–169, 1987.

[20] Chuanhua Lu, Hideaki Uchiyama, Diego Thomas, Atsushi Shimada, and Rin-ichiro Taniguchi. Sparse cost volume for efficient stereo matching. *Remote sensing*, 10(11):1844, 2018.

[21] Zak Murez, Tarrence Van As, James Bartolozzi, Ayan Sinha, Vijay Badrinarayanan, and Andrew Rabinovich. Atlas: End-to-end 3d scene reconstruction from posed images. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16*, pages 414–431. Springer, 2020.

[22] Richard A Newcombe, Shahram Izadi, Otmar Hilliges, David Molyneaux, David Kim, Andrew J Davison, Pushmeet Kohi, Jamie Shotton, Steve Hodges, and Andrew Fitzgibbon. Kinectfusion: Real-time dense surface mapping and tracking. In *2011 10th IEEE international symposium on mixed and augmented reality*, pages 127–136. Ieee, 2011.

[23] Alexander Rich, Noah Stier, Pradeep Sen, and Tobias Höllerer. 3dvnet: Multi-view depth prediction and volumetric refinement. In *2021 International Conference on 3D Vision (3DV)*, pages 700–709. IEEE, 2021.

[24] Mohamed Sayed, John Gibson, Jamie Watson, Victor Prisacariu, Michael Firman, and Clément Godard. Simplerecon: 3d reconstruction without 3d convolutions. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXIII*, pages 1–19. Springer, 2022.

[25] Johannes L Schönberger, Enliang Zheng, Jan-Michael Frahm, and Marc Pollefeys. Pixelwise view selection for unstructured multi-view stereo. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Nether-*

lands, October 11-14, 2016, Proceedings, Part III 14, pages 501–518. Springer, 2016.

[26] Ayan Sinha, Zak Murez, James Bartolozzi, Vijay Badrinarayanan, and Andrew Rabinovich. Deltas: Depth estimation by learning triangulation and densification of sparse points. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16*, pages 104–121. Springer, 2020.

[27] Noah Stier, Baptiste Angles, Liang Yang, Yajie Yan, Alex Colburn, and Ming Chuang. Livepose: Online 3d reconstruction from monocular video with dynamic camera poses. *arXiv preprint arXiv:2304.00054*, 2023.

[28] Noah Stier, Anurag Ranjan, Alex Colburn, Yajie Yan, Liang Yang, Fangchang Ma, and Baptiste Angles. Finerecon: Depth-aware feed-forward network for detailed 3d reconstruction. *arXiv preprint arXiv:2304.01480*, 2023.

[29] Noah Stier, Alexander Rich, Pradeep Sen, and Tobias Höllerer. Vortx: Volumetric 3d reconstruction with transformers for voxelwise view selection and fusion. In *2021 International Conference on 3D Vision (3DV)*, pages 320–330. IEEE, 2021.

[30] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[31] Jiaming Sun, Yiming Xie, Linghao Chen, Xiaowei Zhou, and Hujun Bao. Neuralrecon: Real-time coherent 3d reconstruction from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15598–15607, 2021.

[32] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[33] Kaixuan Wang and Shaojie Shen. Mvdepthnet: Real-time multiview depth estimation neural network. In *2018 International conference on 3d vision (3DV)*, pages 248–257. IEEE, 2018.

[34] Weihan Wang, Bharat Joshi, Nathaniel Burgdorfer, Konstantinos Batsosc, Alberto Quattrini Lid, Philippos Mordohaia, and Ioannis Rekleitisb. Real-time dense 3d mapping of underwater environments. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5184–5191. IEEE, 2023.

[35] Jamie Watson, Oisin Mac Aodha, Victor Prisacariu, Gabriel Brostow, and Michael Firman. The temporal opportunist: Self-supervised multi-frame monocular depth. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1164–1174, 2021.

[36] Silvan Weder, Johannes Schonberger, Marc Pollefeys, and Martin R Oswald. Routedfusion: Learning real-time depth map fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4887–4897, 2020.

[37] Silvan Weder, Johannes L Schonberger, Marc Pollefeys, and Martin R Oswald. Neuralfusion: Online depth fusion in latent space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3162–3172, 2021.

[38] Felix Wimbauer, Nan Yang, Lukas Von Stumberg, Niclas Zeller, and Daniel Cremers. Monorec: Semi-supervised dense reconstruction in dynamic environments from a single moving camera. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6112–6122, 2021.

[39] Wenxuan Wu, Zhi Yuan Wang, Zhuwen Li, Wei Liu, and Li Fuxin. Pointpwc-net: Cost volume on point clouds for (self-) supervised scene flow estimation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, pages 88–107. Springer, 2020.

[40] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. Mvsnet: Depth inference for unstructured multi-view stereo. In *Proceedings of the European conference on computer vision (ECCV)*, pages 767–783, 2018.

[41] Kyle Yee and Ayan Chakrabarti. Fast deep stereo with 2d convolutional processing of cost signatures. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 183–191, 2020.

[42] Peng Yin, Lingyun Xu, Ziyue Feng, Anton Egorov, and Bing Li. Pse-match: A viewpoint-free place recognition method with parallel semantic embedding. *IEEE Transactions on Intelligent Transportation Systems*, 23(8):11249–11260, 2021.

[43] Weihao Yuan, Xiaodong Gu, Heng Li, Zilong Dong, and Siyu Zhu. Monocular scene reconstruction with 3d sdf transformers. *arXiv preprint arXiv:2301.13510*, 2023.