Cycle-Consistent Generative Rendering for 2D-3D Modality Translation

Tristan Aumentado-Armstrong, Alex Levinshtein, Stavros Tsogkas, Konstantinos Derpanis, and Allan Jepson

Samsung AI Centre Toronto

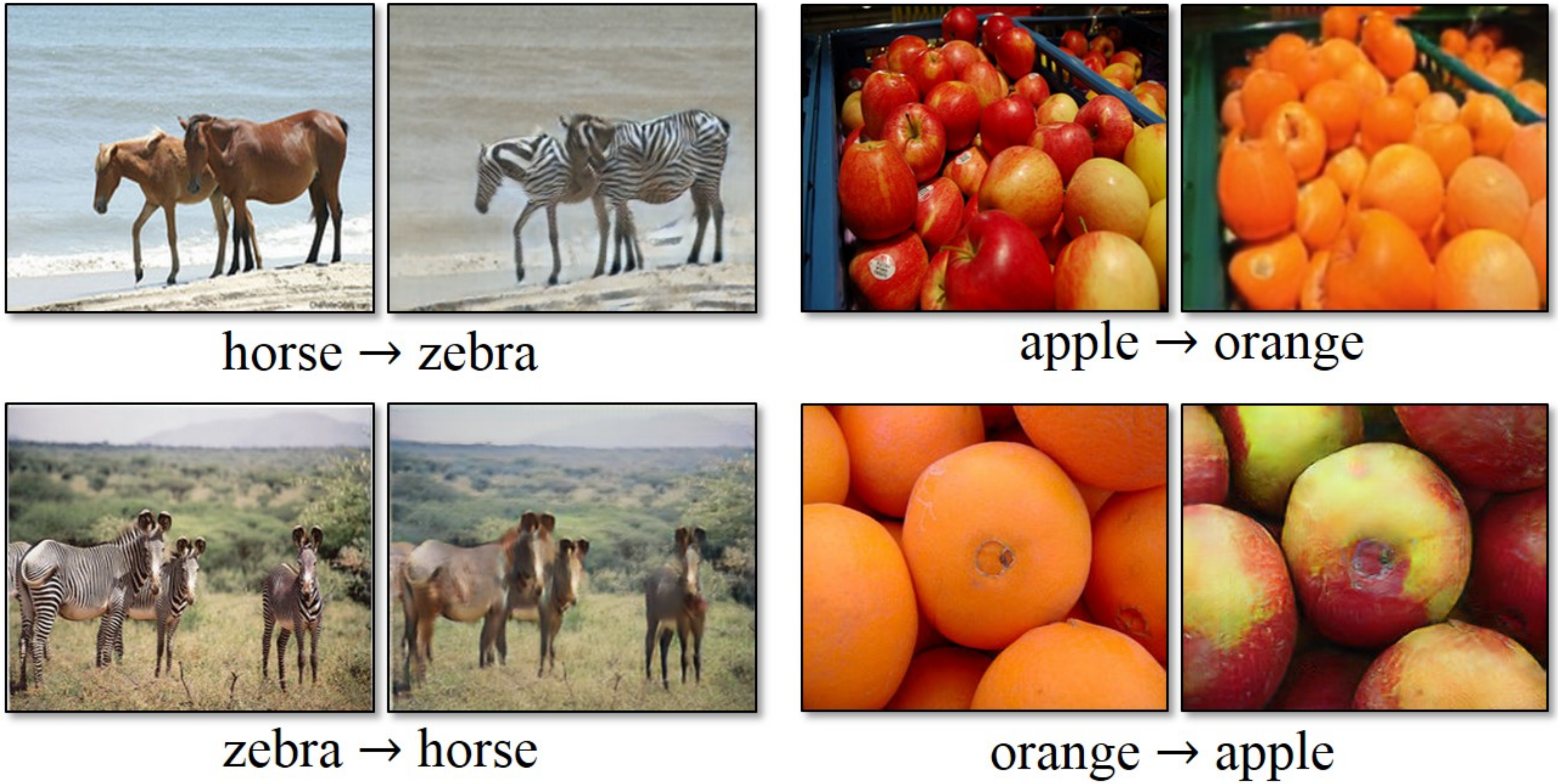
SAMSUNG

AI CENTRE

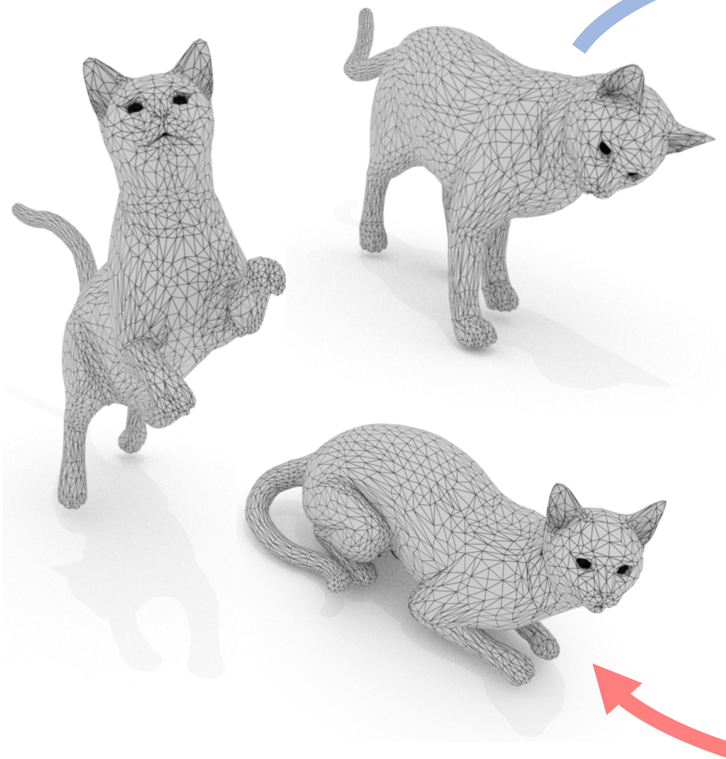
CycleGAN: Unpaired Domain Translation

Allows translating (mapping) between image domains

Requires only unpaired data



Can we learn a CycleGAN between the 2D and 3D object *modalities* in the same way?



Shape-to-Image Generation

(*Graphics problem*)

?

Image-to-Shape Inference

(*Vision problem*)

Motivation

2D-3D **paired data** is hard and/or costly to obtain

From: Pix3D

E.g., need exact 3D shape, correct rigid alignment, camera, etc…

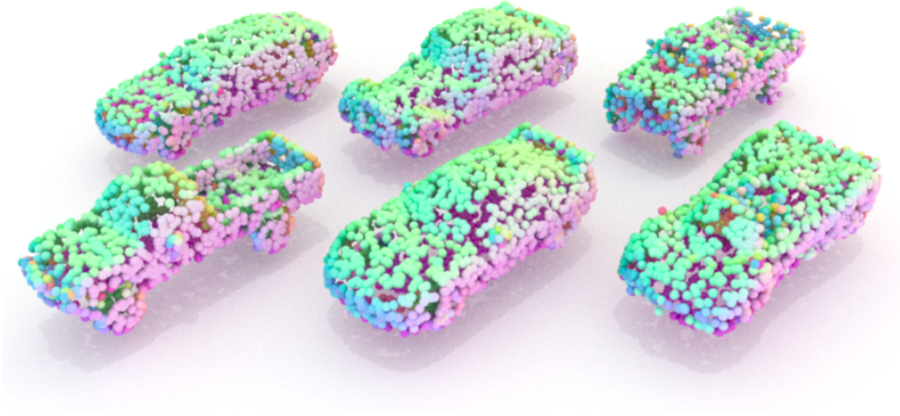
Motivation

But we **do** have large *separate* (unpaired) 3D model datasets and 2D image datasets

Unannotated masked images

Unpaired Data

Untextured 3D shapes



This is sufficient to train our cycle-consistent model

Goal

Learn a 2D-3D modality translator with:

- **No** paired data requirement
- A **generative model** of paired data
- Weakly supervised **textured 3D mesh inference**

Shape-to-Image Generation (2D → 3D)

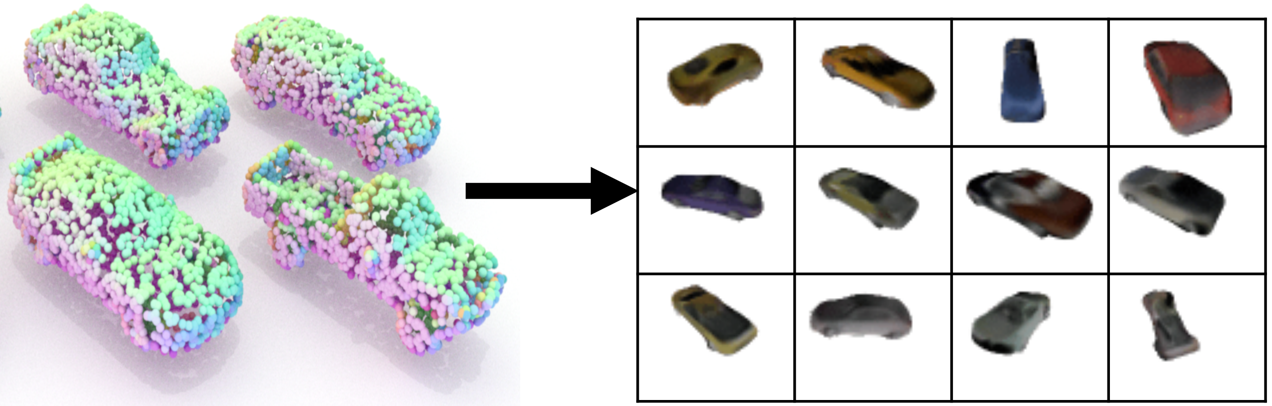
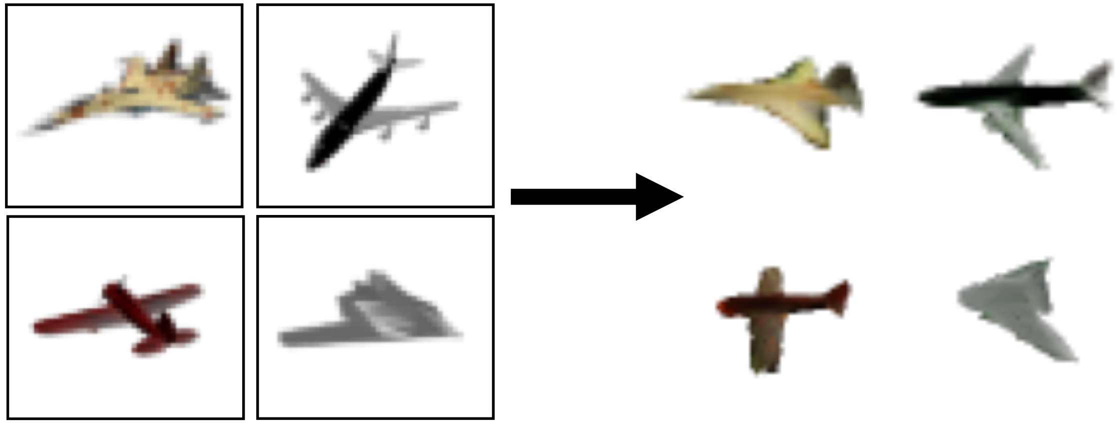
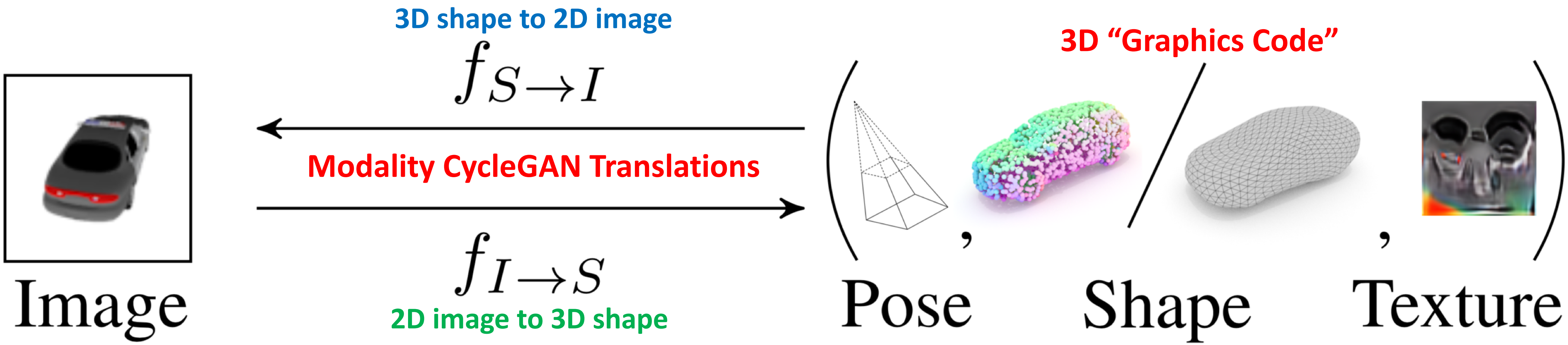


Image-to-Shape Inference (2D → 3D)

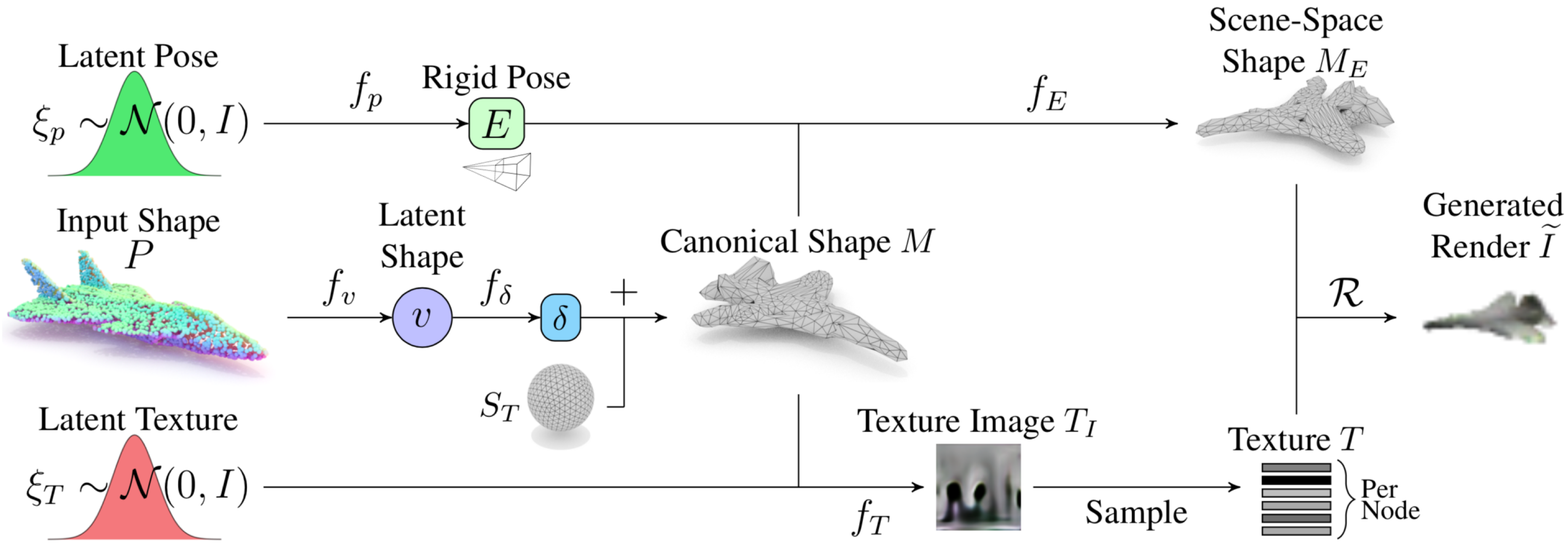


Methods: 2D-3D Modality CycleGAN



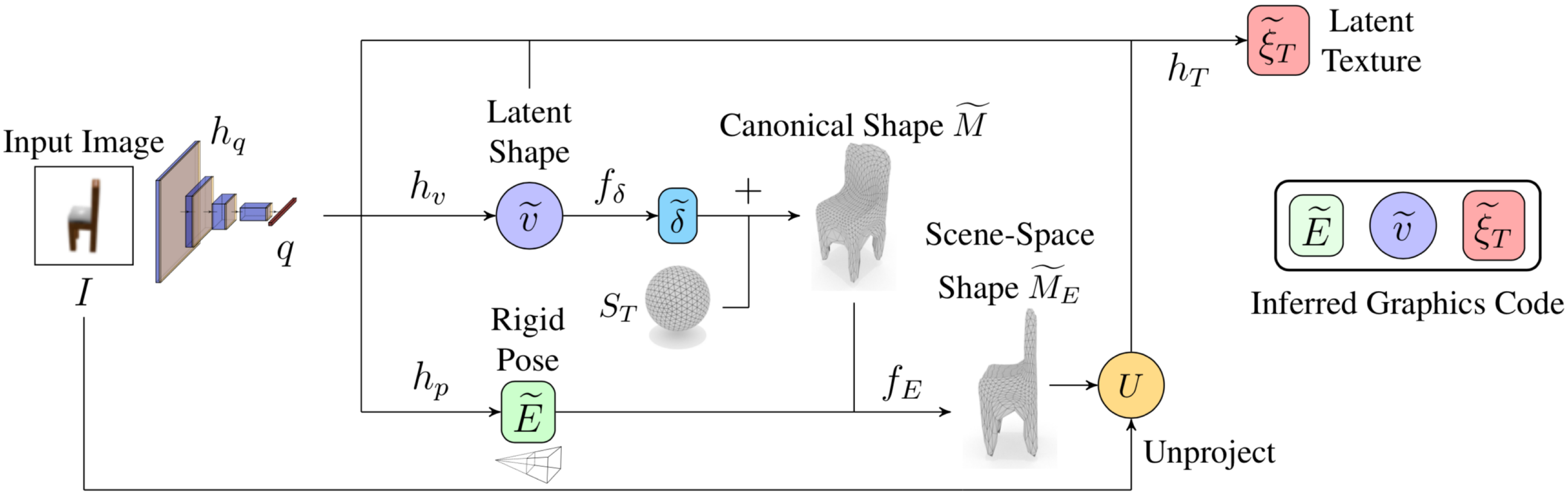
Learn a bidirectional, ~invertible mapping from **graphics code** to 2D rendered image

Methods: Shape-to-Image Translation



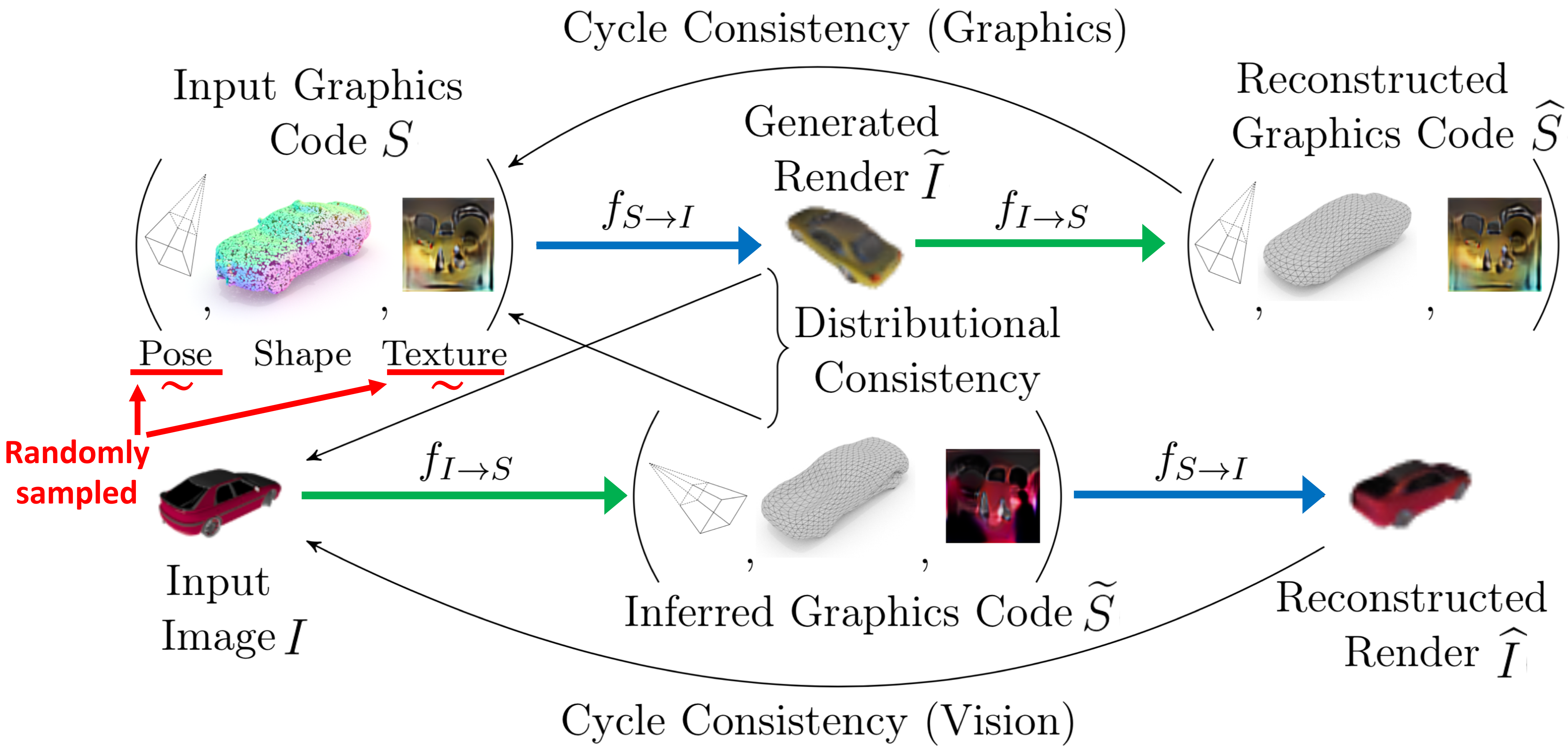
Learned graphics pipeline: generate image from 3D specifications

Methods: Image-to-Shape Translation



Vision as inverse graphics: infer 3D scene parameters from 2D image

Methods: 2D-3D CycleGAN Training Architecture



Methods: Loss Objective

Two main terms per cycle: distribution-matching and cycle-consistency

Distribution-matching losses

Generated images $\widetilde{I}$ and inferred shapes $\widetilde{S}$ should be in-distribution

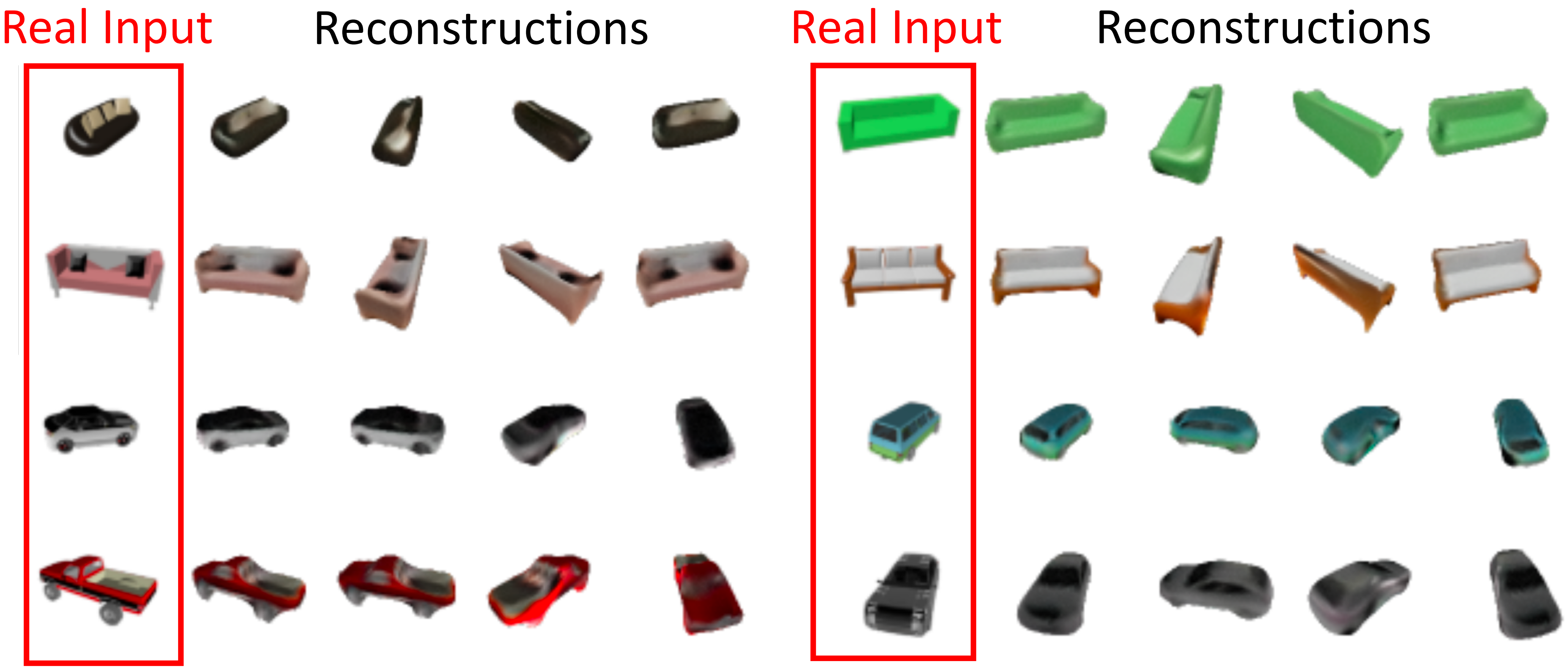
Regularization (e.g., mesh quality)

$$\mathcal{L}_{S,I}(\Phi) = \overbrace{\mathfrak{D}_{\mathcal{S}}\left[S \;||\; \widetilde{S}(I)\right] + \mathfrak{D}_{\mathcal{I}}\left[I \;||\; \widetilde{I}(S)\right]} + \underbrace{\mathcal{C}_{I\to\widetilde{S}\to\widehat{I}}\left[I, \widehat{I}(\widetilde{S})\right] + \mathcal{C}_{S\to\widetilde{I}\to\widehat{S}}\left[S, \widehat{S}(\widetilde{I})\right]} + \overbrace{\mathcal{L}_R}$$

Cyclic losses

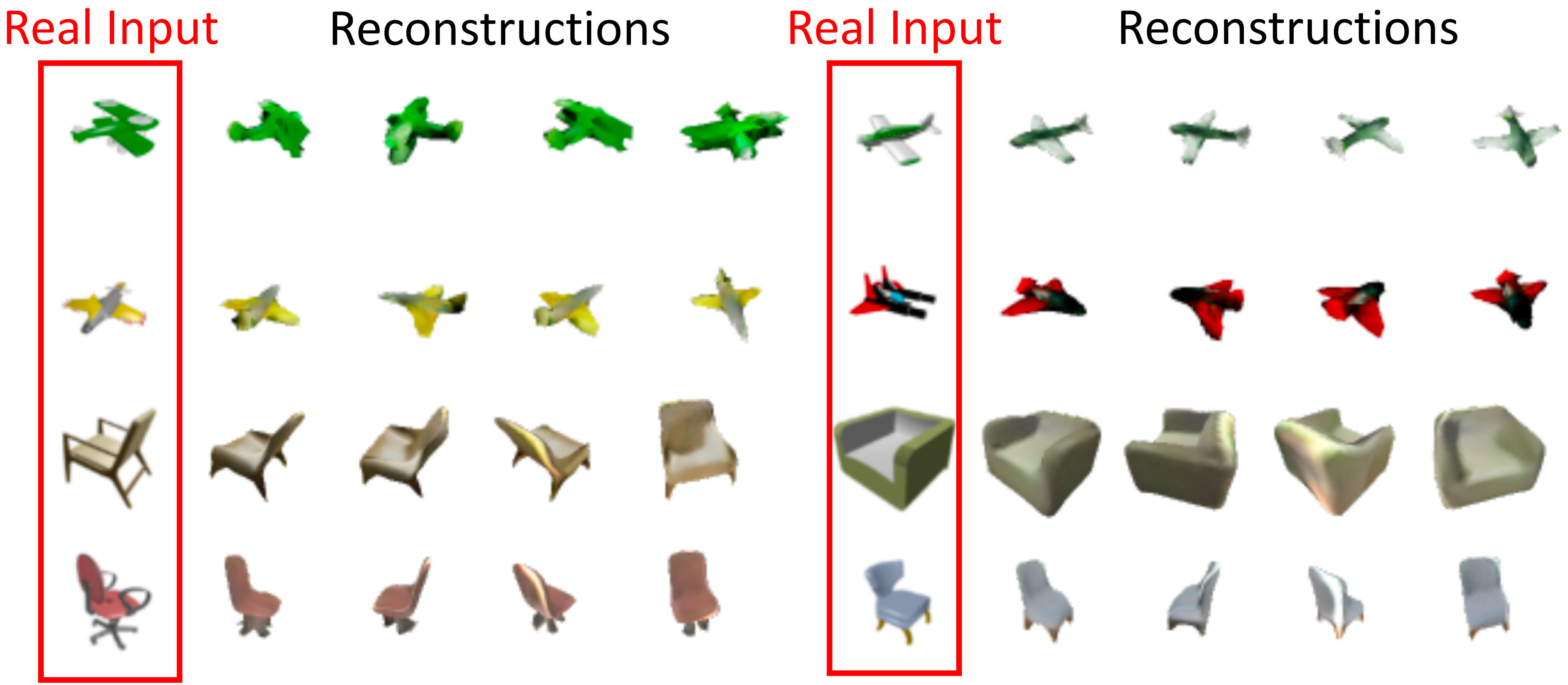
Reconstructed images $\widehat{I}$ and shapes $\widehat{S}$ should equal each cycle's input

Results: Image-to-Shape Translation



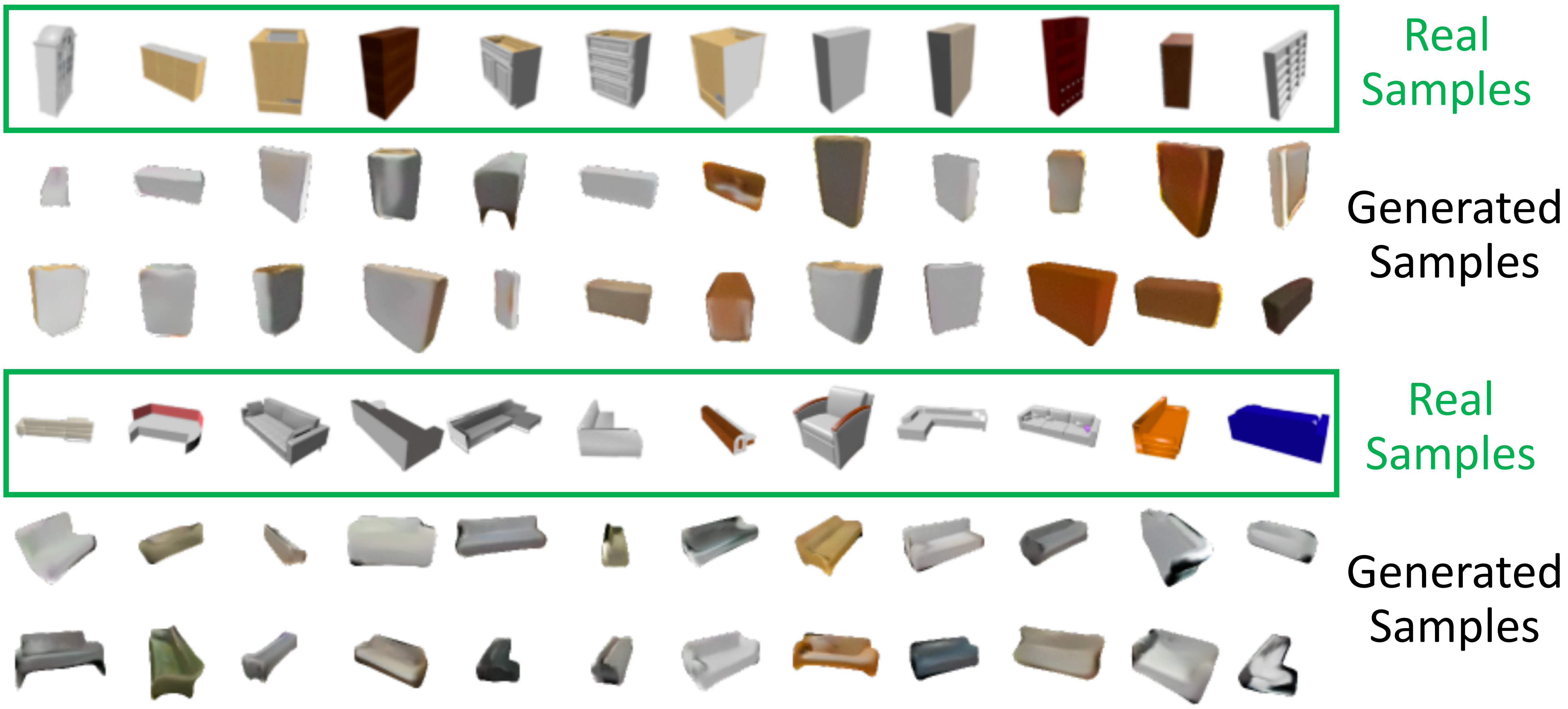
Img2shape translation learns **3D reconstruction**

Results: Image-to-Shape Translation



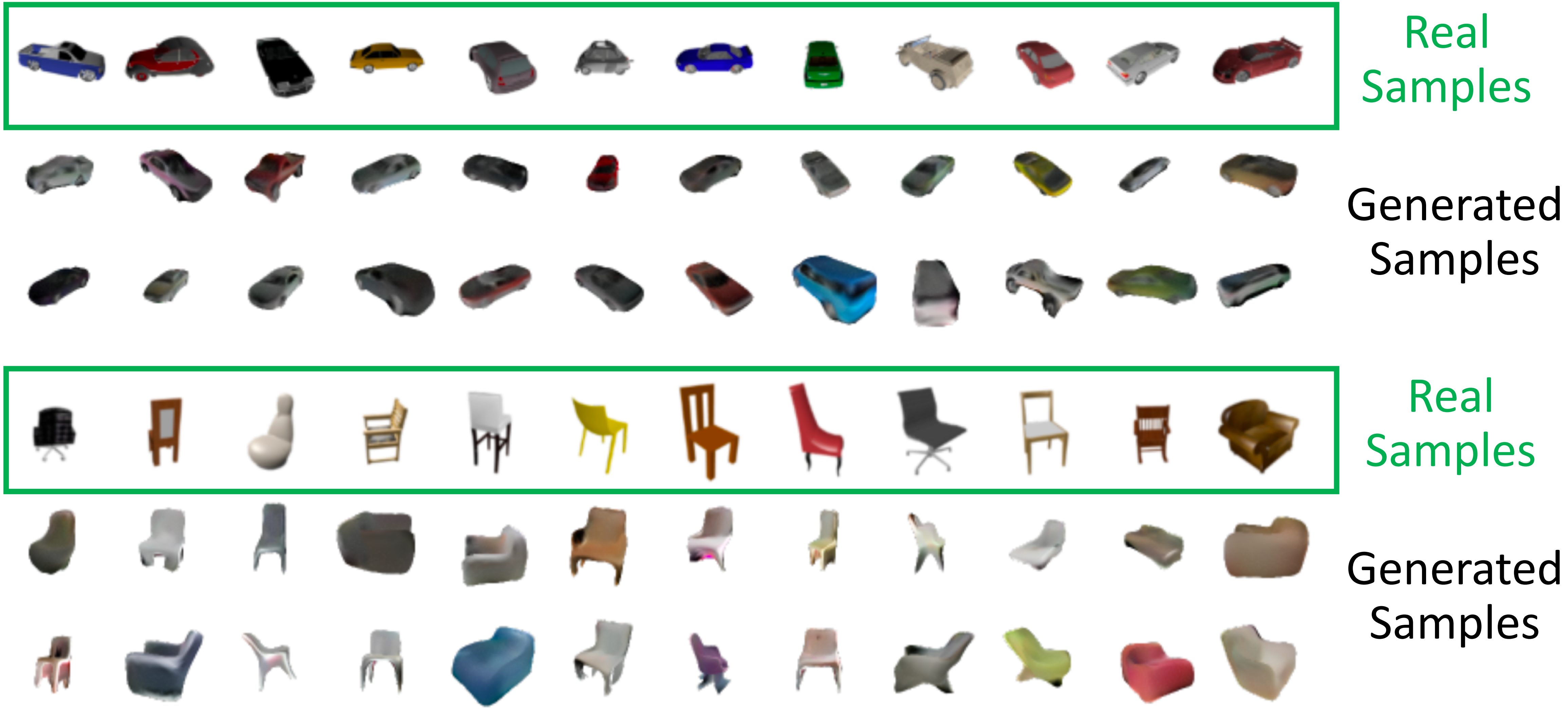
Img2shape translation learns **3D reconstruction**

Results: Shape-to-Image Translation



Shape2Img translation learns **generative image modelling**

Results: Shape-to-Image Translation

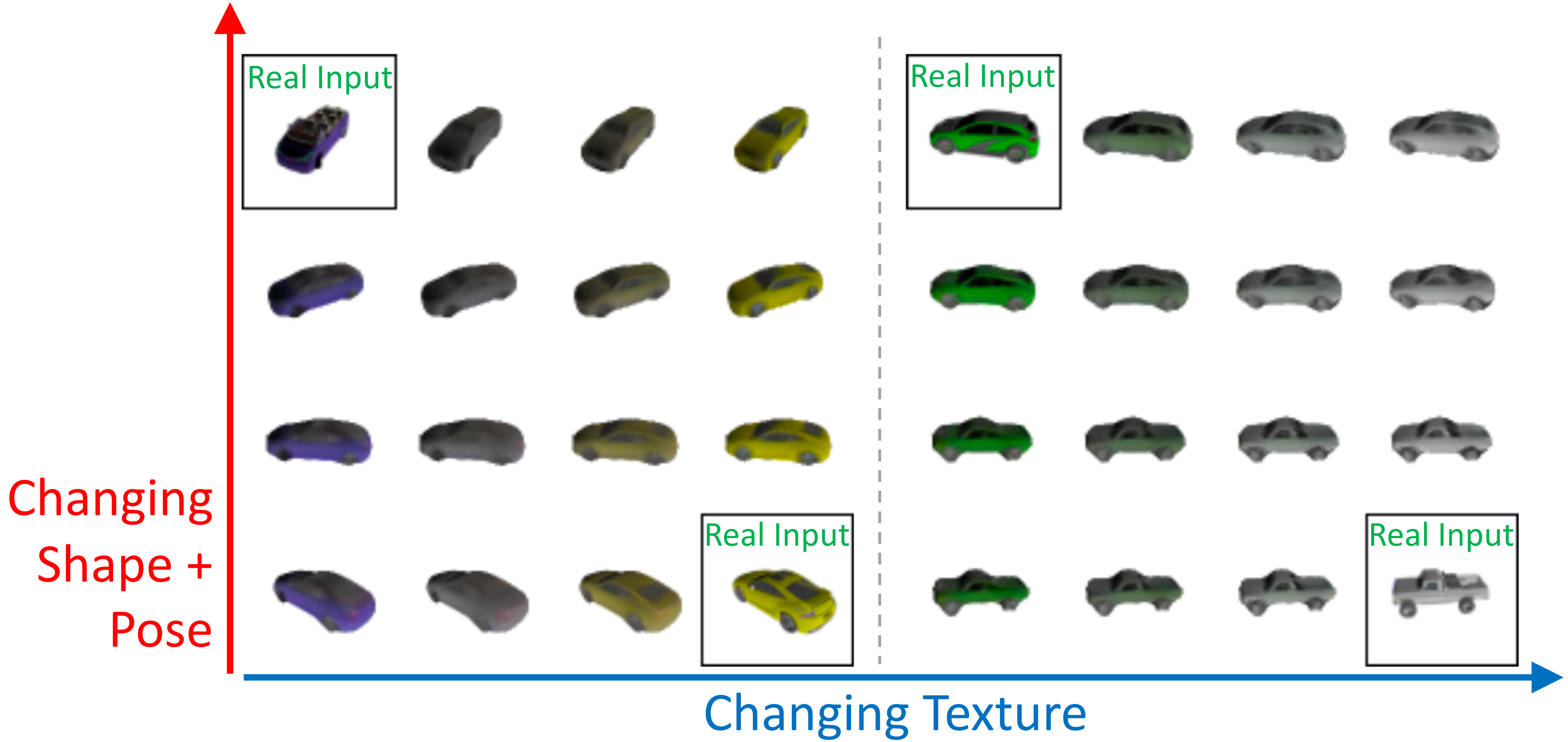


Shape2Img translation learns **generative image modelling**

Results: Unsupervised Aligned Correspondence

Template vertices naturally **correspond** across instances (due to the canonical space) and can be treated as **unsupervised keypoints**

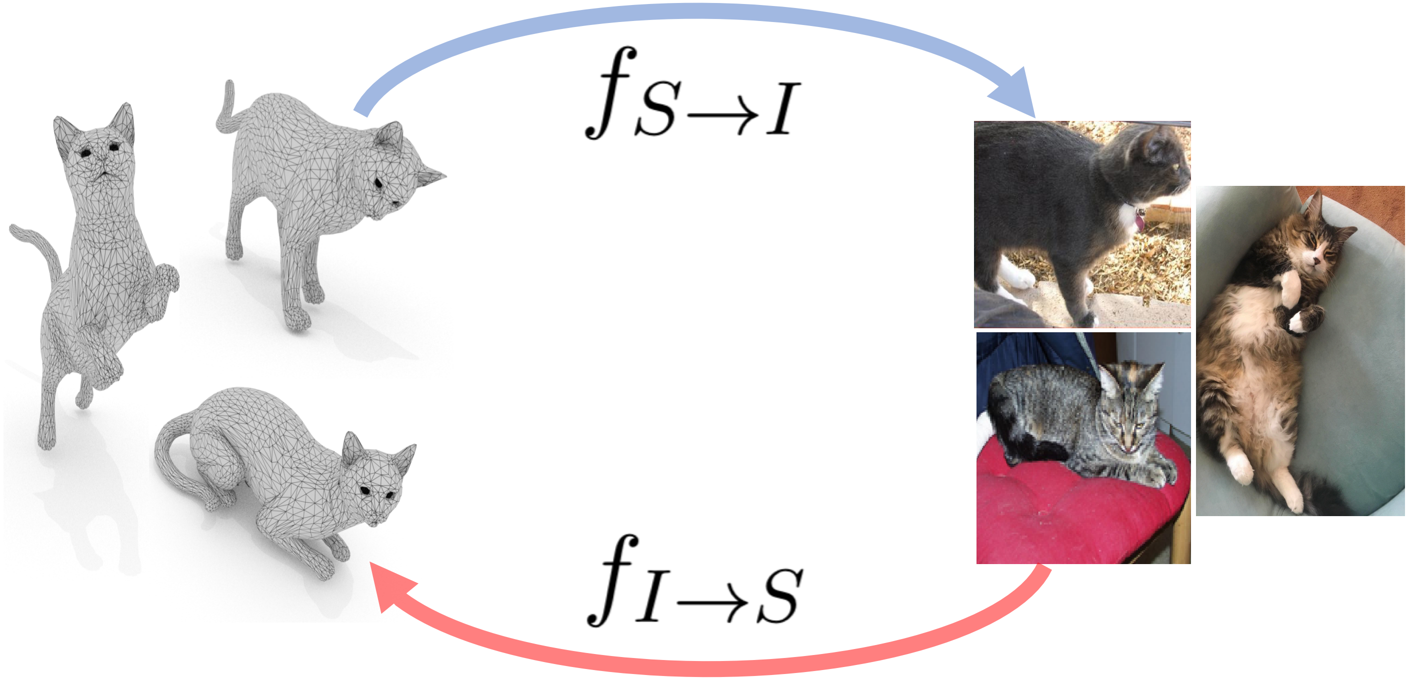
Results: Latent Representation Learning



Enables smooth and disentangled control of shape, pose, and texture

Conclusion

We have shown one can learn a 2D-3D **modality translator** from **un**paired data, capable of **3D reconstruction** and **generative image modelling**.



Still limitations with fine details (shape+texture), topology, lighting, and background.

Thank you for listening