Article

Not peer-reviewed version

Multi-head Attention Refiner For Many View 3d Reconstruction

Kyunghee Lee , Ihjoon Cho , Boseung Yang , Unsang Park *

Posted Date: 10 July 2024

doi: 10.20944/preprints202407.0857.v1

Keywords: multi-view 3D reconstruction; attention mechanism; multi-head attention; refiner; object boundary prediction

 doi:10.20944/preprints202407.0857.v1

Article

Multi-Head Attention Refiner for Many View 3D Reconstruction

Kyunghee Lee [1,†], Ihjoon Cho [1,†], Boseung Yang [1,†] and Unsang Park [1,*]

Department of Computer Science and Engineering, Sogang University, 35 Baekbeom-ro (sinsu-dong) Mapo-gu, Seoul 04107, Republic of Korea; dlrudgml23@sogang.ac.kr (K.L.); dlwns23@sogang.ac.kr (I.C.); didqhtmd@sogang.ac.kr (B.Y.)

* Correspondence: unsangpark@sogang.ac.kr

† These authors contributed equally to this work.

Abstract: Traditional 3D reconstruction models have consistently encountered a challenge: attaining high recall of object edges while preserving precision. In this paper, we introduce a post-processing method Multi-Head Attention Refiner (MA-R) aimed at tackling this challenge by integrating a multi-head attention mechanism within the U-Net style refiner module. The 3D reconstruction model applying this method demonstrates excellence in parsing intricate image details, resulting in significant enhancement of boundary predictions and increased recall rates. In our experiments, our method significantly enhances the reconstruction performance of Pix2Vox++ when multiple images are utilized as input. Specifically, it achieves an enhanced Intersection Over Union score of approximately 1.1% compared to the original Pix2Vox++ model when 16-view images are employed.

Keywords: multi-view 3D reconstruction; attention mechanism; multi-head attention; refiner; object boundary prediction

1. Introduction

In recent decades, the methodologies for creating 3D representations from 2D images have undergone significant changes. The field of 3D reconstruction has evolved from relying on geometry-based computer vision algorithms, such as feature extraction (SIFT) [1], structure estimation using epipolar geometry [2], and model creation via Delaunay triangulation [3], to employing deep learning models. These models offer a notable improvement in precision and detail, marking a paradigm shift in the field of computer vision. The advent of deep learning has particularly revolutionized the refinement of reconstructed shapes, a process once stymied by the complexity of details and textures inherent in 3D objects.

The attention mechanism, inspired by the human visual systems that can focus on salient features within a visual scene, has been successfully integrated into deep learning to enhance performances across various tasks. Studies have shown that attention mechanisms, and particularly multi-head attention, can significantly reduce errors in detail-oriented tasks by focusing on the most relevant aspects of input data [4]. In the realm of 3D reconstruction, our refinement module, which incorporates a multi-head attention mechanism, leverages this principle to enhance the granularity of the reconstruction substantially. As a result, our method not only reduces boundary prediction errors, but also increases the model's fidelity to the original structure. By implementing the multi-head attention refiner, our approach has achieved a higher Intersection over Union (IoU) value in multiview reconstruction, demonstrating the potential of this technique to promote the field of 3D reconstruction models forward. Figure 1 illustrates an example of such 3D reconstruction, showcasing the advanced capabilities of the described methodologies.

 doi:10.20944/preprints202407.0857.v1

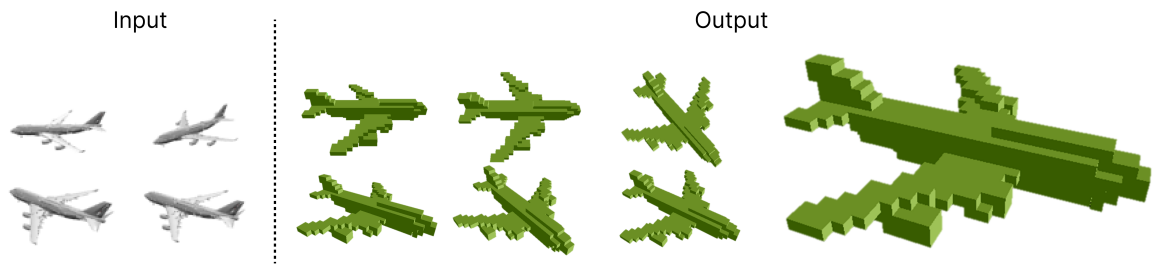

Figure 1. 3D reconstruction of an airplane from multi-view inputs: This figure presents the angle-specific reconstruction results of an airplane, demonstrating the effectiveness of our proposed model in translating multi-view 2D input images into accurate 3D outputs

2. Related Works

2.1. 3D Reconstruction

Comprehensive reviews of 3D object reconstruction approaches can be found in[5]. The 3D reconstruction subtask could be classified by input types such as "single-view 3D reconstruction" and "multiview 3D reconstruction" or classified by output format such as "Voxel," "Point cloud," and "3D mesh". The "Single-view 3D Reconstruction" task is a long-established ill-posed and ambiguous problem. Before the learning-based method, many attempts have been made to address this issue, such as Shape from X [6], where X may represent silhouettes [7], shading [8], or texture [9]. These approaches require strong presumptions and abundant expertise in natural images [10], so they are rarely applied in real-world scenarios. However, after the advent of the learning-based method, many attempts have been successful with strong performances. 3D Variational Autoencoder Generative Adversarial Network (3D-VAE-GAN) [11] uses a generative adversarial network [12] and variational autoencoder [13] to generate 3D objects by taking single-view images as input. Marrnet [14] reconstructs 3D objects by estimating depth, surface normal, and silhouettes of 2D images.

Multiview 3D reconstruction tasks have been studied with algorithm-based methods. Structure-from-motion (SfM) and Simultaneous localization and mapping (SLAM) algorithms require a collection of RGB images. These algorithms have estimated 3D structure by dense feature extraction. and matching [1]. However, these algorithm-based methods have limitations of becoming extremely difficult when multiple viewpoints are separated by a large margin. Furthermore, since inputs are discrete information, it cannot offer full surface of an object which leads to reconstructing incomplete 3D shapes with occluded or hollowed-out areas. With the learning-based method, Pixel2Mesh [15] is the first to reconstruct the 3D shape in a triangular mesh from a single image. Octree Generating Networks (OGN) [16] uses octree to represent high-resolution 3D volumes with a limited memory budget. Matryoshka Networks [17] continuously decomposes a 3D shape into nested shape layers, which outperforms octree-based reconstruction methods. More recently, AttSets [18] used an attentional aggregation module to automatically predict a weight matrix as attention scores for input features. Both 3D Recurrent Reconstruction Neural Network (3D-R2N2) [19] and Learnt Stereo Machines (LSM) [20] are Recurrent Neural Network (RNN) based, resulting in the networks being permutation variants and inefficient for aggregating features from long sequence images.

2.2. Attention Mechanisms

The general form of the attention mechanism is presented below[21].

$$\text{Attention} = f(g(x)|x) \tag{1}$$

Here, $g(x)$ can represent to generate attention which corresponds to the process of paying attention to the discriminate regions. $f(g(x), x)$ means processing input x based on the attention $g(x)$ which is consistent with processing critical regions and getting detailed information. Almost all existing

 doi:10.20944/preprints202407.0857.v1

attention mechanisms can be written into the above formulation. As attention mechanisms have been researched, various attention mechanisms such as "spatial attention [22]: where to pay attention," "Temporal attention [23]: when to pay attention" and "channel attention [24]: what to pay attention" have been proposed. With "self-attention" and "multi-head attention", transformer architecture has been implemented in natural language processing tasks, showing significant results [25].

Several researches used attention mechanism and transformer architecture itself for enhancing 3D reconstruction performance. EVoIT [26] reformulated 3D reconstruction problem as a sequence-to-sequence prediction problem and proposed a 3D Volume Transformer framework inspired from success of transformer. Differently from previous CNN-based networks EVoIT has an advantage by unifying the two stage feature extraction and view fusion in single stage network. Attention mechanism allow them to explore the view-to-view relationships from multi-view input images. Self-attention ONet [27] is an enhanced version of ONet [28] incorporating self-attention mechanism to original 3D object reconstruction model. By employing self-attention mechanism, model could extract global information, ignoring unimportant details and get more consistent meshes. METRO [29] is a mesh transformer framework that reconstruct 3D human pose and mesh from a single input image. By leveraging transformer, it could simultaneously reconstruct 3D human body joints and mesh vertices. VoRTX [30] model for 3D volumetric reconstruction could retain the finer details from fusing multi-view information by performing data-dependent fusion using transformer.

Inspired from these previous works, we introduce multi-head attention refiner that incorporates multi-head attention mechanism to refiner module to recover finer details from coarse volume to dense volume.

3. Proposed Method

3.1. Pix2Vox++ Network Architecture

The Pix2Vox++ network architecture, initially comprising an encoder, decoder, multi-scale context-aware fusion module, and a refiner, has been enhanced in this study. The encoder starts by generating feature maps from input images, which are then processed by the decoder to produce coarse 3D volumes. These volumes are further refined by the multi-scale context-aware fusion module, which selects high-quality reconstructions from all the coarse volumes to create a fused 3D volume.

Our work focuses on improving the existing refiner module, which is crucial in correcting inaccuracies in the fused 3D volume but has limitations in capturing intricate object details and maintaining high recall rates. By introducing a multi-head attention mechanism within the refiner module, we significantly enhanced the model's capability to predict more precise object boundaries, leading to a more accurate and detailed 3D reconstruction. This integration showcases the effectiveness of attention mechanisms in advancing deep learning models for 3D reconstruction tasks.

3.2. Multi-head Attention Refiner

Eqs. 2, 3, and 4 represent the mathematical formulation of the mechanism of self-attention in our Multi-head Attention Refiner. Equation (2) indicates a linear transformation to generate the query vector (Q), the key vector (K), and the value vector (V). Equation (3) represents the computation of attention weights in self-attention. Finally, Equation (4) denotes the final output, which is the product of the attention weights and the value vectors. The architectural update from a standard refiner to one enhanced with multi-head attention is visually summarized in Figure 2, underscoring the refinement process in our proposed model. This integration of attention mechanism enables the refiner to focus on the most informative parts of the 3D volume, resulting in a more accurate reconstruction as shown in our output in Figure 5.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 July 2024 doi:10.20944/preprints202407.0857.v1

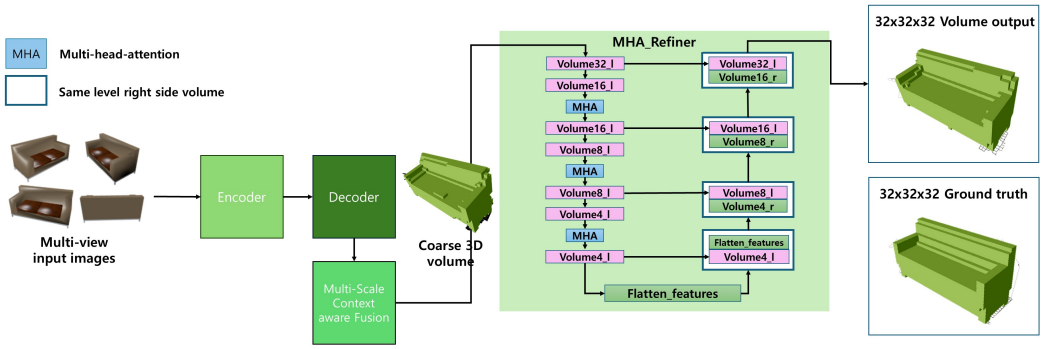

Figure 2. Architectural overview with multi-head self-attention integration in the refiner encoder: This figure demonstrates the integration of Multi-Head Attention within the refiner module of the Pix2Vox++ architecture. The process involves inputting four-view images and generating a refined 3D volume output of $32 \times 32 \times 32$

$$Q, K, V = \text{Linear}(x) \tag{2}$$

$$g(x) = \text{Softmax}(QK) \tag{3}$$

$$f(g(x), x) = g(x)V \tag{4}$$

The architecture of our proposed model is fundamentally based on Pix2Vox++, incorporating Multi-Head Attention (MHA) specifically within the refiner. The multi-view images are generated as a coarse 3D volume through the encoder and decoder stages. This volume is then processed by a refiner integrated with the MHA, i.e. a 3D encoder-decoder with U-net connectivity. The structure of MHA is depicted in Figure 3. Volumes 16-l, Volumes 8-l, and Volumes 4-l served by MHA are skipped and attached to their corresponding Volume 32-r, Volume 16-r, and Volume 8-r, respectively. Volume 8-r, combined with preceding step functions. This process ends with the MHA refiner producing a final volume output of 32^3.

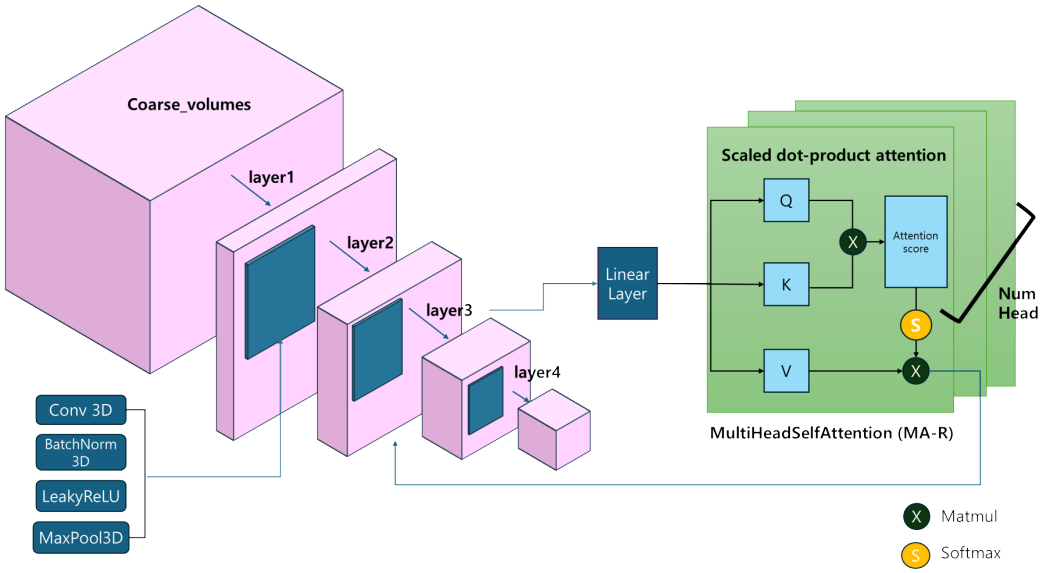

Figure 3. Integration of multi-head self-attention post layer 3: This figure delineates the specific segment within the MA-R refiner where the multi-head self-attention mechanism is employed. It emphasizes the transformation process from the input of the third layer to its output, which then serves as the input for the subsequent fourth layer. This portrays the layered sequential processing and the attention-focused enhancement applied within the refiner's encoder, underlining the effectiveness of the multi-head self-attention in refining 3D reconstruction outputs

 doi:10.20944/preprints202407.0857.v1

Figure 3. Describes how multi-head self-attention is integrated within the encoder section of the refiner. It focuses on the transformation process following layer 3, leading up to the input for layer 4, emphasized by the application of multi-head self-attention. Each layer initially processes the feature map through the convolution 3D layer, batch-normalization 3D layer, LeakyReLu as activation function, and max pooling 3D layer. In layer post-processing, the volume undergoes a linear transformation to set the query (Q), the key (K), and the value (V). The attention score is generated by the multiplication of the matrix of Q and K. These scores, once normalized with a softmax, stabilize the weights, and their multiplication with V produces an output of the original size, which is fed to subsequent layers. A unique feature of this model is the repetition of internal attention that extends over several heads.

4. Experiment

4.1. Datasets

The ShapeNet dataset [31] is a comprehensive and widely-used collection of 3D CAD models, organized based on the WordNet taxonomy. It is renowned for its large scale and diversity, encompassing a wide array of object categories, making it a standard benchmark in the field of 3D object reconstruction. ShapeNet provides richly-annotated 3D models which are crucial for training and evaluating 3D reconstruction algorithms. In this paper, we utilize a subset of the ShapeNet dataset, which includes approximately 44,000 models spanning 13 major categories. This selection is aligned with the datasets used in 3D-R2N2[19], ShapeNetRendering, and ShapeNetVox32. The choice of this subset is driven by our aim to ensure compatibility and facilitate direct comparison with existing studies, particularly those that have employed 3D-R2N2, a well-established framework in multi-view 3D reconstruction. By using this subset, we aim to benchmark our proposed method against established standards in the field, ensuring that our findings are both relevant and comparable within the current research landscape.

4.2. Evaluation Metrics

To evaluate the quality of the proposed methods, we binarized probabilities at a fixed threshold of 0.3 and calculated intersection-over-union (IoU) as a similarity measure between ground truth and prediction.

$$\text{IoU} = \frac{\sum_{i,j,k} I\left(\hat{P}_{i,j,k} > t\right) I\left(\hat{P}_{i,j,k}\right)}{\sum_{i,j,k} I\left[I\left(\hat{P}_{i,j,k} > t\right) + I\left(P_{i,j,k}\right)\right]} \tag{5}$$

More formally, $\hat{P}_{i,j,k}$ and $P_{i,j,k}$ where represent the predicted occupancy probability and the ground truth in (i, j, k), respectively. $\mathrm{I}(\cdot)$ is an indicator function and t denotes a threshold of voxelization. Higher IoU values indicate better reconstruction results.

4.3. Implementation Deatils

We trained the proposed methods with batch size of 64 using 224×224 RGB images as input. The output data are 32^3 voxels. We implemented our networks in PyTorch [32] and trained Pix2Vox++/A using Adam optimizer [33] with β_1 of 0.9 and β_2 of 0.999. The initial learning rate is set at 0.001 and decayed by 2 after 150 epochs. We trained the networks for 250 epochs, while multiscale context-aware fusion does not apply in single-view reconstruction tasks. The environment in the experiment used the A6000 GPU.

4.4. Results

4.4.1. Quantitative Results

Table 1 represents the quantitative evaluation results comparing MA-R with 3D-R2N2, Attsets and Pix2Vox ++ in multiview 3D reconstruction. Mean IoU scores are reported in the table. With the

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 July 2024 doi:10.20944/preprints202407.0857.v1

multi-head attention refiner module, the proposed method acquired superior results when more than five images were used as input. Moreover, the amount of improvements increases as more number of input images are used.

Table 1. Quantitative results comparison

Methods	1view	2view	3view	4view	5view	8view	12view	16view	20view
3D -R2N2	0.560	0.603	0.617	0.625	0.634	0.635	0.636	0.636	0.636
AttSets	0.642	0.663	0.670	0.675	0.677	0.685	0.688	0.692	0.693
Pix2Vox++	**0.670**	**0.695**	**0.704**	**0.708**	0.711	0.715	0.717	0.718	0.719
Ours	0.636	0.681	0.699	**0.708**	**0.713**	**0.721**	**0.726**	**0.729**	**0.730**

4.4.2. Qualitative Results

We also conducted a qualitative evaluation of our proposed method. Figures **??**, Figure 4, and Figure 5 present the results of this evaluation. Figure **??** provides a qualitative assessment of the MA-R performance, specifically focusing on the desk dataset. It compares the 3D volumes with the ground truth, both before and after the application of MA-R, and it is evident that the 3D volume becomes more similar to the ground truth as a result of MA-R.

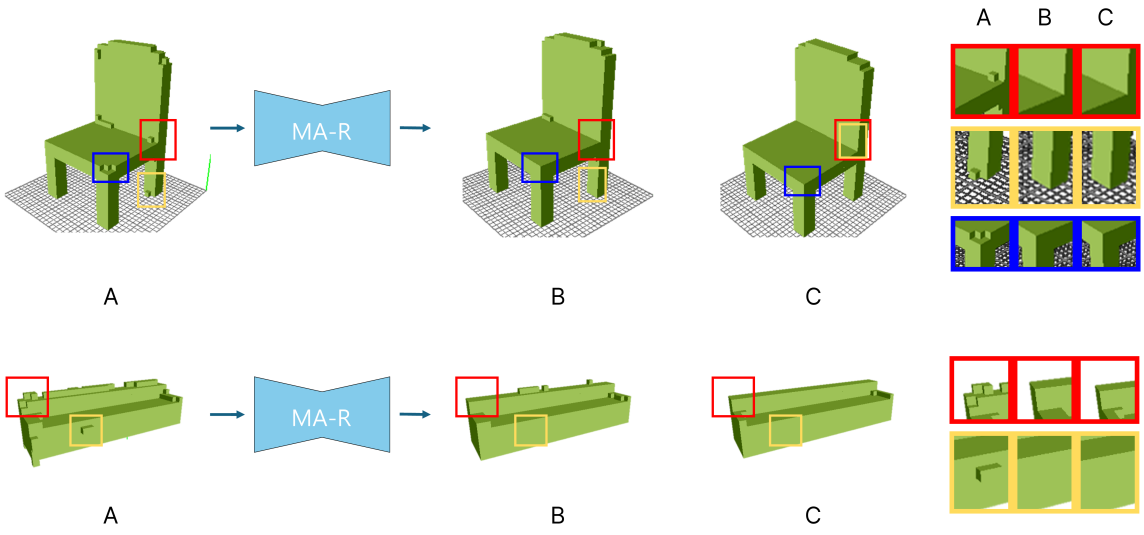

Figure 4. Qualitative demonstration of MA-R performance: Detailed comparison in **??**. Panel A shows the 3D volume before the application of MA-R, Panel B presents the 3D volume after MA-R processing, and Panel C represents the ground truth. The red and yellow boxes highlight areas that were erroneously reproduced in the initial model but have been correctly removed after MA-R, while the blue box indicates a region that was initially missing and has been adequately filled in after MA-R refinement

 doi:10.20944/preprints202407.0857.v1

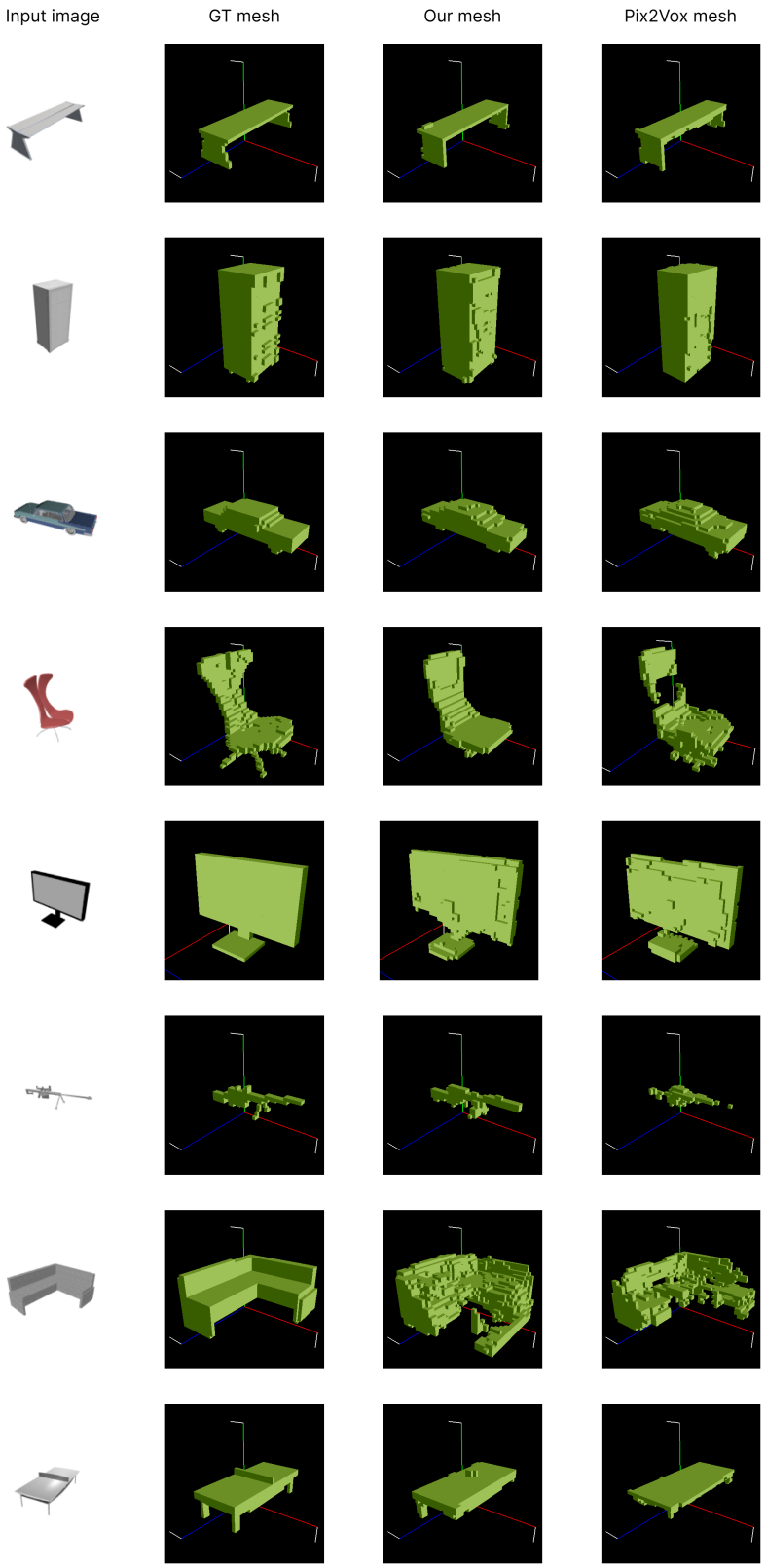

Figure 5. Ground truth meshes and reconstruction results of our model and Pix2Vox++ for various objects: Arranged from top to bottom are the results for bench, cabinet, car, chair, monitor, rifle, sofa and table

 doi:10.20944/preprints202407.0857.v1

In Figure 4, we demonstrate the qualitative evaluation of MA-R performance using the sofa and chair datasets. The improvements made by MA-R are emphasized by using color-coded boxes. The red and yellow boxes indicate that voxels that were present but did not exist in the ground truth have been removed after the application of MA-R. The blue boxes show areas where parts that were missing in the initial model but present in the ground truth have been improved, resulting in a higher fidelity reproduction. This illustrates the superiority of our MA-R in correcting both deficiencies and excesses in the model, thus enhancing overall accuracy.

In Figure 5 we compare our final MA-R output result with the ground truth meshes and the results from Pix2Vox++. In the case of a chair, unlike the results from Pix2Vox++, we did not meet the problem of having holes in the middle of the object. In an additional comparison examples, it is clear that our outputs demonstrate a higher fidelity to the ground truth mesh compared to the outputs obtained from Pix2Vox++.

5. Conclusion

In this study, we proposed the post-processing method for 3D reconstruction models, MA-R (Multi-Head Attention Refiner). This method enhances the model's ability to predict more precise object boundaries and retrieve finer details from coarse volumes. As a result, it facilitates a seamless transition from coarse volumes to dense volumes with remarkable fidelity on target volumes. Our experimental results demonstrate the effectiveness of the proposed MA-R method. Remarkably, we observe improvements of 0.11 in IoU score when employing 16 or 20 input images, highlighting the potential of our method in complex reconstruction scenarios.

However, our approach encounters computational challenges because it aims to enhance the post-processing method to refine the reconstructed volumes. Therefore, there is a need to overcome these limitations by performing light-weighting of the refiner module leveraging techniques such as quantization. Additionally, employing representations like tri-plane instead of volume representation could also be beneficial.

Author Contributions:

Funding:

Conflicts of Interest:

References

1. Lowe, D.G. Distinctive image features from scale-invariant keypoints. *International journal of computer vision* **2004**, *60*, 91–110.
2. Hartley, R.; Zisserman, A. Multiple View Geometry in Computer Vision. In Proceedings of the Cambridge University Press, 2 ed., 2003.
3. Lee, D.T.; Schachter, B.J. Two algorithms for constructing a Delaunay triangulation. *International Journal of Computer and Information Sciences* **1980**, *9*, 219–242. https://doi.org/10.1007/BF00977785.
4. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. In Proceedings of the Advances in Neural Information Processing Systems (NIPS), 2017.
5. Jin, Y.; Jiang, D.; Cai, M. 3D reconstruction using deep learning: A survey. *Communications in Information and Systems* **2020**, *20*, 389–413.
6. Barron, J.T.; Malik, J. Shape, illumination, and reflectance from shading. *IEEE transactions on pattern analysis and machine intelligence* **2014**, *37*, 1670–1687.
7. Dibra, E.; Jain, H.; Oztireli, C.; Ziegler, R.; Gross, M. Human shape from silhouettes using generative hks descriptors and cross-modal neural networks. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 4826–4836.
8. Richter, S.R.; Roth, S. Discriminative shape from shading in uncalibrated illumination. In Proceedings of the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1128–1136.
9. Witkin, A.P. Recovering surface shape and orientation from texture. *Artificial intelligence* **1981**, *17*, 17–45.

 doi:10.20944/preprints202407.0857.v1

10. Zhang, Y.; Liu, Z.; Liu, T.; Peng, B.; Li, X. RealPoint3D: An efficient generation network for 3D object reconstruction from a single image. *IEEE Access* **2019**, *7*, 57539–57549.
11. Wu, J.; Zhang, C.; Xue, T.; Freeman, B.; Tenenbaum, J. Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. *Advances in neural information processing systems* **2016**, *29*.
12. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. *Advances in neural information processing systems* **2014**, *27*.
13. Kingma, D.P.; Welling, M. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* **2013**.
14. Wu, J.; Wang, Y.; Xue, T.; Sun, X.; Freeman, B.; Tenenbaum, J. Marrnet: 3d shape reconstruction via 2.5 d sketches. *Advances in neural information processing systems* **2017**, *30*.
15. Wang, N.; Zhang, Y.; Li, Z.; Fu, Y.; Liu, W.; Jiang, Y.G. Pixel2mesh: Generating 3d mesh models from single rgb images. In Proceedings of the Proceedings of the European conference on computer vision (ECCV), 2018, pp. 52–67.
16. Tatarchenko, M.; Dosovitskiy, A.; Brox, T. Octree generating networks: Efficient convolutional architectures for high-resolution 3d outputs. In Proceedings of the Proceedings of the IEEE international conference on computer vision, 2017, pp. 2088–2096.
17. Richter, S.R.; Roth, S. Matryoshka networks: Predicting 3d geometry via nested shape layers. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 1936–1944.
18. Yang, B.; Wang, S.; Markham, A.; Trigoni, N. Robust attentional aggregation of deep feature sets for multi-view 3D reconstruction. *International Journal of Computer Vision* **2020**, *128*, 53–73.
19. Choy, C.B.; Xu, D.; Gwak, J.; Chen, K.; Savarese, S. 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In Proceedings of the Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14. Springer, 2016, pp. 628–644.
20. Kar, A.; Häne, C.; Malik, J. Learning a multi-view stereo machine. *Advances in neural information processing systems* **2017**, *30*.
21. Guo, M.H.; Xu, T.X.; Liu, J.J.; Liu, Z.N.; Jiang, P.T.; Mu, T.J.; Zhang, S.H.; Martin, R.R.; Cheng, M.M.; Hu, S.M. Attention mechanisms in computer vision: A survey. *Computational visual media* **2022**, *8*, 331–368.
22. Mnih, V.; Heess, N.; Graves, A.; et al. Recurrent models of visual attention. *Advances in neural information processing systems* **2014**, *27*.
23. Xu, S.; Cheng, Y.; Gu, K.; Yang, Y.; Chang, S.; Zhou, P. Jointly attentive spatial-temporal pooling networks for video-based person re-identification. In Proceedings of the Proceedings of the IEEE international conference on computer vision, 2017, pp. 4733–4742.
24. Chen, L.; Zhang, H.; Xiao, J.; Nie, L.; Shao, J.; Liu, W.; Chua, T.S. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 5659–5667.
25. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Advances in neural information processing systems* **2017**, *30*.
26. Wang, D.; Cui, X.; Chen, X.; Zou, Z.; Shi, T.; Salcudean, S.; Wang, Z.J.; Ward, R. Multi-view 3d reconstruction with transformers. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 5722–5731.
27. Salvi, A.; Gavenski, N.; Pooch, E.; Tasoniero, F.; Barros, R. Attention-based 3D object reconstruction from a single image. In Proceedings of the 2020 International Joint Conference on Neural Networks (IJCNN). IEEE, 2020, pp. 1–8.
28. Mescheder, L.; Oechsle, M.; Niemeyer, M.; Nowozin, S.; Geiger, A. Occupancy networks: Learning 3d reconstruction in function space. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 4460–4470.
29. Lin, K.; Wang, L.; Liu, Z. End-to-end human pose and mesh reconstruction with transformers. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 1954–1963.
30. Stier, N.; Rich, A.; Sen, P.; Höllerer, T. Vortx: Volumetric 3d reconstruction with transformers for voxelwise view selection and fusion. In Proceedings of the 2021 International Conference on 3D Vision (3DV). IEEE, 2021, pp. 320–330.

 doi:10.20944/preprints202407.0857.v1

31. Wu, Z.; Song, S.; Khosla, A.; Yu, F.; Zhang, L.; Tang, X.; Xiao, J. 3d shapenets: A deep representation for volumetric shapes. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 1912–1920.
32. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems* **2019**, *32*.
33. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* **2014**.