Supplementary Material for Beyond Viewpoint: Robust 3D Object Recognition under Arbitrary Views through Joint Multi-Part Representation

Linlong Fan, Ye Huang*, Yanqi Ge, Wen Li, and Lixin Duan

Shenzhen Institute for Advanced Study,
University of Electronic Science and Technology of China
fanlinlong703@163.com

The appendix provides: 1) the implementation details of the PANet; 2) additional experimental comparison results; 3) additional experiment analysis; and 4) additional visualization analysis.

1 Implementation details of PANet

We train PANet in two steps similar to MVCNN [4]. Firstly, the pre-trained view feature extractor is fine-tuned on selected multi-view 2D images for classification. We further train the entire architecture end-to-end, including the attention model and the adaptive part refinement module, to generate multiple parts for 3D object recognition. PANet is trained on the target dataset using AdamW, with a learning rate 0.00001 and $\beta = (0.9, 0.98)$. We use the CosineAnnealingLR scheduler with $T_0 = 5$, $T_{mult} = 2$, and utilize label smoothing regularization with a value of 0.2. In order to facilitate fair comparisons, we also adopt the same viewpoint settings as CVR [6] under arbitrary views, with the number of viewpoints ranging from 6 to 20, as shown in Tab. 10, 11. Our model is implemented based on the Pytorch deep learning platform.

2 Additional experimental comparison results

ModelNet40 and ScanObjectNN For a further fair comparison, we follow the viewpoint sampling strategy of the CVR [6] method, where the number of viewpoints is randomly sampled from 6 to 20. PANet improves instance accuracy and class accuracy by 15.34% and 13.2% in ScanObjectNN [5] dataset, it also improves by 5.01% and 4,7% on the ModelNet [9] dataset. Additionally, we do not observe a significant decrease in performance even with fewer views. It's worth noting that we found from the code of CVR [6] that their viewpoint sampling method is not strictly random. The CVR method only achieves random sampling of viewpoint quantities, while the viewpoint positions are continuous. Therefore, we additionally provide the results of random sampling in Tab. 10, 11, where both the number and positions of viewpoints are randomized. We notice that the performance does not significantly change, which also exceeds the CVR method

* Corresponding author

Table 10: Classification on ScanObjectNN. '-': the result was not reported, 'Aligned' means each 3D object is in the same pose. 'Rotated' means randomly rotating 3D objects through different angles, with each 3D object in a different pose.

Method	#Views	Aligned		Rotated	
		Class Acc.	Ins. Acc.	Class Acc.	Ins. Acc.
MVCNN [4]	20	85.71%	87.82%	78.21%	80.62%
GVCNN [2]	20	86.64%	88.68%	82.86%	83.70%
RotationNet [3]	20	84.88%	86.90%	74.68%	76.16%
View-GCN [7]	20	88.67%	90.39%	81.99%	83.50%
CVR [6]	20	88.39%	90.74%	84.70%	85.59%
View-GCN++ [8]	20	89.10%	91.30%	85.80%	86.80%
Ours	20	**93.21%**	**93.99%**	**86.15%**	**87.82%**
MVCNN	6-20	-	-	58.58%	63.29%
GVCNN	6-20	-	-	58.84%	65.35%
CVR	6-20	-	-	68.07%	71.36%
Ours$_{CVR}$	6-20	92.28%	93.31%	83.41%	84.56%
Ours$_{Random}$	6-20	91.50%	93.14%	82.20%	84.56%
Ours	10-20	91.77%	93.48%	84.08%	86.45%

by 14.13% and 13.2%. Additionally, our results under arbitrary views outperform most methods under fixed viewpoints, such as MVCNN [4], GVCNN [2], view-GCN [7], and RotationNet [3], demonstrating the superiority of PANet.

Experiments in the Large-scale Data Scenarios We validate PANet on the large-scale ShapeNetCore55 dataset and compare with the latest research results. As shown in the Tab.12, PANet outperforms MVCFormer [1], demonstrating robustness in handling large-scale datasets.

3 Additional experiment analysis

Influence of Part Feature Representation We investigate how to represent part features, particularly by comparing two approaches: part fusion representation and joint part representation. The part fusion method first concatenates multiple part features and then uses the fused features for training and prediction. On the contrary, the joint part representation involves first predicting the probability for each part and then combining the prediction scores. All experimental results are summarized in Tab. 13 to validate the advantages of the joint part representation.

Time-cost Comparison We provide a time-cost comparison in Tab.14 to validate that our method's advantages. Additionally, we conduct ablation studies on the APR module and discuss how multi-parts is combined.

4 Additional visualization analysis

Firstly, we provide visualization results of view-GCN [7] and view-GCN++ [8] in Fig. 6. It can be observed that the features extracted by view-GCN or view-

Table 11: Classification on ModelNet. '-': the result was not reported, 'Aligned' means each 3D object is in the same pose. 'Rotated' means randomly rotating 3D objects through different angles, with each 3D object in a different pose.

Method	Aligned			Rotated		
	#Views	Class Acc.	Ins. Acc.	#Views	Class Acc.	Ins. Acc.
MVCNN [4]	20	94.30%	96.35%	20	87.95%	88.17%
GVCNN [2]	20	94.46%	96.07%	20	89.69%	88.10%
RotationNet [3]	20	-	97.37%	20	84.74%	85.29%
View-GCN [7]	20	96.50%	97.60%	20	85.90%	88.25%
CVR [6]	20	95.77%	97.16%	20	**91.12%**	92.22%
View-GCN++ [8]	20	96.50%	97.60%	20	90.70%	92.80%
Ours	20	**96.60%**	**97.73%**	20	90.35%	**92.87%**
MVCNN	6-20	-	-	6-20	78.86%	83.20%
GVCNN	6-20	-	-	6-20	80.98%	83.57%
CVR	6-20	-	-	6-20	84.01%	86.91%
Ours$_{CVR}$	6-20	96.15%	97.37%	6-20	89.00%	91.61%
Ours$_{Random}$	6-20	95.86%	97.00%	6-20	89.05%	91.53%
Ours	10-20	95.44%	97.41%	10-20	89.73%	92.26%

Table 12: Classification on ShapeNetCore datasets.

Methods	Type	Class Acc.	Ins. Acc.
View-GCN++ [8]	Fixed	72.9%	86.0%
MVCFormer [1]		72.5%	86.5%
Ours		**74.5%**	**87.2%**
MVCFormer	Arbitrary	71.5%	84.3 %
Ours		**73.2%**	**85.7%**

GCN++ are not entirely focused on the object. Then, Fig. 7, 8, 9, 10 present supplementary results for part distribution in multiple views. In these images, we showcase the heatmap generated by the PANet attention network. These heatmaps display the regions of interest where the model's attention is focused under different views. By observing these heatmaps, we can notice that the regions of attention vary across different views. This provides more candidate features, thus helping to improve the recognition performance and interpretability of objects.

Table 13: Exploring the impact of different part representation methods, 'part fusion' refers to the utilization of fused features to represent 3D objects, and 'joint part' signifies joint multiple part features to represent 3D objects.

Methods	Operations	Class Acc.	Ins. Acc.
part fusion	**mean**	84.69%	86.79%
	max	85.99%	87.14%
	concat	85.65%	86.79%
joint part	**mean**	**86.50%**	**87.31%**

Table 14: Ablation studies on ScanObjectNN in Rotation Views.

Model	Class Acc.	Ins. Acc.	GFLOPs	Inference(ms)
MVCFormer	85.59%	86.45%	174.95	157.86
PANet(w/o APR)	81.18%	84.39%	**82.91**	**6.72**
PANet(Joint Max)	85.87%	87.14%	148.17	8.28
PANet	**86.15%**	**87.82%**	148.17	7.49

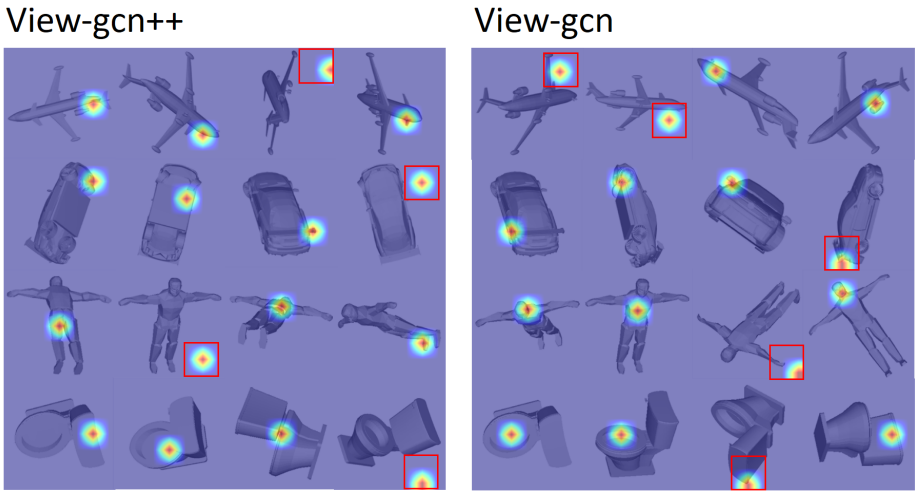

Fig. 6: Left: Visualization results of view-gcn++ and view-gcn. **Right:** Part sampling process.

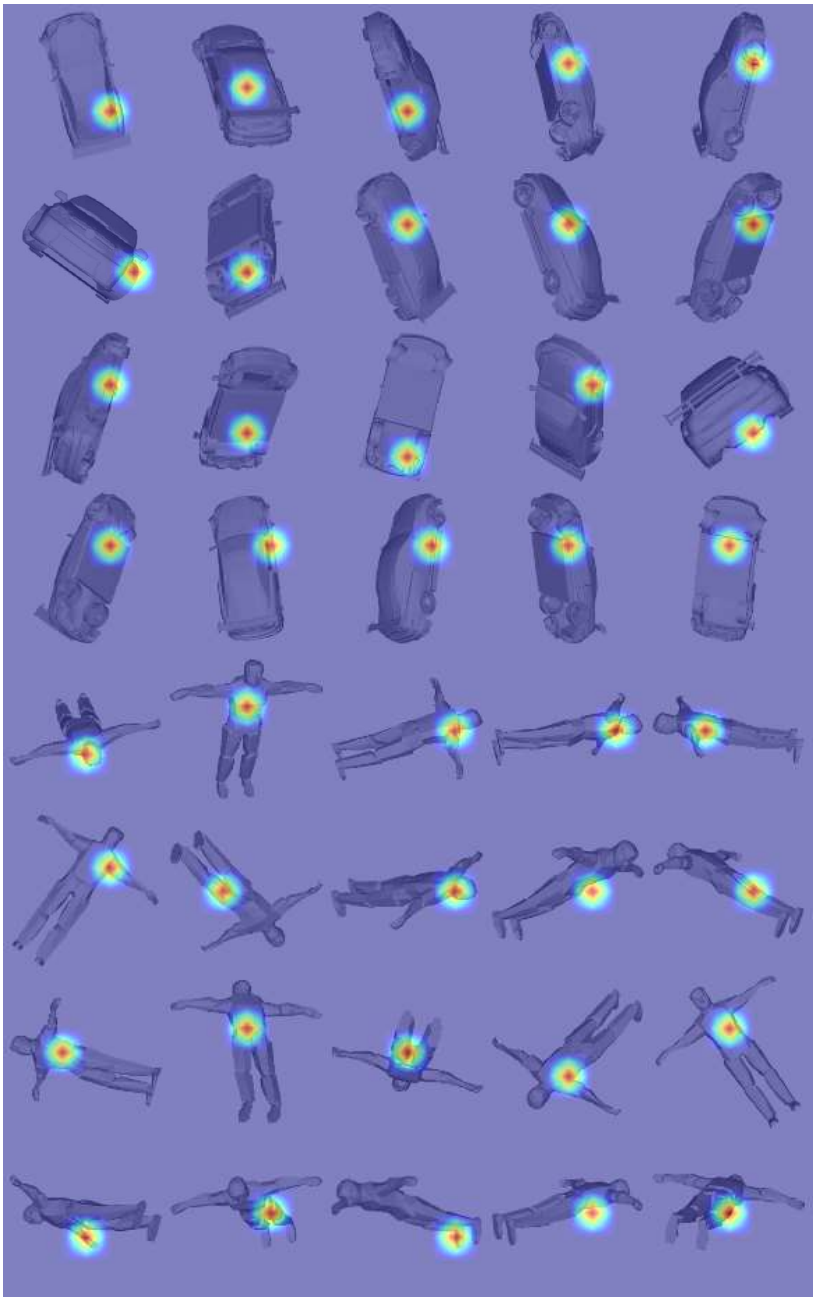

Fig. 7: Addition results for part distribution.

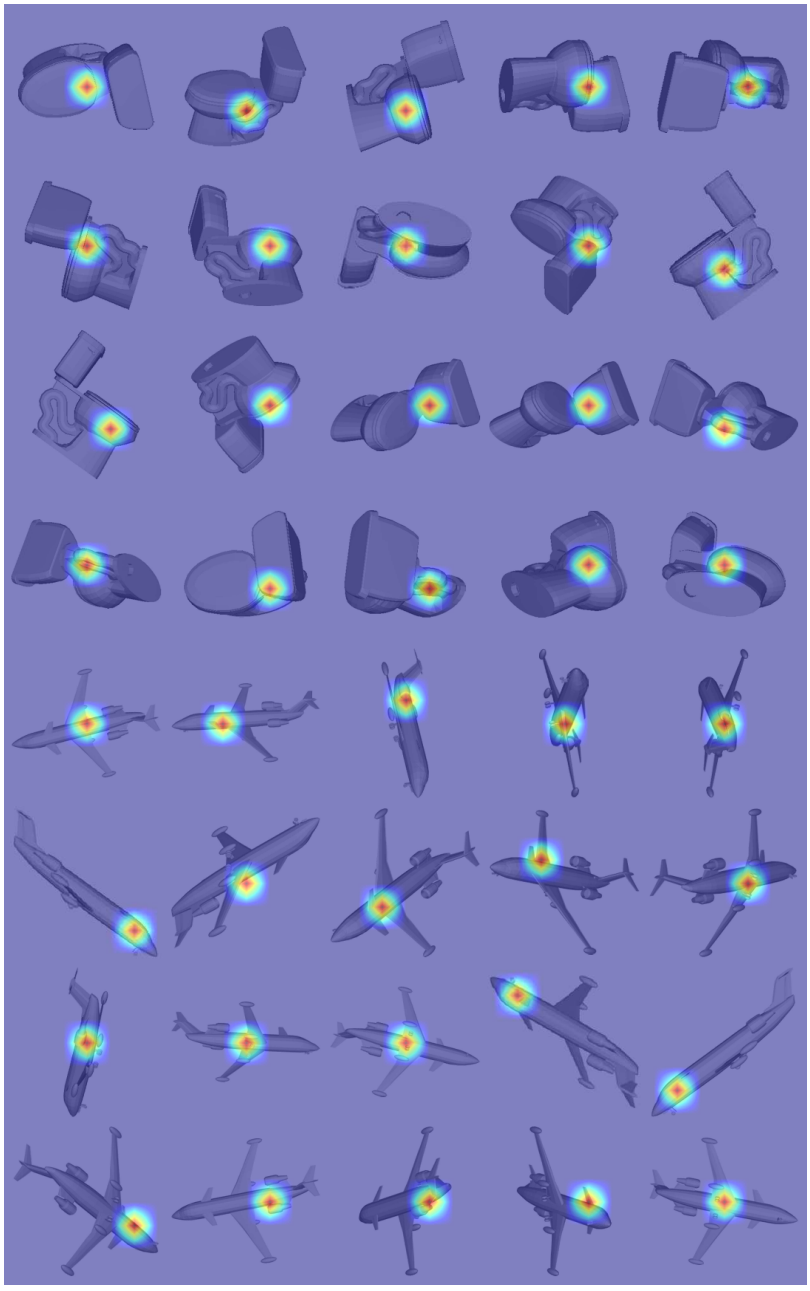

Fig. 8: Addition results for part distribution.

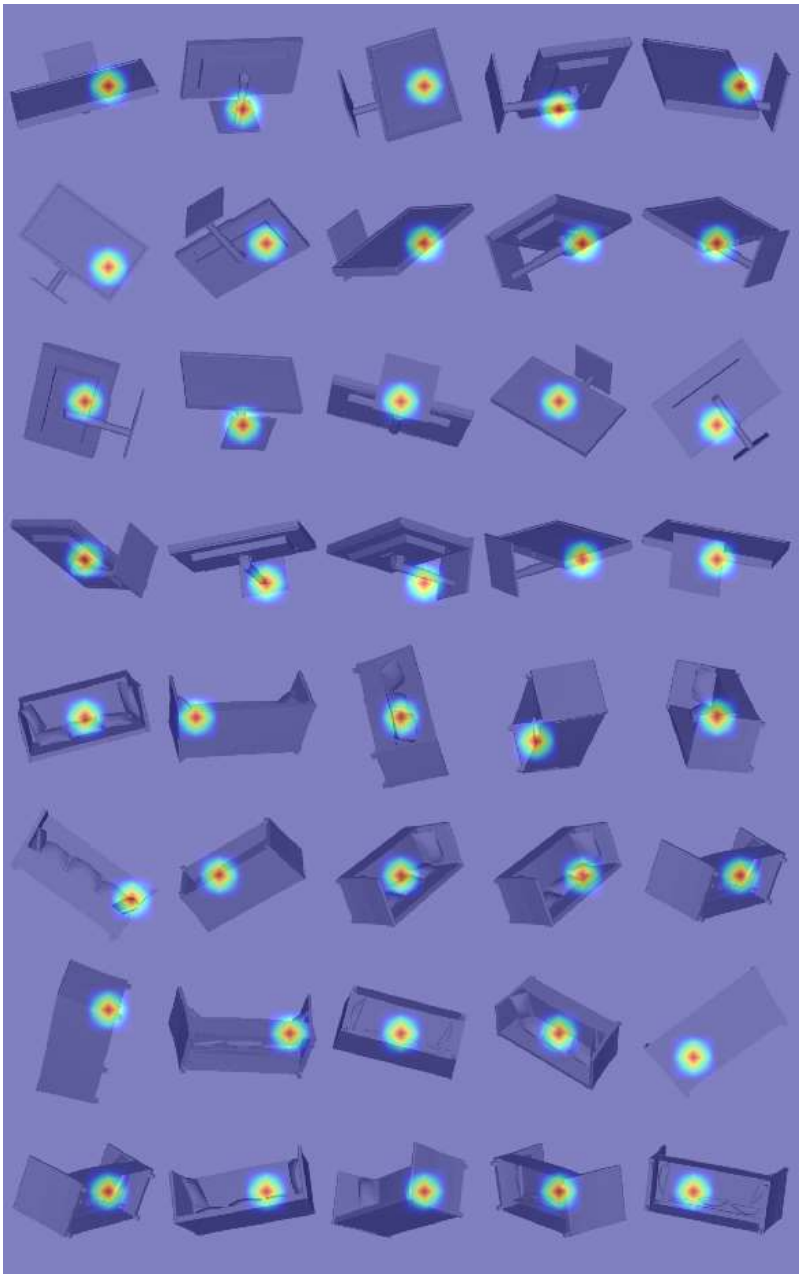

Fig. 9: Addition results for part distribution.

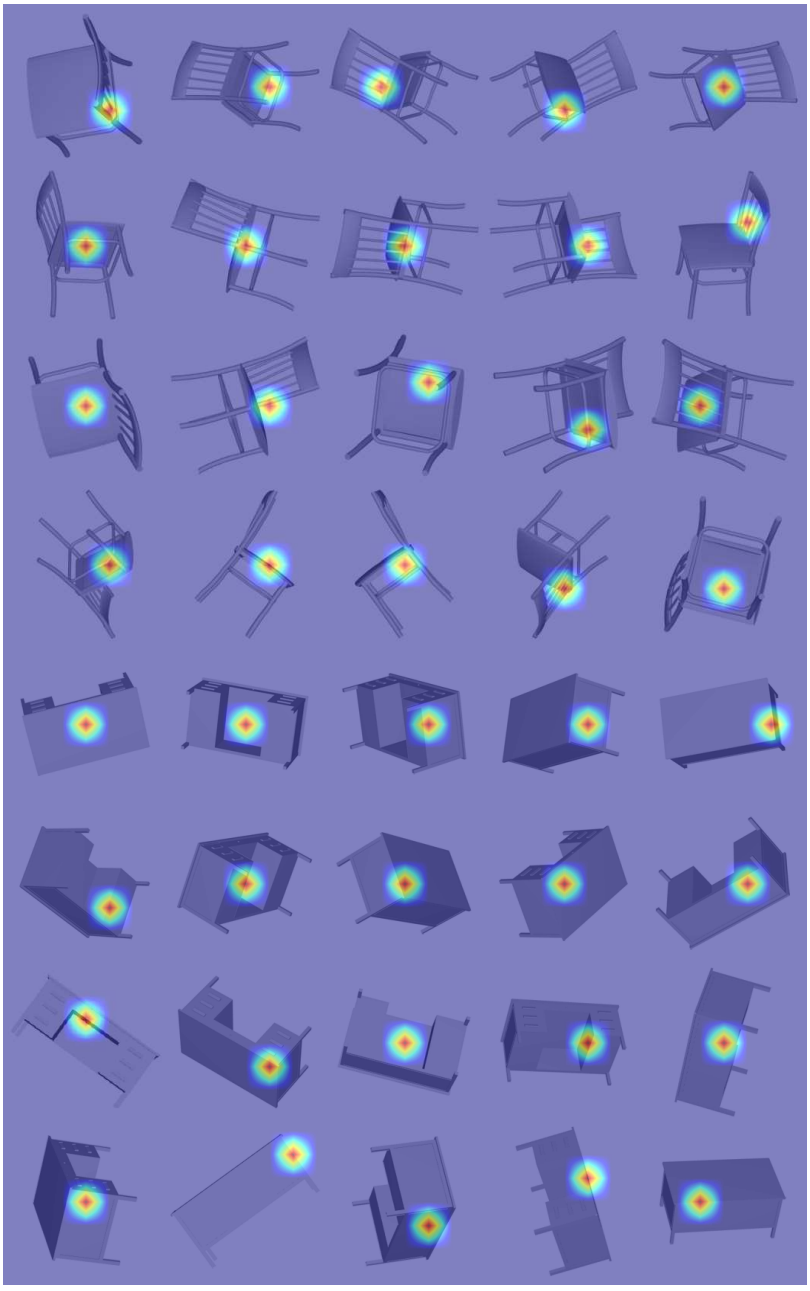

Fig. 10: Addition results for part distribution.

References

1. Fan, L., Ge, Y., Li, W., Duan, L.: Multi-view token clustering and fusion for 3d object recognition and retrieval. In: 2017 IEEE International Conference on Multimedia and Expo (ICME). pp. 1145–1150. IEEE (2023)
2. Feng, Y., Zhang, Z., Zhao, X., Ji, R., Gao, Y.: Gvcnn: Group-view convolutional neural networks for 3d shape recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 264–272 (2018)
3. Kanezaki, A., Matsushita, Y., Nishida, Y.: Rotationnet for joint object categorization and unsupervised pose estimation from multi-view images. IEEE transactions on pattern analysis and machine intelligence **43**(1), 269–283 (2019)
4. Su, H., Maji, S., Kalogerakis, E., Learned-Miller, E.: Multi-view convolutional neural networks for 3d shape recognition. In: Proceedings of the IEEE international conference on computer vision. pp. 945–953 (2015)
5. Uy, M.A., Pham, Q.H., Hua, B.S., Nguyen, T., Yeung, S.K.: Revisiting point cloud classification: A new benchmark dataset and classification model on real-world data. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 1588–1597 (2019)
6. Wei, X., Gong, Y., Wang, F., Sun, X., Sun, J.: Learning canonical view representation for 3d shape recognition with arbitrary views. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 407–416 (2021)
7. Wei, X., Yu, R., Sun, J.: View-gcn: View-based graph convolutional network for 3d shape analysis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1850–1859 (2020)
8. Wei, X., Yu, R., Sun, J.: Learning view-based graph convolutional network for multi-view 3d shape analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence (2022)
9. 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 IEEE conference on computer vision and pattern recognition. pp. 1912–1920 (2015)