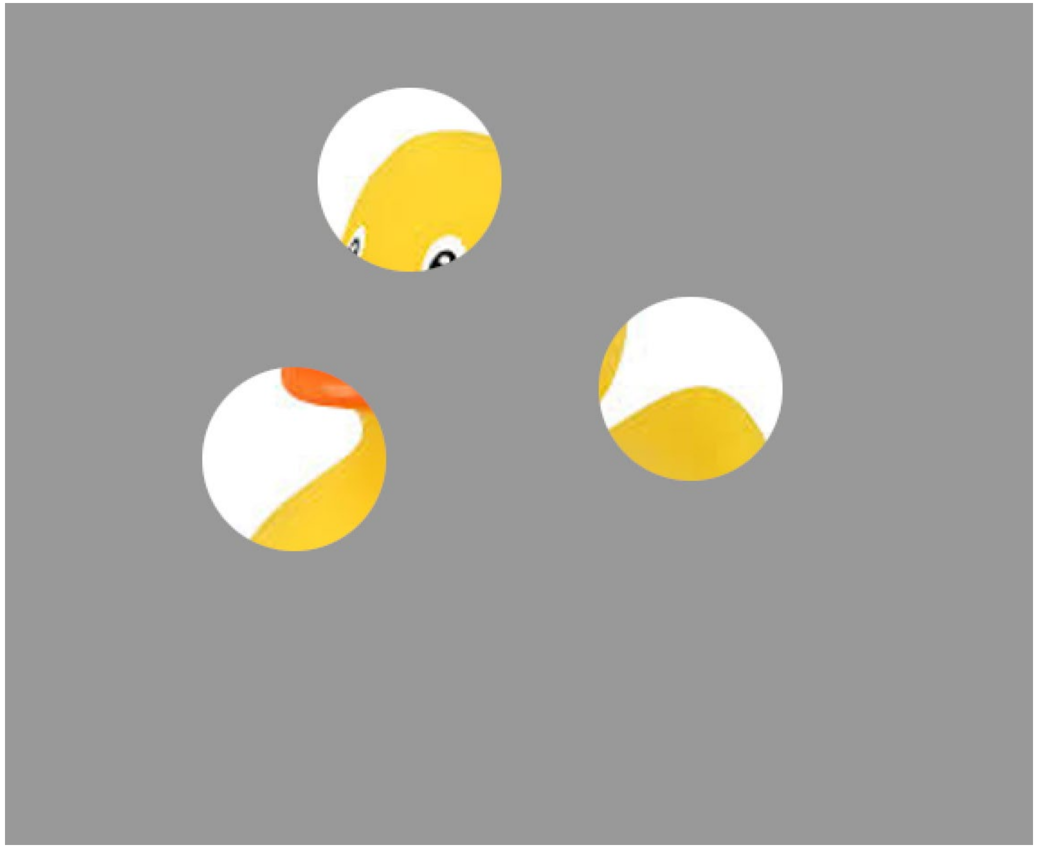
Generative Computer Vision: Robust Generalization with Analysis-by-Synthesis

Adam Kortylewski

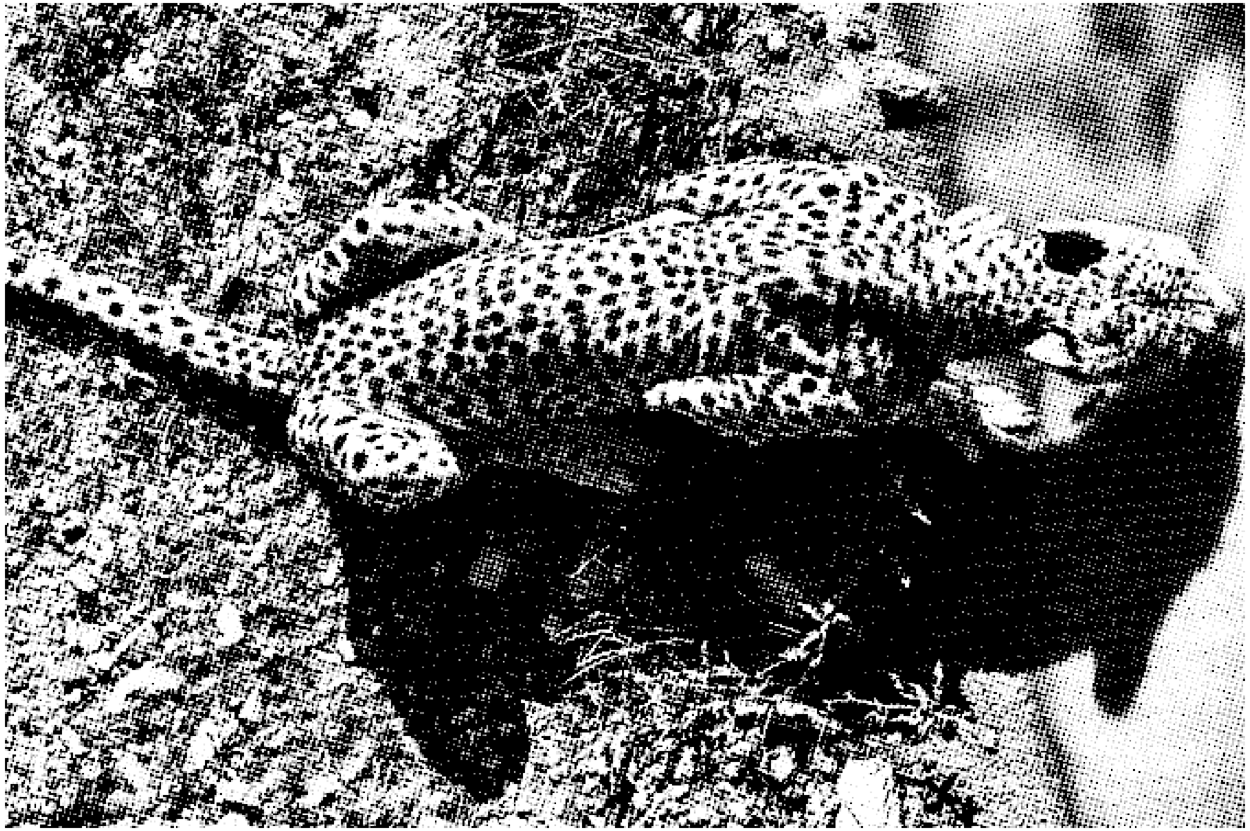
Generative Vision Research Group
University of Freiburg | Max Planck Institute for Informatics

Robust Vision – What object is this?

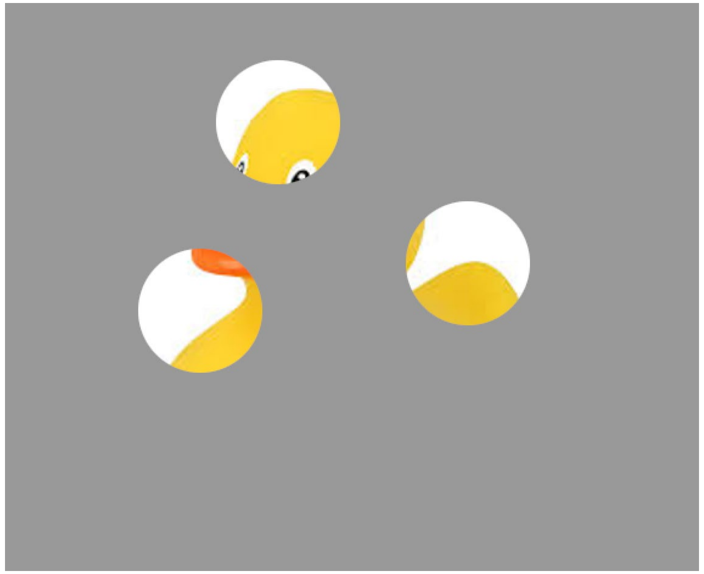
Robust Vision – What object is this?

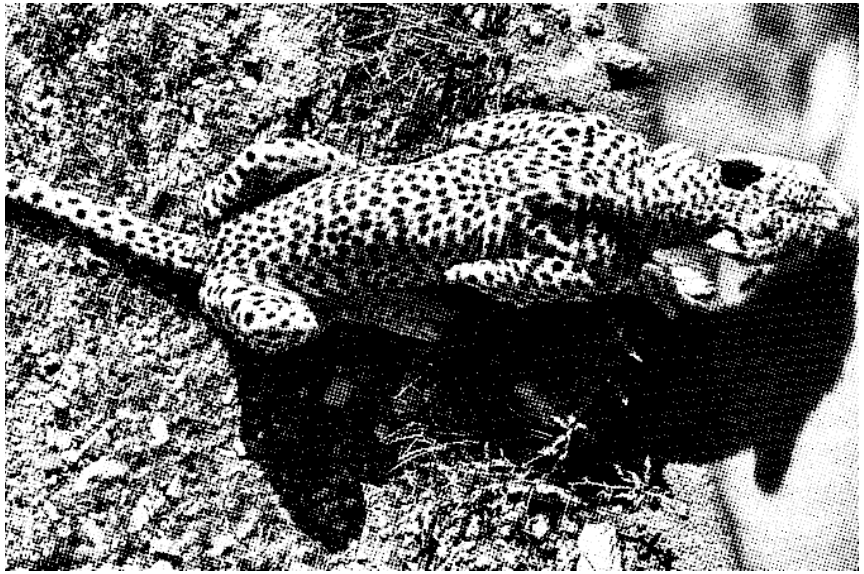


Robust Vision – What object is this?



Robust Vision – Generalization beyond the training data

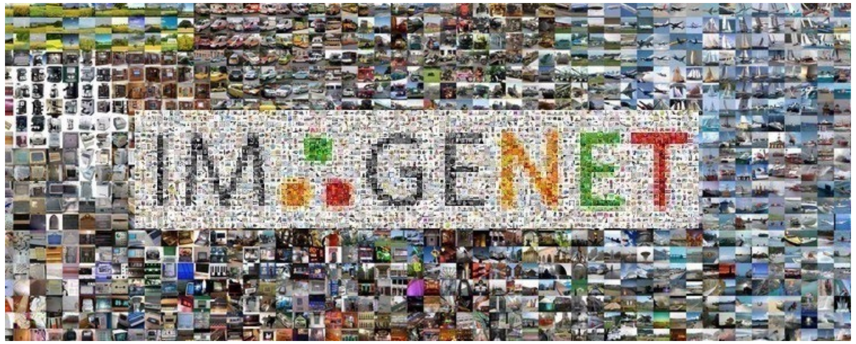




- Human vision is robust in **<u>unseen</u>** viewing conditions
- Important side note: Once you recognize the object, you know pose, parts, shape, …

We love Deep Networks in Computer Vision

Image Classification



>90%   Top-1

Semantic Segmentation

>90% mIoU

Panoptic Segmentation

Human Pose Estimation

<7.5cm MPJPE

Visual Question Answering

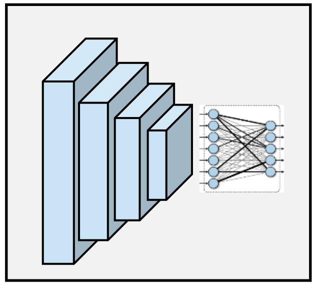
Q: What is the material used to make the vessels in this picture?

But, Deep Nets also have fundamental limitations

Input
3D Pose

[Alcorn, ‘19]

Deep Network



Output

✓ Large-scale visual recognition

× Lack robustness to 3D changes *[Qiu’16,Alcorn’19]*

× Lack robustness to changes of image components *[Rosenfeld’18, Geirhos’19, Michaelis’19, Kortylewski‘20]*

But, Deep Nets also have fundamental limitations

Context

[Rosenfeld, ‘18]

Texture

[Geirhos, ‘19]

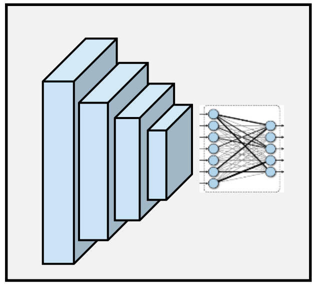
Weather

[Michaelis, ’19]

Occlusion

[Kortylewski, ‘20]

Deep Network



Output

✘

✓ Large-scale visual recognition

✘ Lack robustness to 3D changes *[Qiu’16,Alcorn’19]*

✘ Lack robustness to changes of image components *[Rosenfeld’18, Geirhos’19, Michaelis’19, Kortylewski‘20]*

Why is this relevant?

Open Challenges in Self-driving - Detecting STOP Signs

Andrej Karpathy - AI for Full-Self Driving at Tesla, 2020
https://youtu.be/hx7BXih7zx8

Large variability in:

- Context
- Positions and pose
- Lights
- Occlusion
- Environmental conditions

Detecting STOP signs is **not solved** yet!

The Dawn Project Super Bowl Commercial
https://youtu.be/_ZiSZbWIrzA

STOP signs are explicitly designed to be detectable

Deep Networks do not generalize in out-of-distribution scenarios.

So: What do we need to do?

Is all we need just to collect more data?

Images are combinatorially complex.

So: What do we need to do?

1) Generative computer vision via analysis-by-synthesis

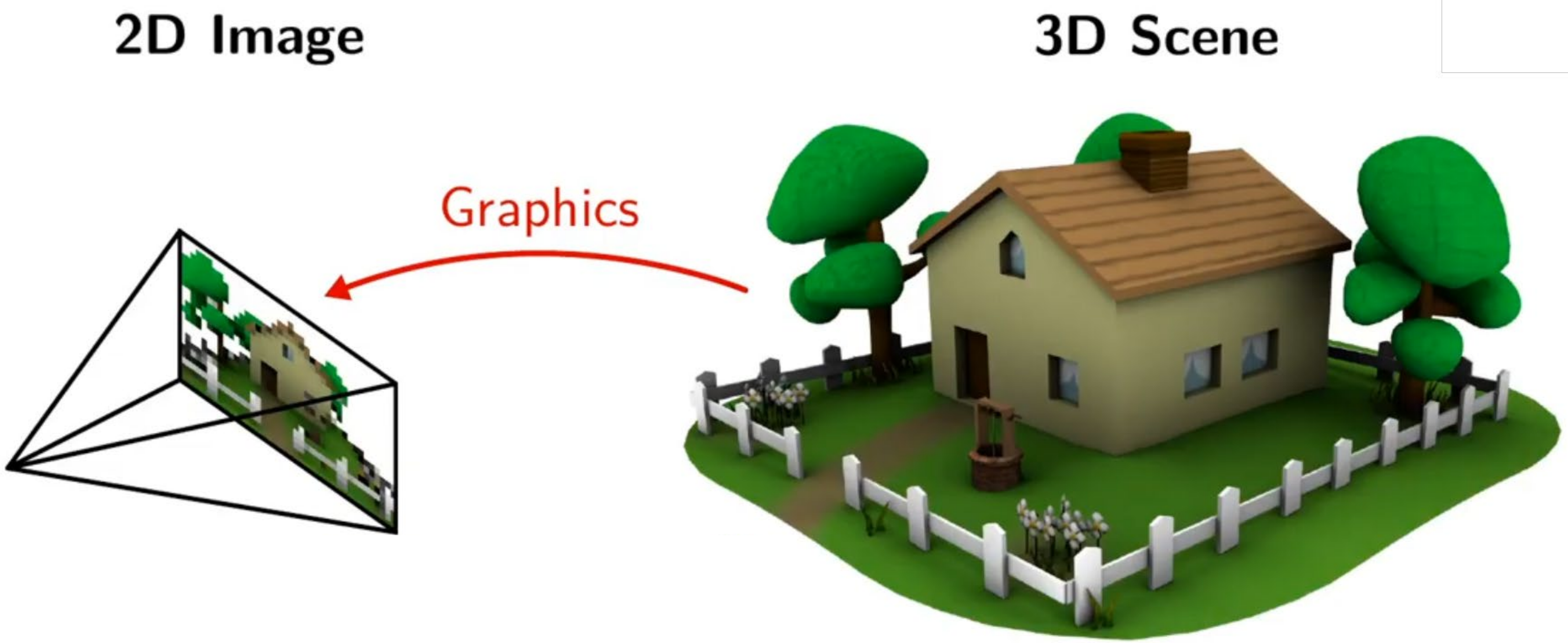
2) Advanced benchmarks that measure out-of-distribution robustness

So: What do we need to do?

1) **Generative computer vision via analysis-by-synthesis**

2) Advanced benchmarks that measure out-of-distribution robustness

Computer Vision via Analysis-by-Synthesis

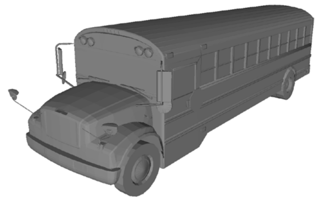


Vision systems that analyze images by synthesizing them.

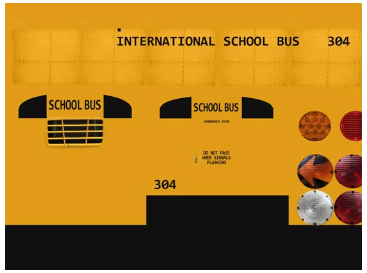
[Grenander1978, Mumford1994, YuilleKersten2006]

Analysis-by-Synthesis (1) - Generative Object Model

Computer Graphics Model



3D shape

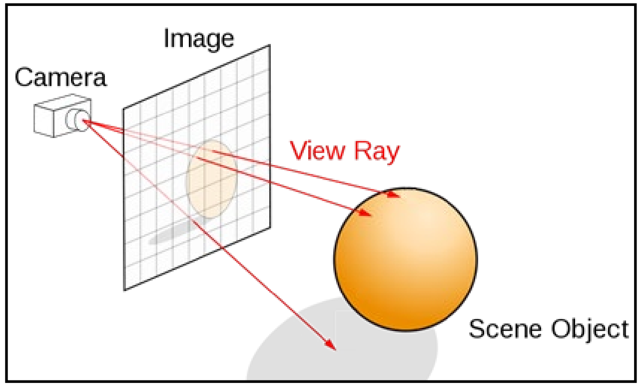


Texture

⋮

Illumination

Render



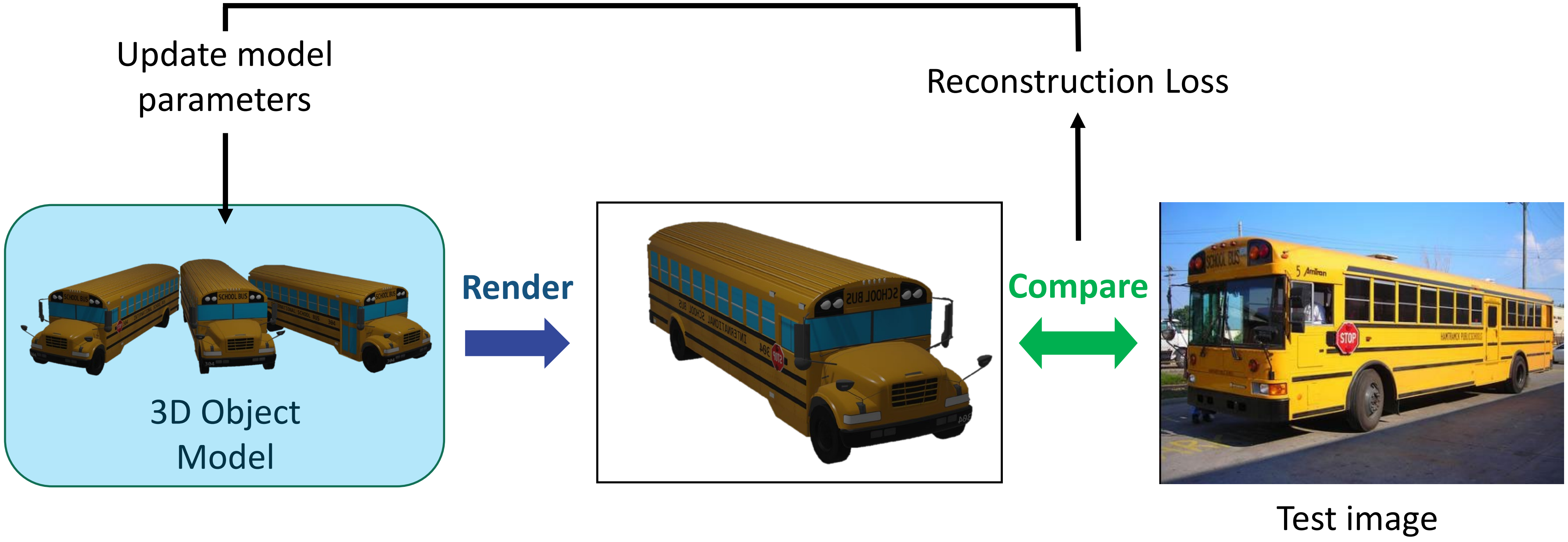
Changing parameter:

Texture

3D Pose

Illumination

Analysis-by-Synthesis (2) – Inverse Rendering



Analysis-by-Synthesis (2) – Inverse Rendering

Update model parameters

Reconstruction Loss

3D Object Model

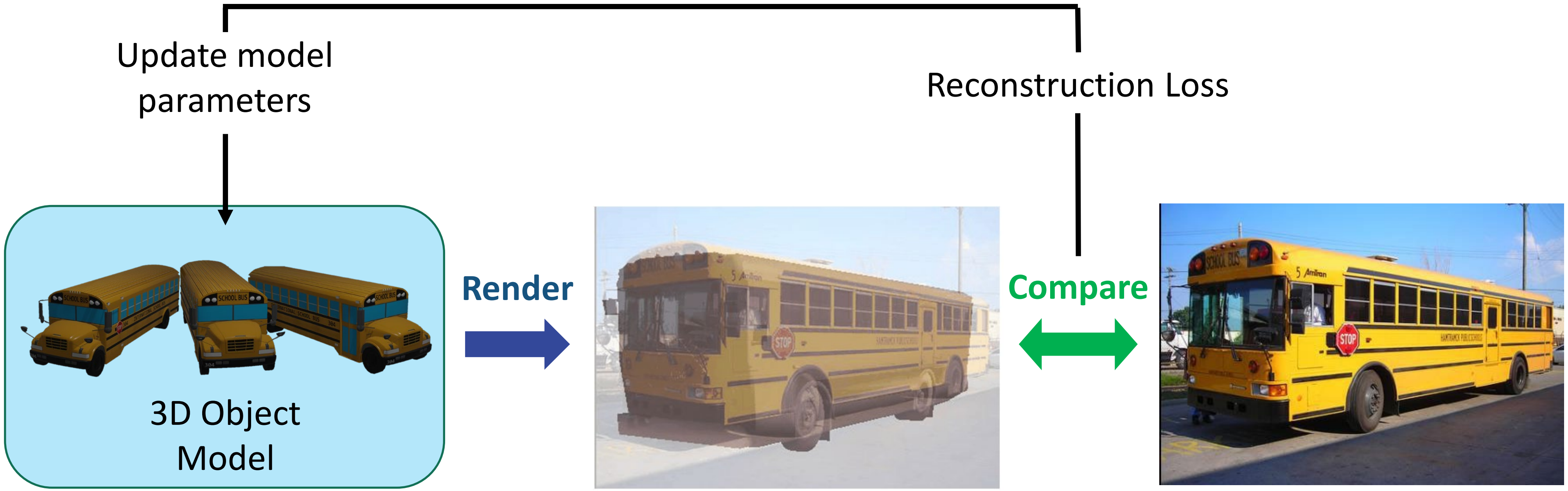
Render

Pose optimization (illustrative example)

Compare

Test image

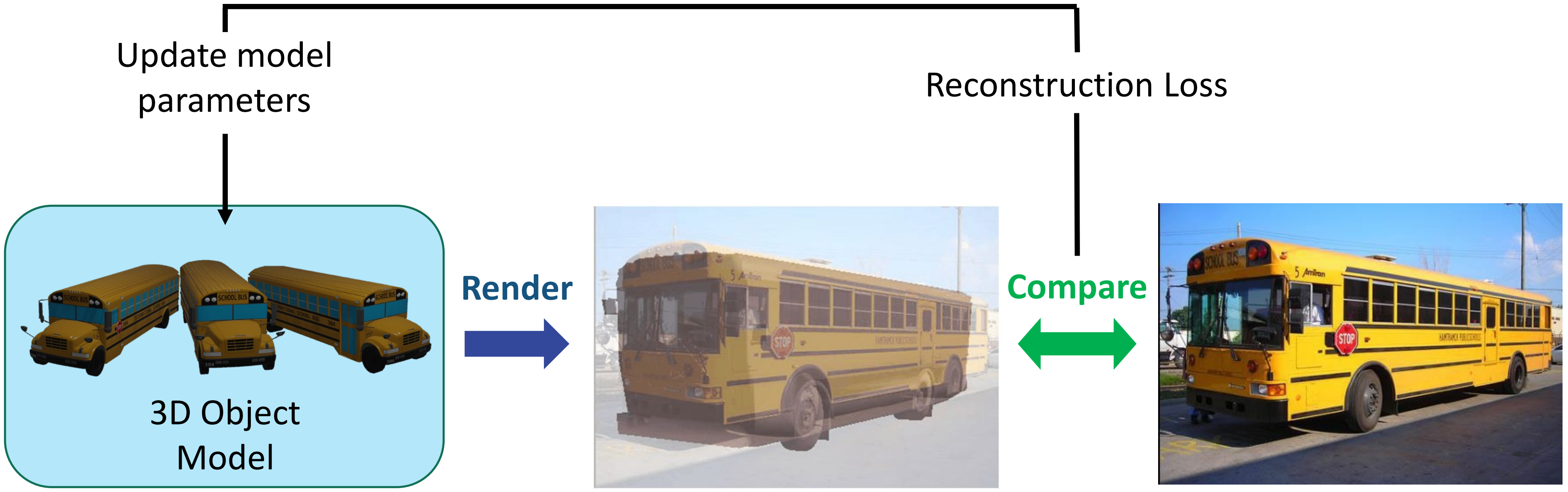
Analysis-by-Synthesis (2) – Inverse Rendering



Advantages over deep networks:

- ✓ **3D-aware** and **compositional**
- ✓ **Robust** (occlusion and unseen poses) [Paysan,'09] [Egger,'18] [Wang,'21]
- ✓ **Multi-tasking**

Analysis-by-Synthesis (2) – Inverse Rendering



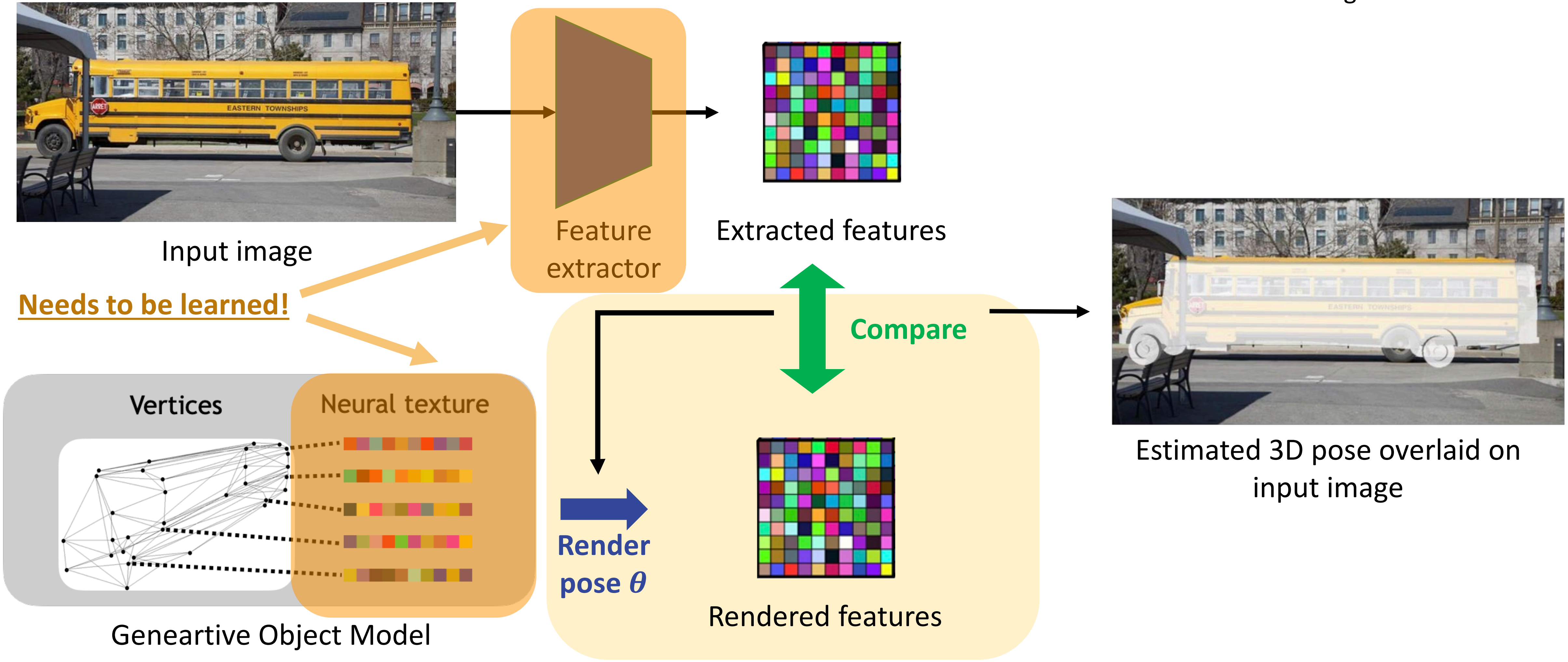
Why is analysis-by-synthesis not widely used in computer vision?

1) Hard to learn the generative object model.

2) Hard to optimize the inverse rendering process.

Neural Analysis-by-Synthesis for 3D Pose Estimation

A. Wang



[Wang, Kortylewski, Yuille, ICLR 2021]

A probabilistic generative model of neural features

- An object category is represented as $O_y = \{M_y, T_y\}$
  - Mesh $M_y = \{v_n \in \mathbb{R}^3\}_{n=1}^N$
  - Neural Texture $T_y = \{t_n \in \mathbb{R}^c\}_{n=1}^N$

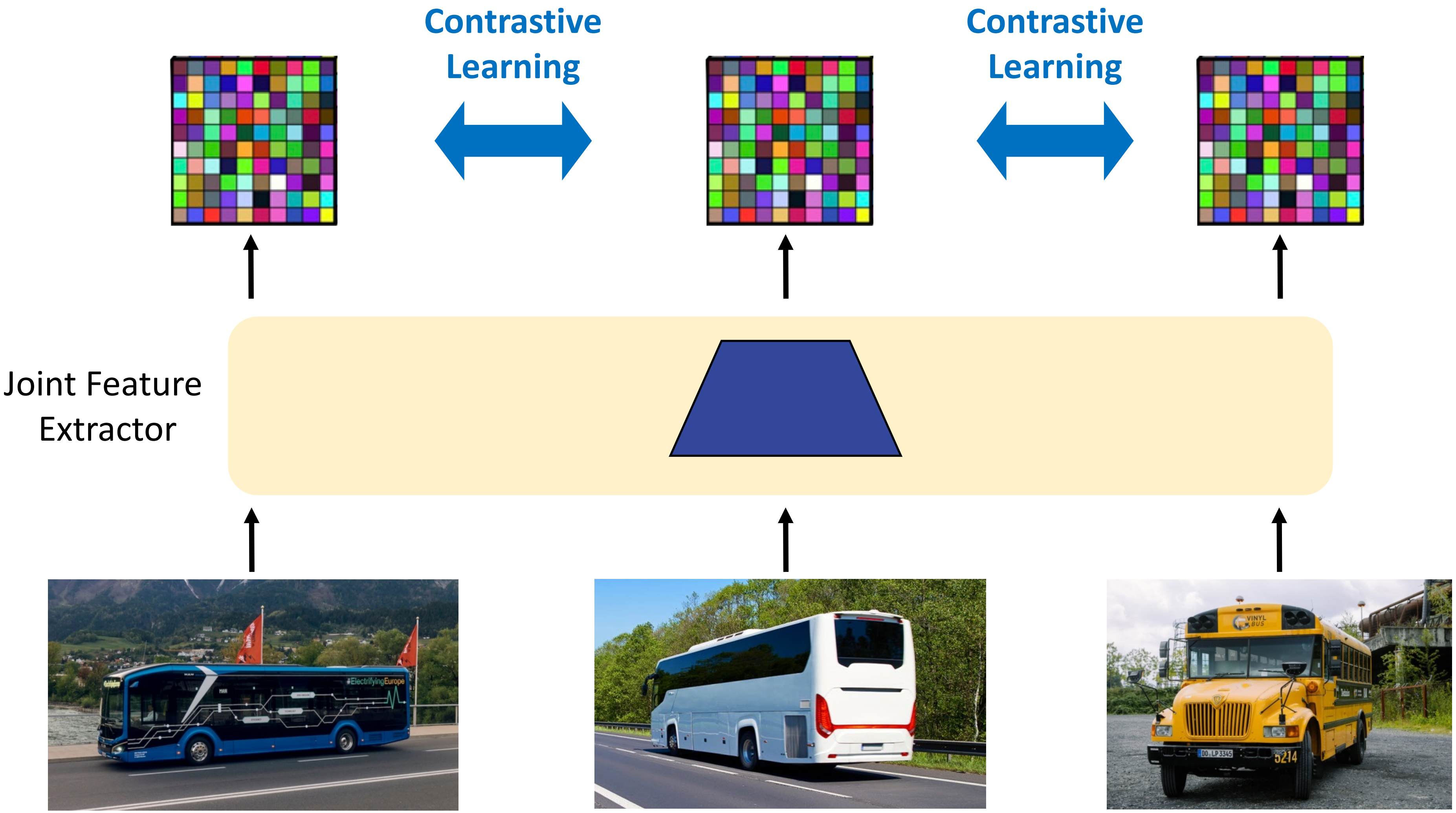
- We formulate a probabilistic generative model

$$p(F|y)=p(F|O_y,\alpha_y,B)=\prod_{i\in\mathcal{FG}} p(f_i|O_y,\alpha) \prod_{i'\in\mathcal{BG}} p(f_i'|B)$$

- Assuming Gaussian likelihoods:

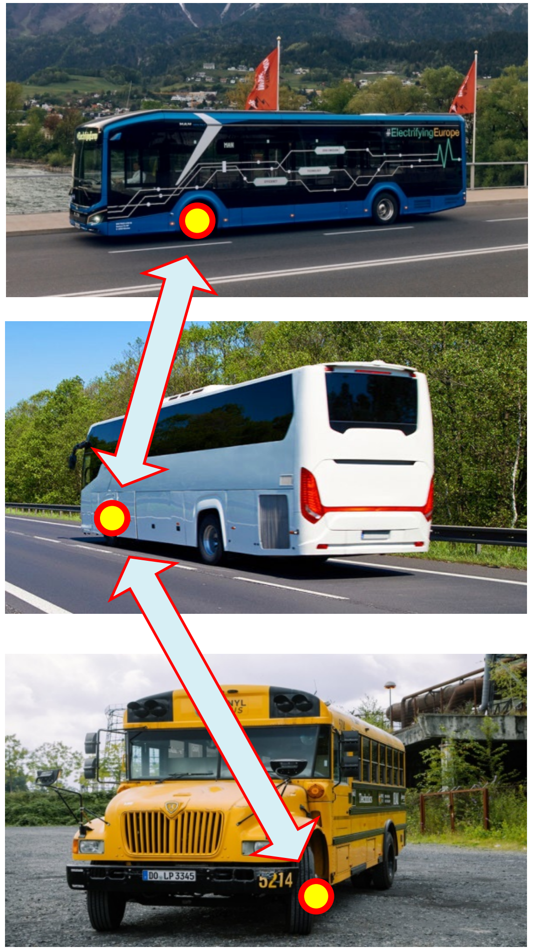
$$\mathcal{L}_{\text{Rec}}(F,O_y,\alpha_y,B) = -\log p(F|y)$$
$$= \sum_{i\in\mathcal{FG}} \|f_i - t_{y,n}\|^2 + \sum_{i'\in\mathcal{BG}} \|f_i' - B\|^2 + const.$$

Neural Analysis-by-Synthesis – Contrastive Learning of Features



[Wang, Kortylewski, Yuille, ICLR 2021]

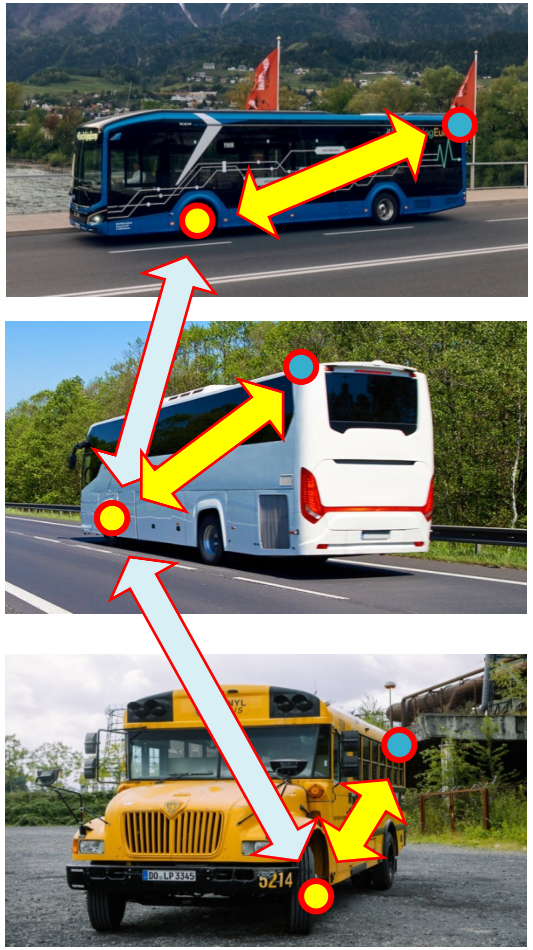
Neural Analysis-by-Synthesis – Contrastive Learning of Features



Contrastive learning of the feature extractor:

1) Features of the **same point** should be **similar**.

Neural Analysis-by-Synthesis – Contrastive Learning of Features



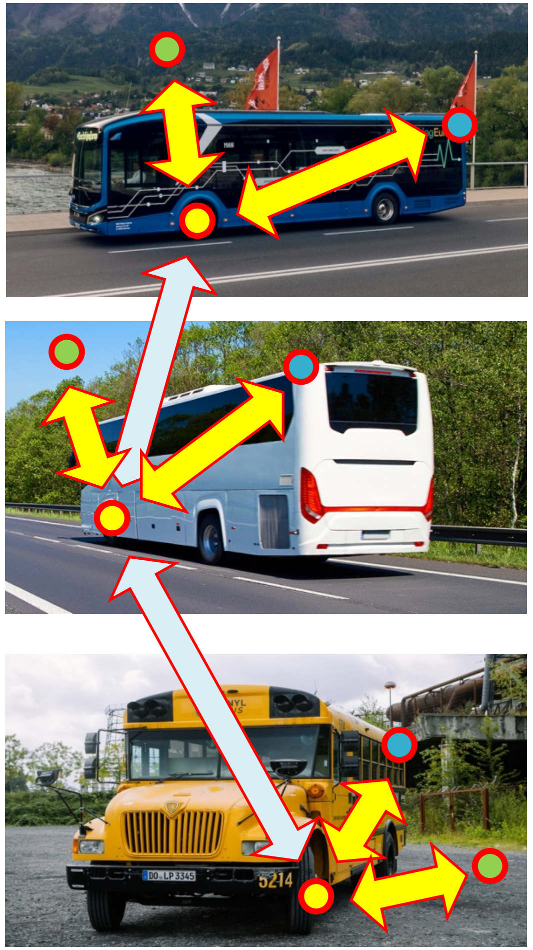
Contrastive learning of the feature extractor:

1) Features of the **same point** should be **similar**.

2) Features of **different points** should be **dissimilar**.

[Wang, Kortylewski, Yuille, ICLR 2021]

Neural Analysis-by-Synthesis – Contrastive Learning of Features



Contrastive learning of the feature extractor:

1) Features of the **same point** should be **similar**.
2) Features of **different points** should be **dissimilar**.
3) Features on the **object** should be **different from background**.

[Wang, Kortylewski, Yuille, ICLR 2021]

Neural Analysis-by-Synthesis for 3D Pose Estimation

Input image

Visualization of pose estimate

Can we extend Neural Analysis-by-Synthesis to classification?

- ✓ E
- ✓ C
- ✓ **F**
- ✓ 3D-aware and compositional

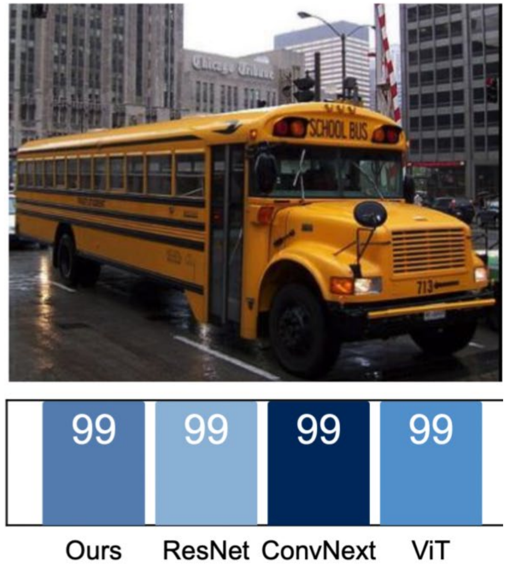
[Wang, Kortylewski, Yuille, ICLR 2021]

Neural Analysis-by-Synthesis for 3D Pose Estimation


A. Jesslen

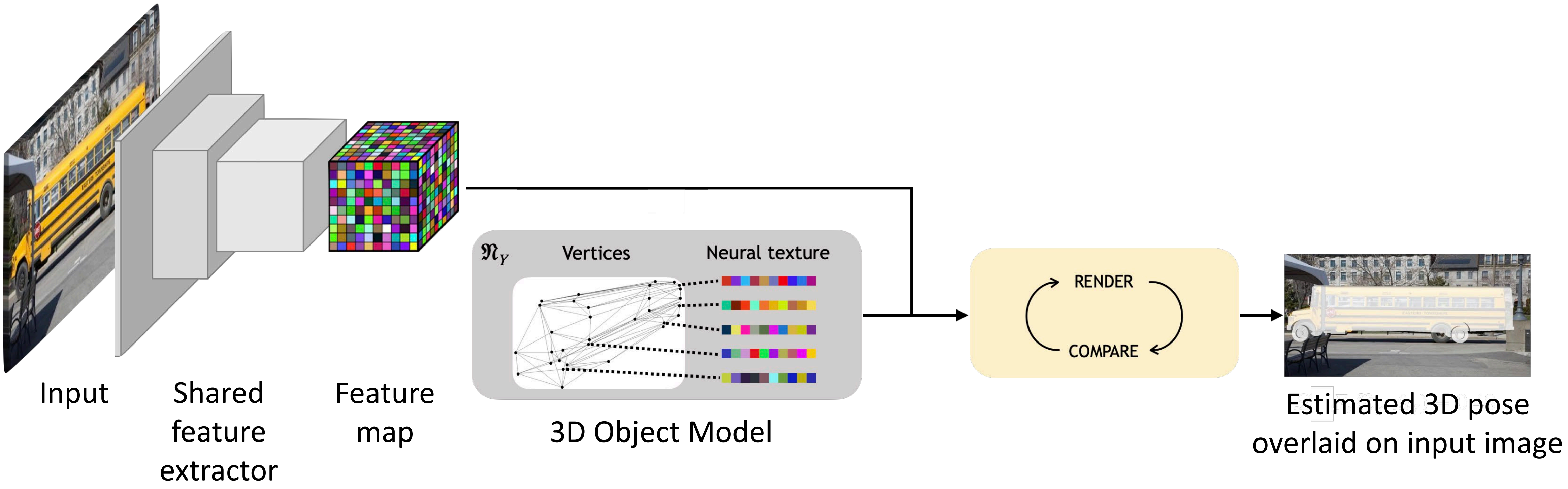

G. Zhang

IID



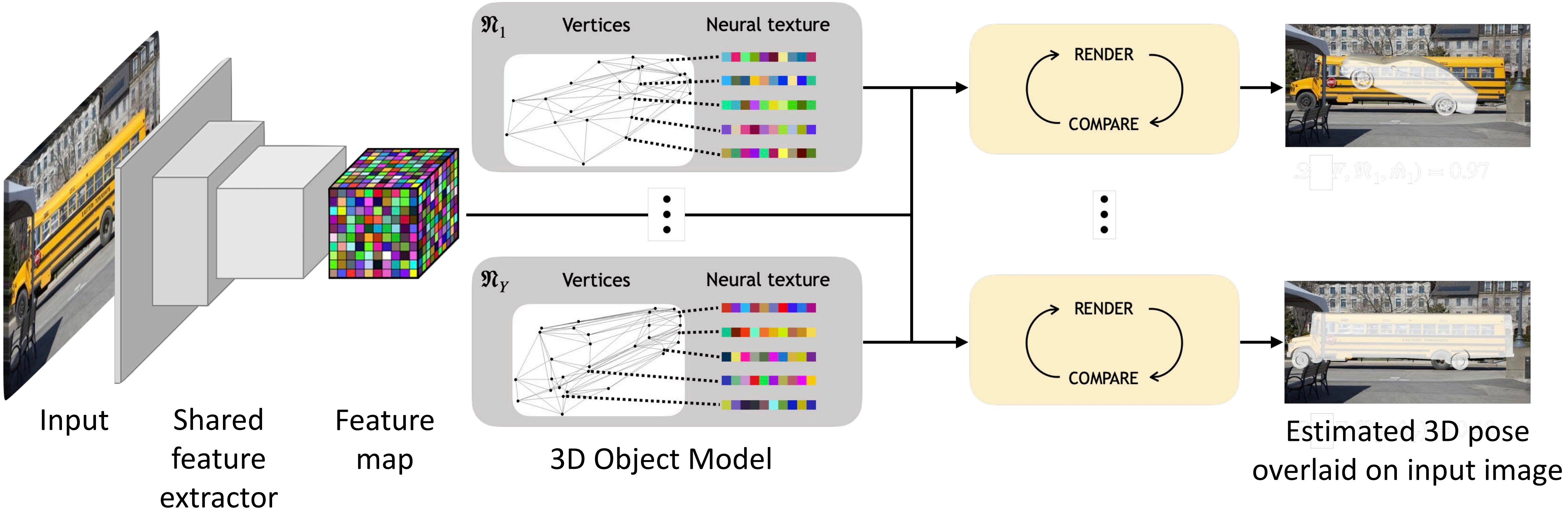
[Jesslen, Zhang, Wang, Yuille, Kortylewski, 2023]

Neural Analysis-by-Synthesis for 3D Pose Estimation



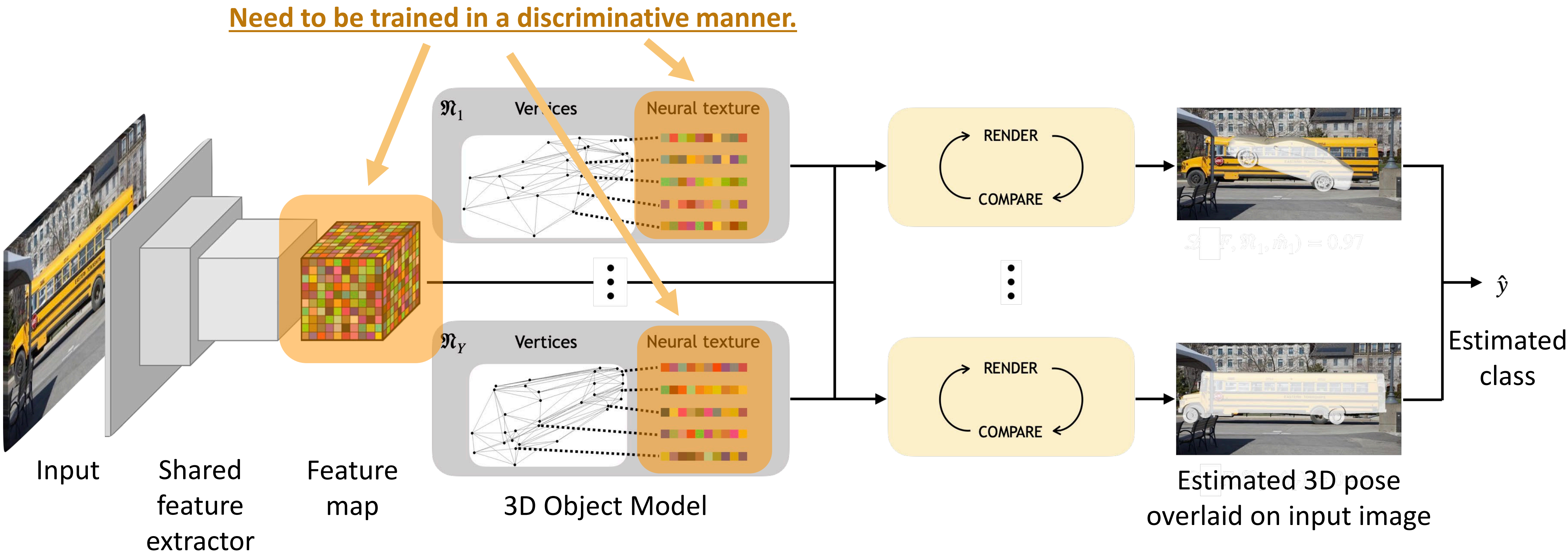
[Jesslen, Zhang, Wang, Yuille, Kortylewski, 2023]

Neural Analysis-by-Synthesis for 3D Pose Estimation



[Jesslen, Zhang, Wang, Yuille, Kortylewski, 2023]

Neural Analysis-by-Synthesis for 3D Pose Estimation



[Jesslen, Zhang, Wang, Yuille, Kortylewski, 2023]

Experiments – Testing Out-of-Distribution Robustness

- CV systems are typically evaluated using average performance on independent and identical distributed (i.i.d.) data

Do we really care about average performance on i.i.d. data?

95%

[Zhao et al. 2022]

Experiments – Testing Out-of-Distribution Robustness

B. Zhao

Training Data

Out-of-Distribution Test Data

Shape

-12%

Pose

-11%

Texture

-8%

Context

-6%

Weather

-16%

Occlusion

-21%

[Zhao et al. ECCV'2022]

Experiments – Results in OOD scenarios

- Image classification

Dataset	P3D+	occluded-P3D+				OOD-CV					
Nuisance		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

- Side note: Our model is trained without data augmentation

[Jesslen, Zhang, Wang, Yuille, Kortylewski, 2023]

Experiments – Results in OOD scenarios

- Even competitive at pose Estimation

Dataset	P3D+	occluded-P3D+	corrupted-P3D+	OOD-CV
Resnet50	39.0	15.8	15.8	18.0
Swin-T	46.2	16.6	15.6	19.8
Convnext	38.9	14.1	24.1	19.9
ViT-b-16	38.0	15.0	21.3	21.5
NeMo	62.9	**30.1**	43.4	21.9
Ours	**65.1**	28.8	**43.9**	**25.5**

[Jesslen, Zhang, Wang, Yuille, Kortylewski, 2023]

What do we need to doto achieve robust generalization?

1) Generative computer vision via analysis-by-synthesis

2) **Advanced benchmarks that measure out-of-distribution robustness**

Why do benchmarks not reflect real-world performance?

- We need to evaluate performance in **unseen** situations

Training Data

Out-of-Distribution Test Data

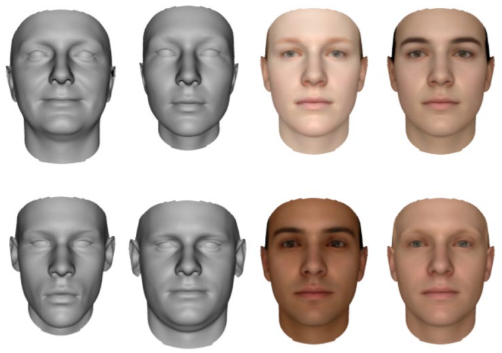
Can we automate the OOD data generation process?

-6%   -16%   -21%

[Zhao et al. ECCV'22]

Collecting and annotating adversarial data is difficult

- Lots of progress in generative models



3D Morphable Models

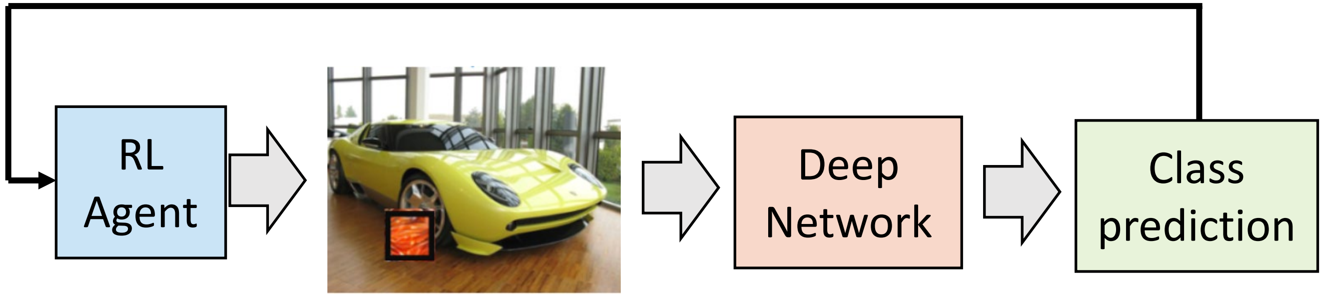
2D GANs

Nerf + 2D Gan

Can generative models help us benchmark CV?

Generative Adversarial Testing of Classifiers

- Find occluders that harm the image classification model

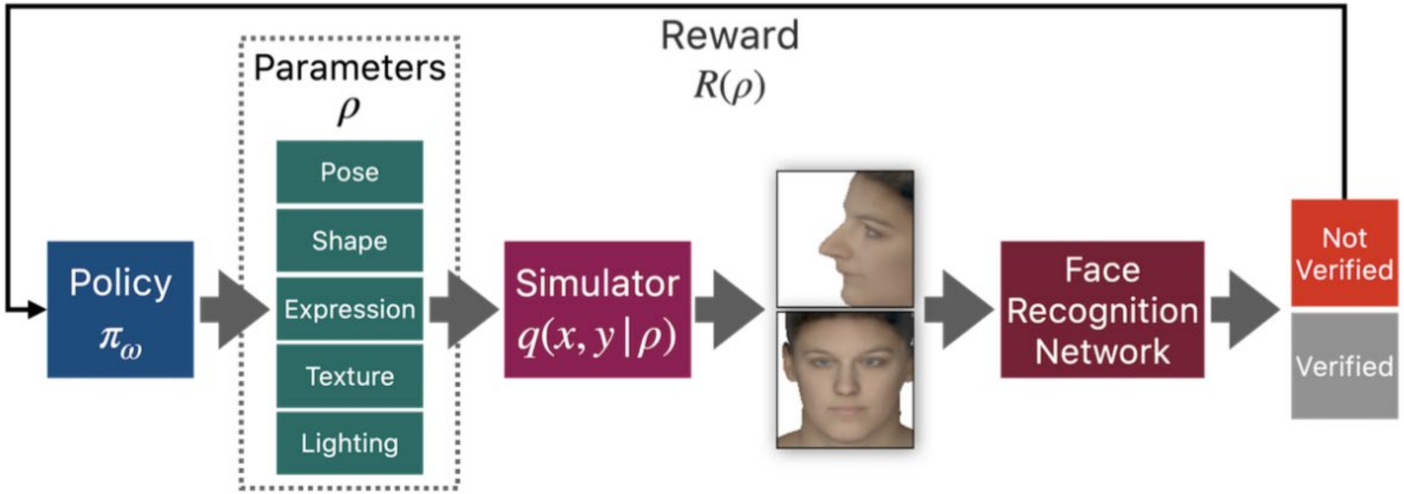


>99% success rate

[Yang et al. ECCV'20]

Generative Adversarial Testing of Face Recognition Models

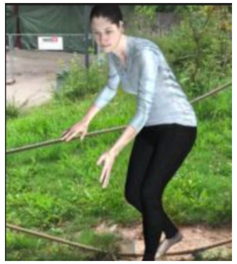
- Use 3DMMs to search for faces that are not recognized correctly



- Discover weaknesses to unusual poses, biases in skin color, exaggerated facial features

[Ruiz et al. CVPR'22]

Generative Adversarial Testing of Human Pose Estimation

Optimizer
$\max_{\theta} |GT\_Pose - Pose'|$ → θ → Data Generator →  → Model → Pose'

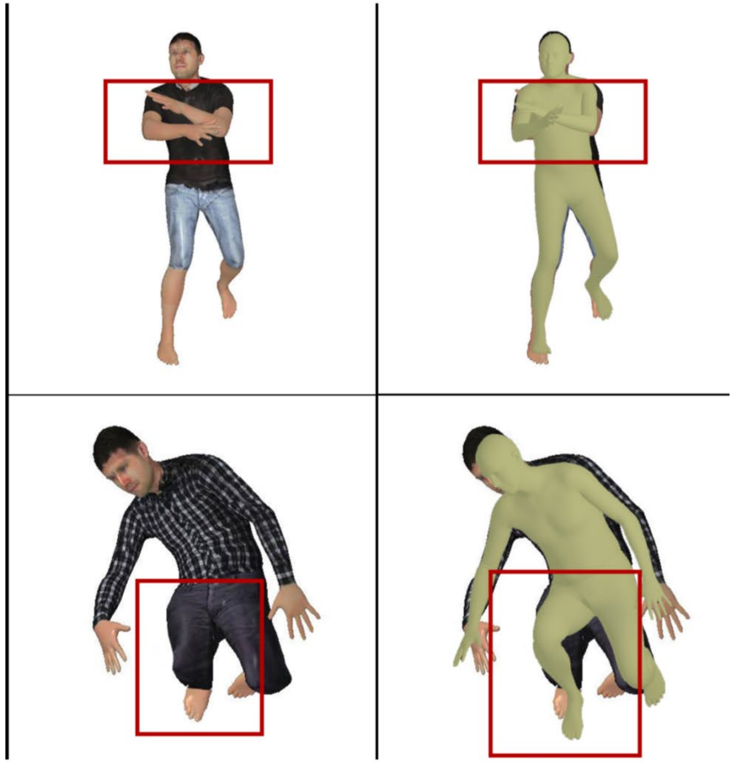
θ controls:
- Pose
- Texture
- Background
- …

- Benchmark occlusion, texture, pose, skin color, etc.
- Discover connected regions in parameter space with large pose error
- Use these to improve pose prediction models → new SOTA

[Liu et al. CVPR 2023]

Generative Adversarial Testing of Human Pose Estimation

- Failure Modes generalize well to real images.



(b) Failure modes found by PoseExaminer

[Liu et al. CVPR 2023]

Conclusion

- Deep Networks do **not generalize robustly**
  - More data is not enough to solve robustness
- We need **more challenging datasets** that "stress test" computer vision models
  - Generative models as parametric datasets that can be searched adversarially
- We need generative models to improve computer vision
  - Deep networks + 3D generative models → Robust Generalization
  - Deep networks **VS** 3D generative models → Generative Adversarial Testing

Generative Models for Computer Vision

CVPR 2023, June 18th

Generative Models


NeRF


2D GAN

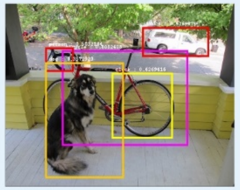

3D GAN

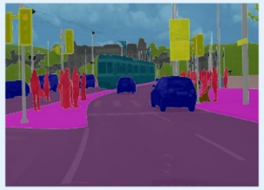

Text-to-Image

?

Computer Vision

Image Classification


Object Detection


Semantic Segmentation


Pose Estimation


Phillip Isola
MIT

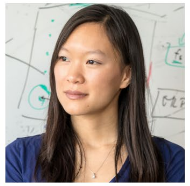

Angjoo Kanazawa
UC Berkeley


Yi Ma
UC Berkeley


Gordon Wetzstein
Stanford University


Shubham Tulsiani
CMU


Ben Mildenhall
Google Research


Angela Dai
TUM


Björn Ommer
LMU


Andrea Tagliasacchi
SFU


Alan Yuille
JHU