Shape-Based Features Complement CLIP Features And Features Learned from Voxels in 3D Object Classification

Zhi Ji
University of Toronto
whitney.ji@mail.utoronto.ca

Michael Guerzhoy
University of Toronto
guerzhoy@cs.toronto.edu

Abstract

Rezanejad et al. [11] recently showed that symmetry-based contour descriptors improve convolutional neural network (CNN) performance on 2D scene categorization, indicating that complex symmetry-based features cannot necessarily be learned and/or represented with CNNs. In this work, we investigate whether there is evidence for a similar phenomenon in 3D visual data. Using $45,949$ *object instances from ScanNet spanning* 440 *classes, we evaluate ten neural architectures across fifteen feature sets, including CLIP [10] embeddings, learned features from voxel, and explicitly computed 3D descriptors: geometric statistics and symmetry-based features extracted with SymmetryNet [14]. We observe that explicit geometric and symmetry-based descriptors consistently provide additional predictive information and improve test classification accuracy. We study the possibility of recovering symmetry-based and geometric features from CLIP embeddings, and we show that they are partially recoverable from CLIP features.*

Our findings extend Rezanejad et al. [11]'s 2D results to 3D, and further demonstrate that symmetry-based and geometric features provide complementary information beyond foundation model embeddings and raw voxel representations. This provides preliminary evidence that global symmetry-based features may be useful in open-world 3D scene understanding.

1. Introduction

Geometric descriptors have long informed visual recognition, even as deep neural networks have come to dominate. Rezanejad et al. [11] showed that 2D symmetry-based contour features derived from medial axis transforms significantly improved CNN scene categorization from contours. Their findings demonstrated that explicit structural descriptors, when combined with learned representations, provide measurable benefit beyond what networks implicitly discover. It remains unclear whether this conclusion extends to 3D, namely whether higher-order 3D shape descriptors provide additional predictive information beyond what modern neural architectures already capture.

In this work, we extend Rezanejad et al. [11]'s 2D findings into 3D. Using $45,949$ ScanNet object instances spanning 440 classes after filtering underrepresented categories, we evaluate explicit geometric statistics and symmetry-based descriptors from SymmetryNet [14] alongside learned embeddings such as CLIP and voxel-learned features. We construct a systematic grid of 10 architectures and 15 input feature sets combinations (150 conditions total), enabling controlled comparison of explicit versus implicit feature utility.

Our results show that symmetry-based and geometric descriptors consistently provide additional information: classification accuracy improves when these features are included, independent of model family. We also demonstrate that CLIP embeddings contain partially recoverable shape information, but remain less reliable than using explicit descriptors directly. Taken together, these findings extend Rezanejad et al. [11]'s 2D results to 3D, showing that explicit computation of symmetry-based and geometry features can complement foundation model features.

Although SymmetryNet features specifically are computed per object, shape-based features such as in Rezanejad et al. [11] can be computed per scene. Our work provides preliminary evidence that shape-based features can be useful in 3D scene understanding and complement foundation model-based features.

2. Background and Related Work

Explicit descriptors and equivariance. Classical shape analysis shows that explicit structural cues can add predictive value: symmetry-based contour descriptors improve CNN scene categorization in 2D [11]. In parallel, E(3)-equivariant networks achieve large gains in data efficiency by hard-wiring geometric symmetries into the architecture [1]. Together, these lines of work suggest that explicitly representing geometry—either as features or as constraints—can complement end-to-end learning [16].

From 2D contours to 3D structure. Transitioning from 2D outlines to 3D shapes introduces pose, occlusion, and volumetric effects that challenge appearance-only pipelines. Explicit 3D symmetry-based descriptors (e.g., SymmetryNet for reflectional symmetries [14]) and compact geometric statistics (e.g., bounding-box ratios, surface/volume surrogates, PCA eigen-structure) provide pose-robust summaries of shape organization. These descriptors are complementary to voxelized occupancy or learned embeddings: the former capture stable global regularities (axes, planes, repetitions), while the latter excel at semantics but may conflate geometry with texture [4, 6, 17, 20].

Limits of foundation models for geometric reasoning. CLIP [10] delivers strong transferable semantics, yet vision models trained on natural images exhibit pronounced texture bias relative to human shape bias [3]. CLIP-style ViT towers also struggle on composition/spatial tests such as Winoground [15], and targeted objectives (e.g., TripletCLIP) improve but do not eliminate these gaps [9]. Moreover, standard ViTs lack built-in rotation/reflection equivariance and positional encodings can disrupt symmetry, motivating the use of explicit geometric signals alongside learned features [2, 5, 7, 8, 12, 13, 18, 19, 21].

Gap and our focus. While there is evidence that explicit geometry helps in 2D [11] and that equivariant designs boost data efficiency [1], there has not been a systematic 3D study that tests across architectures and input sets whether higher-order geometric and symmetry features provide additional predictive information beyond modern learned embeddings. Our work fills this gap via an apples-to-apples evaluation on ScanNet that pairs explicit 3D descriptors (geometry and symmetry) with CLIP embeddings and voxel-learned features, quantifying their complementarity in 3D recognition.

3. Method

We explore whether symmetry-based and geometric features improve classification performance. We then explore to what extent those features can be recovered from CLIP embeddings.

First, we conduct a large-scale classification study on ScanNet object instances, systematically evaluating ten neural architectures across fifteen feature set combinations, yielding 150 experimental conditions. The results (Figure 1) shows how different feature types and model classes contribute to instance-level recognition.

Second, we perform a study where multi-view CLIP embeddings are used to predict explicit symmetry-based and geometric descriptors. This setup explores whether such information can be recovered from vision–language embeddings.

3.1. Dataset

After filtering out classes with fewer than two examples, $45,949$ object instances across 440 valid categories were used in our experiments. We precompute features from meshes, including CLIP embeddings (512D), geometric descriptors (13D), symmetry features (86D), and voxel occupancy grids (32^3). The full class distribution and preprocessing details are provided in our supplementary materials (section 6).

3.2. Input Features

We consider four primary sources of input features:

- **CLIP embeddings (512D).** Extracted from a frozen ViT-B/32 CLIP image encoder applied to 12 rendered views of each ScanNet object instance. Multi-view embeddings are cached to ensure consistency across experiments.
- **Geometric descriptors (13D).** Hand-crafted shape descriptors capturing bounding box aspect ratios, surface-to-volume ratios, and PCA eigenvalue statistics, designed to encode scale- and orientation-independent shape structure.
- **SymmetryNet descriptors (86D).** Learned feature embeddings extracted from a pretrained SymmetryNet encoder. SymmetryNet is a deep network that predicts reflectional and rotational symmetries of 3D objects from single-view RGB-D images, overcoming the limitations of incomplete geometric data. It combines CNN-based appearance features with PointNet geometric features in a multi-task framework that jointly estimates symmetry parameters and dense symmetric correspondences. An optimal assignment mechanism (Hungarian algorithm) enables detection of multiple symmetries, while visibility-based verification handles occlusion. We use the pretrained 86-dimensional SymmetryNet features, which encode global symmetry patterns—capturing reflective, rotational, and translational regularities across diverse 3D object categories.
- **Voxel occupancy grids.** Binary volumetric representations (R^3) of each instance mesh, processed either directly through a 3D CNN or indirectly as precomputed

Table 1. Summary of input feature types used in our ScanNet experiments. Non-voxel features are standardized to zero mean and unit variance. A pre-trained 3D ResNet backbone is used for voxel-related inputs.

Input	**Dim.**	**Constituents**	**Source / Description**
clip	512	CLIP embeddings	Frozen CLIP ViT-B/32 on multi-view renders.
geometric	13	geometry descriptors	Bounding-box ratios, surface/volume stats, PCA eigenvalue ratios, etc.
symmetrynet	86	SymmetryNet features	Symmetry feature vector from SymmetryNet.
geo_clip_concat	$13 + 512 = 525$	geometric + CLIP	Concatenation of geometric descriptors with CLIP embeddings.
sym_clip_concat	$86 + 512 = 598$	symmetry + CLIP	Concatenation of symmetrynet and CLIP embeddings.
sym_geo_concat	$86 + 13 = 99$	symmetry + geometric	Concatenation of symmetrynet and geometric features.
sym_geo_clip_concat	$86 + 13 + 512 = 611$	symmetry + geometric + CLIP	Concatenation of symmetrynet, geometric descriptors, and CLIP.
voxel	32^3 grid	raw voxel grid	End-to-end 3D ResNet on raw occupancy volumes.
geometric_vox_direct_concat	$13 + 512 = 525$	geometric + voxel emb	Fusion: geometric + ResNet3D backbone embedding.
symmetrynet_vox_direct_concat	$86 + 512 = 598$	symmetry + voxel emb	Fusion: symmetry + ResNet3D embedding.
clip_vox_direct_concat	$512 + 512 = 1024$	CLIP + voxel emb	Fusion: clip + ResNet3D embedding.
sym_geo_vox_direct_concat	$99 + 512 = 611$	(sym+geo) + voxel emb	Fusion: sym_geo_concat + ResNet3D embedding.
sym_clip_vox_direct_concat	$598 + 512 = 1110$	(sym+CLIP) + voxel emb	Fusion: sym_clip_concat + ResNet3D embedding.
geo_clip_vox_direct_concat	$525 + 512 = 1037$	(geo+CLIP) + voxel emb	Fusion: geo_clip_concat + ResNet3D embedding.
sym_geo_clip_vox_direct_concat	$611 + 512 = 1123$	(sym+geo+CLIP) + voxel emb	Fusion: sym_geo_clip_concat + ResNet3D embedding.

voxel embeddings (256D).

These inputs are evaluated individually and in concatenated forms, yielding 15 total feature set combinations. Full details of the feature computation are provided in our supplementary materials (section 6), with a summary in Table 1.

3.3. Model Architectures

To probe the interaction between feature type and model class, we instantiate ten neural architectures representing four design families:

1. **Linear Baseline:** A single linear projection (CLIPLINEAR) providing a control for raw feature separability.
2. **Transformers:** (i) CLIPTRANSFORMER, a lightweight 2-layer encoder with learnable [CLS] pooling; (ii) FT-TRANSFORMER, a feature-token transformer adapted for tabular embeddings.
3. **Multi-Layer Perceptrons:** Depth-varying MLPs (MLP1–MLP5) with hidden widths $512 - 768$, ReLU activations, and dropout.
4. **Specialized Models:** (i) MULTIMODAL, a 3-layer fusion MLP for concatenated inputs; (ii) VOXCNN, a 3D CNN for raw voxel grids. Since voxel grids are inherently volumetric data, they are only evaluated with VoxCNN, while all other models are designed for tabular or embedding features rather than raw 3D volumes.

See supplementary materials (section 6) for details and table.

3.4. Complementary Experiments: CLIP → Feature Prediction

In addition to classification, we study whether CLIP embeddings encode sufficient geometric and symmetry information to recover explicit descriptors. For each ScanNet object, we render $V = 12$ views and extract frozen CLIP embeddings (512D each), aggregated into a multi-view token sequence. A ViT-style encoder with a learnable [CLS] token produces a global representation, which is mapped via an MLP head to predict either: (i) Symmetry-based descriptors (1408D), or (ii) Geometric descriptors (13D). Targets are standardized on the training split, and optimization uses a cosine-augmented mean squared error.

4. Experiments and Results

4.1. Instance-Level Classification Results

Figure 1 presents the test classification accuracy across all model architectures and input feature sets on $45,949$ filtered ScanNet instances (See supplementary materials (section 6) for details). Several clear trends emerge:

- **Explicit features improve accuracy.** Incorporating geometric and/or symmetry-based descriptors consistently improves classification over CLIP embeddings alone.
- **Concatenated features dominate.** The strongest results are obtained when CLIP embeddings are combined with voxel/voxel-derived, geometric and/or symmetry-based descriptors.
- **Architectural sensitivity is modest.** While deeper MLPs (MLP2–MLP5) slightly outperform shallower ones, the overall variance across architecture families is smaller than the variance across input feature sets.

These findings demonstrate that explicit geometric and symmetry-based features provide robust additional information for object classification, complementing high-dimensional CLIP embeddings and voxel information.

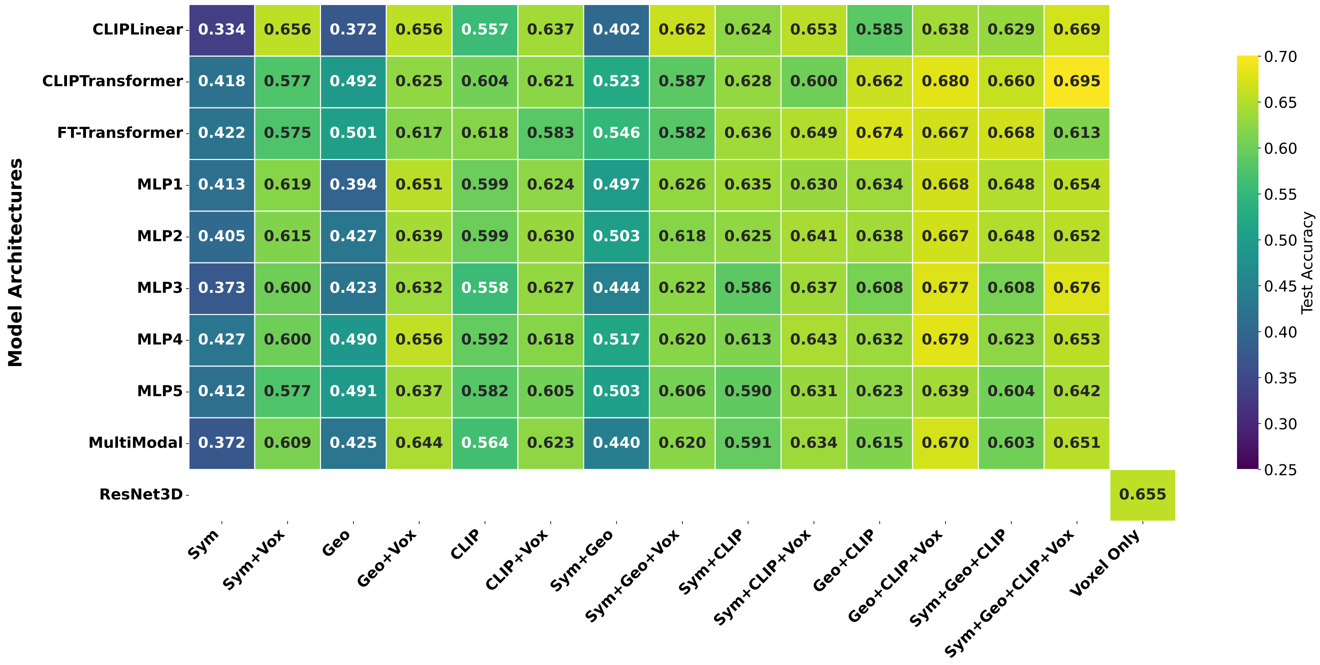

Figure 1. Instance-level classification accuracy on ScanNet across model architectures (rows) and input features(columns). Explicit geometric and symmetry-based features consistently boost performance over CLIP/voxel-only baselines.

4.2. CLIP → Shape-Based Features

To probe whether CLIP embeddings implicitly contain geometric and symmetry-based information, we trained a compact ViT-style 3D Transformer to predict symmetry-based and geometric features.

Cosine similarity is ~ 0.75 when predicting geometric features and ~ 0.68 when predicting symmetry-based features, with MSE around 0.4. These results indicate that geometric and symmetry-based descriptors are partially recoverable from CLIP representations.

5. Conclusions

We studied whether higher-order 3D shape descriptors add predictive value beyond what modern neural architectures implicitly capture. Across $45,949$ ScanNet instances (440 classes), systematic experiments over 10 architectures and 15 input feature sets showed that incorporating geometric or symmetry-based features consistently improves classification performance compared to CLIP embeddings/voxels alone. We show that CLIP embeddings only partially encode geometric and symmetry-based features. These results extend Rezanejad et al. [11]'s 2D findings to 3D, demonstrating that explicit symmetry-based and geometric features complement learned features with additional predictive information. Symmetry-based and geometric features may not be easily learnable or may not be mechanistically representable with our architectures.

Our work provides preliminary evidence for exploring shape-based features in 3D scene understanding. In particular, we provide evidence that as is the case in 2D [11], shape-based features can be useful in 3D as well. Note that the evidence we provide is based on precomputed object meshes. However, shape-based features can be computed for the entire scene as in Rezanejad et al. [11].

6. Supplementary materials

The supplementary materials are available at the following anonymized repository:
`https://anonymous.4open.science/r/opensun3d_sup-8FD3`

References

[1] Simon Batzner, Albert Musaelian, Lixin Sun, Mario Geiger, Jonathan P. Mailoa, Mordechai Kornbluth, Nicola Molinari, Tess E. Smidt, and Boris Kozinsky. E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nature Communications*, 13(1), 2022.

[2] Peijian Ding, Davit Soselia, Thomas Armstrong, Jiahao Su, and Furong Huang. Reviving shift equivariance in vision transformers, 2023.

[3] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel.

Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness, 2022.

[4] Minghao Guo, Bohan Wang, and Wojciech Matusik. Medial skeletal diagram: A generalized medial axis approach for compact 3d shape representation. *ACM Trans. Graph.*, 43(6), 2024.

[5] Nikolai Kalischek, Rodrigo Caye Daudt, Torben Peters, Reinhard Furrer, Jan D. Wegner, and Konrad Schindler. Biasbed - rigorous texture bias evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22221–22230, 2023.

[6] 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. In *Eur. Conf. Comput. Vis.*, 2020.

[7] Katelyn Morrison, Benjamin Gilby, Colton Lipchak, Adam Mattioli, and Adriana Kovashka. Exploring corruption robustness: Inductive biases in vision transformers and mlp-mixers, 2021.

[8] Muzammal Naseer, Kanchana Ranasinghe, Salman Khan, Munawar Hayat, Fahad Khan, and Ming-Hsuan Yang. Intriguing properties of vision transformers. In *Advances in Neural Information Processing Systems*, 2021.

[9] Maitreya Patel, Abhiram Kusumba, Sheng Cheng, Changhoon Kim, Tejas Gokhale, Chitta Baral, and Yezhou Yang. Tripletclip: Improving compositional reasoning of clip via synthetic vision-language negatives, 2024.

[10] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

[11] Morteza Rezanejad, John Wilder, Dirk B. Walther, Allan D. Jepson, Sven Dickinson, and Kaleem Siddiqi. Shape-based measures improve scene categorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(4):2041–2053, 2024.

[12] Renan A Rojas-Gomez, Teck-Yian Lim, Minh N Do, and Raymond A Yeh. Making vision transformers truly shift-equivariant. *arXiv preprint arXiv:2305.16316*, 2023.

[13] David W. Romero and Jean-Baptiste Cordonnier. Group equivariant stand-alone self-attention for vision, 2021.

[14] Yifei Shi, Junwen Huang, Hongjia Zhang, Xin Xu, Szymon Rusinkiewicz, and Kai Xu. SymmetryNet: Learning to predict reflectional and rotational symmetries of 3d shapes from single-view rgb-d images, 2020.

[15] Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *CVPR*, 2022.

[16] Duomin Wang, Yu Deng, Zixin Yin, Heung-Yeung Shum, and Baoyuan Wang. Progressive disentangled representation learning for fine-grained controllable talking head synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17979–17989, 2023.

[17] Pengyuan Wang, Takuya Ikeda, Robert Lee, and Koichi Nishiwaki. Gs-pose: Category-level object pose estimation via geometric and semantic correspondence, 2023.

[18] Zehan Wang, Sashuai Zhou, Shaoxuan He, Haifeng Huang, Lihe Yang, Ziang Zhang, Xize Cheng, Shengpeng Ji, Tao Jin, Hengshuang Zhao, et al. Spatialclip: Learning 3d-aware image representations from spatially discriminative language. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 29656–29666, 2025.

[19] Renjun Xu, Kaifan Yang, Ke Liu, and Fengxiang He. $e(2)$-equivariant vision transformer, 2023.

[20] Zhu Yu, Runmin Zhang, Jiacheng Ying, Junchen Yu, Xiaohai Hu, Lun Luo, Si-Yuan Cao, and Hui-Liang Shen. Context and geometry aware voxel transformer for semantic scene completion, 2024.

[21] Yihan Zeng, Chenhan Jiang, Jiageng Mao, Jianhua Han, Chaoqiang Ye, Qingqiu Huang, Dit-Yan Yeung, Zhen Yang, Xiaodan Liang, and Hang Xu. Clip^2: Contrastive language-image-point pretraining from real-world point cloud data, 2023.