Supplementary Material of "Long-Range Grouping Transformer for Multi-view 3D Reconstruction"

1. 3D Representation

The current commonly used 3D representation includes mesh [1, 2], point clouds [3, 4], voxel [5, 6] and implicit representation [7, 8]. Our method uses voxel representation without any extra information. However, methods using implicit representation [7, 8] require camera parameters, and mesh-based methods [1, 2] rely on an ellipsoid mesh as an assistive tool.

2. More Experimental Details

2.1. Details of Experiments in Table 1

Table 1 in the main paper shows the performance comparison of our proposed methods (LRGT, LRGT+) and the SOTA methods of multi-view 3D reconstruction [5, 9, 6, 10, 11, 12, 13, 14] on the ShapeNet dataset [15]. Table 1 shows the maximum number of views used to train these methods. Among them, Pix2Vox++ [6] and EVolT [10] do not provide relevant information. According to our experiment, more view inputs during training achieve better performance for processing heavy amount of inputs, however, cost more training consumption. All of these methods adopt not less than 3 views as the maximum number of training inputs and even [9] trains the model with 24 views. LRGT performs well in multi-view reconstruction with only a few inputs during training. It verifies that our method is indeed effective. In addition, we train LRGT on 2 Tesla V100 for 1 day and LRGT+ on 8 Tesla V100 for 2 days.

2.2. Details of Experiments in Figure 5

Figure 5 in the main paper shows the performance comparison of different encoder strategies. Please note that the model with full-range attention (FRA) is different from the full-range method mentioned in Figure 4 of the main paper. Here, FRA is used on all encoder transformer blocks but not only parts of them. In addition, the separated-stage strategy uses the encoder architecture from 3D-RETR [12] and the blended-stage strategy uses the encoder architecture from Legoformer [11].

Number of Views during Training			
3D-R2N2 [5]	**AttSets [9]**	**Pix2Vox++ [6]**	**EVolT [10]**
5	24	-	-
GARNet [14]	**GARNet+**	**Legoformer [11]**	**3D-RETR [12]**
3	8	8	3
UMIFormer [13]	**UMIFormer+**	**LRGT**	**LRGT+**
3	8	3	8

Table 1: The maximum number of view inputs during training each method. "-" indicates the relevant information is not provided in their paper.

3. Supplementary Experiments

3.1. Results on 24-View Reconstruction

As a multi-view reconstruction algorithm, we are concerned about the performance of the model when facing a large number of inputs. Therefore, we compare the performance of LRGT and LRGT+ with the SOTA methods [6, 10, 13] when input 24-view on the ShapeNet dataset and the reconstruction results for each category are shown in Table 2. LRGT and LRGT+ outperform the other methods in all categories.

3.2. Computational Complexity and Inference Time

Computational Complexity. Assume the size of view image is t^2, the token dimension is d, and the number of views and groups are v and g. In our long-range grouping attention (LGA), we have $v \ll g$. The computational complexity of FRA is $O(v^2t^4d)$. Using the standard transformer block to process each view is $O(vt^4d)$. Using our LGA can reduce the complexity to $\boldsymbol{O(v^2\frac{t^4}{g}d)}$.

Inference Time. We compare the inference time of the methods using different grouping strategies in Figure 1. For each model, the architecture setting is the same as Section 4.5 in the main paper. We obtain the average inference time on ShapeNet test set using a single Tesla V100 device when facing a different number of views.

As we mentioned in the main paper, FRA processes tokens from all views at one time. Compared with other grouping strategies, the method with FRA spends the most inference time for multi-view input. When using the basic

	24-view IoU						24-view F-Score@1%					
Category	**Pix2Vox++ [6]**	**EVolT[10]**	**GARNet [14]**	**GARNet+**	**LRGT**	**LRGT+**	**Pix2Vox++**	**EVolT**	**GARNet**	**GARNet+**	**LRGT**	**LRGT+**
airplane	0.729	0.741	0.724	0.739	0.778	**0.793**	0.614	0.636	0.606	0.628	0.678	**0.696**
bench	0.686	0.707	0.698	0.707	0.753	**0.768**	0.522	0.548	0.536	0.551	0.591	**0.607**
cabinet	0.829	0.832	0.841	0.840	0.869	**0.881**	0.456	0.464	0.473	0.505	0.498	**0.520**
car	0.883	0.894	0.888	0.894	0.900	**0.904**	0.598	0.624	0.608	0.623	0.633	**0.643**
chair	0.647	0.681	0.674	0.683	0.722	**0.744**	0.341	0.373	0.369	0.384	0.408	**0.428**
display	0.613	0.674	0.668	0.665	0.750	**0.767**	0.335	0.403	0.386	0.396	0.466	**0.485**
lamp	0.493	0.520	0.516	0.513	0.580	**0.611**	0.351	0.366	0.366	0.369	0.422	**0.456**
speaker	0.762	0.772	0.773	0.772	0.825	**0.839**	0.326	0.339	0.338	0.346	0.399	**0.418**
rifle	0.686	0.711	0.697	0.709	0.763	**0.783**	0.624	0.653	0.634	0.647	0.711	**0.734**
sofa	0.782	0.800	0.807	0.810	0.834	**0.846**	0.454	0.478	0.489	0.500	0.521	**0.536**
table	0.666	0.675	0.693	0.692	0.737	**0.748**	0.419	0.431	0.449	0.452	0.476	**0.485**
telephone	0.849	0.867	0.871	0.879	0.895	**0.898**	0.666	0.687	0.698	0.716	0.719	**0.726**
watercraft	0.668	0.693	0.693	0.696	0.734	**0.747**	0.460	0.494	0.494	0.504	0.550	**0.571**
Overall	0.720	0.738	0.737	0.742	0.779	**0.793**	0.473	0.497	0.493	0.505	0.536	**0.552**

Table 2: Comparison of the performance for 24-view reconstruction on the test set of ShapeNet using IoU and F-Score@1%. The best score for each category is highlighted in bold.

method (baseline) and the remaining four grouping attention methods, it costs relatively close time consumption. Among them, token-range grouping attention (TGA) has the most efficient by a narrow margin, however, it brings a significant loss in performance (referring to Figure 4 in the main paper). In contrast, it is more appropriate to employ LGA exhibits superior performance at similar inference time.

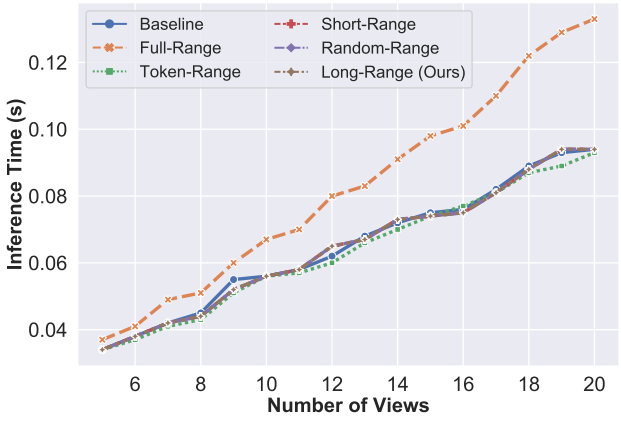

Figure 1: Comparison of the average inference time between different encoder strategies on the test set of ShapeNet. To control the variables, the experiments utilize the same decoder as LRGT.

4. Supplementary Visualizations

4.1. Reconstruction Examples of Table 3

In Figure 2, we supplement multi-view reconstruction examples generated by two models from Table 3 in the main paper. In our decoder, the progressive upsampling reduces reconstruction difficulty. Moreover, with the help of skip connection, HR basic units combine high-level semantic information and low-level semantic information. Two types of semantic information are complementary to each other, further improving the final reconstruction quality.

Figure 2: Two examples on the test set of ShapeNet to compare the qualitative results whether using skip connection in HR basic units of the decoder.

4.2. Visualization of Grouping Diversity

Figure 3 supplements a relatively complete visualization for Figure 6 in the main paper, which provides attention weight maps from various groups in an LGA. Obviously, the complete visualization demonstrates that LGA assem-

bles rich and diverse features from all groups for a reliable representation.

4.3. Multi-View Reconstruction Examples

We supplement more multi-view reconstruction examples generated by LRGT, LRGT+, and other methods [5, 9, 6, 11, 12, 14] on the test set of ShapeNet in Figure 4 and Figure 5.

References

[1] Nanyang Wang, Yinda Zhang, Zhuwen Li, Yanwei Fu, Wei Liu, and Yu-Gang Jiang. Pixel2mesh: Generating 3d mesh models from single rgb images. In *Proceedings of the European conference on computer vision (ECCV)*, pages 52–67, 2018. 1

[2] Chao Wen, Yinda Zhang, Zhuwen Li, and Yanwei Fu. Pixel2mesh++: Multi-view 3d mesh generation via deformation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1042–1051, 2019. 1

[3] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017. 1

[4] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017. 1

[5] Christopher B Choy, Danfei Xu, JunYoung Gwak, Kevin Chen, and Silvio Savarese. 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In *European conference on computer vision*, pages 628–644. Springer, 2016. 1, 3

[6] Haozhe Xie, Hongxun Yao, Shengping Zhang, Shangchen Zhou, and Wenxiu Sun. Pix2vox++: Multi-scale context-aware 3d object reconstruction from single and multiple images. *International Journal of Computer Vision*, 128(12):2919–2935, 2020. 1, 2, 3

[7] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 1

[8] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4578–4587, 2021. 1

[9] Bo Yang, Sen Wang, Andrew Markham, and Niki Trigoni. Robust attentional aggregation of deep feature sets for multi-view 3d reconstruction. *International Journal of Computer Vision*, 128(1):53–73, 2020. 1, 3

[10] Dan Wang, Xinrui Cui, Xun Chen, Zhengxia Zou, Tianyang Shi, Septimiu Salcudean, Z Jane Wang, and Rabab Ward. Multi-view 3d reconstruction with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5722–5731, 2021. 1, 2

[11] Farid Yagubbayli, Alessio Tonioni, and Federico Tombari. Legoformer: Transformers for block-by-block multi-view 3d reconstruction. *arXiv preprint arXiv:2106.12102*, 2021. 1, 3

[12] Zai Shi, Zhao Meng, Yiran Xing, Yunpu Ma, and Roger Wattenhofer. 3d-retr: End-to-end single and multi-view 3d reconstruction with transformers. In *British Machine Vision Conference (BMVC)*, 2021. 1, 3

[13] Zhenwei Zhu, Liying Yang, Ning Li, Chaohao Jiang, and Yanyan Liang. Umiformer: Mining the correlations between similar tokens for multi-view 3d reconstruction. *arXiv preprint arXiv:2302.13987*, 2023. 1

[14] Zhenwei Zhu, Liying Yang, Xuxin Lin, Lin Yang, and Yanyan Liang. Garnet: Global-aware multi-view 3d reconstruction network and the cost-performance tradeoff. *Pattern Recognition*, 142:109674, 2023. 1, 2, 3

[15] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015. 1

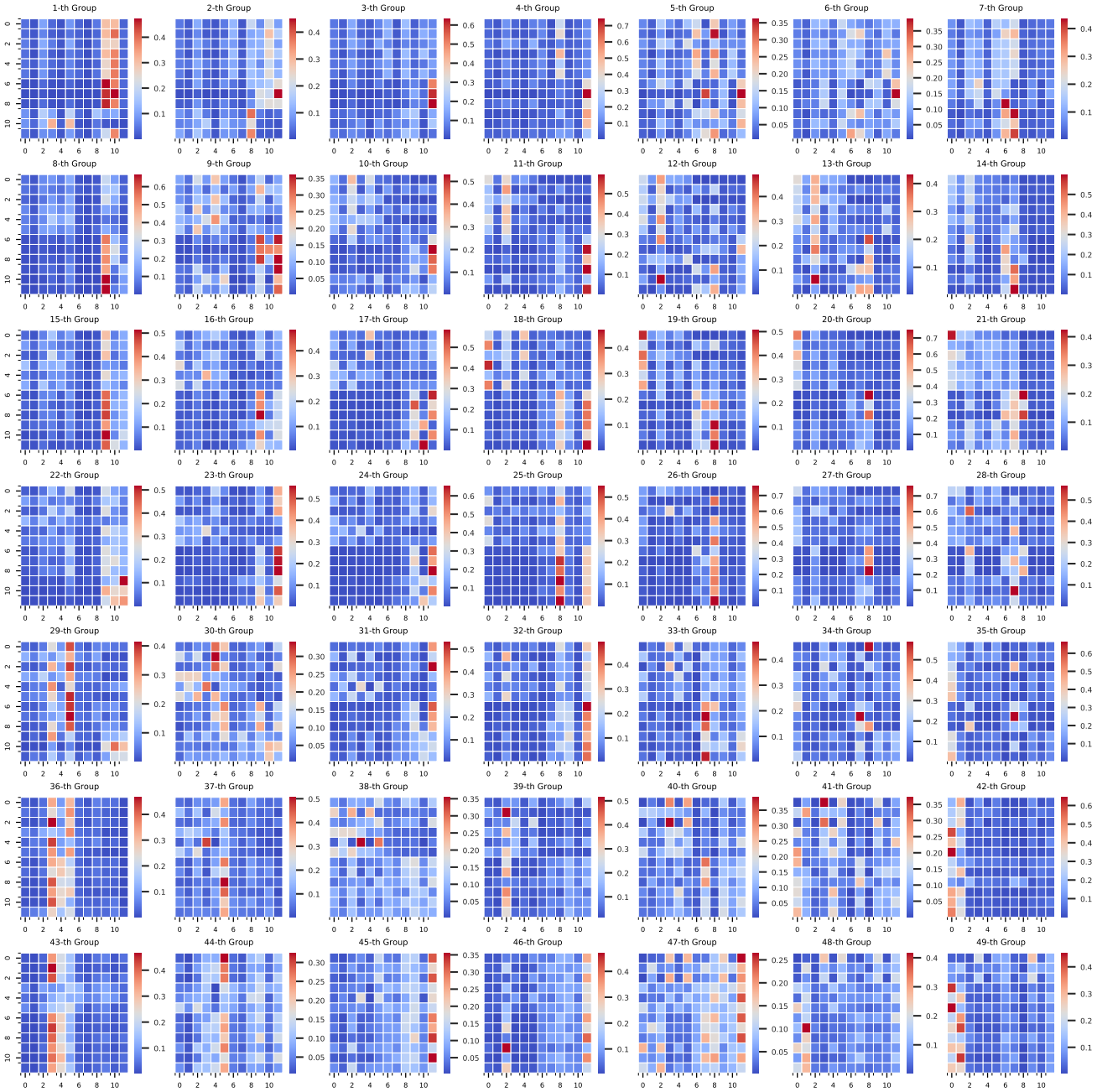

Figure 3: More visualization results of the attention weight maps from different groups in the 1-st head of the 2-nd LGA when processing 3-view input. There is a significant difference in the regions concerned by the attention operations between the groups, which ensures a diversity of the overall features.

Figure 4: Qualitative reconstruction results when facing 5 views, 10 views, 15 views and 20 views as input for telephone, table, and chairs.

Figure 5: Qualitative reconstruction results when facing 5 views, 10 views, 15 views and 20 views as input for chair, cabinet, airplane, and lamp.