Neural fields for 3D Vision (a MAP perspective)

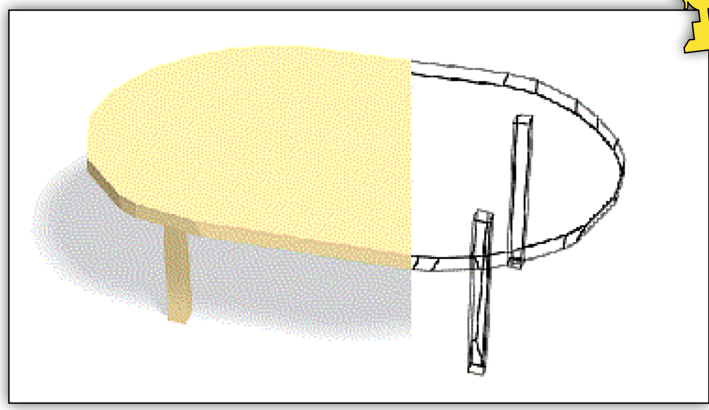
Andrea Tagliasacchi (@taiyasaki)

Associate Professor – Simon Fraser University
Staff Research Scientist – Google Deepmind

Google AI SFU

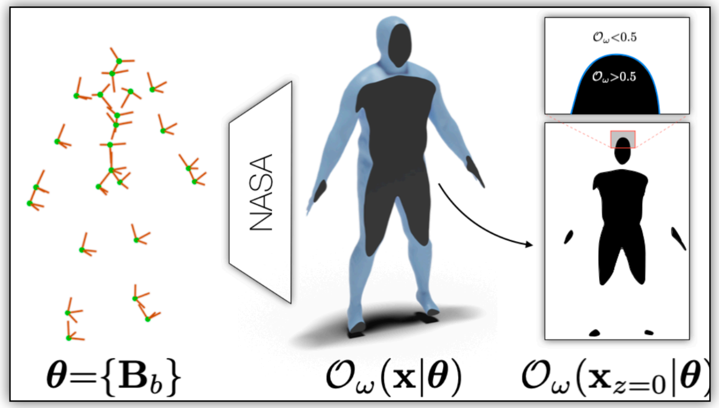
Recent: bridge graphics to vision

BSPNet @ CVPR'20



…represent meshes as fields

NASA @ ECCV'20

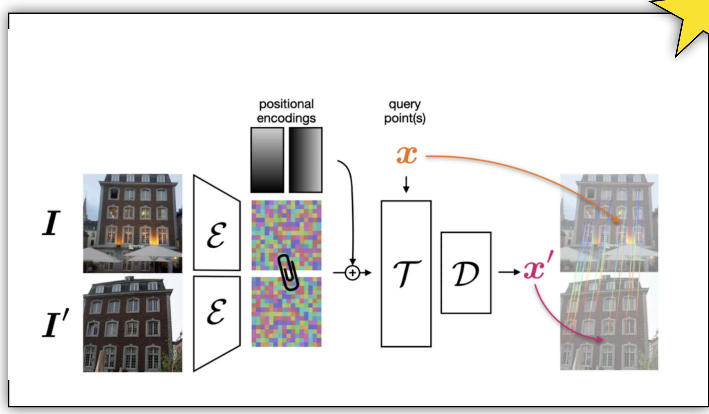


…represent humans as fields

Capsules @ NIPS'21

…canonicalize data

COTR @ ICCV'21



…image correpondences as fields

LoLNeRF @ CVPR'22

…multi-identity face field

Urban Fields @ CVPR'22

…leverage multi-sensor training data

CVPR 23: fields, fields, fields

MobileNeRF

what if I don't have a 4090?

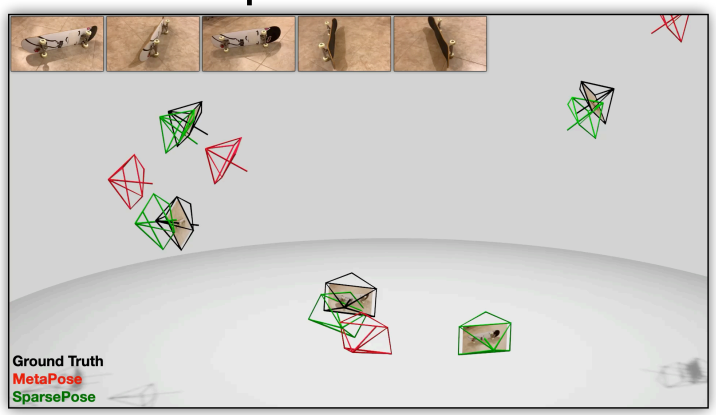
RobustNeRF

…if scene is not static?

BlendFields

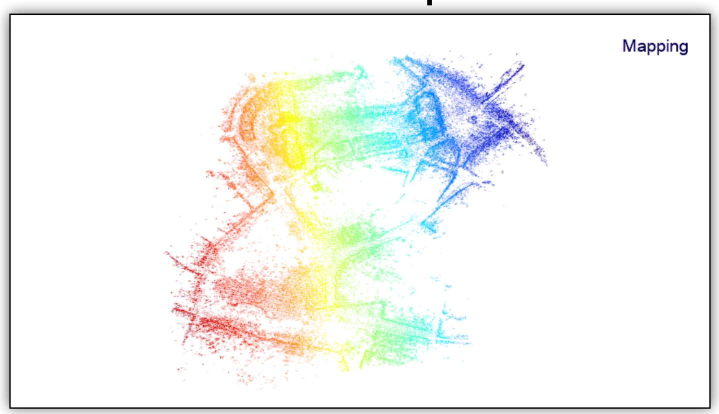
…if I have few-shot data?

SparsePose



calibration… when COLMAP breaks

NeuMap



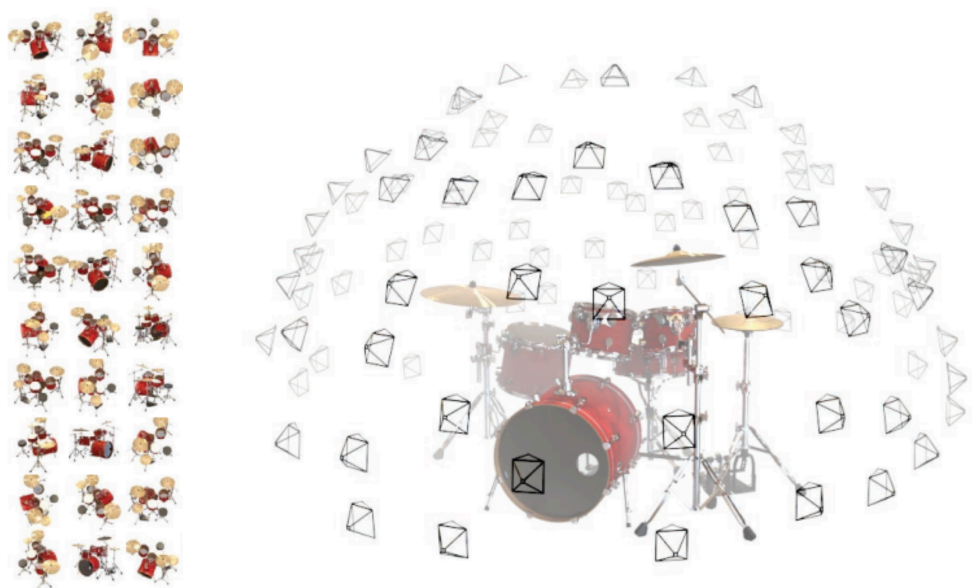
…localize myself?

Continuous Upsampling

…super-resolute an image?

Neural Radiance Fields

- “NeRF freed us from the shackles of synthetic data”

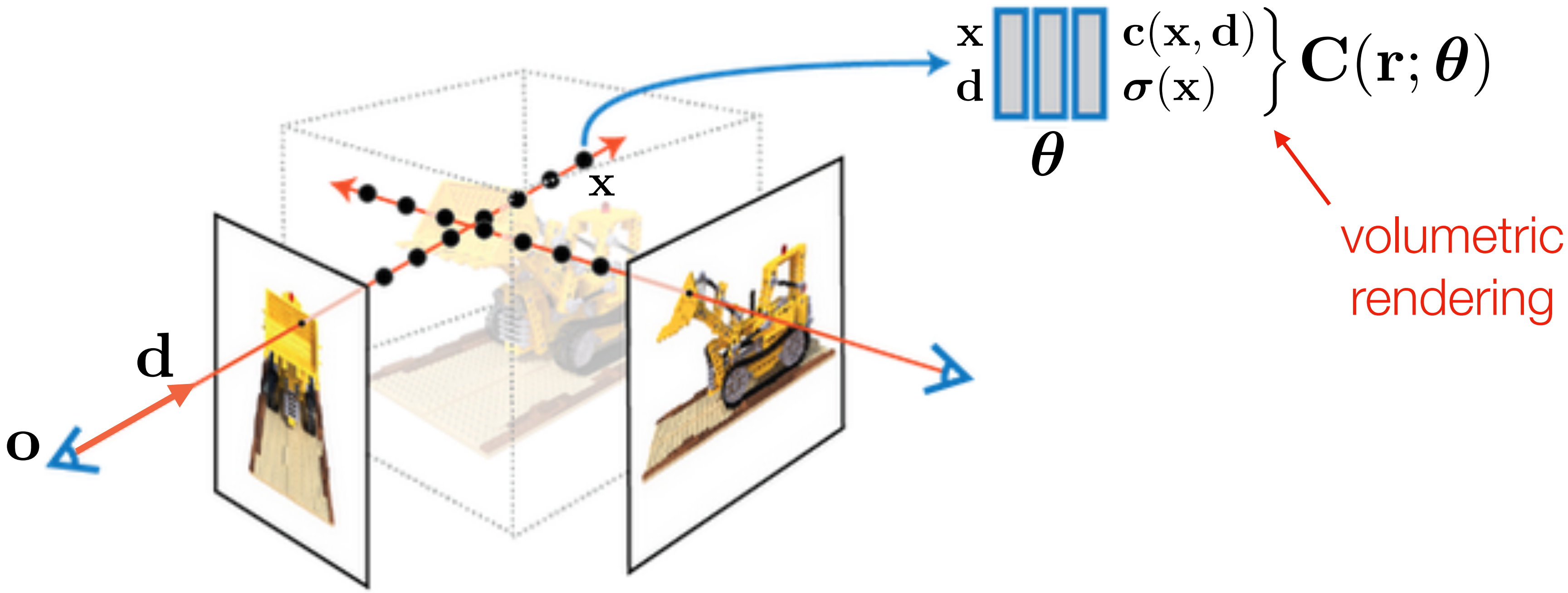


set of posed images

novel view synthesis

Mildenhall et al. «Representing Scenes as Neural Radiance Fields for View Synthesis» ECCV 2020

Basics of NeRF



$$\mathcal{L}(\boldsymbol{\theta}) = \underbrace{\mathbb{E}_{\mathbf{C}\sim\{\mathbf{C}_i\}}}_{\text{select image}} \underbrace{\mathbb{E}_{\mathbf{r}\sim\mathbf{C}}}_{\text{select pixel/ray}} ||\underbrace{\mathbf{C}(\mathbf{r}; \boldsymbol{\theta})}_{\text{predicted color}} - \underbrace{\mathbf{C}_i(\mathbf{r})}_{\text{ground truth}}||_2^2$$

NeRF as MAP (maximum a posteriori)

data model

$$\arg\max_{\mathcal{M}} \; p(\mathcal{D}, \mathcal{M}) \equiv \arg\max_{\mathcal{M}} \; p(\mathcal{D}|\mathcal{M}) \cdot p(\mathcal{M})$$

Bayes' rule

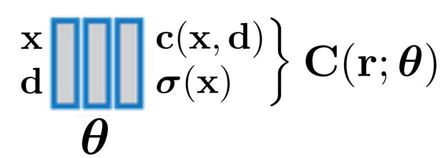
$$\equiv \arg\max_{\mathcal{M}} \sum_i \underbrace{p(\mathbf{d}_i|\mathcal{M})}_{likelihood} \cdot \underbrace{p(\mathcal{M})}_{prior}$$

i.i.d. pixels

photometric loss

$$||\mathbf{C}(\mathbf{r};\boldsymbol{\theta}) - \mathbf{C}_i(\mathbf{r})||_2^2$$

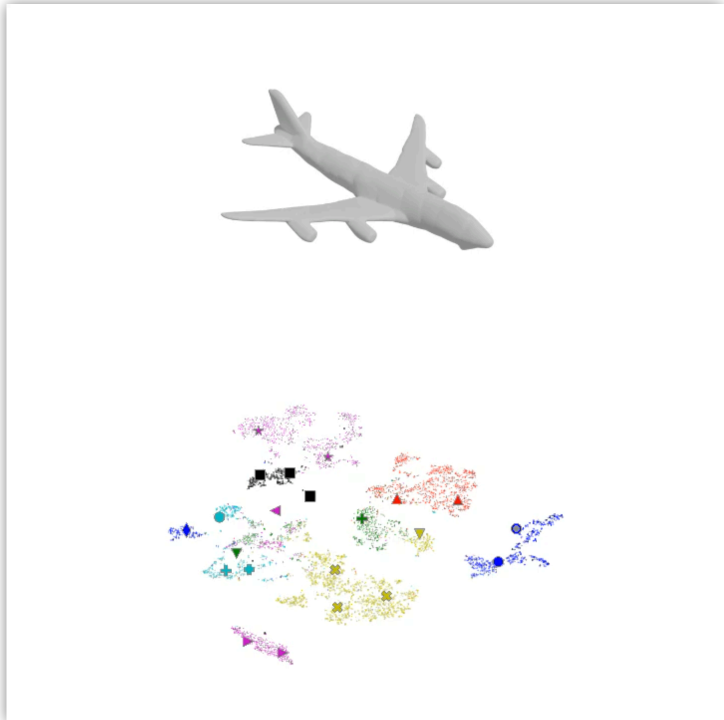
spectral bias of MLPs



A historical perspective

$$\arg\max_{\mathcal{M}} \; p(\mathcal{D}|\mathcal{M}) \cdot p(\mathcal{M})$$

encoder-decoder

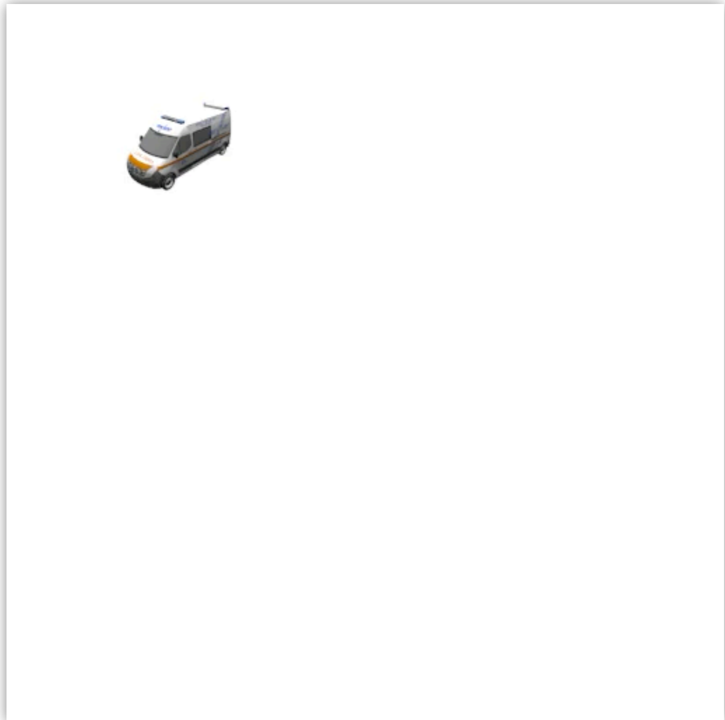


CvxNet @ CVPR'20

auto-decoder

LoLNeRF @ CVPR'22

meta-learning



3DiM @ ICLR'23

A historical perspective

$$\arg\max_{\mathcal{M}} \; p(\mathcal{D}|\mathcal{M}) \cdot p(\mathcal{M})$$

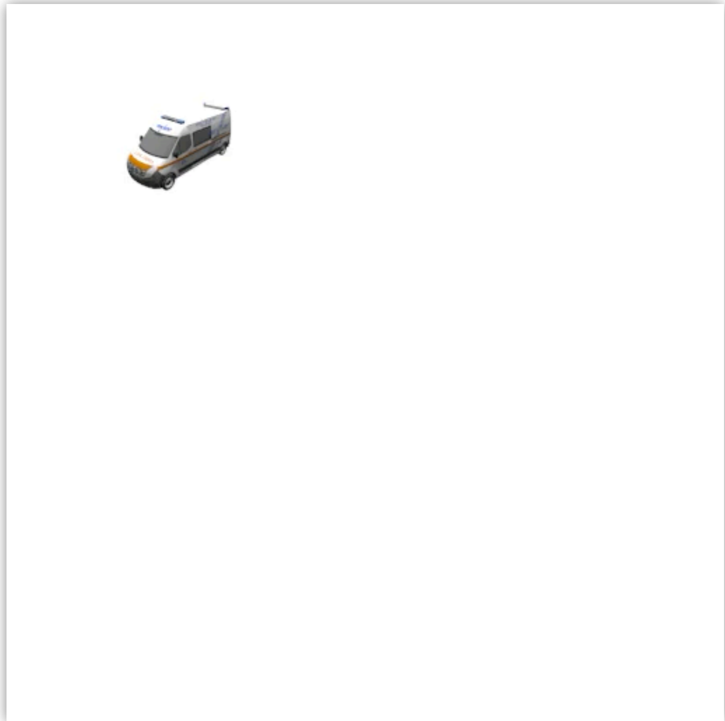
encoder-decoder

CvxNet @ CVPR'20

auto-decoder

LoLNeRF @ CVPR'22

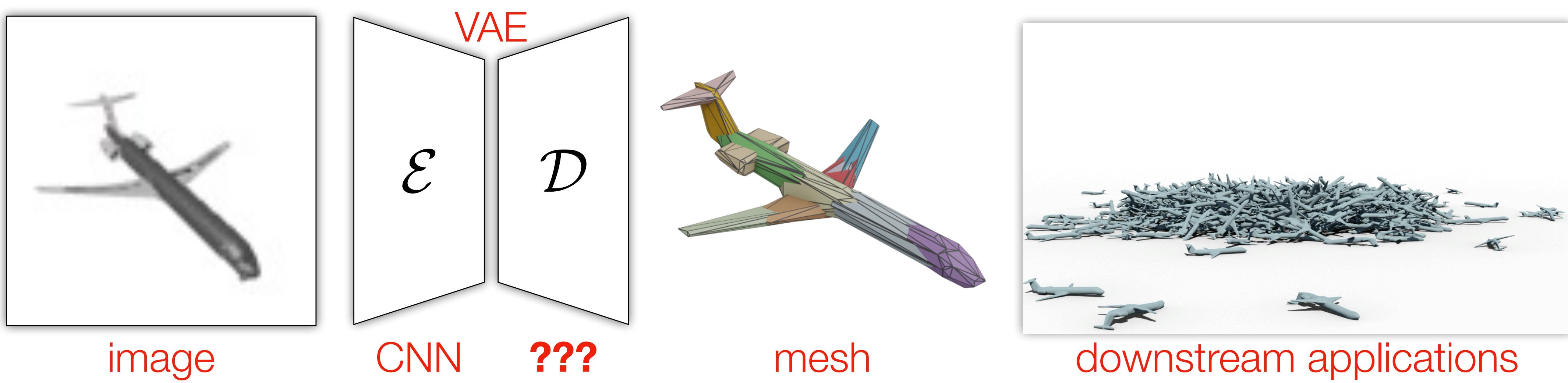
meta-learning



3DiM @ ICLR'23

Problem Statement

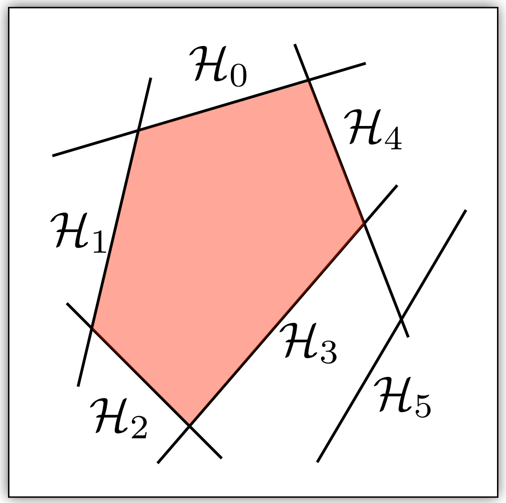
- What type of decoder should we use? …a 3D CNN?



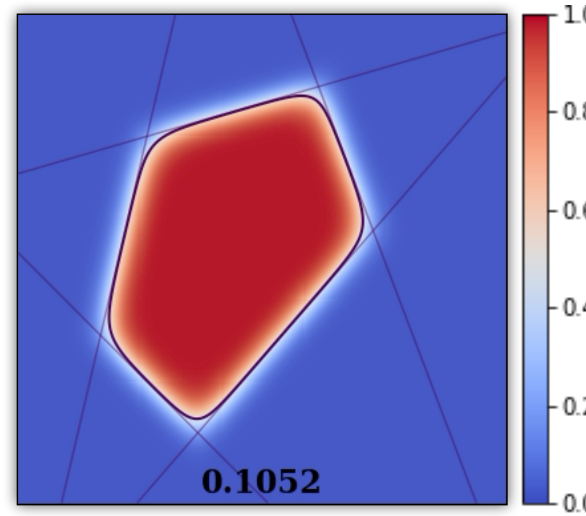
Deng et al. «CvxNet: Learnable Convex Decomposition» CVPR 2020

Convexes as differentiable fields

polygonal mesh **vs.** implicit function



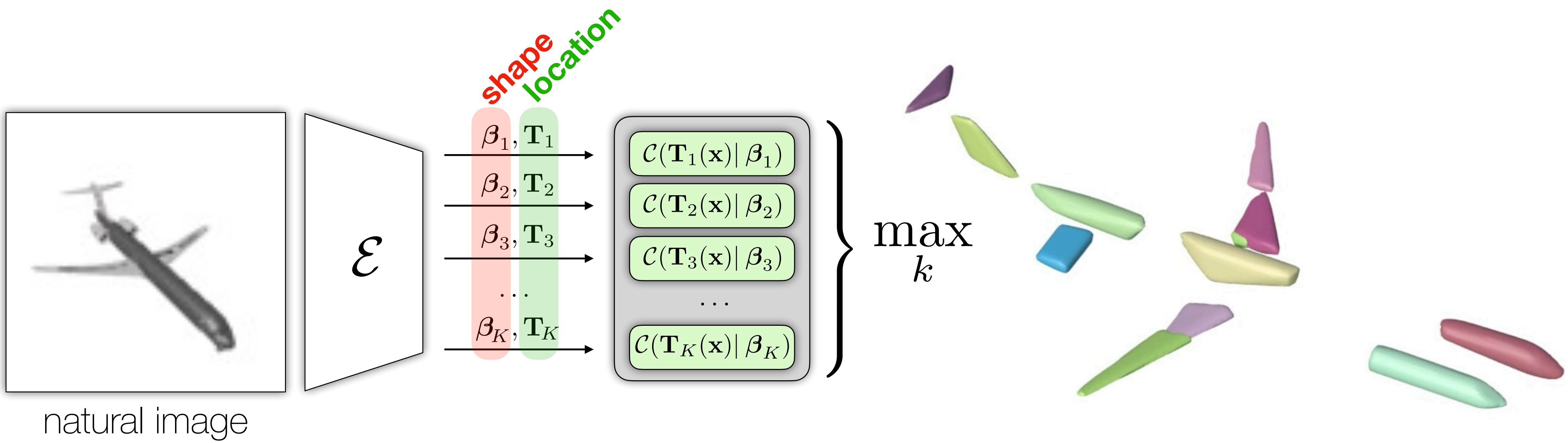
$$\mathcal{C}(\mathbf{x}|\{\mathcal{H}_h\}, \sigma, \delta)$$



$$\overbrace{\mathrm{Sigmoid}(-\sigma \mathrm{LogSumExp}\{\delta(\mathbf{n}_h \cdot \mathbf{x} + d_h)\})}$$

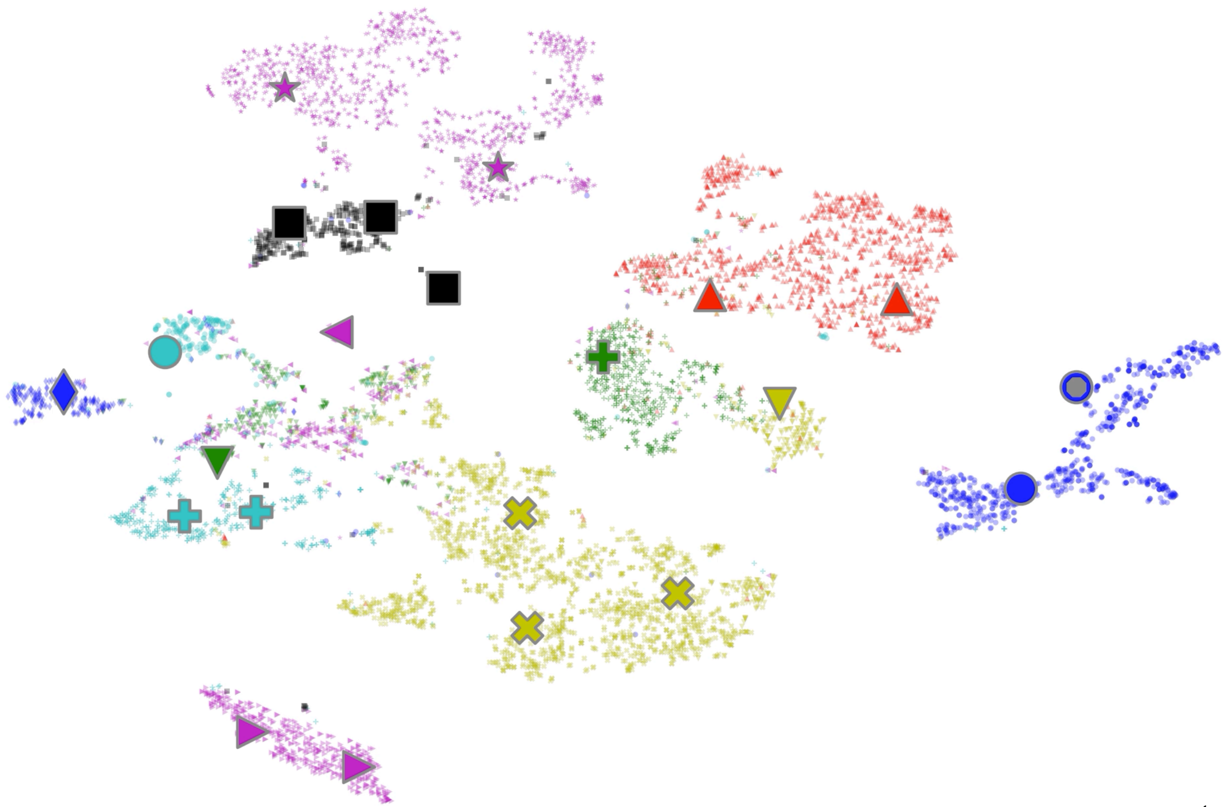
Deng et al. «CvxNet: Learnable Convex Decomposition» CVPR 2020

Convexes for generative modeling



Deng et al. «CvxNet: Learnable Convex Decomposition» CVPR 2020

Navigating the latent space



$$\arg\max_{\mathcal{M}} \; p(\mathcal{D}|\mathcal{M}) \cdot \boxed{p(\mathcal{M})}$$

So why this path “dried up”?

- Required 3D supervision and paired image/3D
- Shapes lack crisp details
  - Latent code causes representation bottleneck
  - NeRF positional encoding… does not help

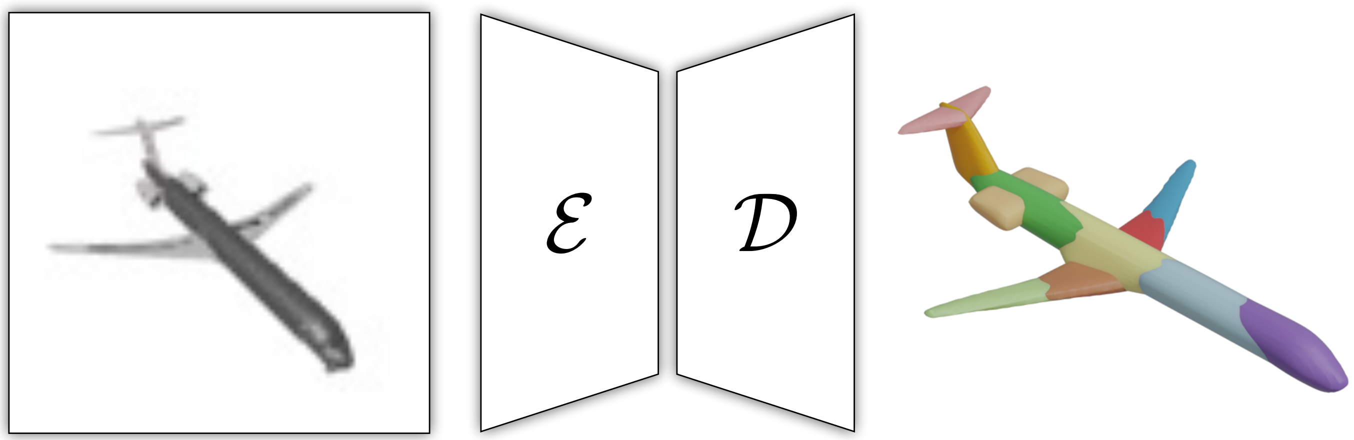
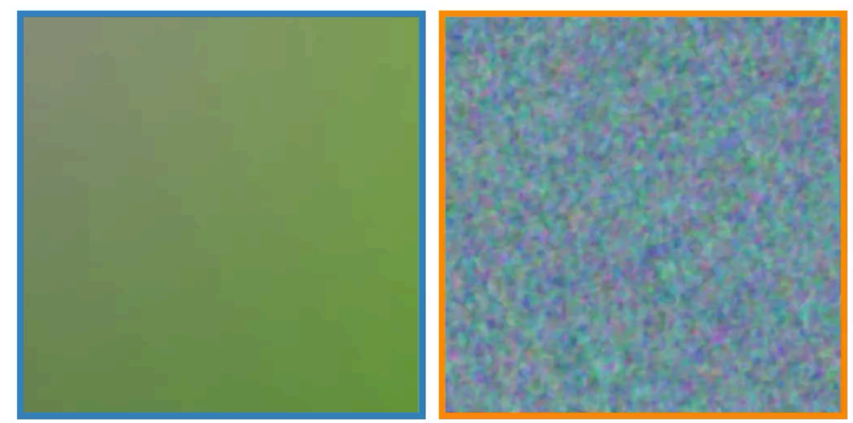


image CNN field (smooth) mesh

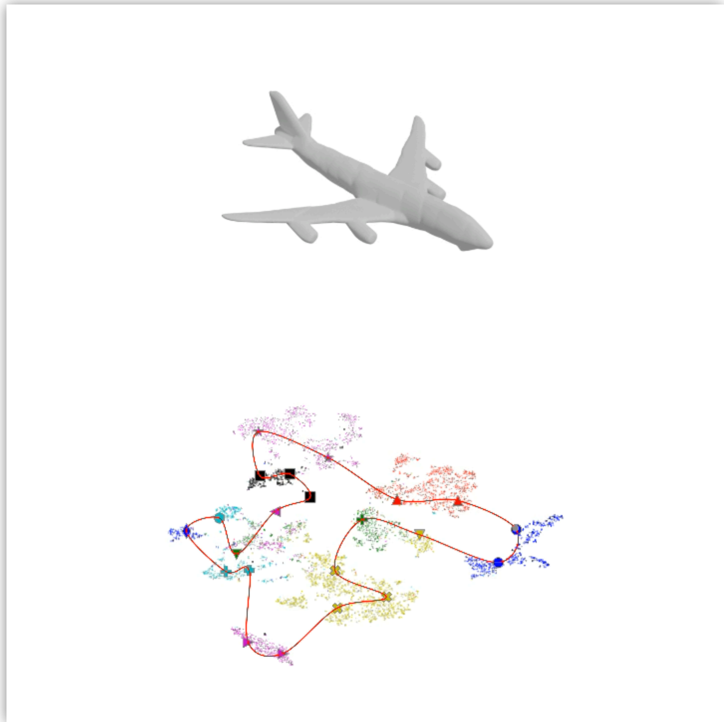


w/o posenc w/ posenc

A historical perspective

$$\arg\max_{\mathcal{M}} \; p(\mathcal{D}|\mathcal{M}) \cdot p(\mathcal{M})$$

encoder-decoder

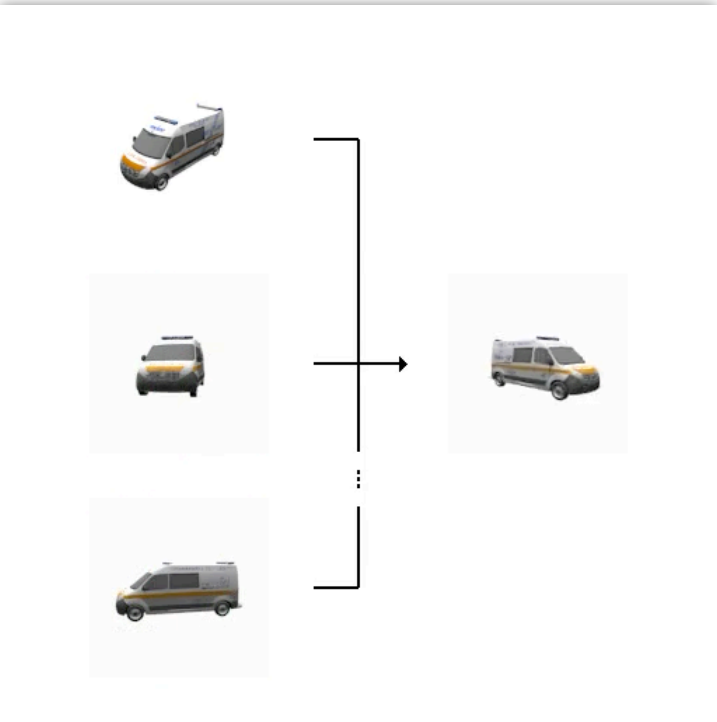


CvxNet @ CVPR'20

auto-decoder

LoLNeRF @ CVPR'22

meta-learning



3DiM @ ICLR'23

Problem Statement

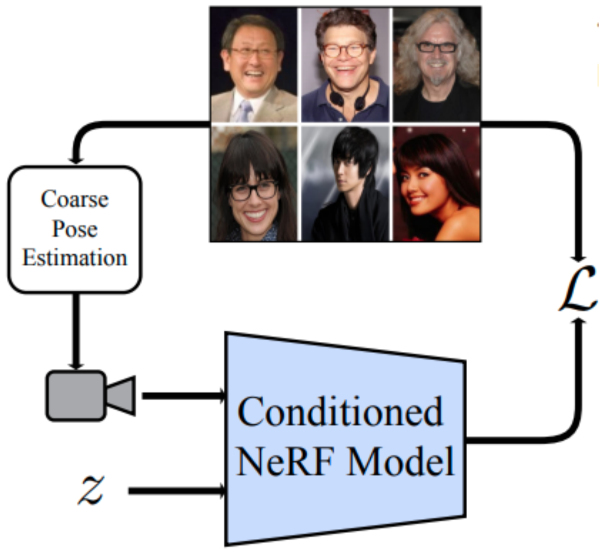
many images / scene
one scene

vs.

single image / scene
many scenes

Conditional Fields

- Lift 3D out 2D – via a large collection of 2D images (e.g. CelebA-HQ)
- Avoid relying on GANs to enforce 3D consistency (e.g. pi-GAN)



conditioning latent

$$\mathcal{L}_{\mathrm{rgb}}^{\mathbf{r},i}(\mathbf{z}_i, \boldsymbol{\theta}) = ||\mathbf{C}(\mathbf{r}; \mathbf{z}_i, \boldsymbol{\theta}) - \mathbf{C}_i(\mathbf{r})||_2^2$$

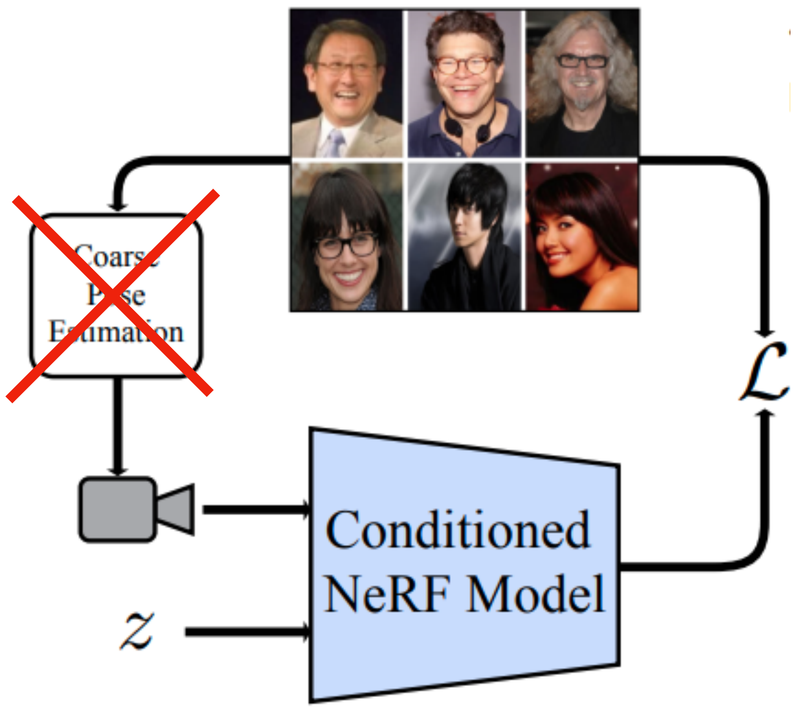
$$\mathcal{L}_{\mathrm{rgb}}(\boldsymbol{\theta}, \{\mathbf{z}_i\}) = \sum_i \mathbb{E}_{\mathbf{r}\sim\mathbf{C}_i}\left[\mathcal{L}_{\mathrm{rgb}}^{\mathbf{r},i}(\mathbf{z}_i, \boldsymbol{\theta})\right]$$

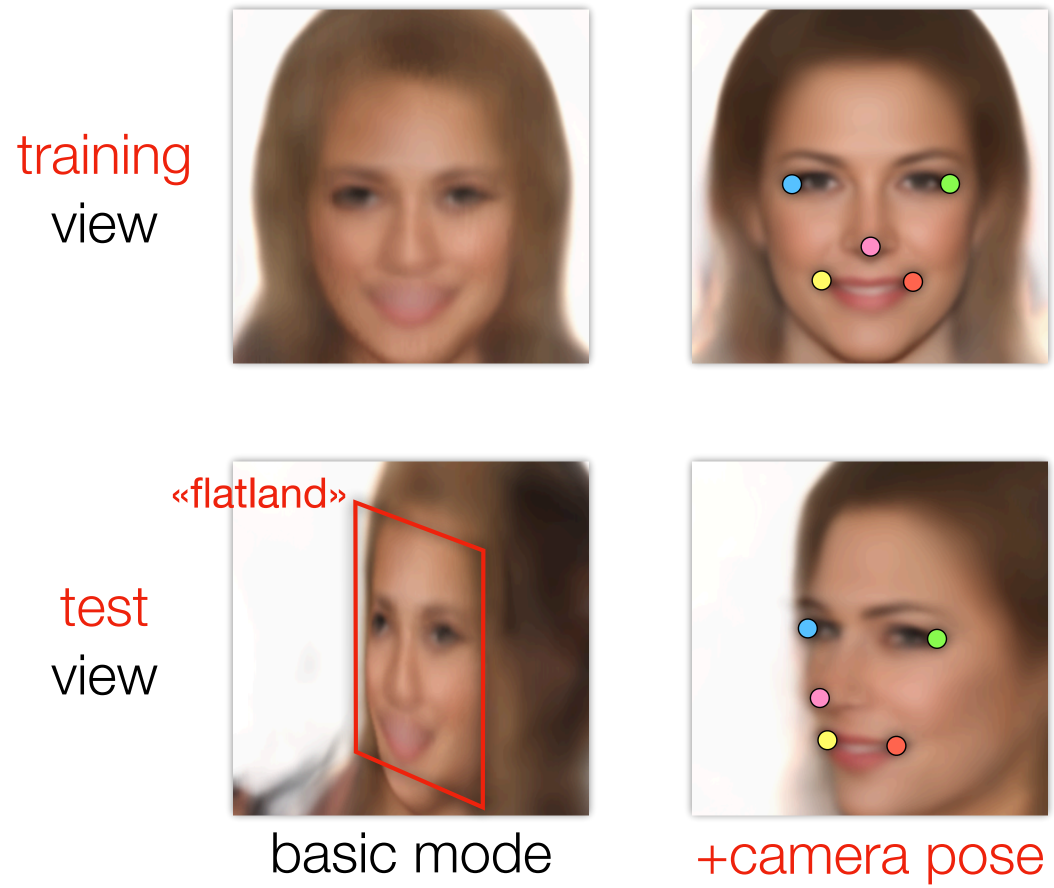
$$\arg\min_{\{\mathbf{z}_i\},\boldsymbol{\theta}} \mathcal{L}_{\mathrm{rgb}}(\{\mathbf{z}_i\}, \boldsymbol{\theta})$$

auto-decoder

Flatland mode collapse

- What happens if no camera pose is given?



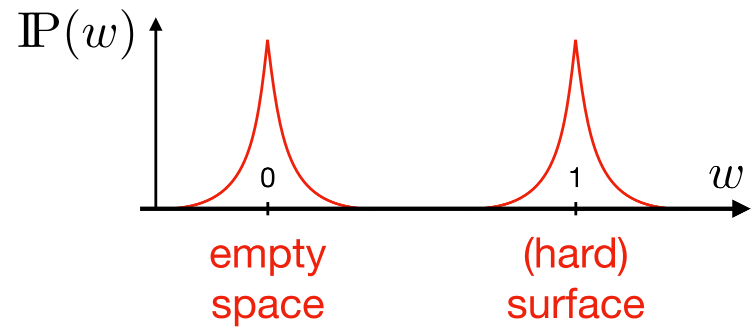


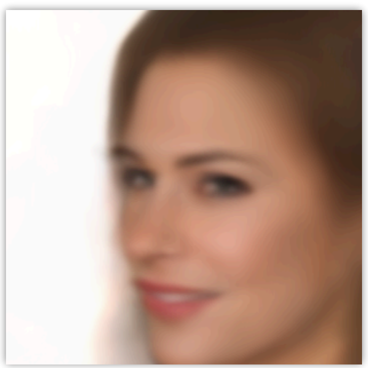
{empty | solid} prior model

- What happens if most cameras are frontal?
- How to regularize this behavior?
  - skin ~ hard / solid media

$$\mathbf{C}(\mathbf{r}) = \int_{t_n}^{t_f} w(\mathbf{x}) \cdot \mathbf{c}(\mathbf{x}, \mathbf{d})\, dt$$

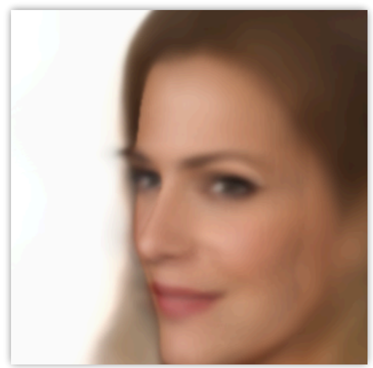
$$\mathcal{L}_{\text{hard}} = -\log(\underbrace{e^{-|w|} + e^{-|1-w|}}_{\mathbb{P}(w)})$$





"pinocchio"

skin≠smoke



w/ hard loss

regularizes output geometry

Fitting new identities

- Seek the latent-code that minimizes the losses (i.e. frozen weights)

Input Image (FFHQ)

piGAN fit (CelebA@128^2)

LoLNeRF fit (CelebA@256^2)

testing dataset

CelebA

FFHQ

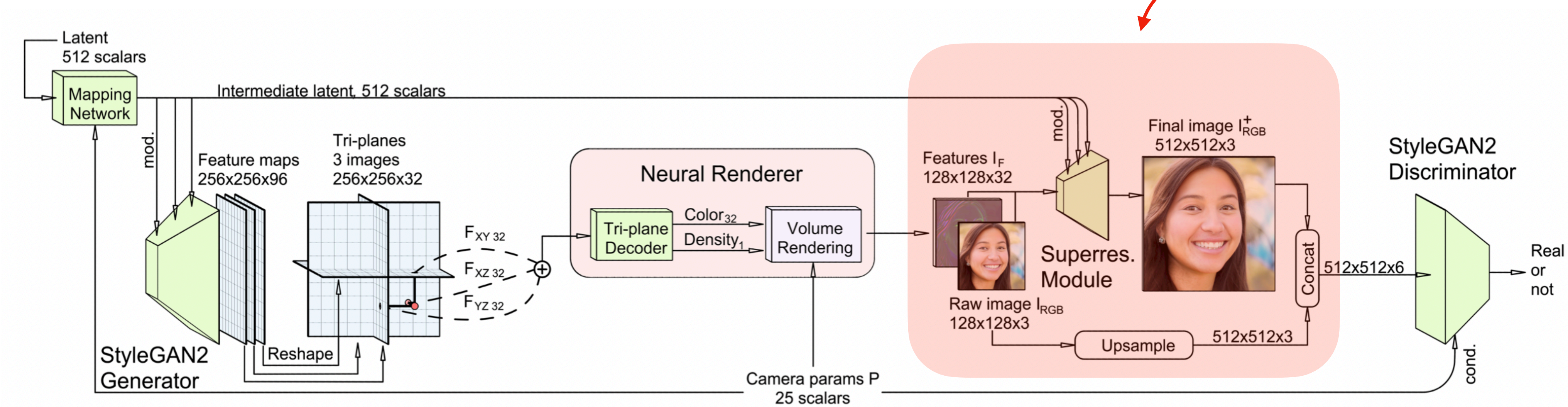
Method	PSNR↑	SSIM↑	LPIPS↓	Res.
π-GAN [11] (CelebA)	21.8	0.796	0.412	256^2
Ours (CelebA-HQ)	**26.2**	**0.856**	**0.363**	
π-GAN [11] (CelebA)	20.9	0.795	0.522	512^2
Ours (CelebA-HQ)	25.1	0.831	0.501	
Ours (FFHQ)	**25.3**	**0.836**	**0.491**	

testing dataset

Can we move away from 2D?

- Hybrid of NeRF / GAN models
    - supervised by adversarial 2D losses 😢😭

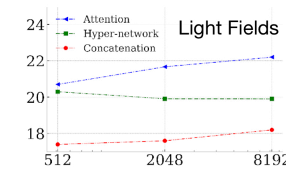
sweep: after/before upsampling



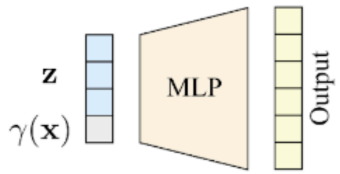
Chan et al. “EG3D: Efficient Geometry-aware 3D Generative Adversarial Networks” CVPR 2022

More powerful conditioning?

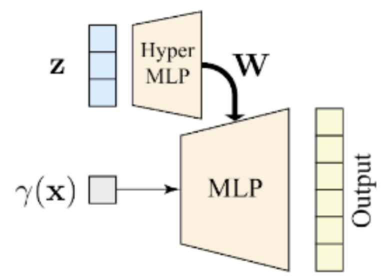
- Are MLPs architectures for conditional fields the way to go?



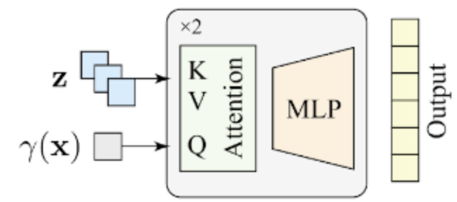
concatenate



hyper-network



attention



+2dB

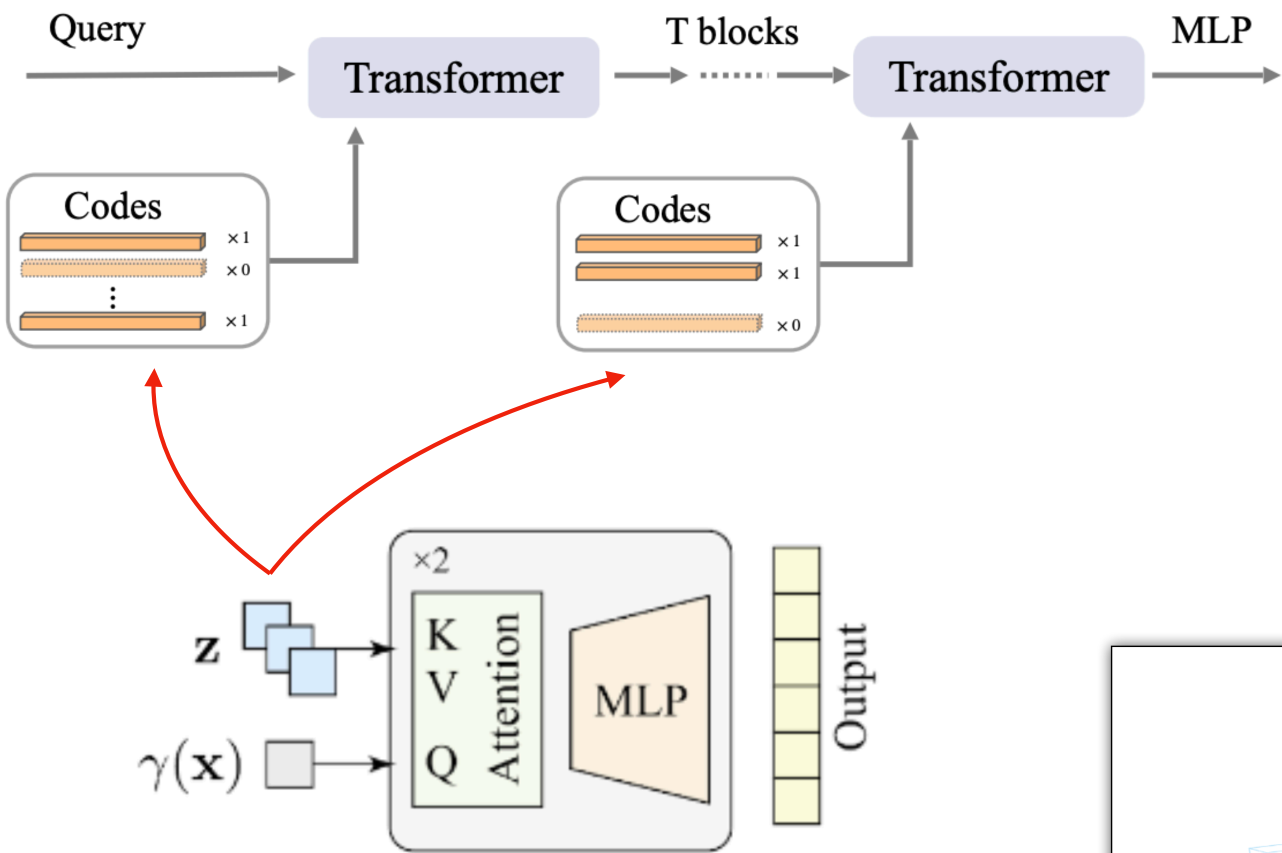
+2dB

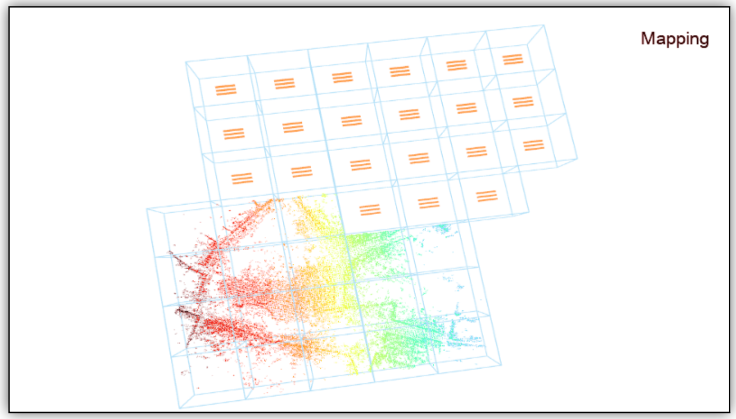
Rebain et al. «ABC: Attention Beats Concatenation for Conditioning Neural Fields» TMLR 2023

Hierarchical Neural Field

hierarchical transformer field decoder

transformer field decoder



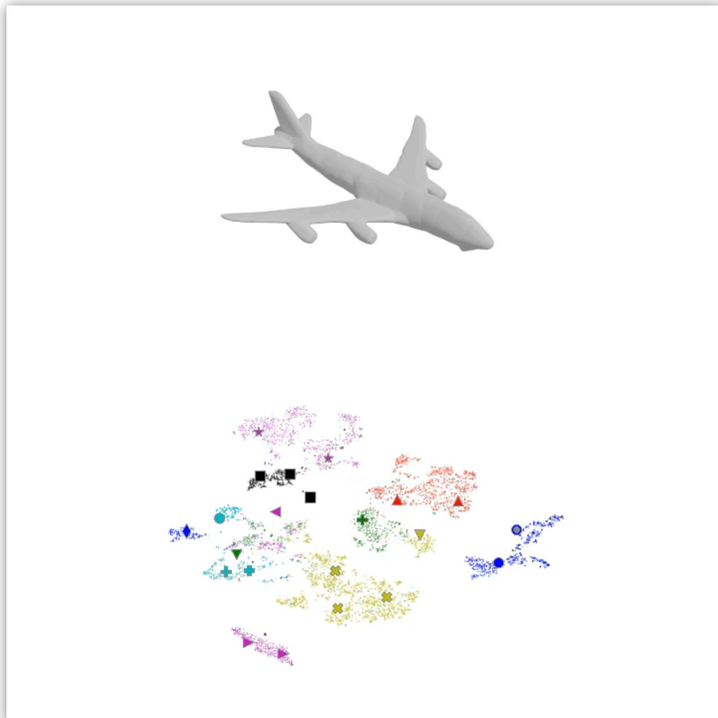


Tang et al. «NeuMap: Neural Coordinate Mapping by Auto-Transdecoder for Camera Localization» CVPR 2023

A historical perspective

$$\arg\max_{\mathcal{M}} \; p(\mathcal{D}|\mathcal{M}) \cdot p(\mathcal{M})$$

encoder-decoder

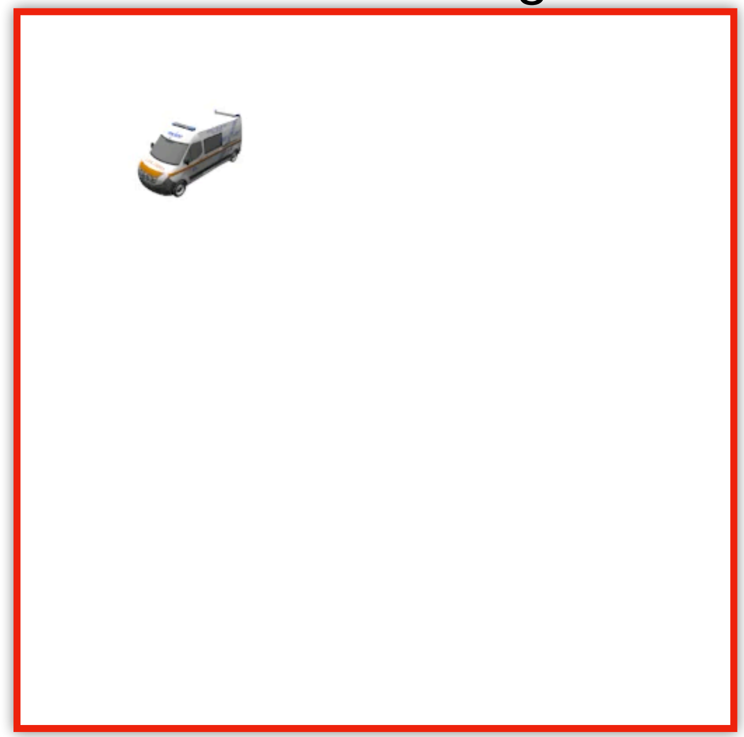


CvxNet @ CVPR'20

auto-decoder

LoLNeRF @ CVPR'22

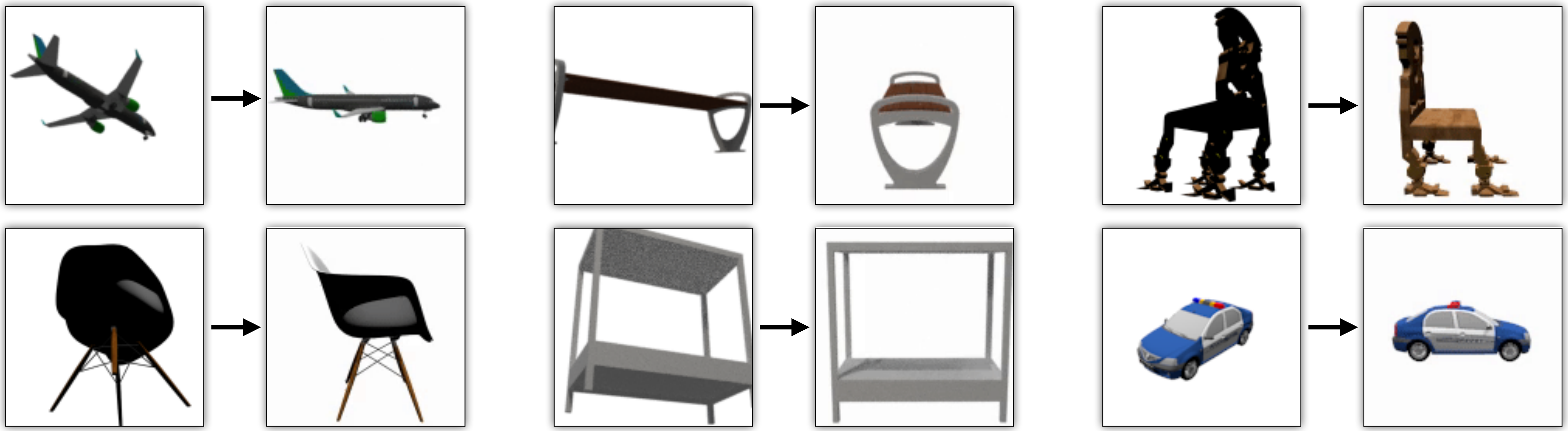
meta-learning



3DiM @ ICLR'23

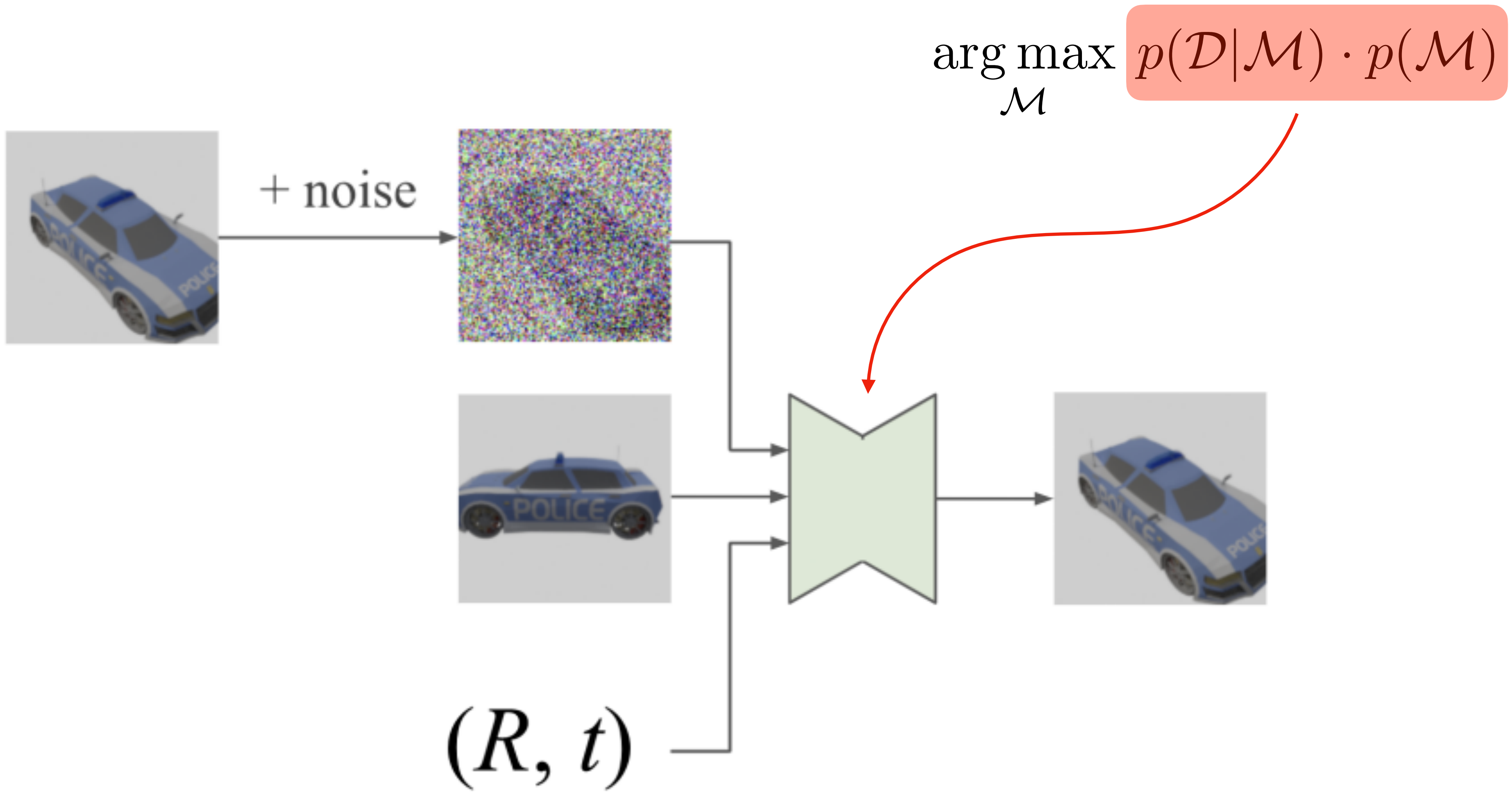
Problem Statement

- Input: single image
- Output: novel view given relative camera
- Objective: can diffusion models learn 3D?   …without NeRF!



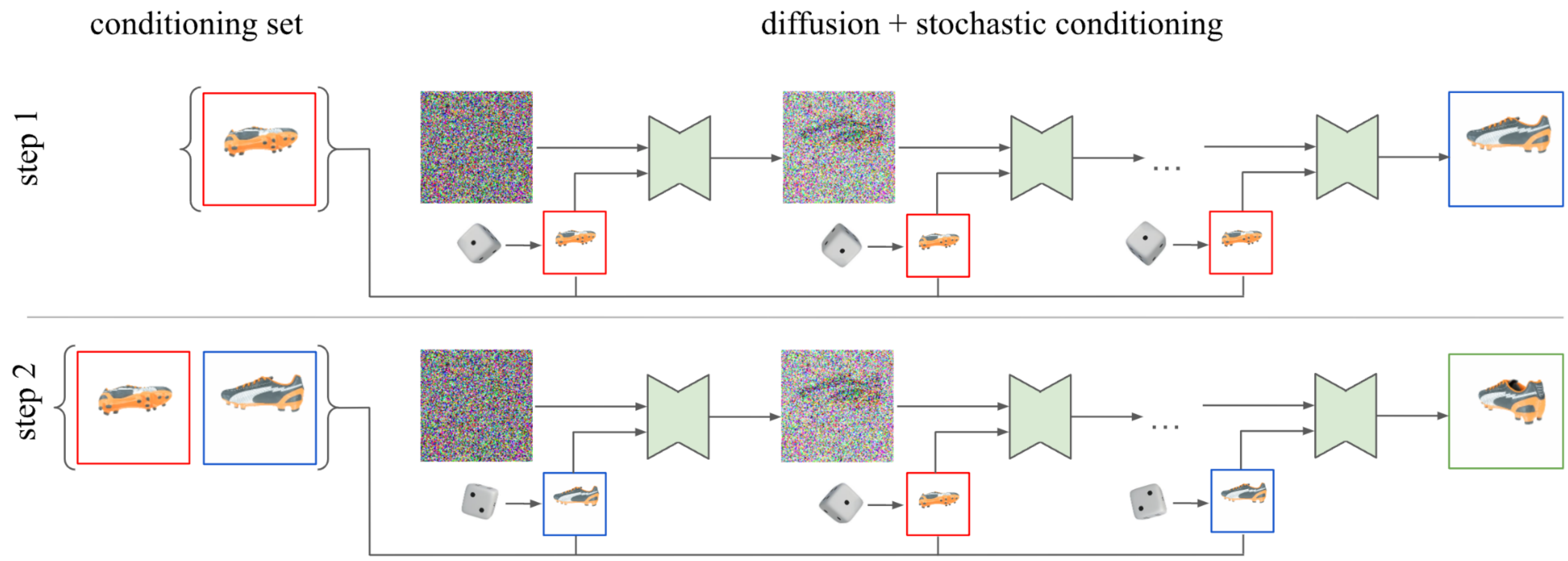
Watson et al. «Novel View Synthesis with Diffusion Models» ICLR 2023

Novel View Synthesis by Diffusion



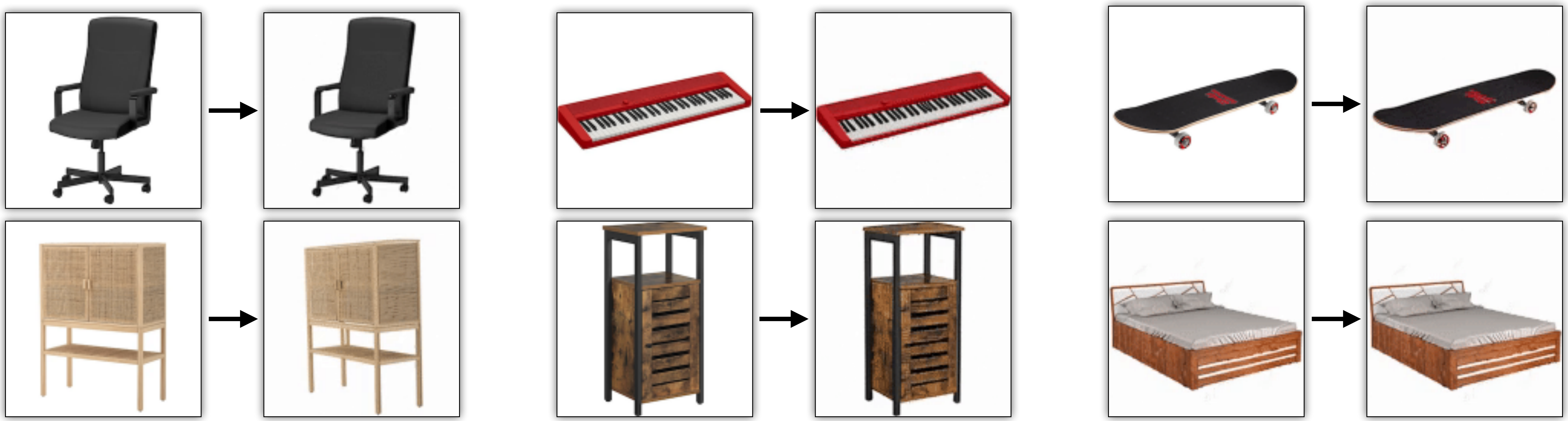
3DiM – Stochastic Conditioning

- Question: how to make sure (n+1)-th image consistent with {0, 1, …, n}?



In-the-wild generalization

- Trained on ShapeNet renderings (Kubric), tested on in-the-wild images
- Outcome: diffusion model “learnt 3D” from raw data



Greff et al. «Kubric: a scalable dataset generator» CVPR 2022

Imagen to 3D

"A mug with '3DiM' written on it, white background, no shadow."

"An upright piano, at an angle, white background, no shadow."

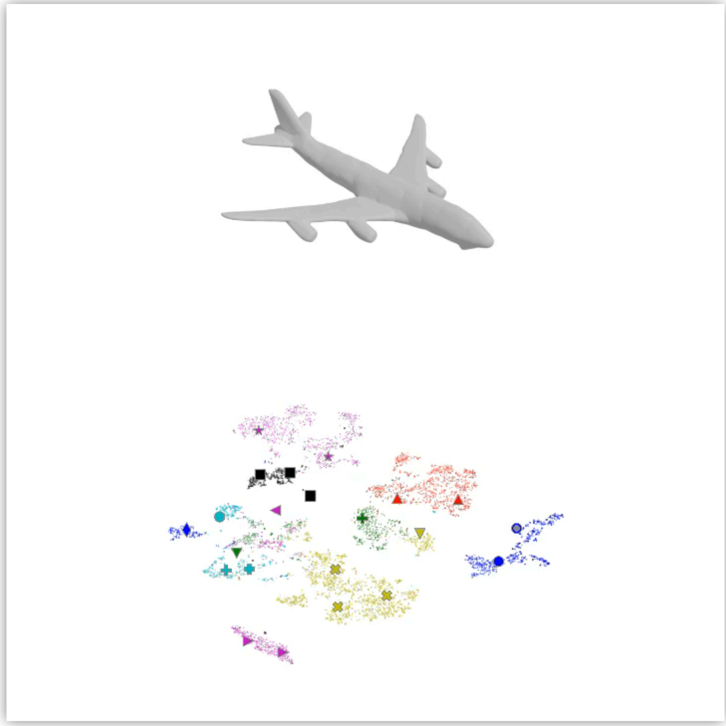
"A bed made of stone, white background, no shadows, at an angle."

Saharia et al. «Photorealistic text-to-image diffusion models with deep language understanding» NeurIPS 2022

Foundation Models

$$\underset{\mathcal{M}}{\arg\max}\ \underbrace{p(\mathcal{D}|\mathcal{M})}_{\text{NeRF}} \cdot \underbrace{p(\mathcal{M})}_{\text{???}}$$

encoder-decoder

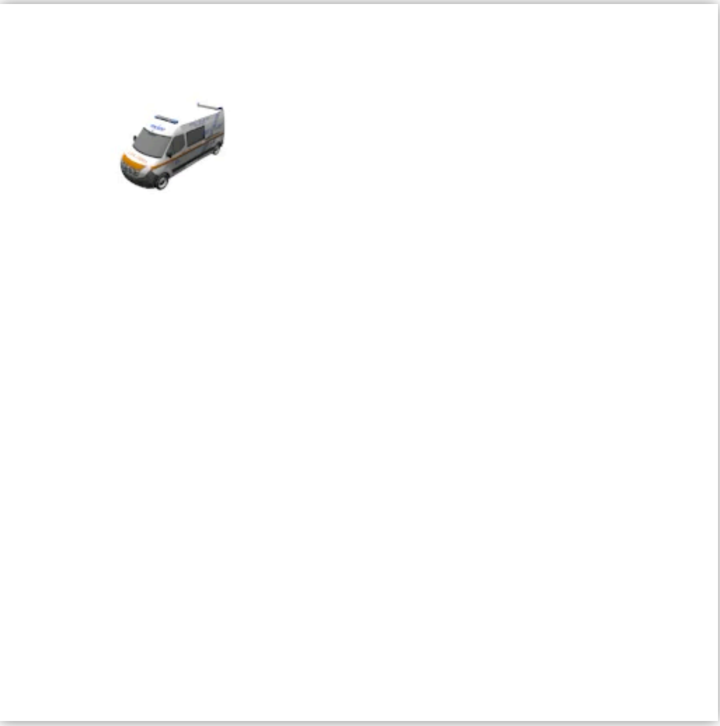


CvxNet @ CVPR'20

auto-decoder

LoLNeRF @ CVPR'22

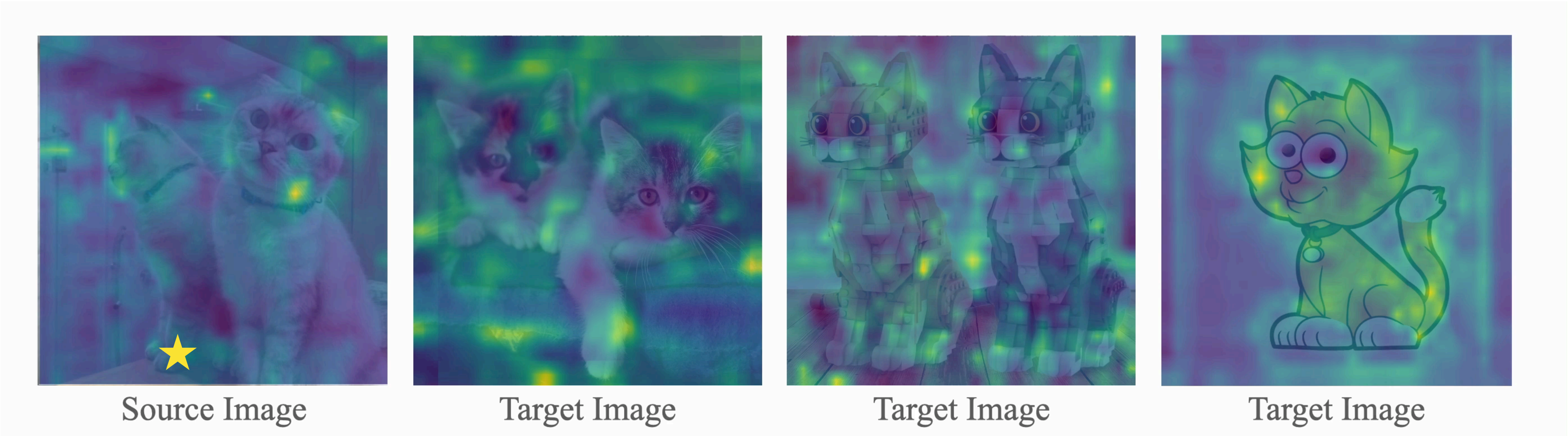
meta-learning



3DiM @ ICLR'23

Exciting new ideas

- Distill structural understanding from pre-trained image models
  - can these empower unsupervised 3D object discovery?



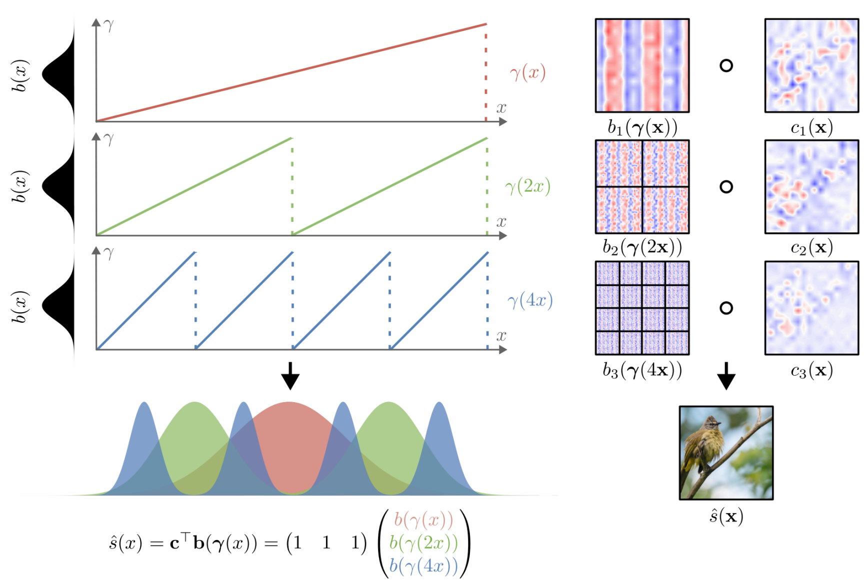
Hedlin et al. «Unsupervised Semantic Correspondence Using Stable Diffusion» preprint 2023

Exciting new ideas

- Write (unobserved) fields as weighted combinations of (local) functions

dictionary

$$\arg\max_{\mathcal{M}} \; p(\mathcal{D}|\mathcal{M}) \cdot \boxed{p(\mathcal{M})}$$

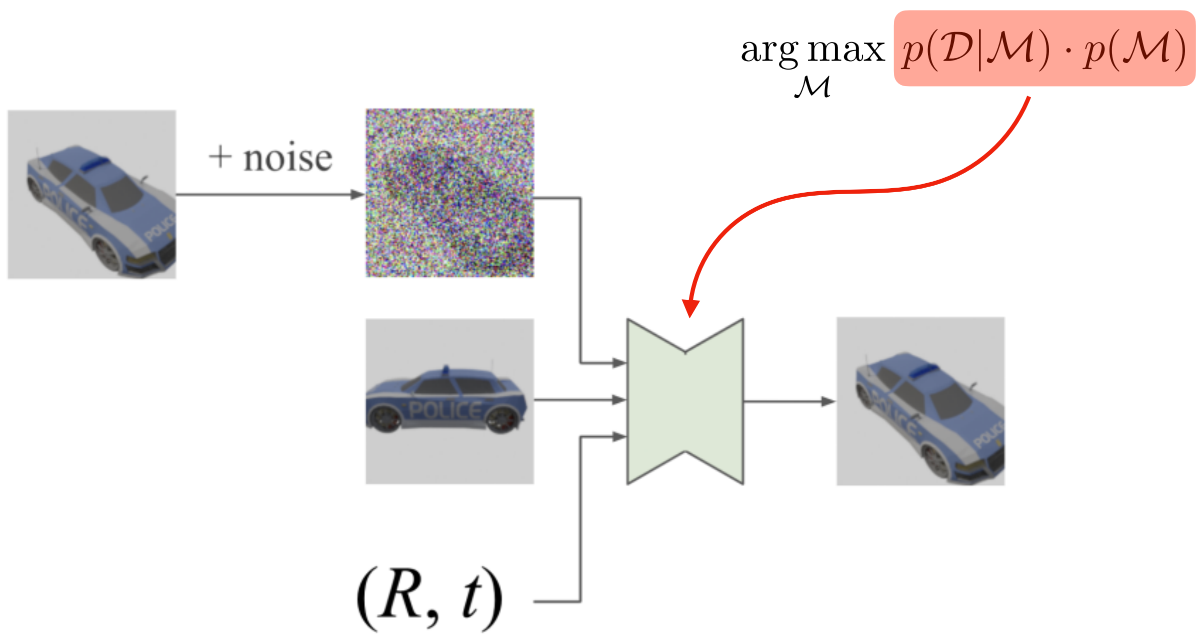


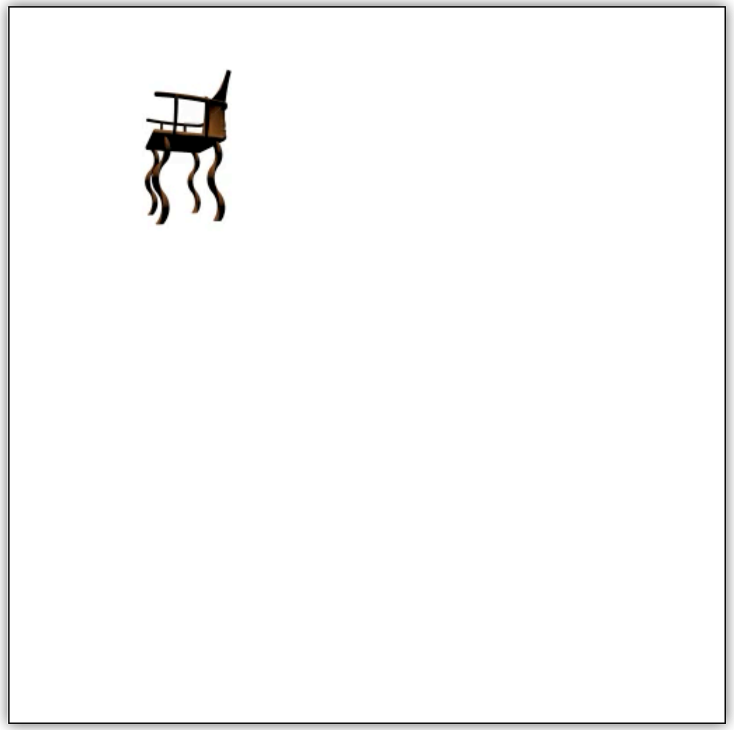
few shot generalization

Chen et al. «Dictionary Fields: Learning a Neural Basis Decomposition» SIGGRAPH 2023

Exciting new ideas

- …or perhaps learning from video at scale will "solve" 3D vision
  - is all we need large video datasets with calibrated cameras?





Neural fields for 3D Vision

Andrea Tagliasacchi (@taiyasaki)

Associate Professor – Simon Fraser University
Staff Research Scientist – Google Deepmind

Boyang Deng

Daniel Rebain

Daniel Watson

...

many others