UNIVERSITY OF TORONTO

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

NEURAL INFORMATION PROCESSING SYSTEMS

Zhi (Whitney) Ji, Michael Guerzhoy

whitney.ji@mail.utoronto.ca, guerzhoy@cs.toronto.edu

University of Toronto

Abstract

Understanding how explicit geometric and symmetry information complements learned visual embeddings is important for advancing both 2D and 3D recognition. We investigate this question using large-scale datasets, including ScanNet [1] and MIT67 [2]. Each 3D mesh is represented through multi-view CLIP [3] embeddings, symmetry features extracted with SymmetryNet [4], and explicit geometric descriptors such as PCA statistics. These features are evaluated individually and in combination with voxel embeddings from a pre-trained 3D ResNet. In 2D, we compute contour-based shape maps capturing separation, parallelism, taper, and mirror symmetry, and test all combinations with VGG-16 and CLIP embeddings. **In both 2D and 3D, explicit geometric and symmetry features improve classification accuracy beyond foundation model embeddings alone.** In 3D, the fusion of CLIP, voxel, geometric, and symmetry representations achieves the best performance. Our findings demonstrate that **shape features provide complementary information beyond foundation model embeddings and raw voxel representations**, offering preliminary evidence that global symmetry-based features improve both 2D and 3D object recognition.

Background

- Shape-Based Measures Improve Scene Categorization [5]

Rezanejad et al. note that deep neural networks tend to rely on color and texture features, whereas humans can categorize scenes from outlines, so they introduce medial-axis-based algorithms to detect contour cues such as separation, parallelism, taper, and mirror symmetry and score them using Gestalt grouping rules . They show that weighting contours with these shape-based measures boosts scene categorization accuracy for both human observers and CNNs compared with unweighted contours, indicating that current CNNs do not naturally extract these structural cues.

- SymmetryNet [4]

SymmetryNet is a deep neural network that identifies reflectional and rotational symmetries of 3D objects from a single RGB-D image, using multi-task learning to estimate symmetry axes and point-wise correspondences; this design allows it to detect multiple symmetries per object and achieve strong generalization on a new benchmark.

(preliminary) 2D Experiments and Results

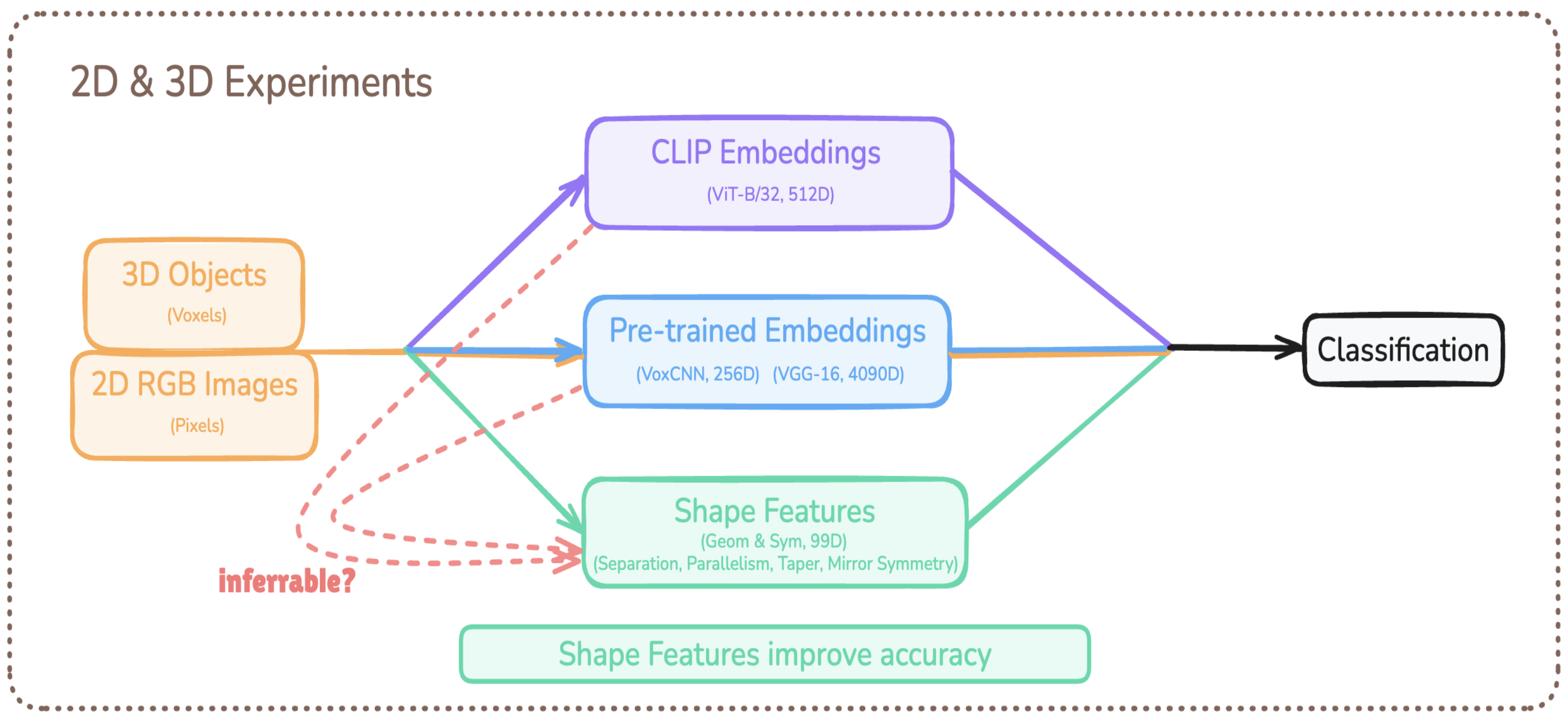
- Validation and test performance across different VGG16 fusion experiments on scene classification.
- Using the MIT67 dataset [2], which contains 67 Indoor categories, and a total of 15620 images. The number of images varies across categories, but there are at least 100 images per category.

Experiment	Validation	Test
VGG16_RGB_only	0.6702	0.6829
VGG16_Shape_only	0.0705	0.0828
VGG16_CLIP_only	0.8885	0.8946
VGG16_RGB_Shape	0.7766	0.7746
VGG16_RGB_CLIP	0.9244	0.9317
VGG16_Shape_CLIP	0.9022	0.9031
VGG16_RGB_Shape_CLIP	0.9432	0.9496

- CLIP [3] embeddings are the most informative features taken by themselves, but shape feature provide additional information.

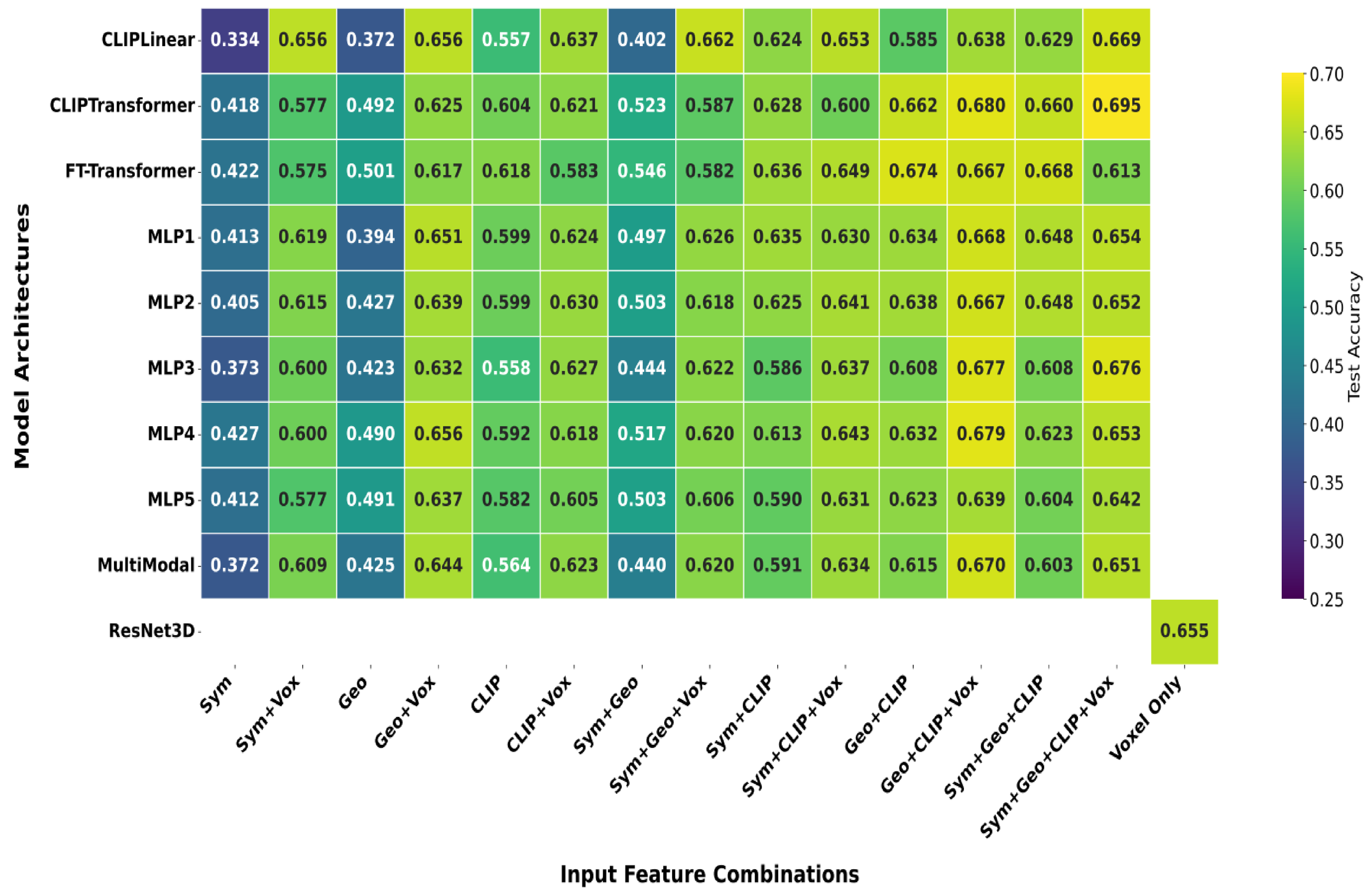
Are shape features recoverable?

- Try predicting symmetry-based and geometric features using CLIP embeddings
- 2D: models do not converge
- 3D: trained a compact ViT-style 3D Transformer, 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 in 3D, although they are derived from multi-view renderings



3D Experiments and Results

- Using the ScanNet [1], an RGB-D dataset containing 2.5 million views in more than 1500 scans, annotated with 3D camera poses, surface reconstructions, and instance-level semantic segmentations.
- Instance-level classification accuracy on 45, 949 filtered ScanNet [1] instances across 10 model architectures (rows) and 15 input features (columns).



- Explicit geometric and/or symmetry-based 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.

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.

Table 2: Model architectures used in our 3D experiments.

Model	Architecture type	Core design / depth	Inputs
CLIPLinear	Linear classifier	Single fully connected layer (LinearHead) to logits; dropout 0.10	Tabular
CLIP Transformer	TinyTransformer	Project to $d = 192$; prepend learnable [CLS]; 2 encoder layers ($n_{\text{head}} = 6$, FF=384); dropout 0.10	Tabular
FT-Transformer	Feature-token Transformer	Tokenize to $d = 256$; [CLS] pooling; 2 encoder layers ($n_{\text{head}} = 8$, FF=512); dropout 0.10	Tabular
MLP1	Fully connected (ReLU, Dropout)	Depth = 1; hidden = 512; dropout 0.10	Tabular[1]
MLP2	Fully connected (ReLU, Dropout)	Depth = 2; hidden = 640; dropout 0.10	Tabular[1]
MLP3	Fully connected (ReLU, Dropout)	Depth = 3; hidden = 768; dropout 0.10	Tabular[1]
MLP4	Fully connected (ReLU, Dropout)	Depth = 4; hidden = 768; dropout 0.10	Tabular[1]
MLP5	Fully connected (ReLU, Dropout)	Depth = 5; hidden = 768; dropout 0.10	Tabular[1]
MultiModal	MLP for tabular concatenations	Depth = 3; hidden = 768; dropout 0.10	Tabular[1]
ResNet3D	3D ResNet backbone	Pretrained r3d_18 (default; or mc3_18); input voxels 32^3 with depth as time; $1 \to 3$ channel repeat; global avg pool $\to$ linear head	Voxel (raw)

[1] For any * vox direct concat column, the same heads (Linear/Transformer/MLP) are used but preceded by a ResNet3D backbone. The voxel grid is encoded to a 512-D embedding (vox_emb_dim=512), concatenated with tabular features, and trained end-to-end.

References

[1] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser and M. Nießner, "ScanNet: Richly-Annotated 3D Reconstructions of Indoor Scenes," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, 2017, pp. 2432-2443.

[2] A. Quattoni and A. Torralba, "Recognizing indoor scenes," *2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Miami, FL, USA, 2009, pp. 413-420.

[3] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning Transferable Visual Models From Natural Language Supervision. *International Conference on Machine Learning (ICML).*

[4] Yifei Shi, Junwen Huang, Hongjia Zhang, Xin Xu, Szymon Rusinkiewicz, and Kai Xu. 2020. SymmetryNet: learning to predict reflectional and rotational symmetries of 3D shapes from single-view RGB-D images. ACM Trans. Graph. 39, 6, Article 213 (December 2020), 14 pages.

[5] 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.

Please see our paper here →

Workshop on Symmetry and Geometry in Neural Representations @ NeurIPS 2025, San Diego, CA