# CVRecon: Rethinking 3D Geometric Feature Learning For Neural Reconstruction
# Supplementary Materials

This document contains the supplementary materials for "CVRecon: Rethinking 3D Geometric Feature Learning For Neural Reconstruction".

## 1. Additional Qualitative Comparisons

Fig 2 shows some additional qualitative comparisons of our method and state-of-the-art volumetric methods Atlas [2] and VoRTX [5]. The only difference between our method and the VoRTX [5] is that we use our proposed 3D Ray-contextual Compensated Cost Volume ($RCCV$) as the source of volumetric features instead of the widely-used back-projected 2D features. Our $RCCV$ contributes to more complete geometries and more clear fine details.

Fig 1 shows additional qualitative comparisons of our method and the state-of-the-art depth-based method SimpleRecon [4]. One of the major limitations of the depth-based methods is the inter-frame inconsistency. Since the depth maps of different frames are predicted separately, the fluctuation of their scales is a fundamental problem and will result in artifacts on surfaces like floors and walls. In contrast, our method holistically reconstructs the scene and generates much more clear and more coherent geometries.

## 2. Additional Discussions

### 2.1. 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 [1] dataset testing split.

As shown in Fig 3, 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 [4]. 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 [3] 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.

### 2.2. Use of Reference Frames

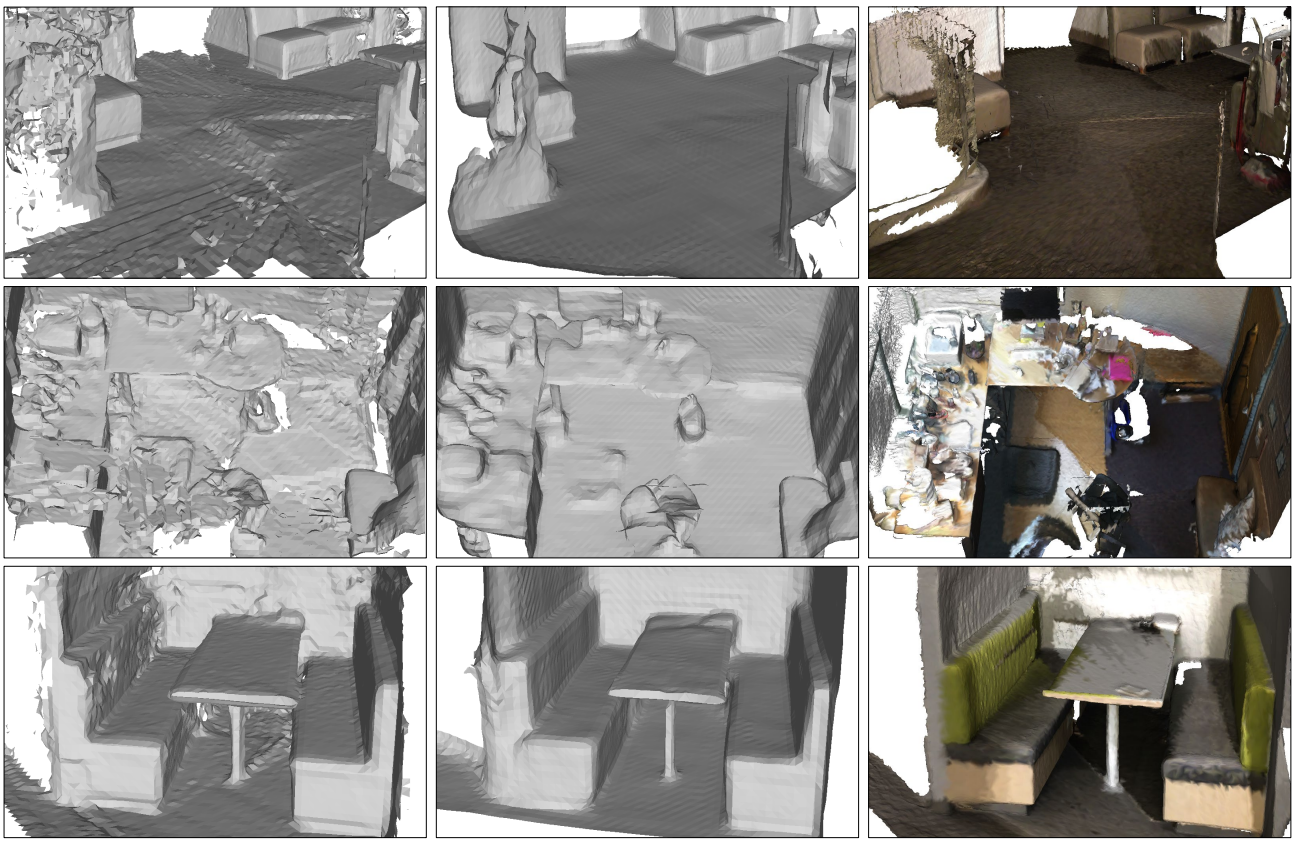
The construction of our keyframe cost volumes requires reference image frames. While most of these reference images come from the keyframe pool, our method may utilize slightly more image frames than some volumetric baselines, depending on the chosen frame selection strategy. To determine if our improved 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 [2] baseline. Both the baseline and our modified Atlas were using all available image frames, ensuring a fair comparison. (2) We evaluated our baseline VoRTX [5] 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.

### 2.3. 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.

### 2.4. 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.



SimpleRecon Ours Ground Truth

Figure 1. **Additional Qualitative Comparison with Depth-based Method.** We compare our method with the state-of-the-art depth-based method SimpleRecon [4]. The lack of translation parallax in narrow spaces and the texture-less floors will lead to inter-frame inconsistency and degrade the performance of the depth-based methods. In contrast, our holistic prediction generates a much clear and coherent reconstruction.

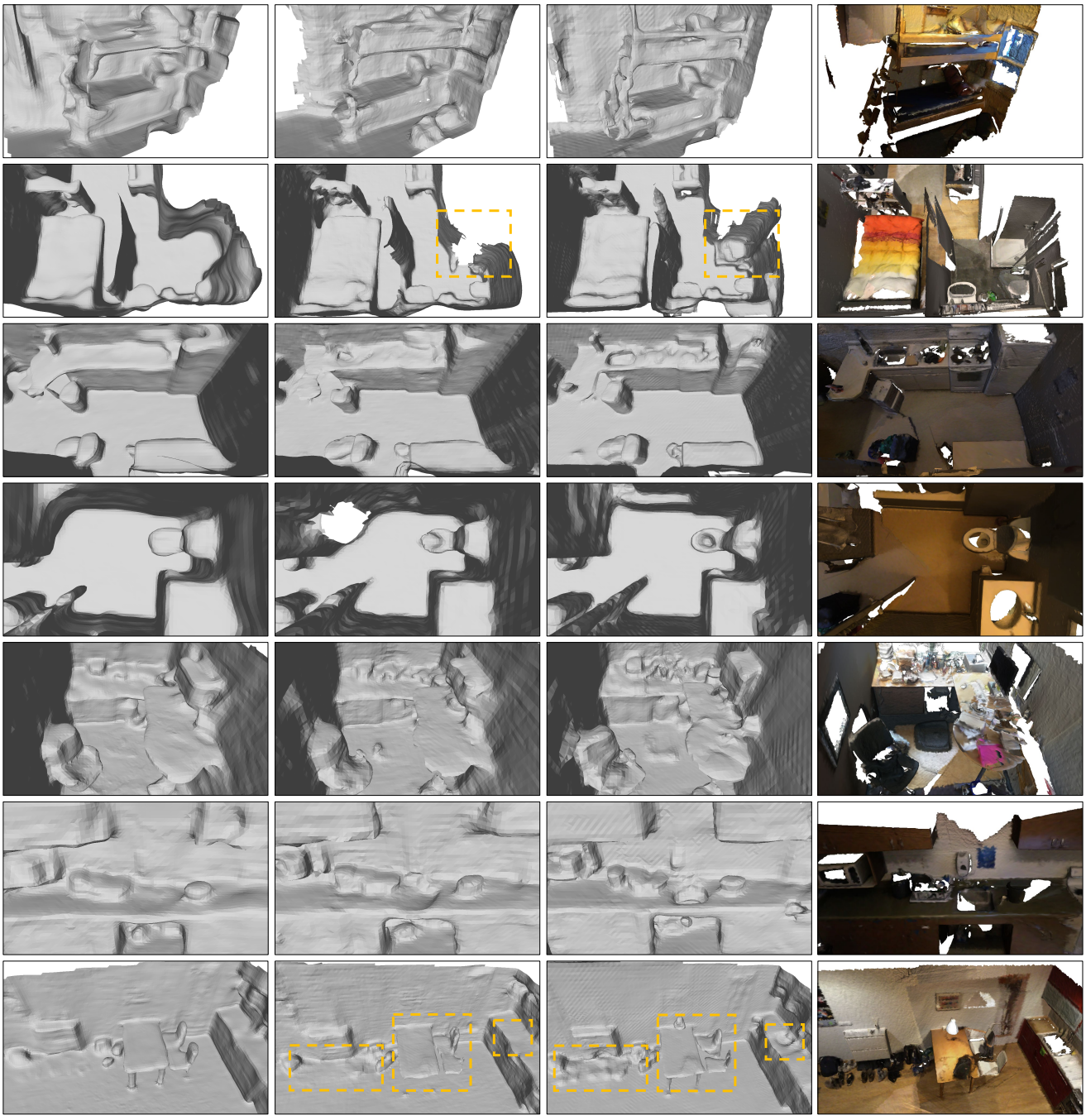
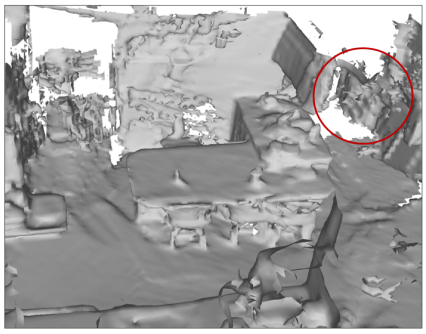
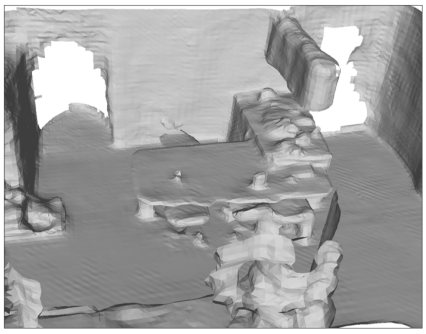


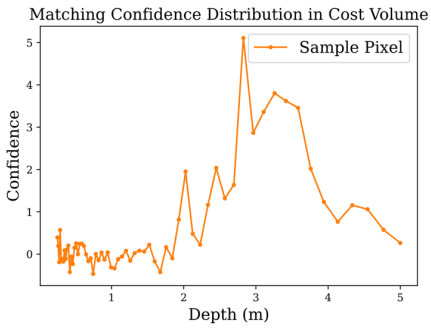

Figure 2. **Additional Qualitative Comparison with Volumetric Methods.** Additional comparison of our method with Atlas [2] and the current state-of-the-art VoRTX [5] on the ScanNet2 [1] dataset test split. The only difference with the VoRTX [5] is that we use our $RCCV$ as the 3D geometric feature representation, leading to significantly clear geometry details.



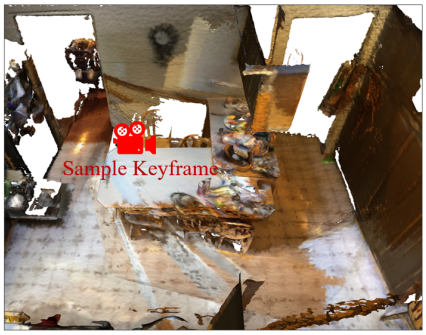
SimpleRecon Mesh



Our Mesh




SimpleRecon Point Cloud




Ground Truth Mesh


Sample Keyframe

Figure 3. **Information Lost of Depth-based Methods.** This is a sample scene from the ScanNet2 [1] test split. For the sample pixel with a ground truth depth of 3.3m, the state-of-the-art depth-based method SimpleRecon [4] predicted a depth of 2.92m, only keeping the information around the highest matching confidence in the cost volume. The downstream TSDF Fusion [3] is unable to filter out this outlier and generates floating artifacts. In contrast, we preserve the overall cost volume distribution of all keyframes and holistically predict a more clear geometry.

## References


[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.