CNN vs. CvT for Geometric Scene Understanding: Evaluating Architecture and Iterative Weight-Sharing

Mateusz Juszczak, Anna Zalesinska

Poznań University of Technology, Computer Vision

Abstract

We compare CNN, CvT, and Recurrent CvT on joint prediction of shape, color, material, rotation, position, and scale using 50k synthetic Blender scenes and a **3-rule OOD test split**.

Key Findings:

- **CNN leads classification:** **96.3%** shape accuracy vs. 92–93% for Transformers
- **Geometry parity:** All three tie on rotation (MAE 0.63–0.65)
- **Efficiency:** Recurrent CvT matches geometry with **28% fewer params**
- **Failure mode:** Rotation error tracks object visibility, not novelty

Introduction & Hypothesis

Research Question: *"Does architectural complexity (attention, recurrence) improve multi-task attribute prediction and OOD generalization compared to a CNN of equal parameter budget?"*

We hypothesize that convolutional inductive bias provides superior classification features, while geometric regression can be solved iteratively without deep unshared layers.

Related Work

- **CvT** (Wu et al., ICCV 2021): Conv Q/K/V combines local priors with global attention
- **ViT** (Dosovitskiy et al., ICLR 2021): CNNs outperform ViT on small datasets
- **Circular loss** (Zhou et al., CVPR 2019): $\mathcal{L}_{\text{rot}} = 1 - \cos(\hat{\theta} - \theta)$
- **Multi-task learning** (Caruana, 1997): shared backbone over heterogeneous tasks
- **CLEVR** (Johnson et al., CVPR 2017): synthetic visual reasoning benchmarks

Data

50,000 Blender scenes, 256×256 px, single object per image.

Attribute	Range / Classes
Shape	Cut Cone, Cut Cube, L, Macaroni
Color	8 distinct classes
Material	Rubber, Metal
Position	$x, y \in [-3, 3]$
Rotation	$\theta \in [0°, 360°)$
Scale	$s \in [0.30, 1.00]$

OOD Split – 3 Withheld Domains:

Rule	Condition	Tests
Combo	4 shape–color pairs	Attribute binding
Size	$s<0.35$ $s>0.95$	or Scale extrapolation
Pos	$\lvert x\rvert>2.2$ $\lvert y\rvert>2.2$	or Boundary clipping

Source Code

https://github.com/Gienkooo/wk-proj

Methods

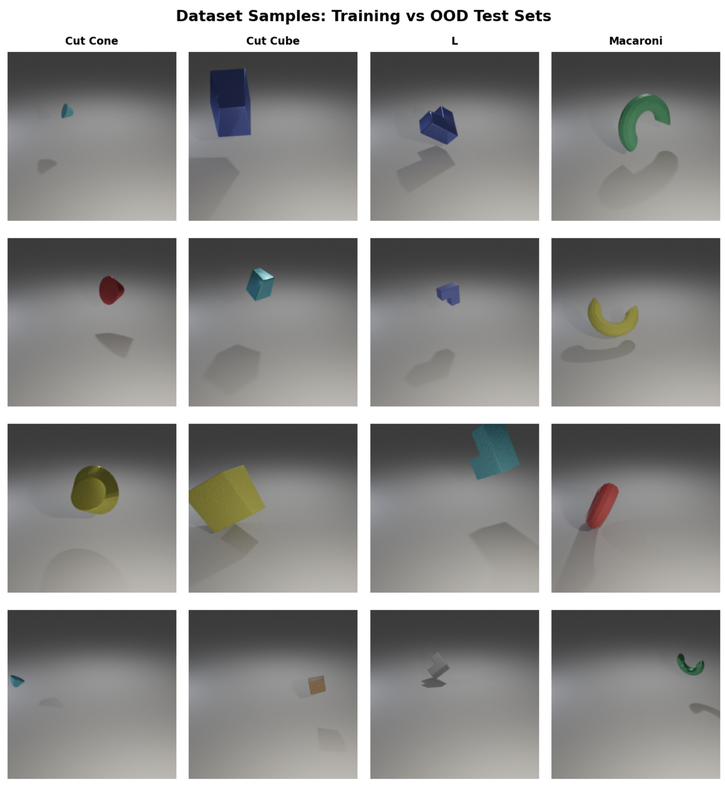
All architectures trained with **identical data, optimizer, and loss.**

1. CNN (∼560k): ResNet-style, 4 residual blocks, BatchNorm+ReLU, AdaptiveAvgPool, per-task linear heads.

2. CvT (∼523k): 4-stage conv tokenizer ($256 \to 8$ px spatial), depthwise Conv3×3 Q/K/V, positional embedding at 8×8 only, **3 independent blocks** at stage 4, spatial-attention pooling: $\mathbf{z} = \sum_i \mathrm{softmax}(\mathbf{w})_i \cdot \mathbf{f}_i$

3. Recurrent CvT (∼404k): identical tokenizer, stage 4 replaced with **1 shared block × 4 passes:** $\mathbf{h}_t = \mathrm{CvTBlock}(\mathbf{h}_{t-1}),\ t{=}1 \ldots 4$

Loss: $\mathcal{L} = \sum_k \lambda_k \mathcal{L}_k$, $\mathcal{L}_{\text{rot}} = 1 - \cos(\hat{\theta} - \theta)$, AdamW (lr=$10^{-4}$), early stop p=15.



Results

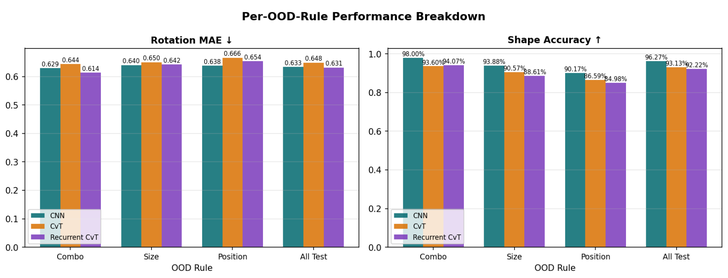
Model	Par.	Shape↑	Rot↓	Pos↓	Scale↓
CNN	560k	**96.3%**	**0.633**	**0.152**	0.040
CvT	523k	93.1%	0.648	0.157	**0.039**
Rec. CvT	404k	92.2%	**0.631**	0.187	0.048

CNN wins classification; all three tie on rotation. Recurrent CvT achieves this with the fewest parameters.

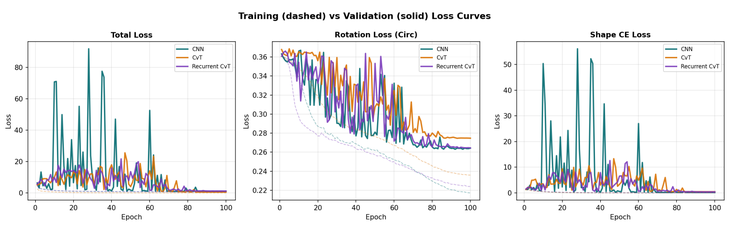
Scale Extrapolation (Size OOD only):

Subset	CNN	CvT	Rec. CvT
Tiny ($s{<}0.35$, $n{=}354$)	0.034	0.035	0.037
Large ($s{>}0.95$, $n{=}393$)	0.054	**0.051**	0.078

Recurrent CvT degrades on large objects – shared weights cannot specialise across passes when activations shift out of training range.



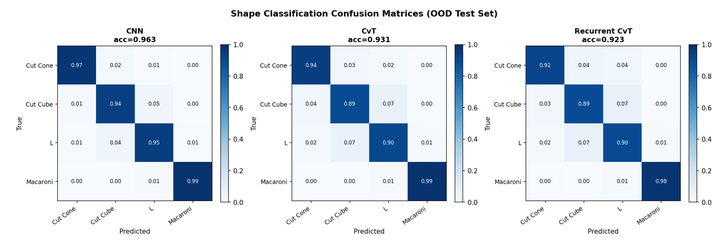
Combo OOD = best rotation (full object visible). Pos OOD = worst rotation (silhouette clipped at boundary).



Periodic val-loss spikes in all models – circular-loss gradient variance, not architecture-specific.

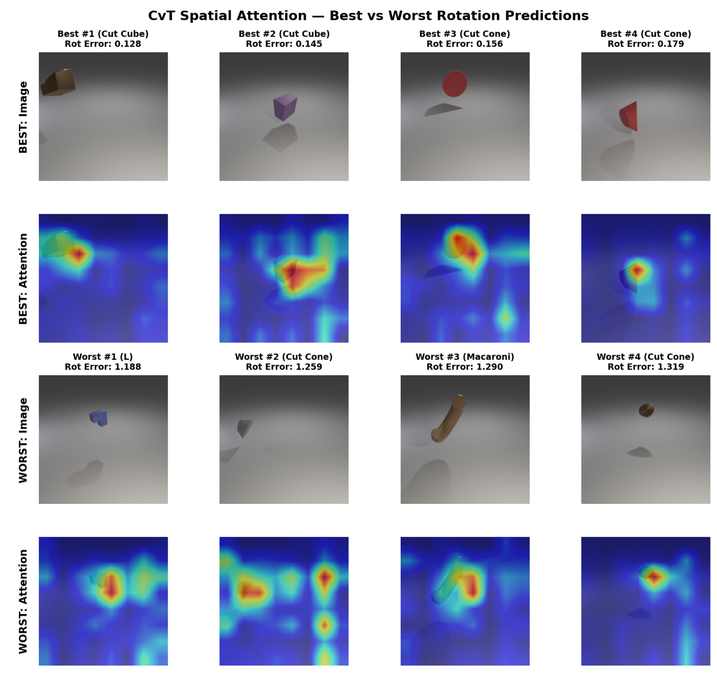
Error Analysis

Shape Confusion Matrices:



CNN confuses Cut Cone ↔ Macaroni under Size OOD (edges blur at tiny scale). Transformers show wider confusion – weaker classification inductive bias.

CvT Spatial Attention – Best vs. Worst Rotation:



Correct predictions focus on the object's principal axis. Pos OOD failures scatter attention toward the clipped boundary – directly visualising the visibility–error relationship.

Conclusion

- **No rotation advantage:** All models tie once circular loss is applied correctly.
- **CNN wins classification:** Consistent 4–5% shape accuracy advantage across all OOD rules.
- **Recurrent efficiency:** Weight-sharing beats independent depth – same geometry at 28% fewer parameters.
- **Geometric visibility:** Rotation error tracks object completeness (Combo best; Pos worst), not attribute novelty.
- **Scale extrapolation:** Only condition where architecture matters for regression – RecCvT degrades on large objects (MAE 0.078 vs. 0.051 CvT).

Limitations: Rotation-range OOD not tested. Macaroni 180° symmetry creates irreducible error floor. CvT vs. RecCvT differ only in stage 4 – not fundamentally different architectures.

Future work: Multi-object scenes; withheld rotation ranges; real-world domain transfer.

References

- Wu et al., *CvT: Introducing Convolutions to Vision Transformers*, ICCV 2021.
- Dosovitskiy et al., *An Image is Worth 16x16 Words*, ICLR 2021.
- Zhou et al., *On the Continuity of Rotation Representations*, CVPR 2019.
- Johnson et al., *CLEVR: A Diagnostic Dataset*, CVPR 2017.
- PyTorch, MLflow, Blender – implementation and tracking.
- Perplexity AI – architectural debugging and literature guidance.