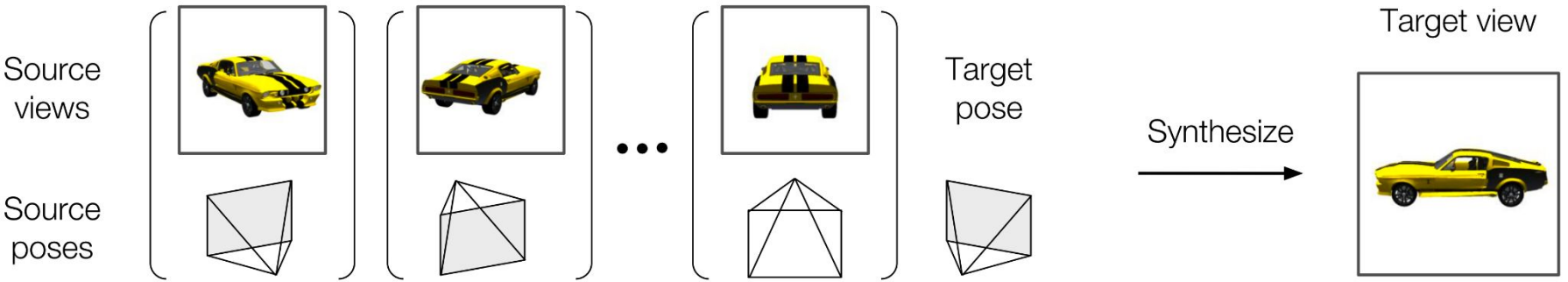
Novel View Synthesis with Diffusion Models

Nikita Morozov
Centre of Deep Learning and Bayesian Methods
HSE University

Novel view synthesis

Given a number of scene views and respective camera poses, reconstruct views from new poses.



Common approach

Take some parameterized scene representation and a differentiable rendering algorithm, then optimize using reconstruction loss.

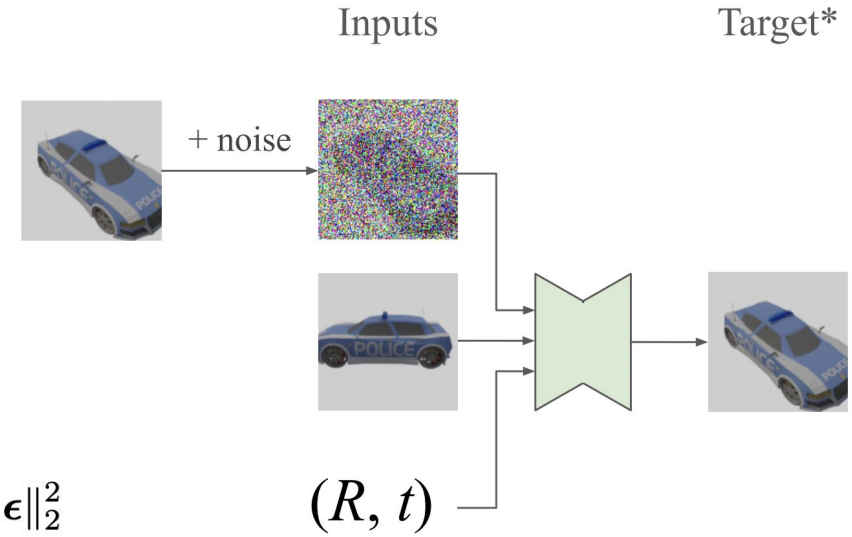
- Volumetric representations (NeRF)
- Signed distance functions
- Point-based approaches

Problems:

- One model is trained per scene
- Struggle when there are only few input views

Pose-conditional diffusion model

Directly generate a novel view conditioned on one clean input view and the respective camera poses.

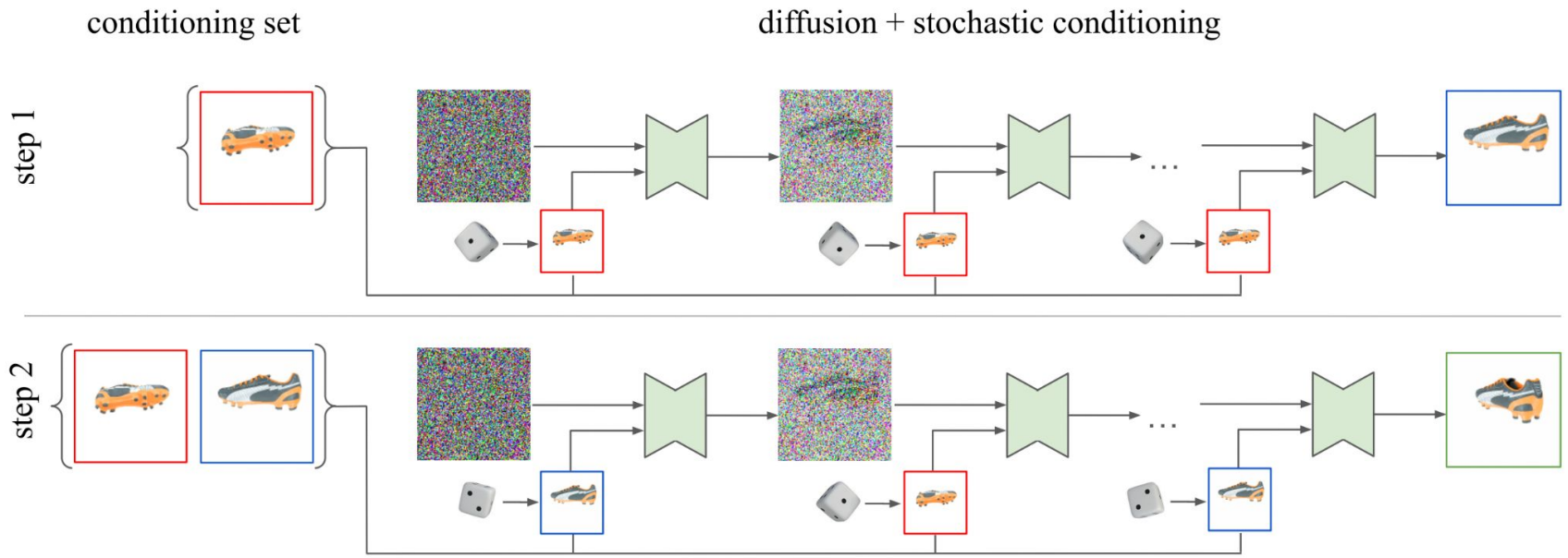


Training objective:

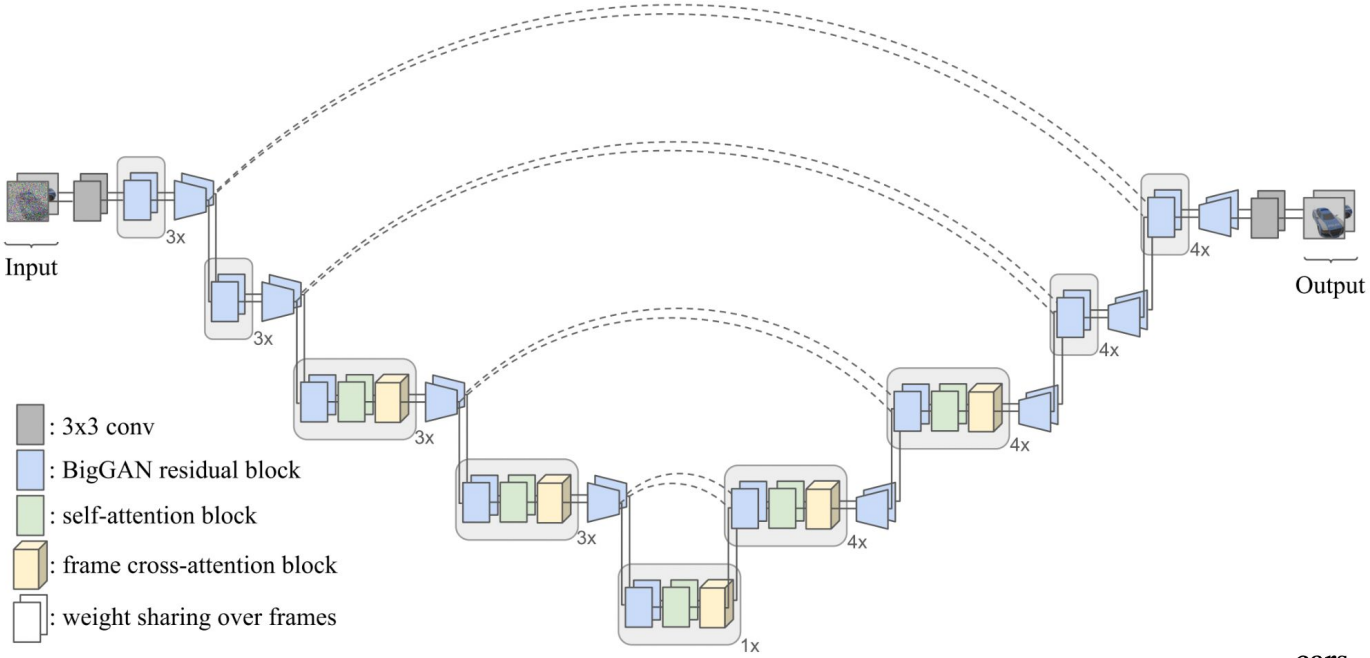
$$L(\theta) = \mathbb{E}_{q(\boldsymbol{x}_1, \boldsymbol{x}_2)} \; \mathbb{E}_{\lambda, \boldsymbol{\epsilon}} \; \|\boldsymbol{\epsilon}_\theta(\boldsymbol{z}_2^{(\lambda)}, \boldsymbol{x}_1, \lambda, \boldsymbol{p}_1, \boldsymbol{p}_2) - \boldsymbol{\epsilon}\|_2^2$$

Inference stage: stochastic conditioning

At each denoising step, condition on a random view from the set of available views.

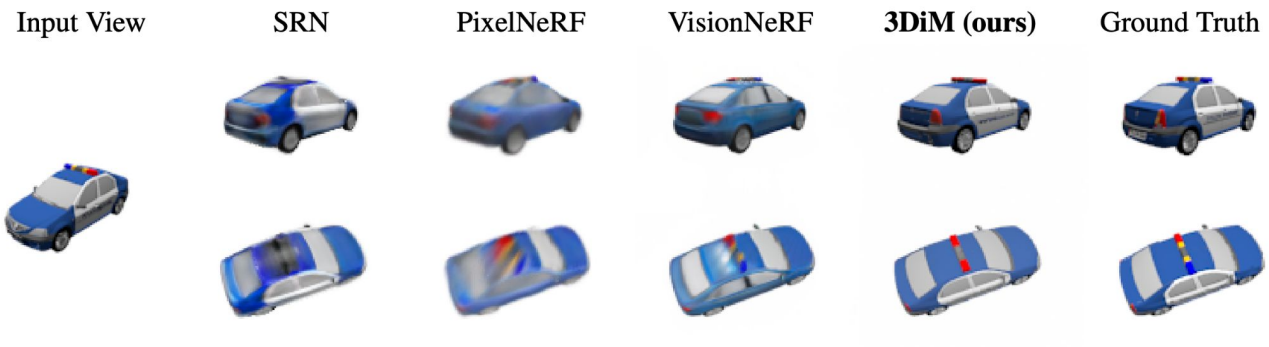


X-UNet architecture



	cars		
	PSNR (↑)	SSIM (↑)	FID (↓)
Concat-UNet	17.21	0.52	21.54
X-UNet	**21.01**	**0.57**	**8.99**

Comparisons



	SRN cars			SRN chairs		
	PSNR (↑)	SSIM (↑)	FID (↓)	PSNR (↑)	SSIM (↑)	FID (↓)
Geometry-aware						
SRN	22.25	0.88	41.21	22.89	0.89	26.51
PixelNeRF	23.17	0.89	59.24	23.72	0.90	38.49
VisionNeRF	22.88	0.90	21.31	**24.48**	**0.92**	10.05
CodeNeRF	**23.80**	***0.91**	–	23.66	*0.90	–
Geometry-free						
LFN	22.42	*0.89	–	22.26	*0.90	–
ENR	22.26	–	–	22.83	–	–
3DiM (ours)	21.01	0.57	**8.99**	17.05	0.53	**6.57**

3D consistency metric

Generate a number of views using the model, then train and evaluate NeRF on the generated views. The less consistent the views are, the worse will NeRF perform.

Training view source	SRN cars			SRN chairs*		
	PSNR (↑)	SSIM (↑)	FID (↓)	PSNR (↑)	SSIM (↑)	FID (↓)
Original data (3D consistent)	28.21	0.96	10.57	24.87	0.93	17.05
3DiM (∼1.3B params)	28.48	**0.96**	29.55	**22.90**	**0.86**	**58.61**
3DiM (∼471M params)	**28.53**	**0.96**	**22.09**	18.84	0.79	98.78
+ no stochastic conditioning	25.78	0.94	30.51	17.61	0.75	116.16