SINGLE-VIEW SHAPE RECONSTRUCTION VIA IMAGE-CONDITIONED 3D DIFFUSION

Cristian Sbrolli

Advisor: Prof. Matteo Matteucci

Co-Advisor: Paolo Cudrano, Matteo Frosi

POLITECNICO MILANO 1863

HONOURS PROGRAMME HP-SR in Information Technology

AIRLAB ARTIFICIAL INTELLIGENCE AND ROBOTICS LAB

OUTLINE

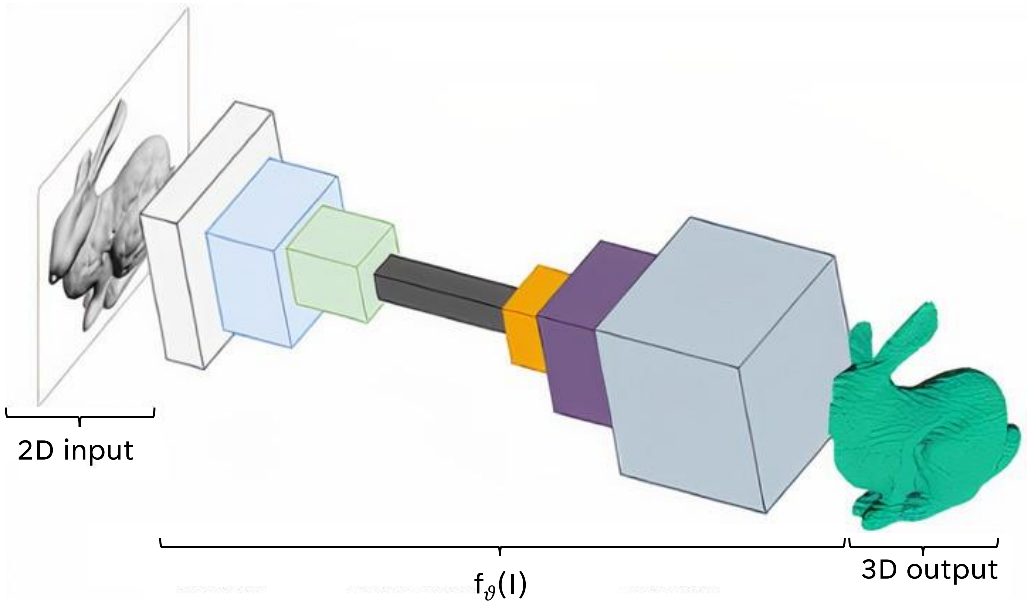
- Introduction
- Problem Statement/Our Idea
- Denoising Diffusion Probabilistic Models
- CISP
- IC3D

WHAT IS SINGLE-VIEW 3D RECONSTRUCTION

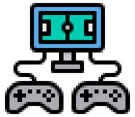
Input
Image of an object I

Predictor $f_{\vartheta}(I)$

Output
Predicted 3D Shape of the represented object $\overline{S}$

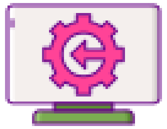


WHY IS IT IMPORTANT?

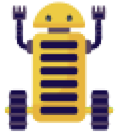


Videogames

Medical Imaging

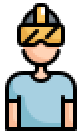


Reverse Engineering



Robotic Mapping

Cultural Heritage



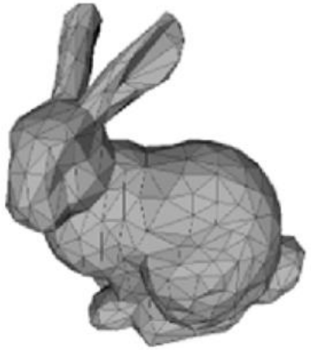
VR & Metaverse

HOW TO REPRESENT 3D SHAPES

- ✓ Relatively easy to collect
- ✓ Exact representation
- ✗ Often not directly used
- ✗ Do not model connectivity

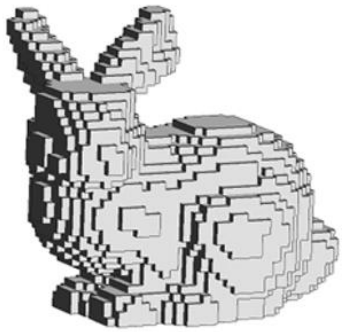
POINT CLOUDS

- ✓ Easy to render and transform
- ✓ Computers optimized for it
- ✗ Curved objects approximated
- ✗ Don’t hold up in all resolutions



SURFACE MESHES

- ✓ Direct pixel extension
- ✓ Can have high resolutions
- ✗ Memory consumption
- ✗ Manhattan world bias

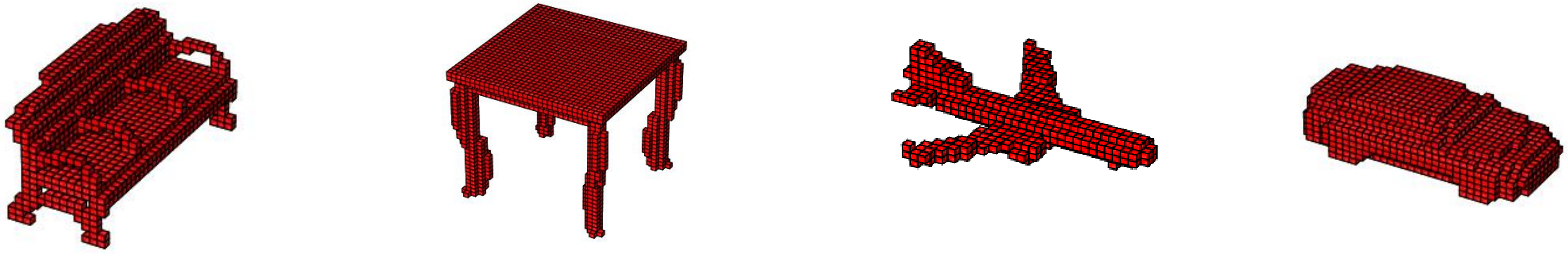


VOXELS

DATASET

ShapeNet subset:

- 13 categories of voxelized objects and corresponding renderings
- 44k models
- 32^3 resolution

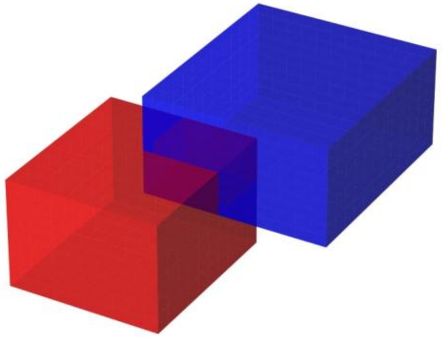


Following the literature for 3D diffusion models, we use mainly the aeroplane, car, chair categories.

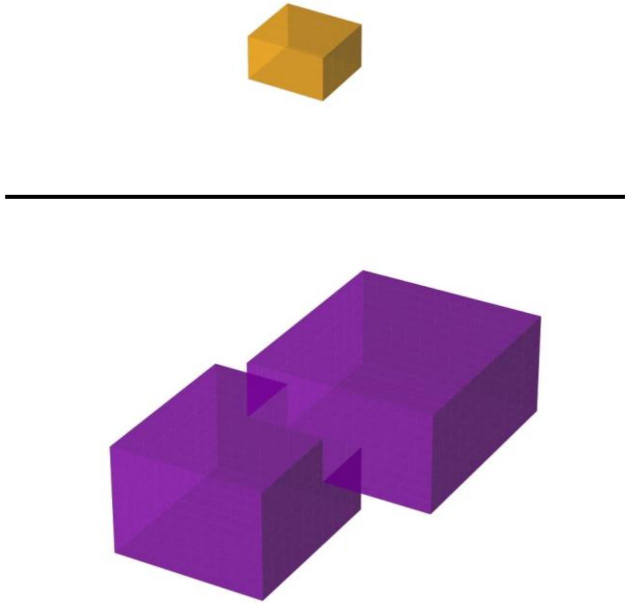
INTERSECTION OVER UNION

Intersection over Union (IoU): $IoU(A,B) = \frac{|A \cap B|}{|A \cup B|}$

shapes A, B



IoU:

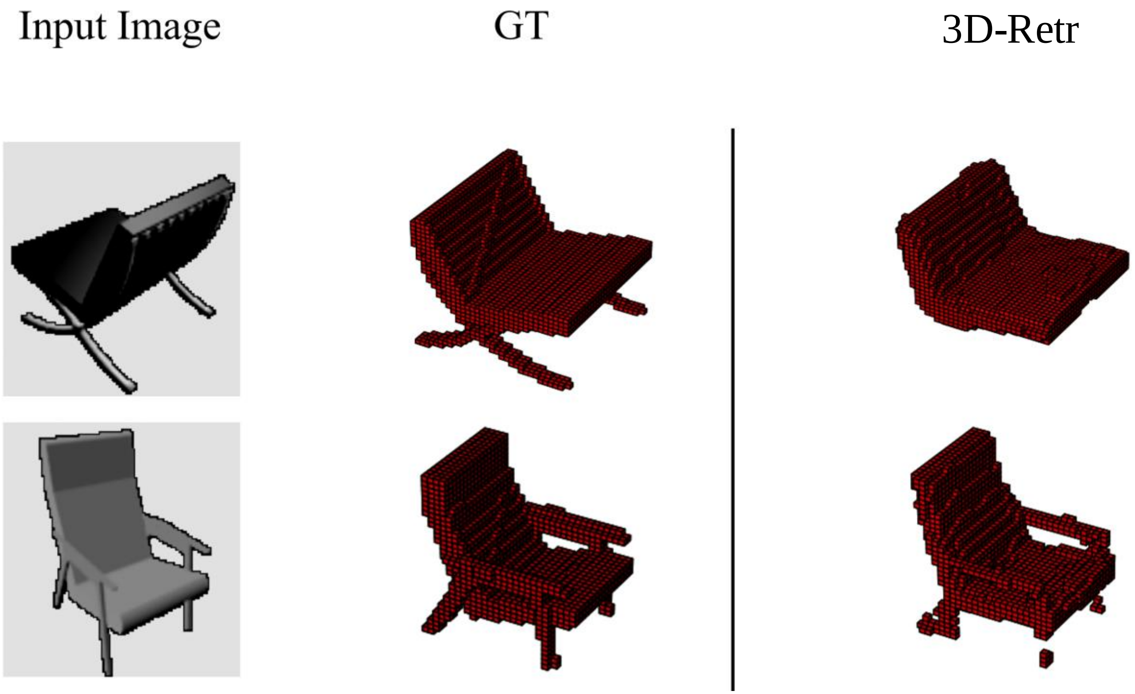


OUTLINE

- Introduction
- Problem Statement/Our Idea
- Denoising Diffusion Probabilistic Models
- CISP
- IC3D

ISSUES OF 3D RECONSTRUCTION NETWORKS

- ✅ SoTA 3D reconstruction models reach **impressive scores**.
- ❌ **Realism, integrity and structural correctness** are not considered.



- ⚠️ Scores are **not heavily impacted**
- ❌ **Unusable** in many applications!

OUR OBJECTIVE

Some applications → **may** not need exact reconstructions
Some applications → **require** realistic and structurally correct objects.

Generative approaches learn the structural semantics of the training data.

Generative models may solve the presented issues.

OUR AIM

Develop a **3D image-driven generative model** able to both capture **realism aspects** while **respecting the features of the object in the image**

OUTLINE

DENOISING DIFFUSION PROBABILISTIC MODELS

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I})$$

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \underbrace{\boxed{\mu_\theta(\mathbf{x}_t, t)}}_{\text{NN}}, \sigma_t^2\mathbf{I})$$

Forward process: add noise at each step

Backward process: denoise until step 0

Training

Model learns the backward process by predicting the noise added w.r.t. prev step.

Generation

Start from random noise.

Denoise for n steps → sample from t=0.

GUIDANCE

DDPMs can be conditioned to generate samples respecting some **additional information *y***:

$$\mu(x_t, t) \rightarrow \mu(x_t, t \mid y)$$

The conditioning token y is an embedding vector/tensor that can represent:

- A class/category
- Information from a different domain

Approaches

classifier guidance

classifier-free guidance

2D DIFFUSION

Diffusion models obtained impressive results in image generation.
In particular, **text-driven image generation** models as:

OpenAI DALL·E 2

GOOGLE IMAGEN AI

stability.ai stable diffusion

"A robot couple fine dining with Eiffel Tower in the background."

3D DIFFUSION

What about 3D diffusion in the literature?

REPRESENTATION

Limited to point clouds

Voxels

CONDITIONING

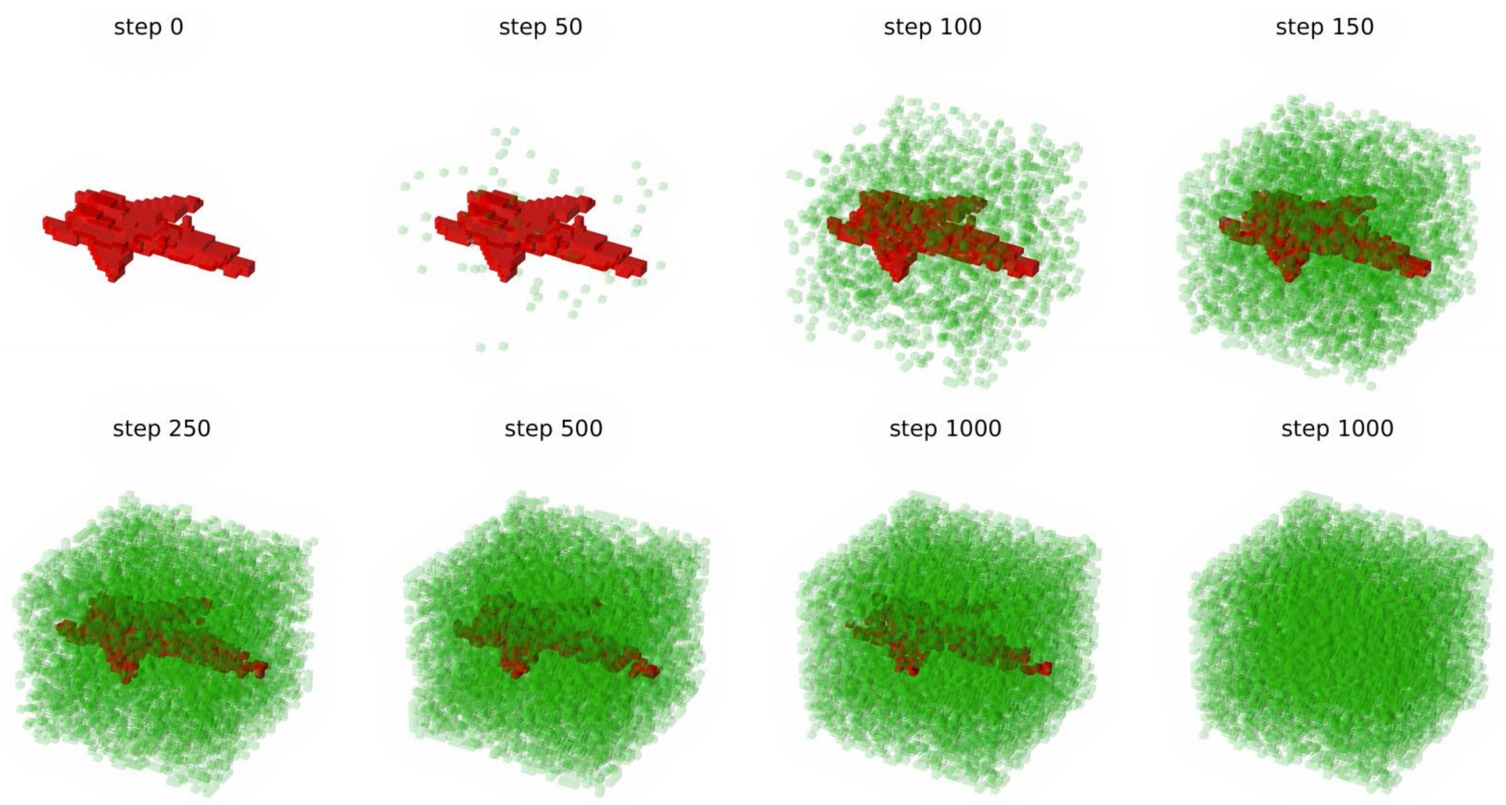
unguided
class-guided

Image-guided

RESULTS

SoTA in shape generation

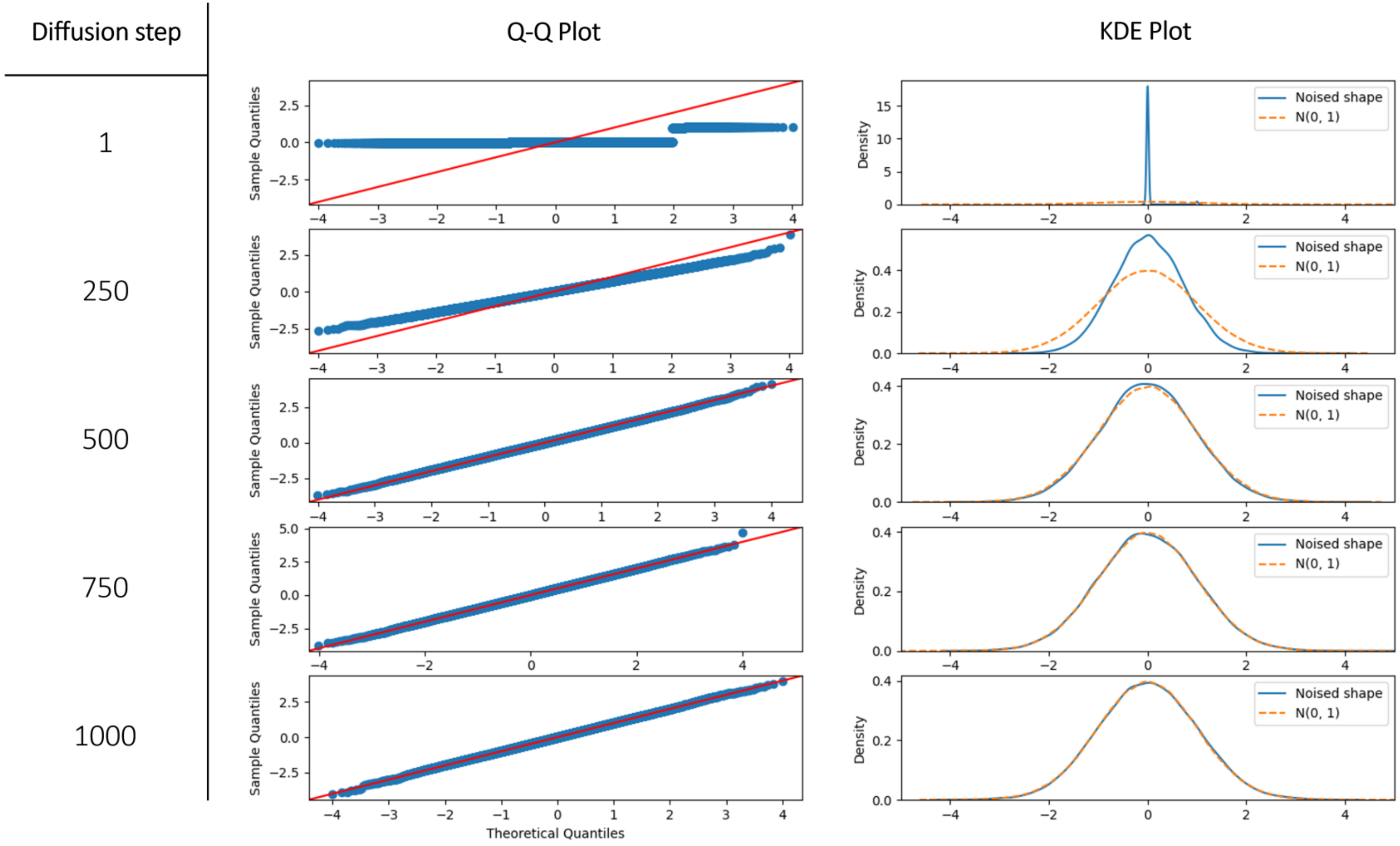
ANALYZING VOXEL DIFFUSION



Visualization of voxel diffusion, data is thresholded at 0.5.

The last step is shown both with original shape highlighted and without.

ANALYZING VOXEL DIFFUSION



The distribution of data is progressively transformed into a Standard Gaussian distribution.

3D DIFFUSION

What about 3D diffusion in the literature?

REPRESENTATION

Limited to point clouds

Voxels

CONDITIONING

unguided
class-guided
shape-latents

Image-guided

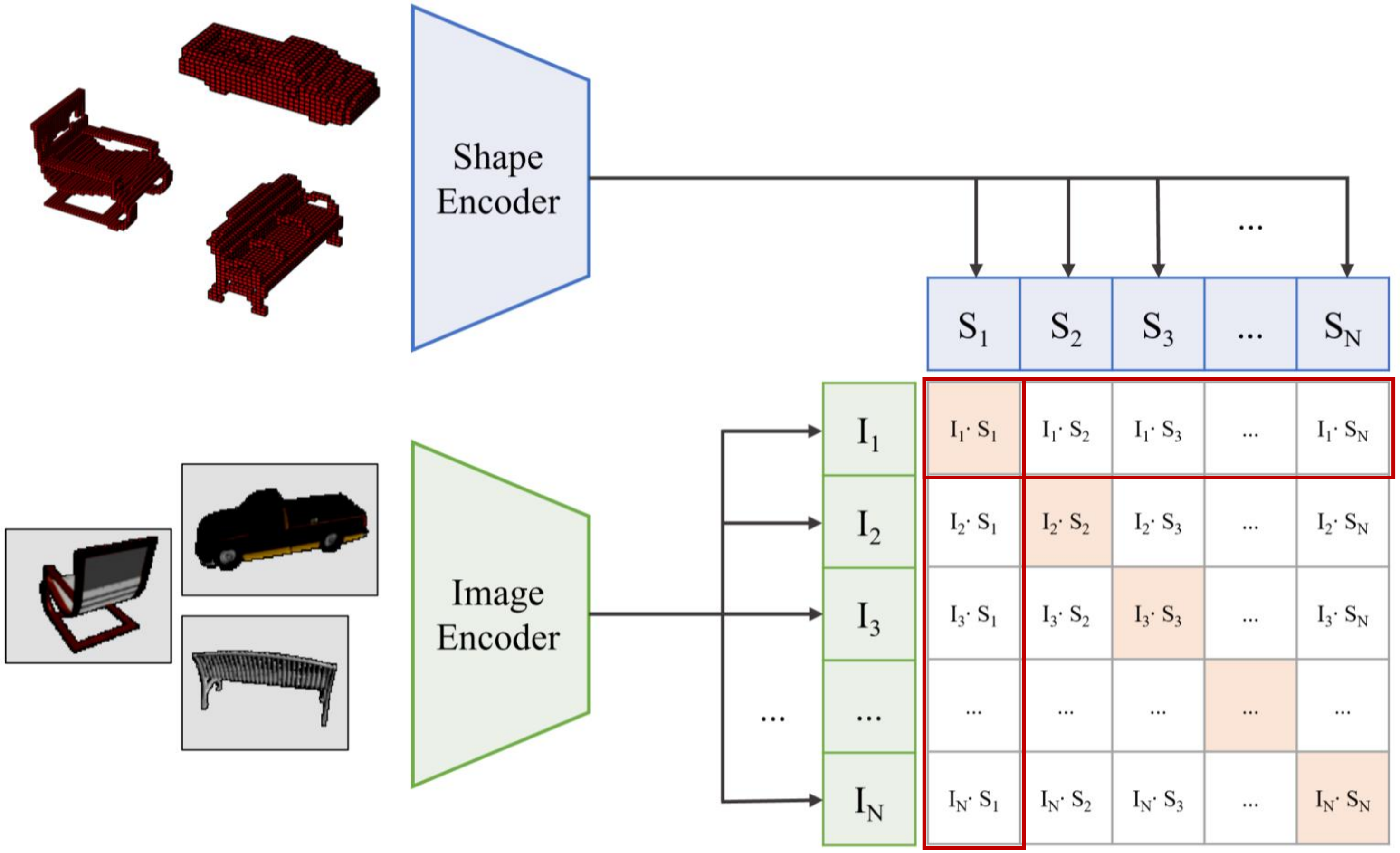
RESULTS

SoTA in shape generation

OUTLINE

- Introduction
- Problem Statement/Our Idea
- Denoising Diffusion Probabilistic Models
- CISP
- IC3D

CISP:
CONTRASTIVE IMAGE-SHAPE PRETRAINING



Build a joint image-shape space by learning to associate shapes and images

Training

batches of (image,shape) pairs

cosine similarity matrix

Cross Entropy over rows and columns

CISP CONFIGURATIONS

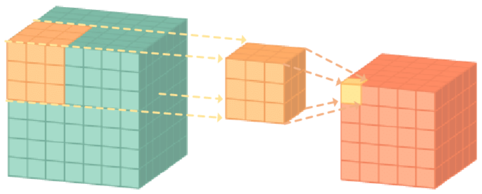
Image Encoder

Vision Transformer

Shape Encoder

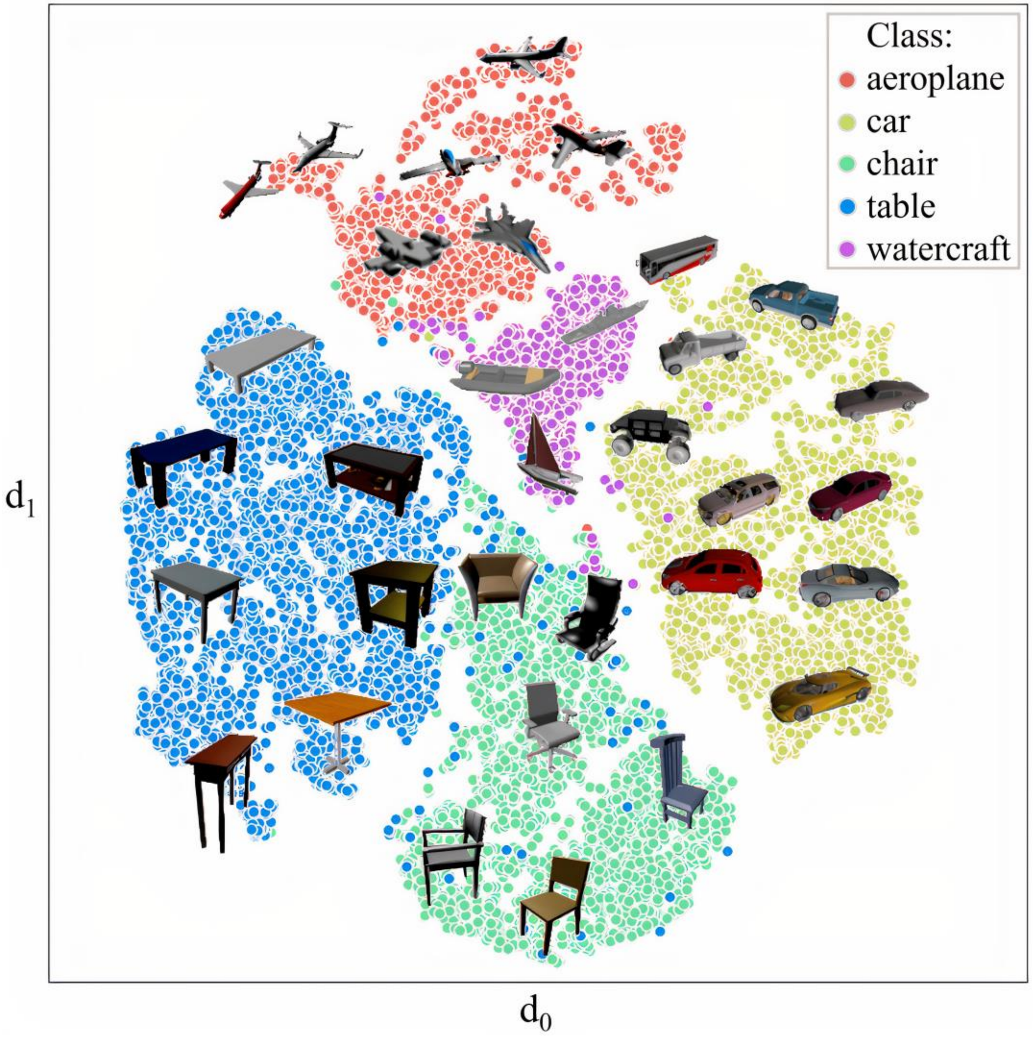
2 configuration tested

CNN



Transformer:

CISP EMBEDDING SPACE ANALYSIS

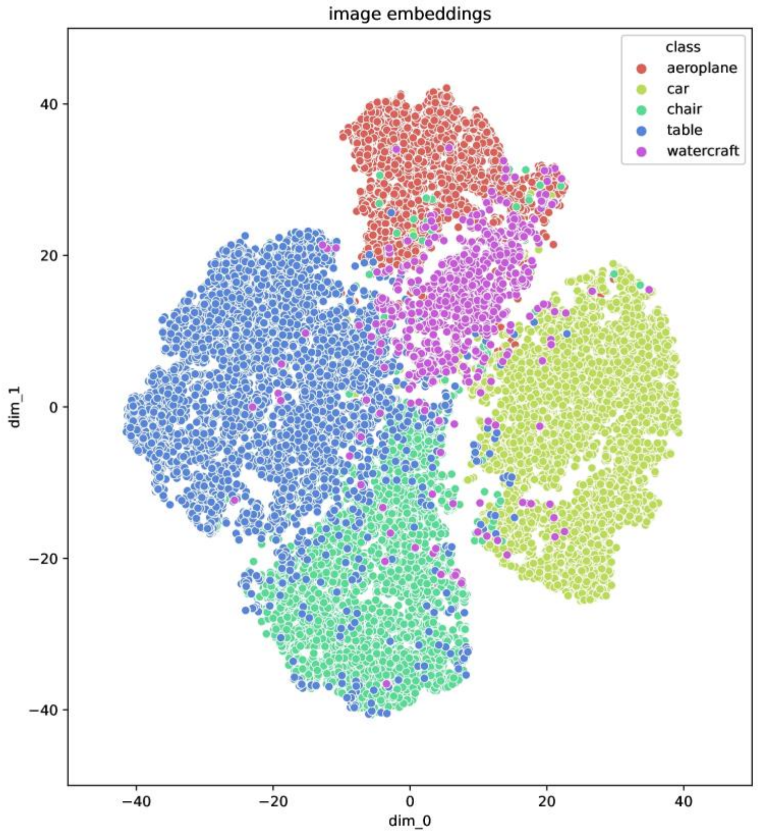


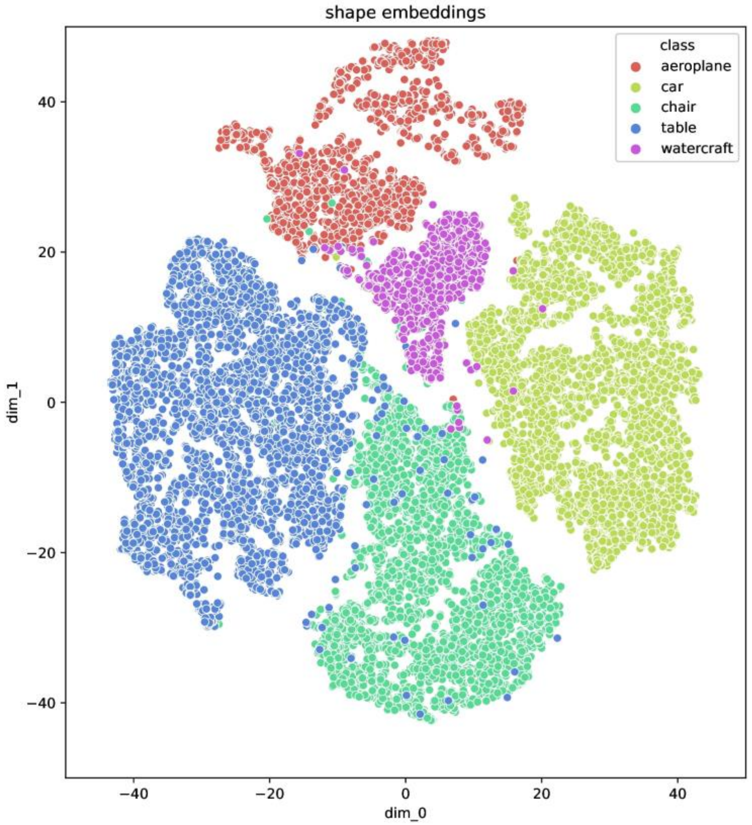
The model captures details and subcategories.

For example:

- Airplanes: two main clouds, combat and line airplanes.
- Tables: higher tables with lower d_1. Shelves are added with higher d_0.
- Cars: bigger cars increasing d_1, with sports cars and trucks/buses on the extremes.

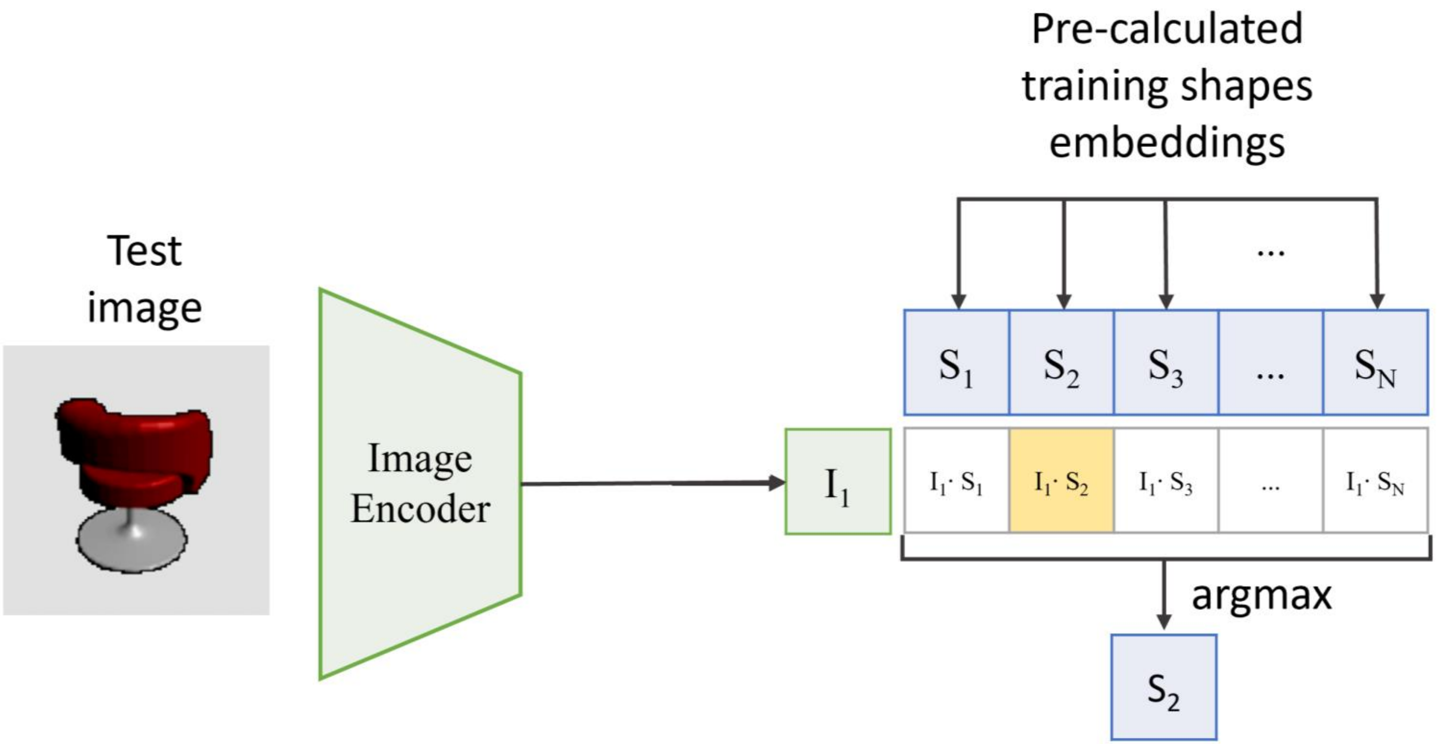
VISUALIZATION OF CISP EMBEDDING SPACE





*The space shown here is from the best transformer configuration

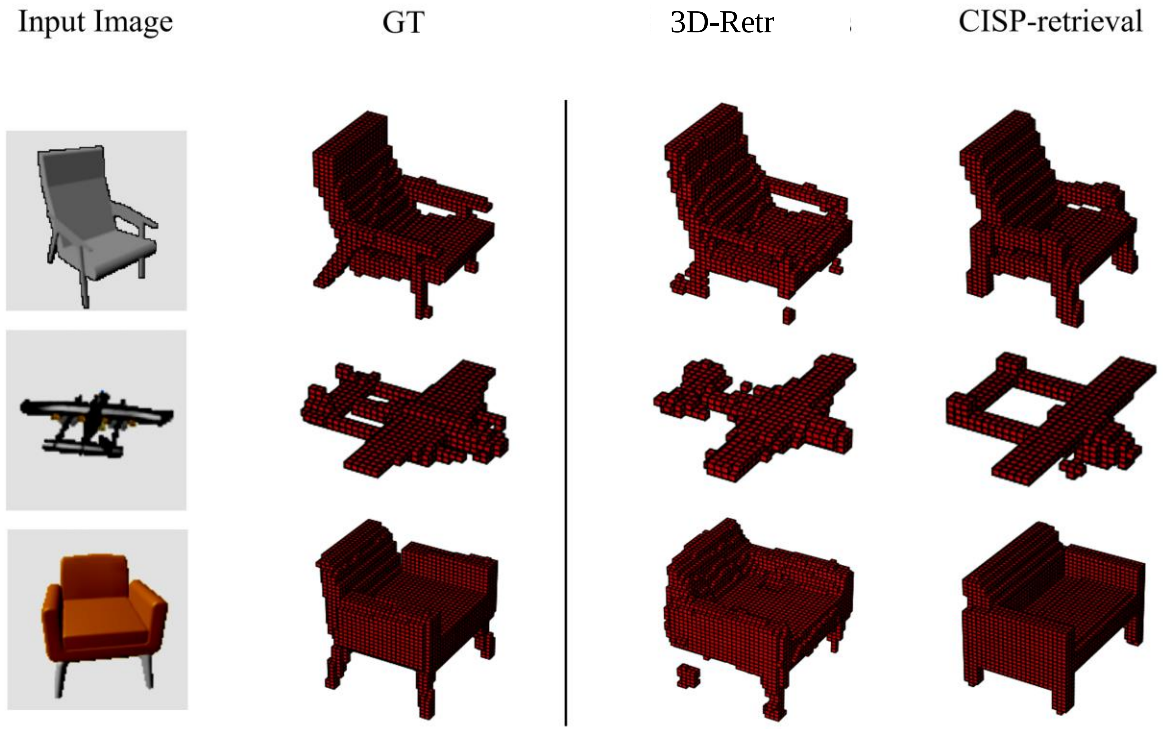
CISP APPLICATIONS: RECONSTRUCTION BY RETRIEVAL METHOD



Given a database of shapes:

1. Project test image
2. Calculate similarity w.r.t. each database shape
3. Find argmax
4. Return the corresponding shape

CISP AS A ZERO-SHOT MODEL: RECONSTRUCTION BY RETRIEVAL RESULTS



	CISP-retrieval
Aeroplane	0.645
Car	0.76
Chair	0.412
Table	0.436
Watercraft	0.458
Overall	0,542

test set IoU

Shape selected from dataset ›› Realistic and structurally correct

Good results in terms of coherence to the image.

3D DIFFUSION

What about 3D diffusion in the literature?

REPRESENTATION

Limited to point clouds

Voxels

CONDITIONING

unguided
class-guided
shape-latents

Image-guided

RESULTS

SoTA in shape generation

AGENDA

Introduction

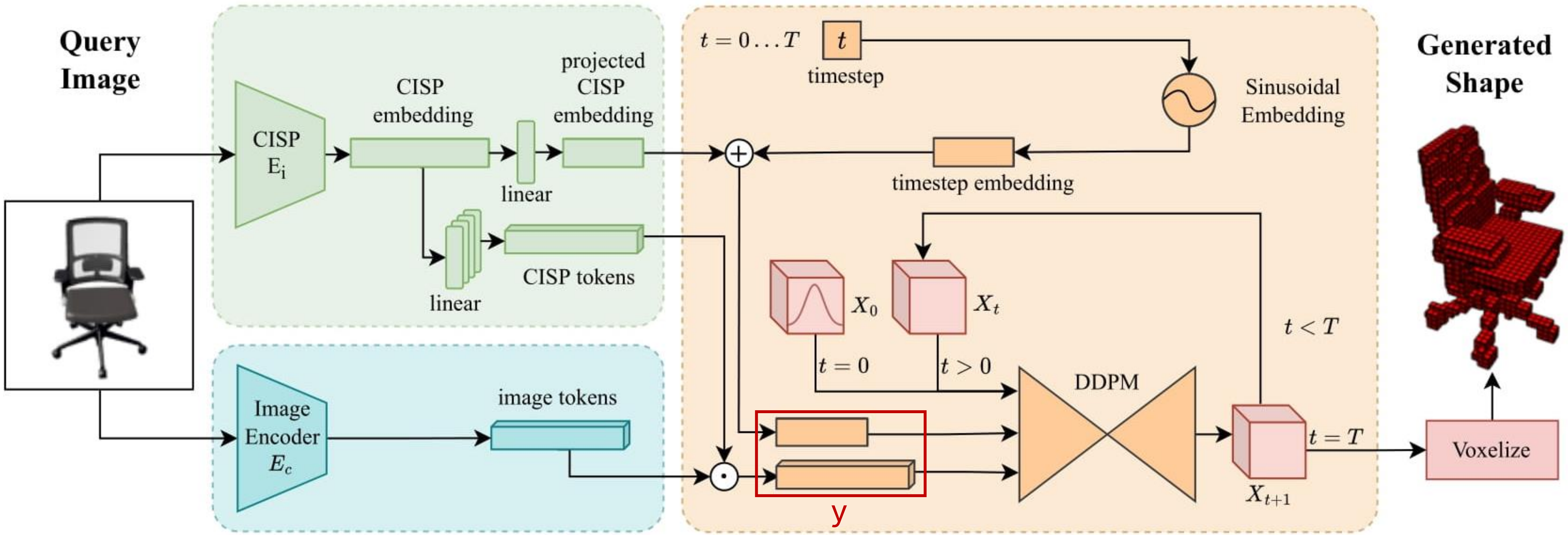
Problem Statement/Our Idea

Denoising Diffusion Probabilistic Models

CISP

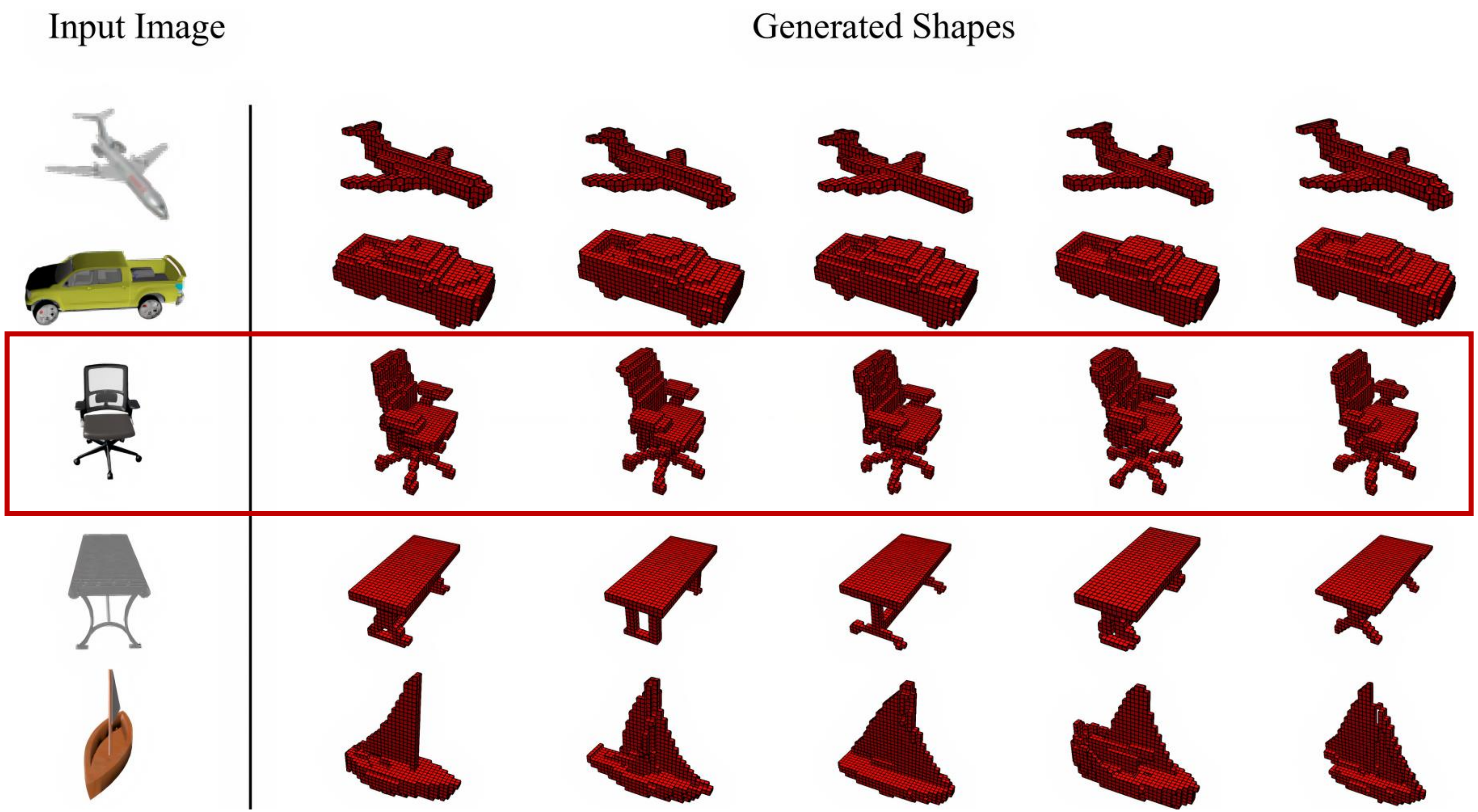
- IC3D

IC3D PIPELINE



Pipeline of our image-driven 3D diffusion model

QUALITATIVE RESULTS



QUANTITATIVE RESULTS

Shape	Model	1-NNA(%) CD	1-NNA(%) EMD
Airplane	PointFlow	75.68	70.74
	SoftFlow	76.05	65.80
	DPF-Net	75.18	65.55
	Shape-GF	80.00	76.17
	luo et al.	62.71	67.14
	PVD	73.82	64.81
	Ours	**57.64**	**53.89**
Car	PointFlow	58.10	56.25
	SoftFlow	64.77	60.09
	DPF-Net	62.35	54.48
	Shape-GF	63.20	56.53
	luo et al.	-	-
	PVD	54.55	53.83
	Ours	**52.44**	**51.68**
Chair	PointFlow	62.84	60.57
	SoftFlow	59.21	60.05
	DPF-Net	62.00	58.53
	Shape-GF	68.96	65.48
	luo et al.	62.08	64.45
	PVD	56.26	53.32
	Ours	**53.58**	**51.73**

1-NNA measures the accuracy of a 1-NN classifier in distinguish real and generated samples.

Optimal score is 50%

1-NNA measures both quality and diversity.

SINGLE VIEW 3D RECONSTRUCTION RESULTS

As the model is **probabilistic**, we display the maximum scores obtained when sampling an **increasing amount of shapes**.

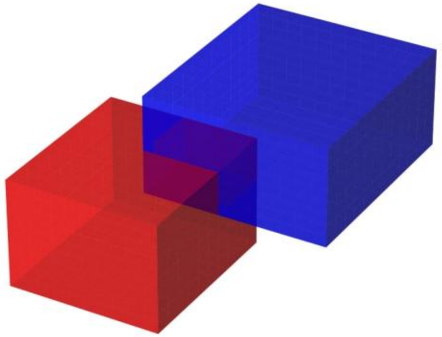
	Baselines					SoTA models.					
	3D-R2N2	OGN	Pixel2Mesh	AttSets	Pix2Vox++/F	3D-Retr	TMV-Net	Ours(1)	Ours(5)	Ours(10)	Ours(15)
aeroplane	0,512	0,587	0,508	0,594	0,607	0,704	0,691	0,540	0,600	0,620	0,630
car	0,798	0,828	0,67	0,844	0,841	0,861	0,87	0,790	0,8237	0,8328	0,838
chair	0,466	0,483	0,484	0,559	0,548	0,592	0,721	0,407	0,476	0,494	0,506
overall	0,592	0,633	0,554	0,666	0,665	0,719	0,761	0,579	0,633	0,649	0,658

As expected, increasing the number of samples, the maximum IoU score increases.

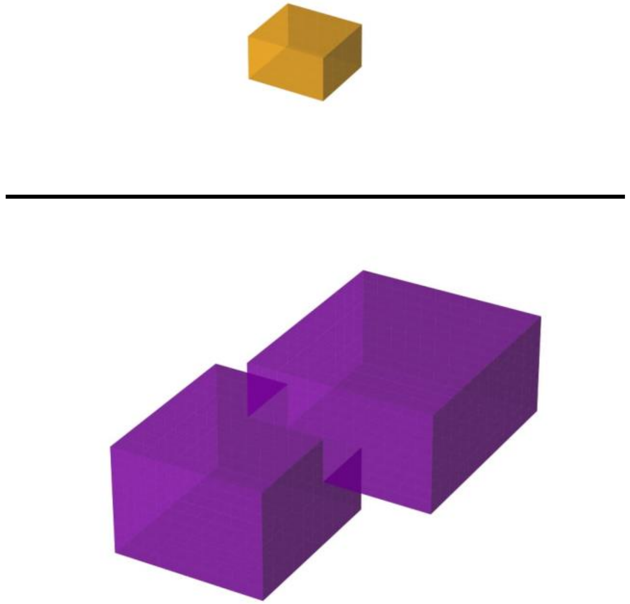
INTERSECTION OVER UNION

Intersection over Union (IoU): $IoU(A,B) = \frac{|A \cap B|}{|A \cup B|}$

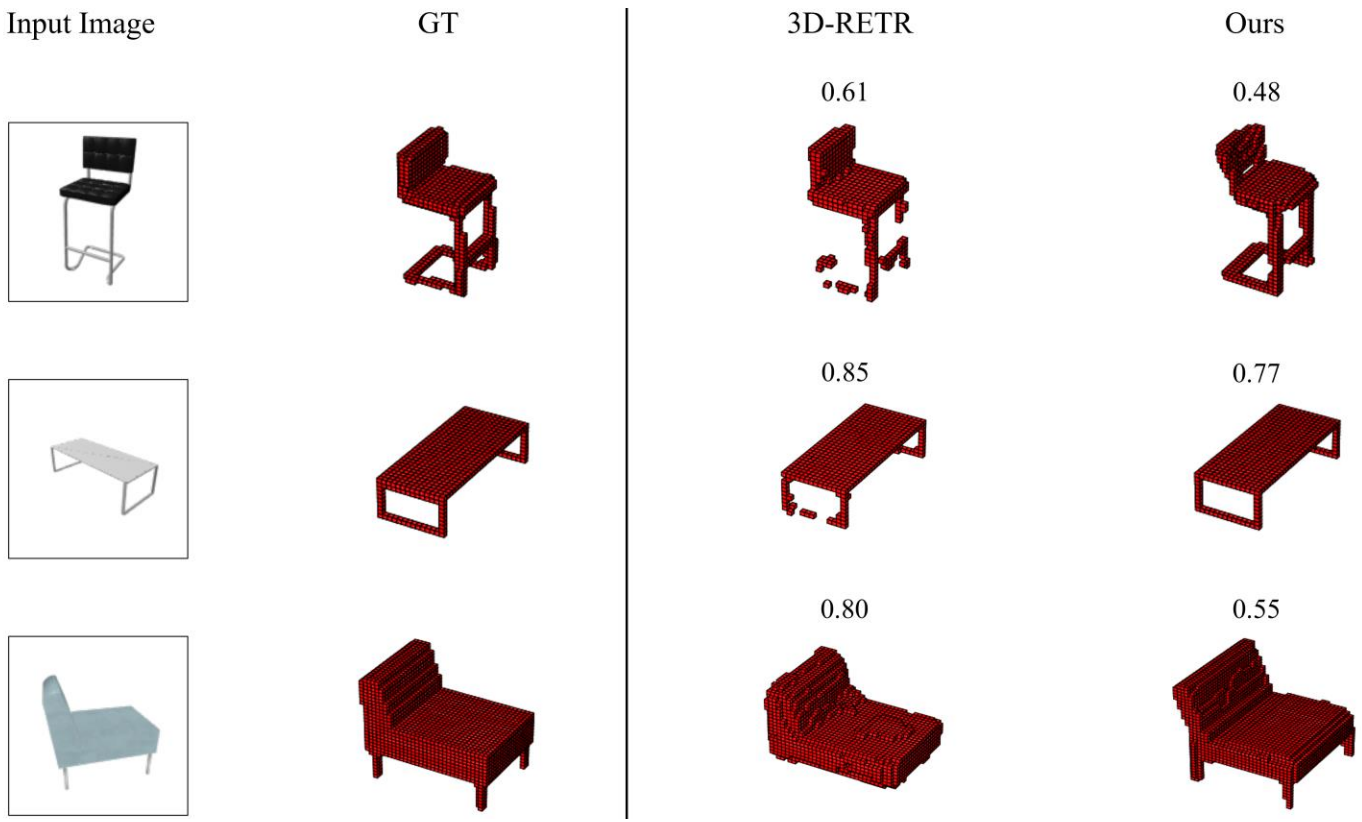
shapes A, B



IoU:



IOU FLAWS



IoU measures exact correspondence, thus preferring correct but unrealistic models.

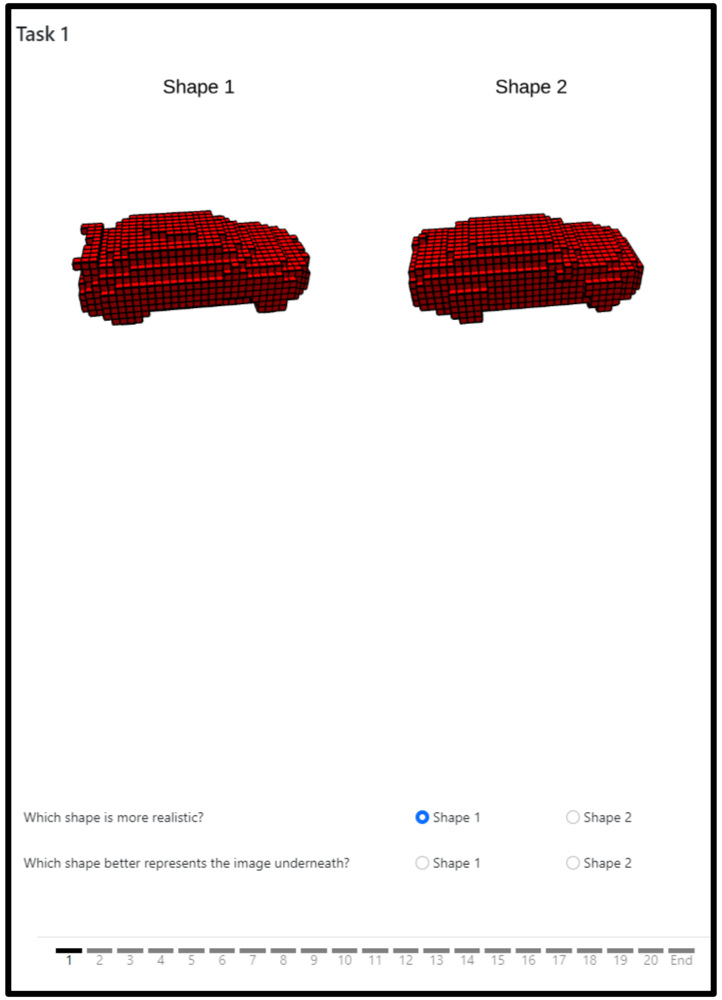
How can we measure coherence to the image with other metrics?

SIDE-BY-SIDE HUMAN EVALUATION

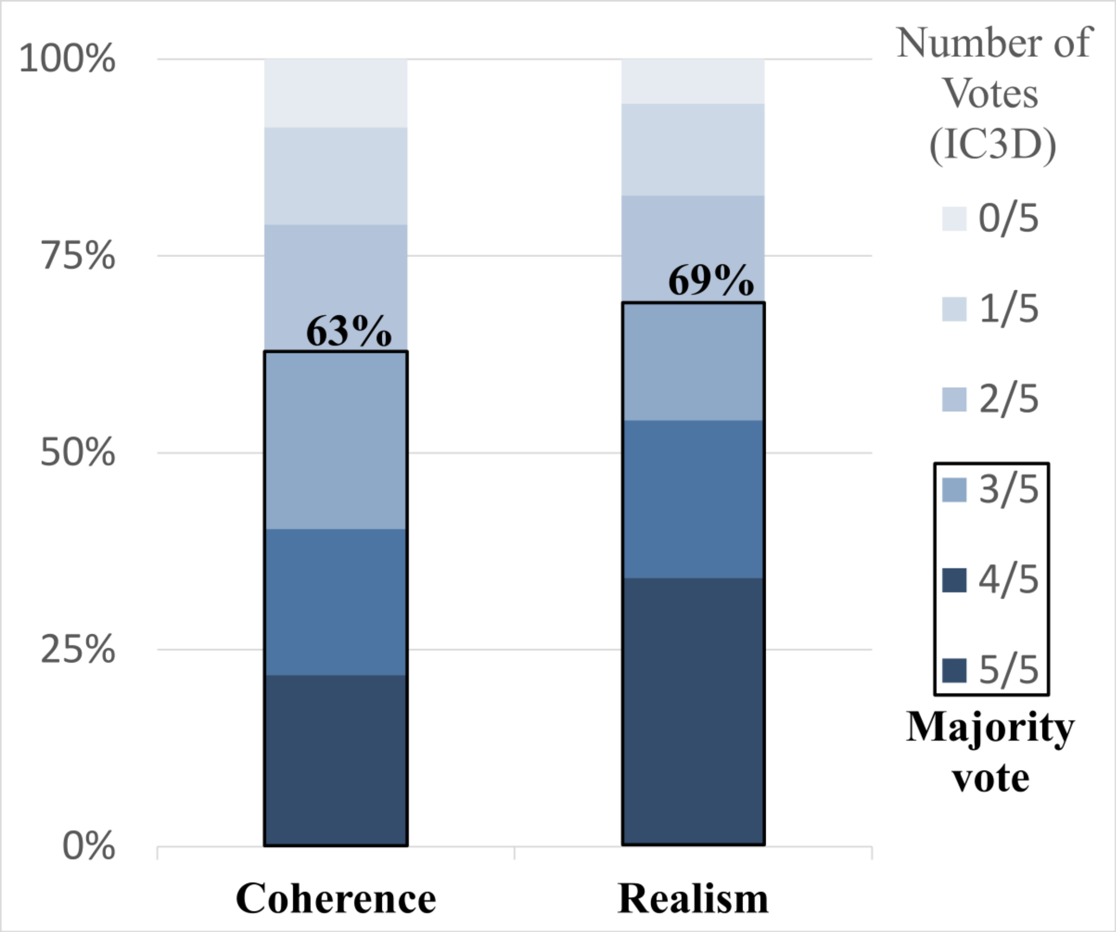
150 evaluators

600 total questions, 20 per form

Each form is shown to exactly 5 evaluators



HUMAN EVALUATION RESULTS



Human evaluation results

The majority of the evaluators prefer our model for realism in 69% of the questions.

Our model solves the realism issues arising in the 3D reconstruction approach.

It is also preferred for coherence, showing the **effectiveness of the guidance**.

PER-CLASS HUMAN EVALUATION RESULTS

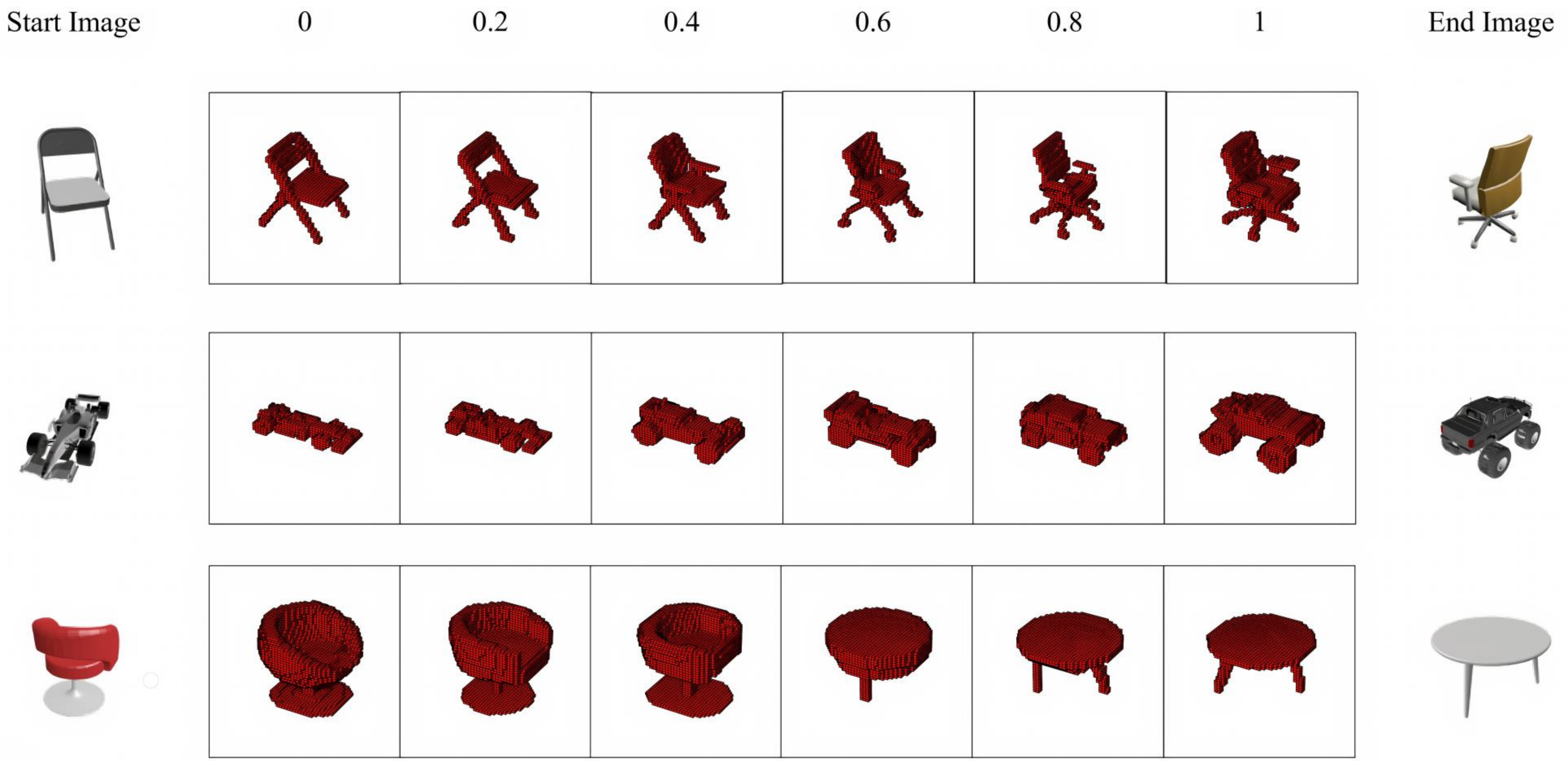
	0/5	1/5	2/5	3/5	4/5	5/5	3/5 or higher
aeroplane	6,00%	16,50%	16,00%	24,50%	19,50%	17,50%	61,50%
car	12,50%	9,50%	19,00%	23,50%	19,50%	16,00%	59,00%
chair	7,50%	11,00%	13,00%	20,00%	16,50%	32,00%	68,50%
overall	8,67%	12,33%	16,00%	22,67%	18,50%	21,83%	63,00%

coherence per-class results

	0/5	1/5	2/5	3/5	4/5	5/5	3/5 or higher
aeroplane	3,50%	12,50%	19,00%	16,50%	21,00%	27,50%	65%
car	9,50%	18,50%	20,50%	18,50%	19,00%	14,00%	52%
chair	4,00%	4,00%	1,50%	9,50%	20,00%	61,00%	91%
overall	5,67%	11,67%	13,67%	14,83%	20,00%	34,17%	69%

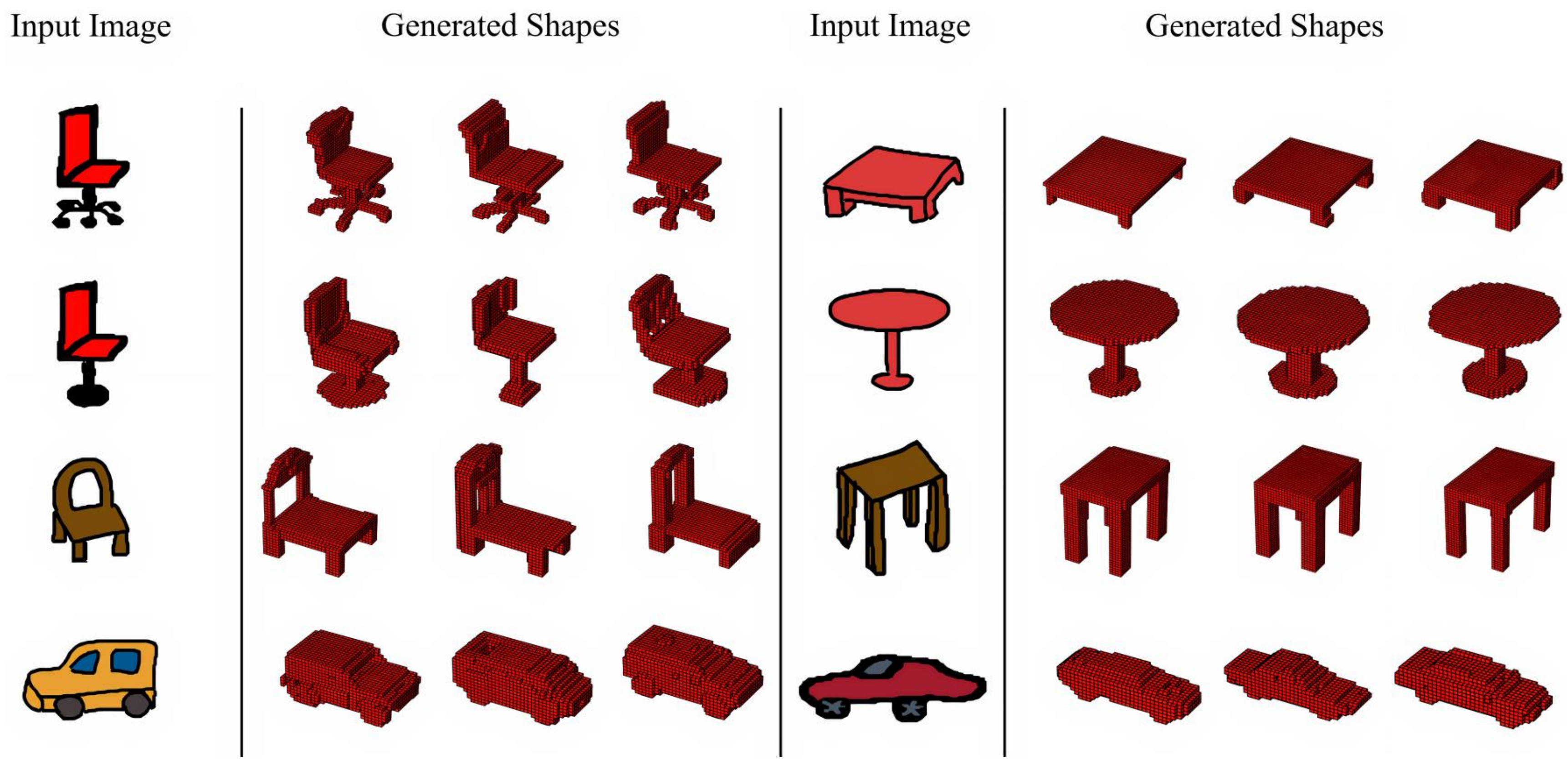
realism per-class results

INTERPOLATIONS



Examples of intra- and inter-class interpolations. CISP embeddings are interpolated by spherical linear interpolation (Slerp) with a 0.2 step.

HAND-DRAWN SHAPES



Thanks to CISP embeddings, we can also use handmade drawings of objects as query images. IC3D produces relevant and high-quality shapes even in this case.

LIMITATIONS AND FUTURE WORKS

Limitation	Solution
Low sampling speed (108s/sample)	New techniques for DDPMs
Scalability of the model	Explore other 3D representations
Generalization on shape categories	Train on more/vaster datasets
Single view conditioning	Combine CISP embeddings

CONCLUSIONS

CISP

Joint image-shape embeddings

IC3D

Image-Driven Voxel diffusion

SoTA generation results

Coherent to the query image

Solve realism & structural issues

SINGLE-VIEW SHAPE RECONSTRUCTION VIA IMAGE-CONDITIONED 3D DIFFUSION

Cristian Sbrolli

Advisor: Prof. Matteo Matteucci

Co-Advisor: Paolo Cudrano, Matteo Frosi

POLITECNICO MILANO 1863

HONOURS PROGRAMME HP-SR in Information Technology

AIRLAB ARTIFICIAL INTELLIGENCE AND ROBOTICS LAB