Approximate Analysis by Synthesis and Robustness to Out of Distribution Data

Alan Yuille

Bloomberg Distinguished Professor

Depts. Computer Science and Cognitive Science

The need for OOD Testing of AI algorithms

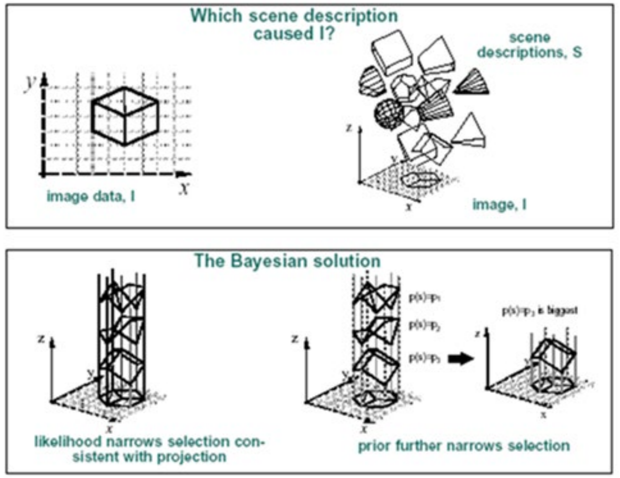
- We typically test AI performance on IID data which comes from the same source as the training data.
- ***But it is more interesting, and insightful, to test algorithms on Out-of-Distribution (OOD) data which comes from a different source than the training data.***
- This ensures that the algorithms have deeper understanding of the data and are more likely to perform well in real world situations.
- ***Algorithms which perform well on IID data are like intelligent parrots. Feedforward deep networks are parrots, admittedly very intelligent in some cases.***

An Old Idea: Analysis by Synthesis

- Analysis by Synthesis (Ulf Grenander) proposes that to “solve” vision requires inverting the image formation process.
- ***We should study synthesis -- how images are generated– in order to perform analysis and determine what visual scene is most likely to have generated the images.***
- This can be formulated as Bayesian inference using a likelihood P(I|W) and a prior P(W).
- A modern version of this theory is inverse computer graphics.
- ***This relates to classic theories of visual perception by Helmholtz and Gregory.***

Bayesian Formulation: P(I|W), P(W)

- Why do we perceive a cube? The likelihood P(I|W) constrains the world state W and is supplemented by prior world knowledge P(W). (P. Sinha).
- ***Why do we perceive a woman on a flying carpet? Because we reason about where the shadow comes from (making a mistake) and make an interpretation of the 3D world. (A. Yuille and D. Kersten).***



Analysis by Synthesis and Human Perception

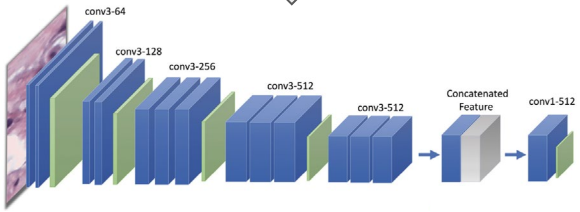
- Studies of human perception suggest that humans do “approximate and as-needed analysis by synthesis” (D.Kersten and A Yuille). This may be done by exploiting the feedforward and feedback pathways in the visual cortex (D.B. Mumford).

- ***This is consistent with studies of human cognition (J.B. Tenenbaum et al.) where the goal of vision is to construct 3D representations of objects and scenes which can be used for higher level cognition with commonsense reasoning (social knowledge and intuitive physics). This world knowledge is learnt during development using multi-modal cues.***

Deep Neural Networks vs. Approximate Analysis by Synthesis

Deep neural networks

Feedforward classifier



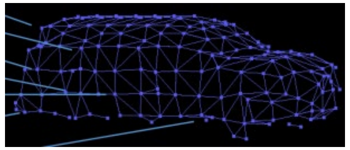
- Designed for **specific** visual **tasks**
- **Vulnerable** to **occlusions** and **domains**
- Only work on a **single viewpoint**

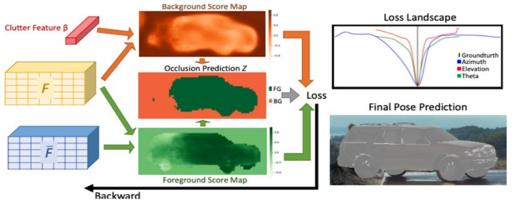
Approximate Analysis by Synthesis

3D generative models

Render-and-Compare

+





+ Work with **multiple** visual **tasks**
+ **Robust** to **occlusions** and **domains**
+ Generalization to **novel viewpoints**

Approximate Analysis by Synthesis: Render and Compare

We specify a generative model which renders the neural features conditioned on the 3D object and its 3D pose. Angtian Wang et al. Wufei Ma et al. These are for objects in the Pascal 3D+ dataset.

This generative model is factorizable which enables rapid inference. The backbone is trained so that the neural features are invariant to viewpoint.

Ideal Analysis by Synthesis would use non-factorizable generative models of images.

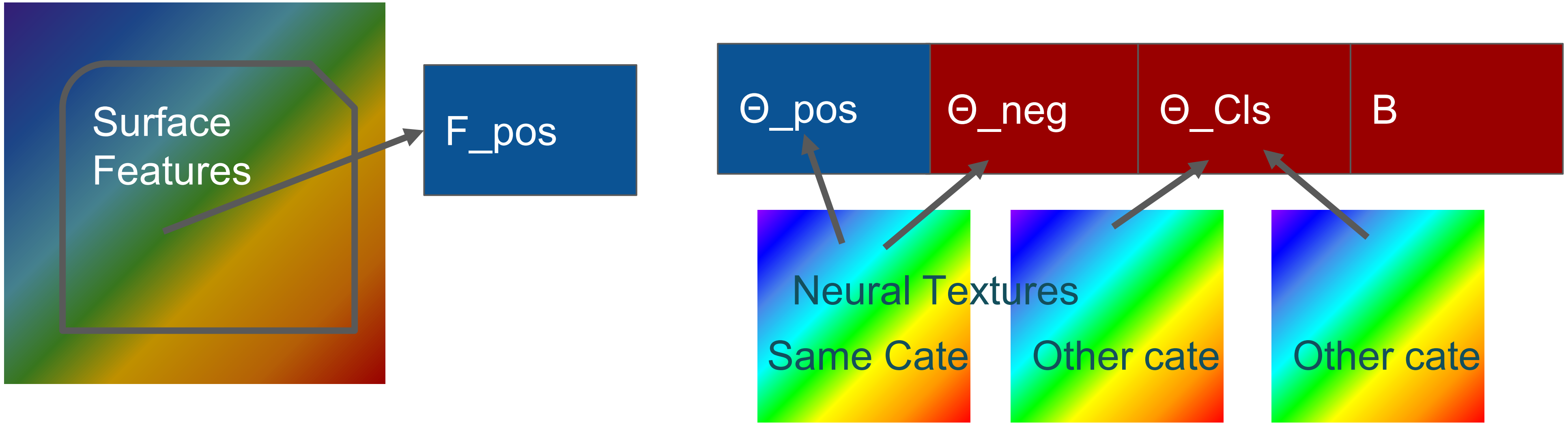
The likelihood of the features F conditioned on object class c and pose m:

$$p(F|\Gamma, \Theta, c, m, B) = \prod_{i \in FG} p(f_i|\Gamma, \Theta, c, m) \prod_{i \in BG} p(f_i|B)$$

$$p(f_i|\Gamma, \Theta, c, m) = \sum_b \frac{\alpha_b}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}\|f_i - \theta_{u,v,c,b}\|^2\right)$$

$$p(f_{i'}|B) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}\|f_{i'} - \beta\|^2\right)$$

Contrastive learning of neural features and backbone.



$$\mathcal{L}(F,\Gamma,\Theta,m,B) = \mathcal{L}_{ML}(F,\Gamma,\Theta,m) + \mathcal{L}_{Object}(F,\Theta_{\mathcal{S}}) + \mathcal{L}_{Class}(F,\Theta_{\mathcal{S}}^{\mathbf{C}}) + \mathcal{L}_{Back}(F,\Theta_{\mathcal{BG}})$$

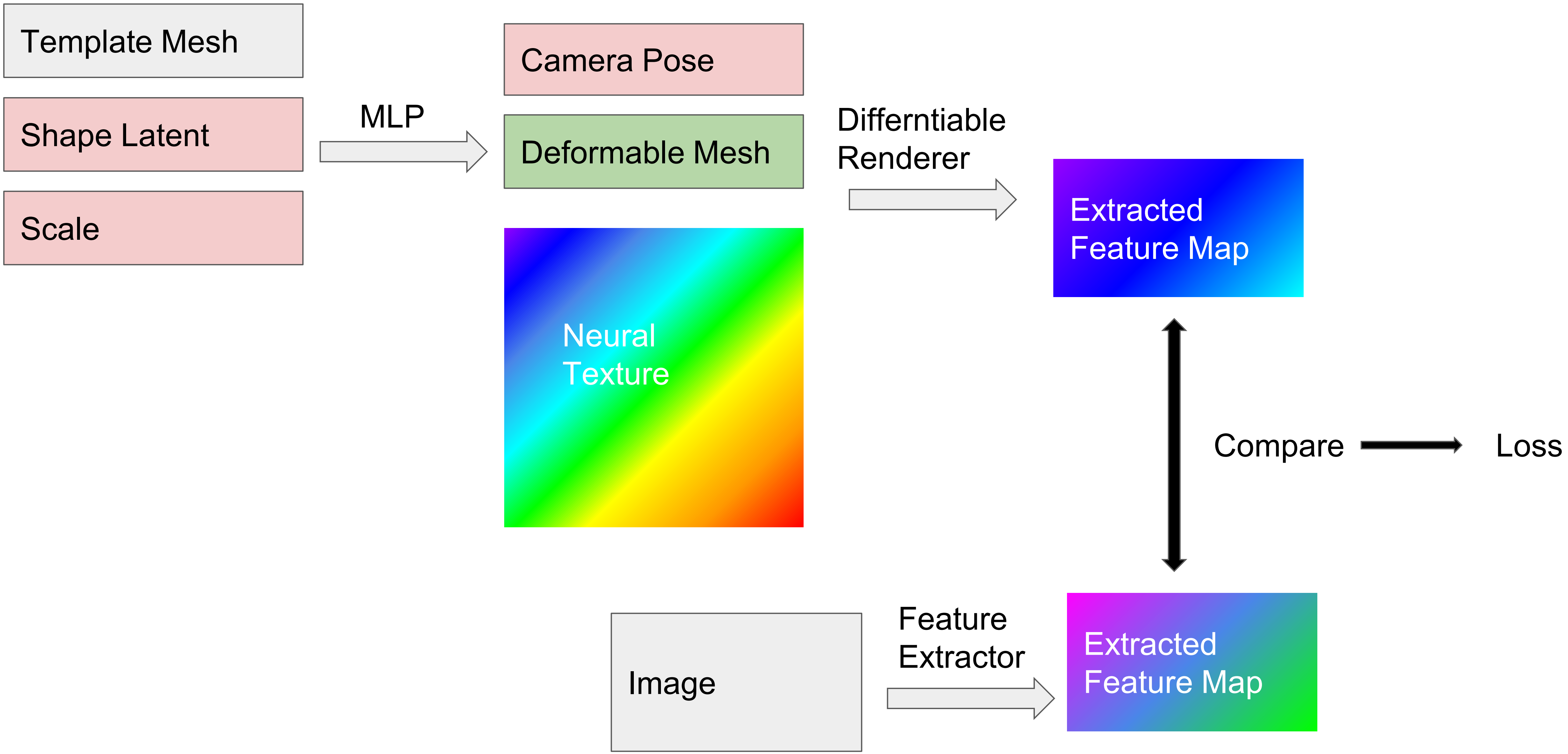
$$\mathcal{L}_{ML}(F,\Gamma,\Theta,m) = -\ln p(F|\Gamma,\Theta,m)$$
$$= \epsilon - \sum_{(u,v)\in\mathcal{S}} -\frac{1}{2\sigma^2}\|f^c_{u,v} - \theta^c_{u,v}\|^2$$

$$\mathcal{L}_{Class}(F,\Theta_{\mathcal{S}}^{\mathbf{C}}) = -\sum_{c\in\mathbf{C},c'\neq c}\ \sum_{(u,v)\in\mathcal{P}}\ \sum_{(u',v')\in\mathcal{M}} \|f^c_{u,v} - \theta^{c'}_{u',v'}\|^2$$

$$\mathcal{L}_{Object}(F,\Theta_{\mathcal{S}}) = -\sum_{(u,v)\in\mathcal{P}}\ \sum_{(u',v')\in\mathcal{N}} \|f^c_{u,v} - \theta^c_{u',v'}\|^2$$

$$\mathcal{L}_{Back}(F,\Theta_{\mathcal{BG}}) = -\sum_{(u,v)\in\mathcal{P}}\ \sum_{j\in\mathcal{BG}} \|f^c_{u,v} - \theta^c_j\|^2.$$

Inference by Render and Compare: (recent alternatives)



Robustness of 3D Generative Models: Occlusion and Pose

- 3D Generative models are robust to occlusion because they incorporate an outlier process and exploit knowledge of the 3D structure of objects. I.e. the intensity/features in the image can be generated by the object or by some other object (e.g., an occluder). The object models and the outlier process cooperate and compete to explain the images.

- *Note: studies show that humans are much more robust than Deep Nets to occluders and generative models are much better (e.g., H. Zhu et al. 2019).*

- The 3D Generative Models have 3D models of objects which makes it fairly straightforward for them to generalize to novel poses.

3D Generative Models for classification and 3D pose

occ Level	Clean	L1	L2	L3
Resnet50-general	**99.30**	93.81	77.78	45.21
NeMo	88.04	72.51	49.31	22.27
Ours	98.79	**95.81**	**85.63**	**57.91**

Table 1. P3D+ Classification Result

Metric	$ACC_{\frac{\pi}{6}} \uparrow$				$ACC_{\frac{\pi}{18}} \uparrow$			
occ Level	L0	L1	L2	L3	L0	L1	L2	L3
Res50	82.4	65.6	45.1	21.4	38.8	24.7	14.5	5.1
NeMo	82.4	62.1	39.1	13.3	60.1	38.9	21.1	5.2
Ours	**85.1**	**72.9**	**53.1**	**29.9**	**60.4**	**41.5**	**24.8**	**10.5**

Table 3. P3D+ 3D-Classification

L1: 20-40%
L2: 40-60%
L3: 60-80%

- **Normal classification**: 3D generative models achieve **similar results** to deep neural networks

- **Occlusion classification:** 3D generative models perform much **better** than deep neural networks

Reference

[1] Zhang, Guofeng, et al. “Inverse Rendering of Discriminative Neural Textures of 3D-aware Image Classification.” In review 2023.

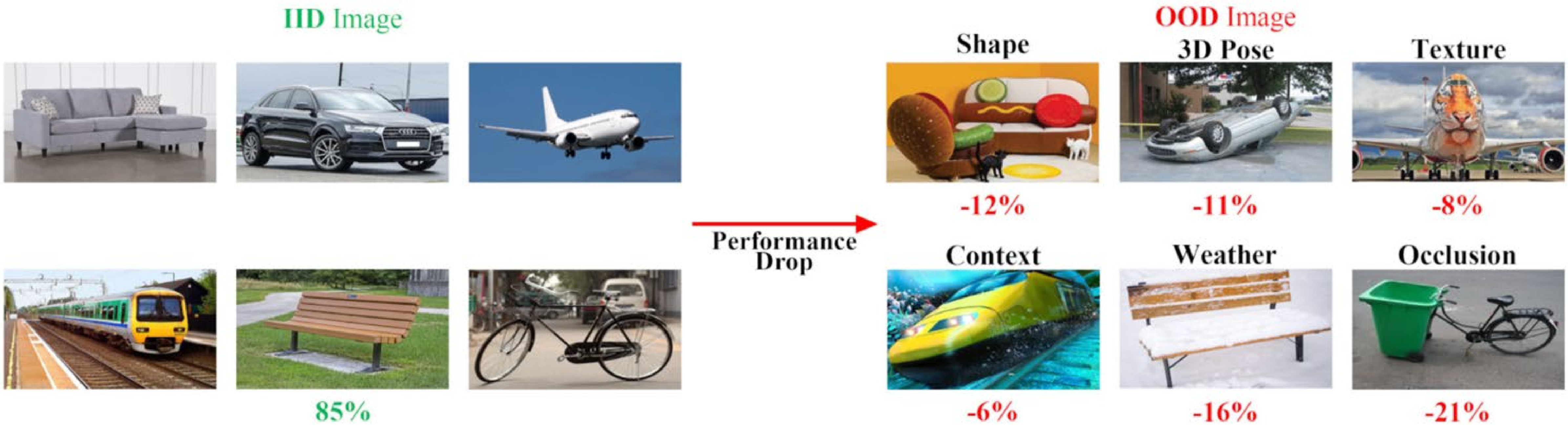
Generalization to Out-of-Distribution (OOD) Data

- Generalization to out-of-distribution (OOD) data is a challenge for current deep network algorithms. There are image benchmarks for evaluating this generalization. We developed OOD-CV (Binchen Zhao ECCV 2022).

- 3D Generative Models performed well on OOD-CV for object classification compared to SOTA alternatives.

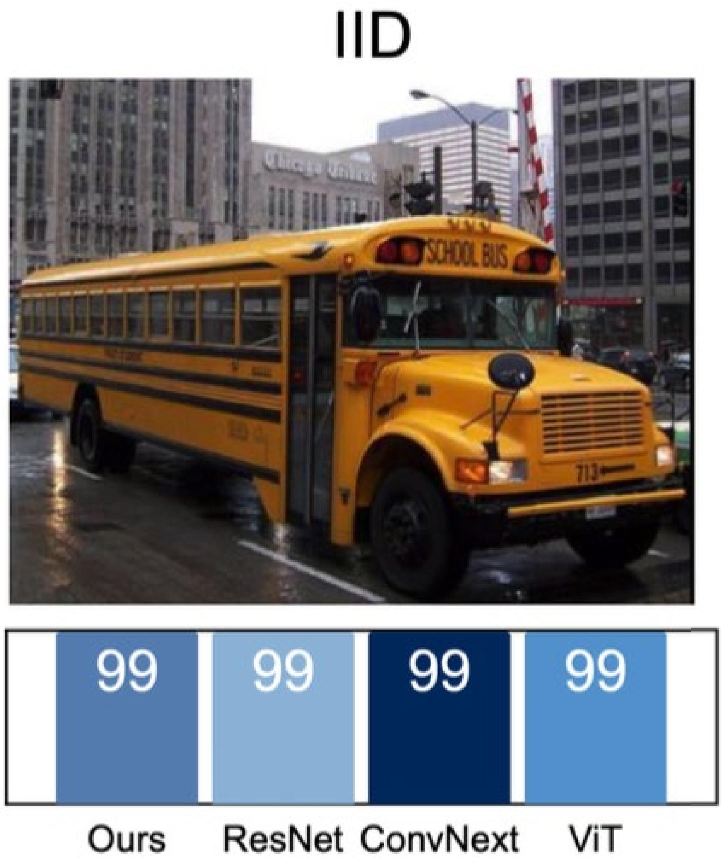
- 3D Generative Models also perform well on Synthetic Data (for similar reasons).

Testing on IID data vs. OOD data: standard deep networks have a big drop in performance



Reference
[1] Zhao, Bingchen, et al. "OOD-CV: A Benchmark for Robustness to Individual Nuisances in Real-World Out-of-Distribution Shifts." ICML 2022 Shift Happens Workshop. 2022.

Background: deep networks work well on IID data

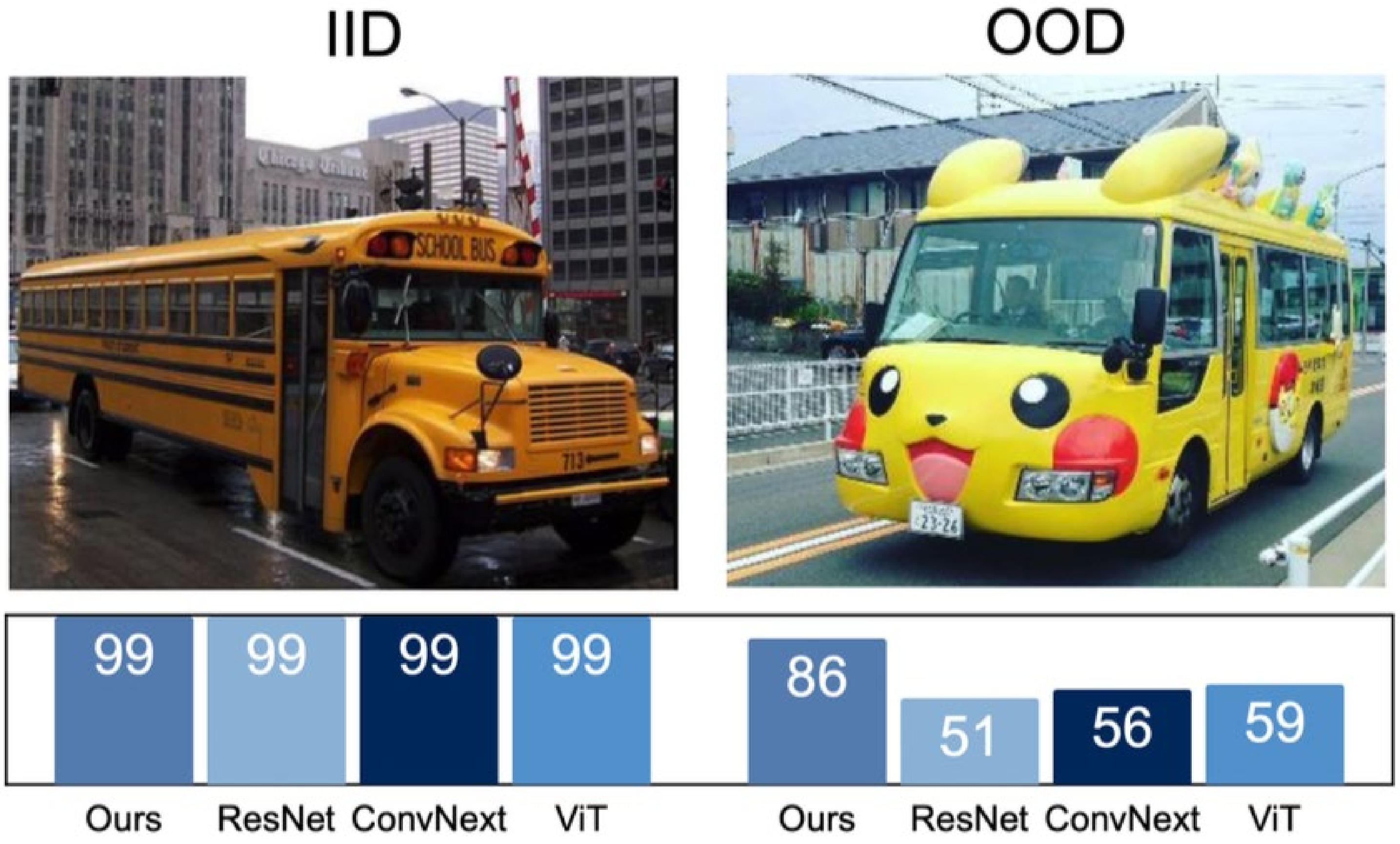


Results show object classification accuracy obtained on the OOD-CV dataset [1].

Reference

[1] Zhao, Bingchen, et al. "OOD-CV: A Benchmark for Robustness to Individual Nuisances in Real-World Out-of-Distribution Shifts." ICML 2022 Shift Happens Workshop. 2022.

Deep networks don't work well on OOD data

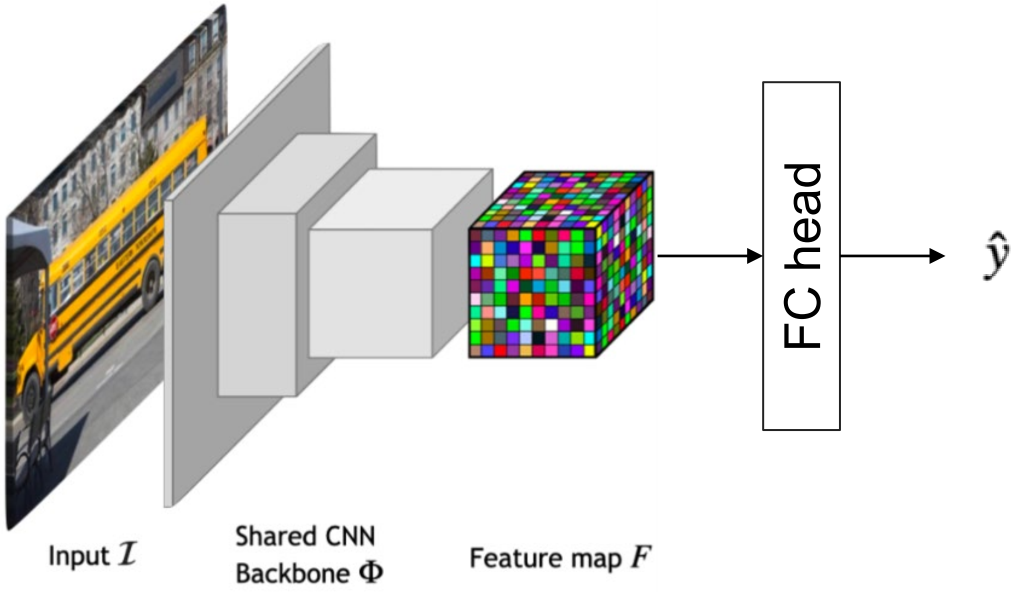


Results show object classification accuracy obtained on the OOD-CV dataset [1].

Reference
[1] Zhao, Bingchen, et al. "OOD-CV: A Benchmark for Robustness to Individual Nuisances in Real-World Out-of-Distribution Shifts." ICML 2022 Shift Happens Workshop. 2022.

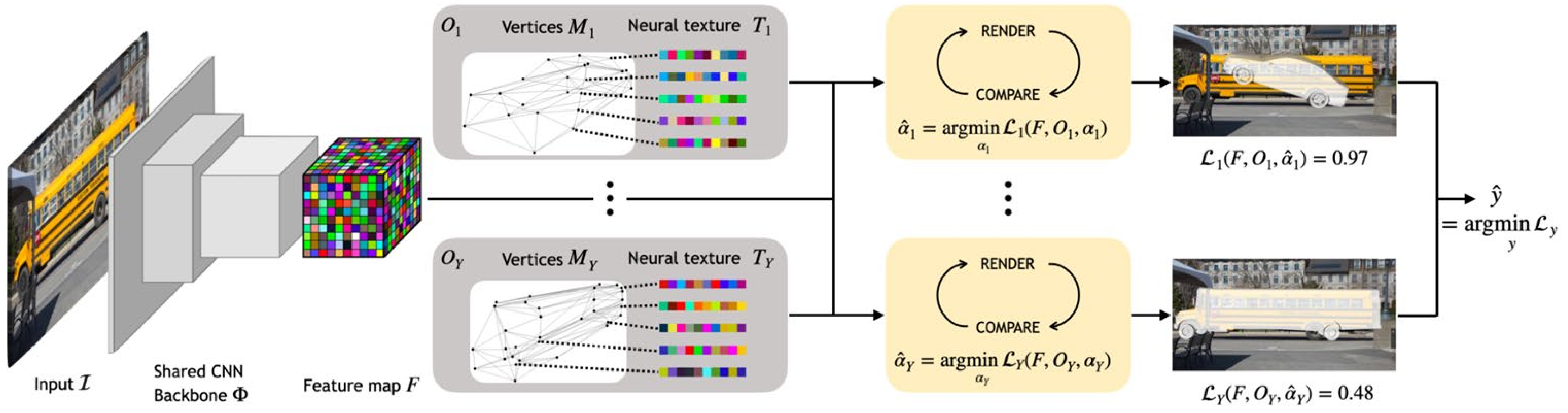
Why deep networks don't work well on OOD data?

Reason: traditional deep networks do not use **the 3D structure of objects**



3D Generative models perform better deep networks

3D Generative models exploit the **3D structure** of the object



Classification results:
3D Generative Models have SOTA performance on OOD and occluded data

Dataset	P3D+	occluded-P3D+				OOD-CV					
Nuisance	L0	L1	L2	L3	Mean	Context	Pose	Shape	Texture	Weather	Mean
Resnet50	99.3	93.8	77.8	45.2	79.6	45.1	61.2	55.2	48.3	47.3	51.4
Swin-T	**99.4**	93.6	77.5	46.2	79.7	63.0	71.4	65.9	61.4	59.6	64.2
Convnext	**99.4**	95.3	81.3	50.9	81.8	53.6	61.2	60.8	57.2	47.1	56.0
ViT-b-16	99.3	94.7	80.3	49.4	80.9	57.8	67.3	61.0	54.7	54.5	59.0
Ours	99.1	96.1	86.8	59.1	85.3	**85.2**	**88.2**	**84.6**	**90.3**	82.4	**86.0**
Ours++	**99.4**	**96.8**	**87.2**	**59.2**	**85.7**	85.1	88.1	84.1	88.5	**82.7**	85.4

Classification results:
3D Generative Models have high performance on corrupted data

Dataset	P3D+	corrupted-P3D+												
Nuisance	L0	defocus blur	glass blur	motion blur	zoom blur	snow	frost	fog	brightness	contrast	elastic transform.	pixelate	jpeg	mean
Resnet50	99.3	67.6	41.4	73.5	87.5	84.4	84.3	93.9	98.0	90.0	46.4	82.1	95.5	78.7
Swin-T	**99.4**	60.7	37.1	70.9	81.3	88.5	91.6	95.4	97.9	92.1	56.3	79.2	95.3	78.9
Convnext	**99.4**	70.1	58.7	76.5	**90.0**	**92.3**	92.9	**98.5**	**99.2**	**98.4**	67.6	84.2	**98.7**	85.6
ViT-b-16	99.3	64.5	**78.1**	80.3	88.2	91.2	**94.1**	90.5	98.7	85.1	84.8	**96.9**	**98.7**	87.6
Ours	99.1	**90.1**	66.9	86.8	84.9	81.3	88.1	98.2	97.9	96.8	**96.7**	**96.9**	98.1	90.2
Ours++	**99.4**	89.9	66.4	**87.3**	87.2	83.3	89.8	98.4	98.0	96.9	96.5	96.7	98.4	**90.4**

Pose estimation results:
3D Generative Models are SOTA on OOD-CV

Dataset		P3D+	occluded-P3D+				OOD-CV					
Nuisance		L0	L1	L2	L3	Mean	Context	Pose	Shape	Texture	Weather	Mean
$ACC_{\frac{\pi}{18}} \uparrow$	Resnet50	33.8	22.4	15.8	9.1	15.8	15.5	12.6	15.7	22.3	23.4	18.0
	Swin-T	29.7	23.3	15.6	10.8	16.6	18.3	14.4	16.9	21.1	26.3	19.8
	Convnext	38.9	22.8	12.8	6.6	14.1	18.1	**14.5**	16.5	21.7	26.6	19.9
	ViT-b-16	38.0	23.9	13.7	7.4	15.0	24.7	13.8	15.6	25.0	28.3	21.5
	NeMo	62.9	**45.0**	**30.7**	**14.6**	**30.1**	21.9	6.9	19.5	34.0	30.4	21.9
	Ours	61.6	42.8	27.0	11.6	27.2	**23.6**	10.4	**22.7**	37.5	**35.5**	**25.5**
	Ours++	**65.1**	**45.0**	28.7	12.5	28.8	23.5	9.8	22.3	**37.9**	34.5	24.8
$ACC_{\frac{\pi}{6}} \uparrow$	Resnet50	82.2	66.1	53.1	42.1	53.8	**57.8**	34.5	50.5	**61.5**	60.0	51.8
	Swin-T	81.4	58.5	47.3	38.8	48.2	52.3	41.1	45.7	50.1	64.9	50.9
	Convnext	82.4	63.7	47.9	36.4	49.3	51.7	**43.4**	44.8	48.0	**65.9**	50.7
	ViT-b-16	82.0	65.4	49.5	37.6	50.8	54.7	34.0	49.5	59.1	59.0	51.3
	NeMo	87.4	75.9	**63.9**	**45.6**	**61.8**	50.3	35.3	49.6	57.5	52.2	48.0
	Ours	86.1	74.8	59.2	37.3	57.1	54.3	38.0	**53.5**	60.5	57.3	**51.9**
	Ours++	**89.8**	**78.1**	62.6	39.0	59.9	55.4	36.7	53.1	60.0	57.0	51.4

Table S3: Pose Estimation results on (occluded)-PASCAL3D+, and OOD-CV dataset. Pose accuracy is evaluated for error under two thresholds: $\frac{\pi}{6}$ and $\frac{\pi}{18}$ separately. Noticeably, RCNet has equivalent performances to current SOTA for 3D-pose estimation event hough it has not been specifically designed for this task.

Qualitative results on Occluded PASCAL3D+ and OOD-CV

(a) L2 Occluded, Car

(b) Context OOD, Bicycle

(c) Weather OOD, Aeroplane

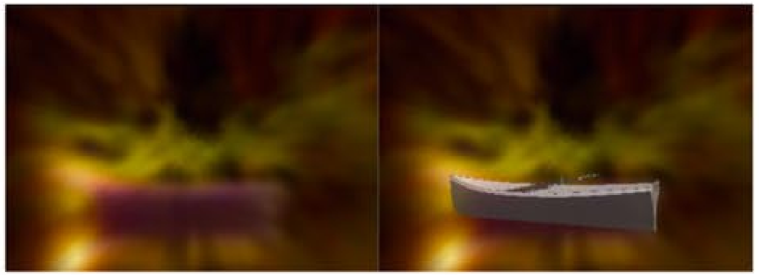
(d) Texture OOD, Sofa

(e) Shape OOD, Bus

(f) Pose OOD, Motorbike

3D Generative Models are **robust** to **occluded** data and **OOD** data

Qualitative results of on Corrupted PASCAL3D+



(g) Zoom Blur Corruption, Boat

(h) Motion Blur Corruption, Table

(i) Frost Corruption, Tvmonitor

(j) Snow Corruption, Chair

(k) Brightness Corruption, Bottle

(l) Fog Corruption, Bus

Generative 3D models are **robust** to **corrupted data**

Conclusion

- It is important to test AI on OOD data. Algorithms that perform well on IID data can degrade badly on OOD.

- Approximate Analysis by Synthesis is an old idea which has a lot of potential for computer vision and has been proposed as a theory for human visual perception.

- 3D generative models and occluder processes enables robustness to occluded data.

- They can also generalize to out-of-distribution (OOD) data needing minimal modification.

- This is only the starting point. 3D generative models work only on a limited number of objects. There are many technical ways to improve them.

- *Work with A. Kortylewski, A. Jesslen, A. Wang, W. Ma, G. Zhang and many others.*