Building Multimodal Datasets for Immersive 3D Neural Fields

Srinath Sridhar

BIRS Workshop on 3D Generative Models

July 2023

two hands holding small yellow mallets and playing a simple tune on a toy glockenspiel with rainbow-colored bars

BROWN

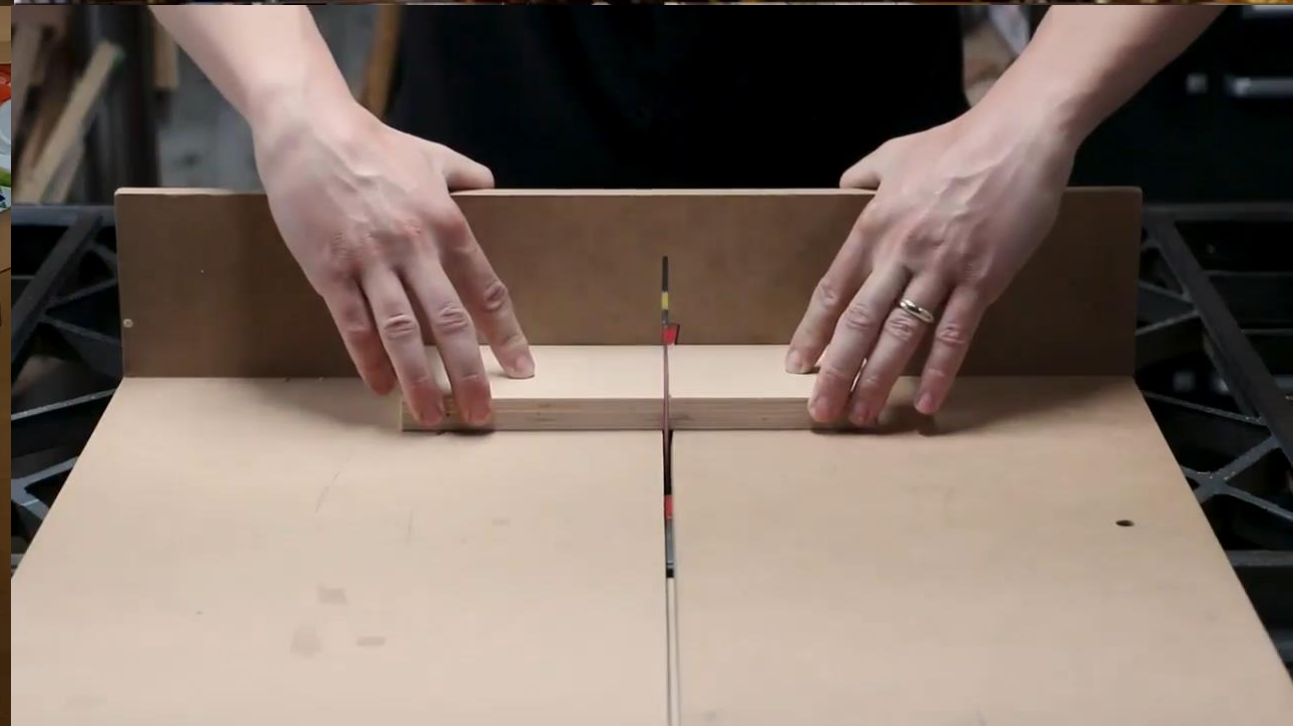
Brown Interactive 3D Vision and Learning Lab (IVL)

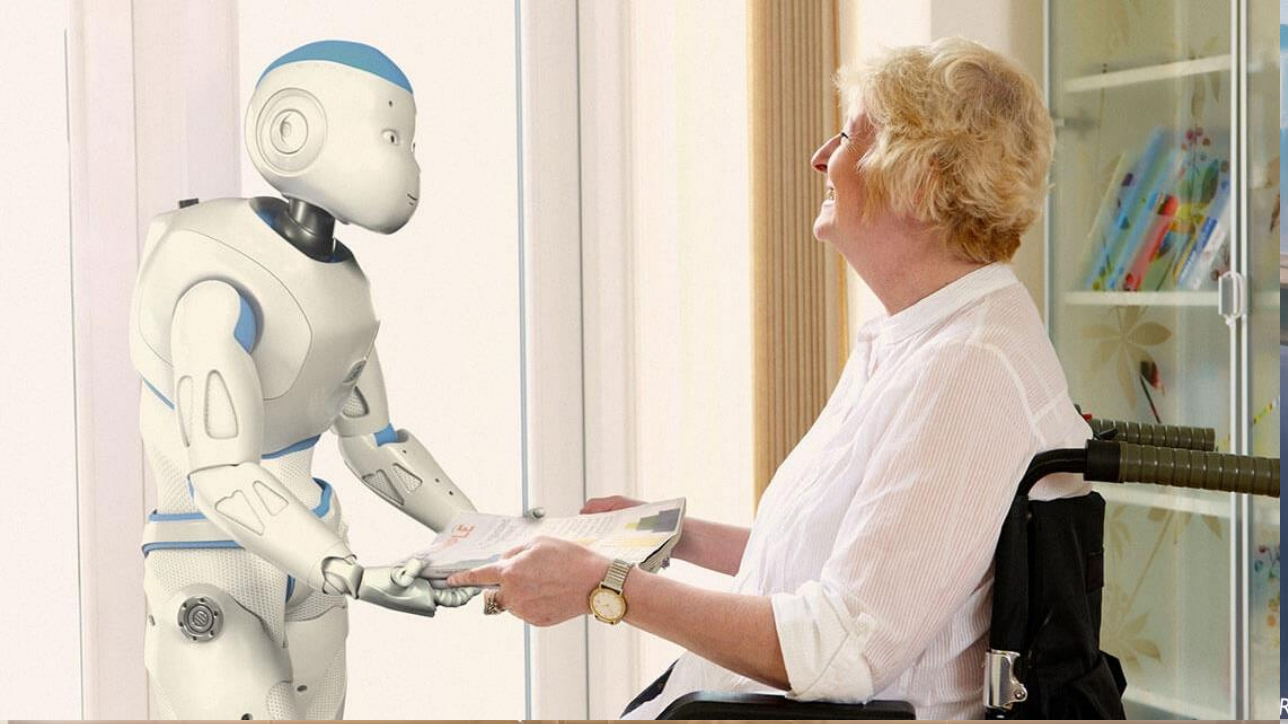
Rohith Agaram | Kefan Chen | Yiwen Chen | Arnab Dey | Jacob Frausto | Rao Fu | Dylan Hu | Iris Huang | Filip Kierzenka | Cheng-You Lu | Rugved Mavidipalli | Theo McArn | Chandradeep Pokhariya | Rahul Sajnani | Qiuhong Wei | Angela Xing | Xiao Zhan | Peisen Zhou

https://ivl.cs.brown.edu

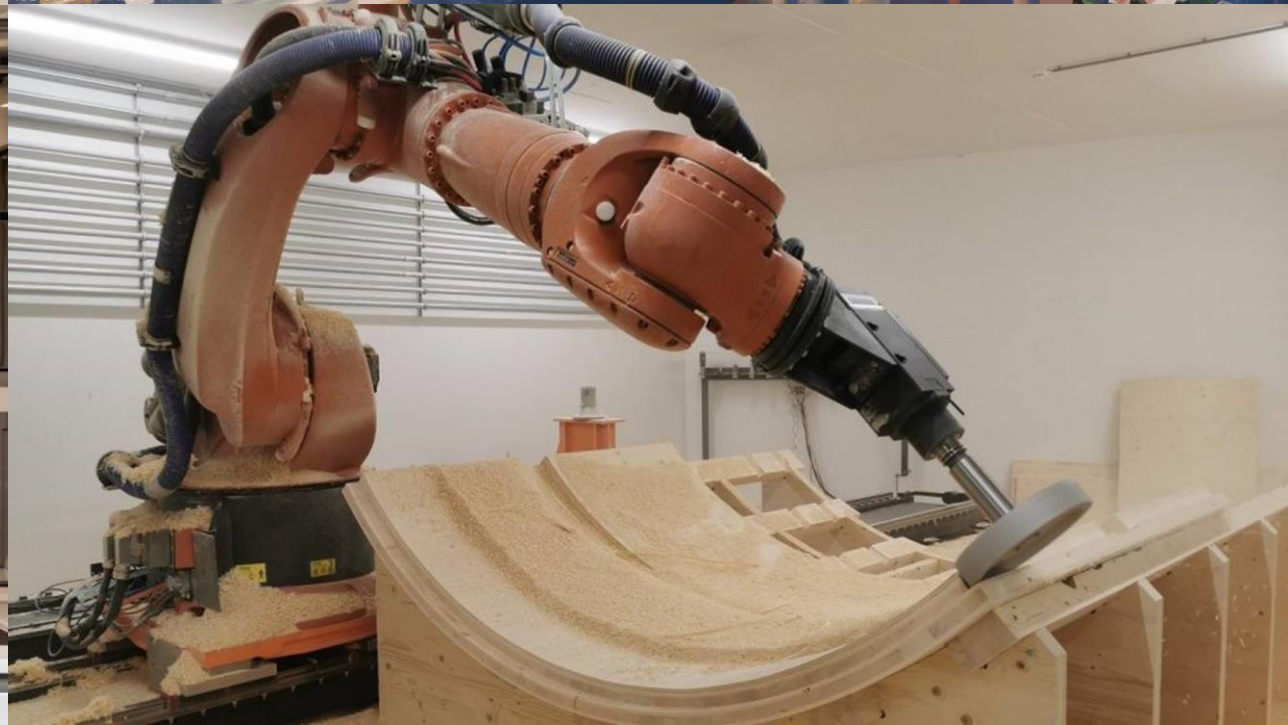

HOVIS
SEED SENSATION
Seven Seeds
alpro
ALMOND

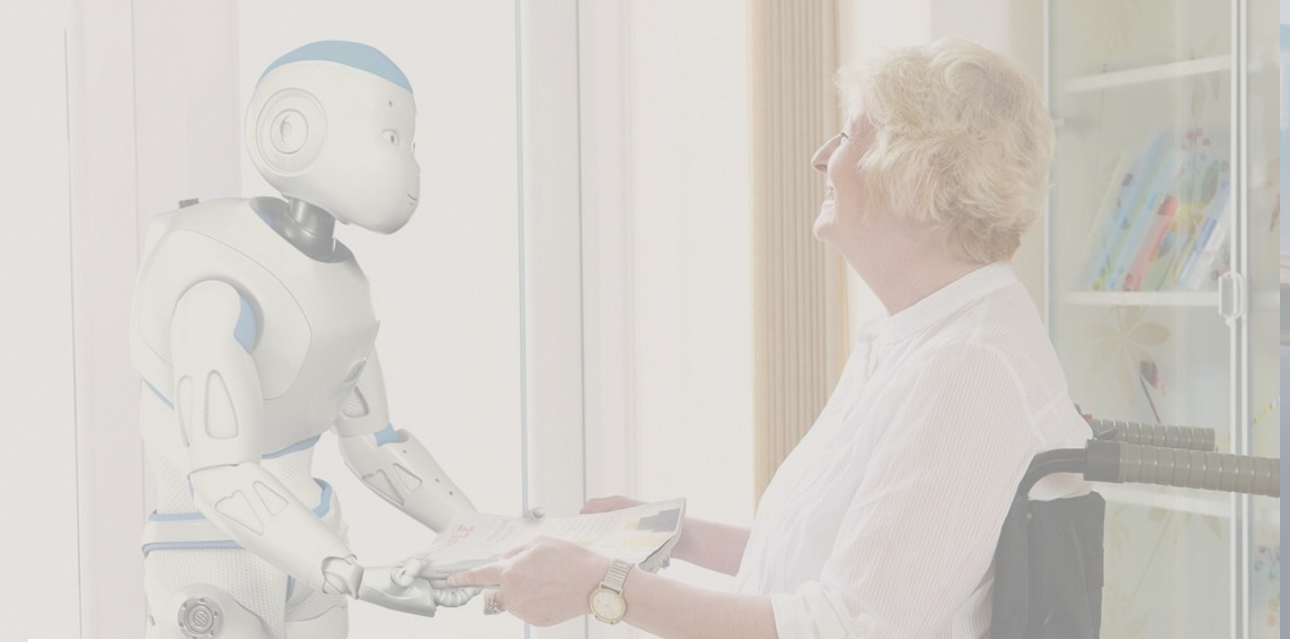

HOVIS
SEED SENSATION
Seven Seeds
alpro
ALMOND






Roberto Parada

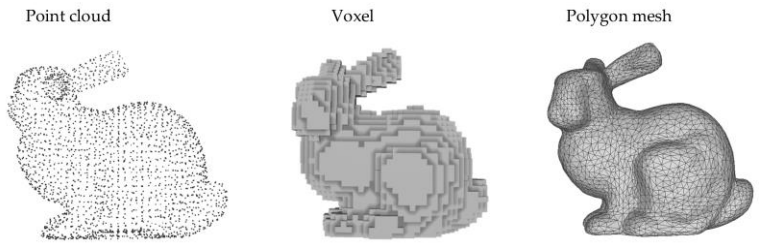


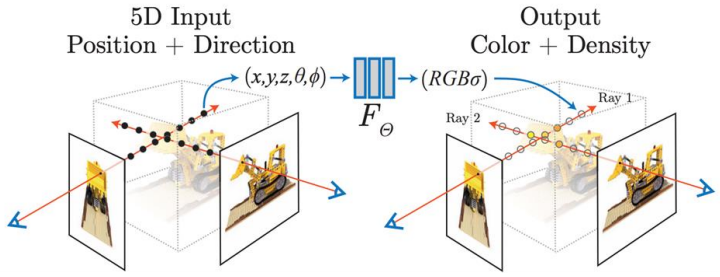


Robots need **multimodal dynamic 3D** understanding.

Current Progress in Multimodal Dynamic 3D Understanding

Current Progress in Multimodal Dynamic 3D Understanding

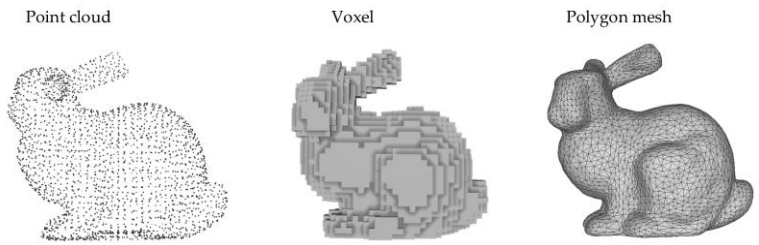


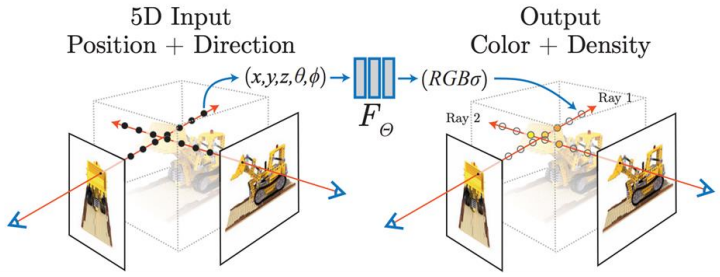


NeRF, Mildenhall et al. 2020

3D Representations

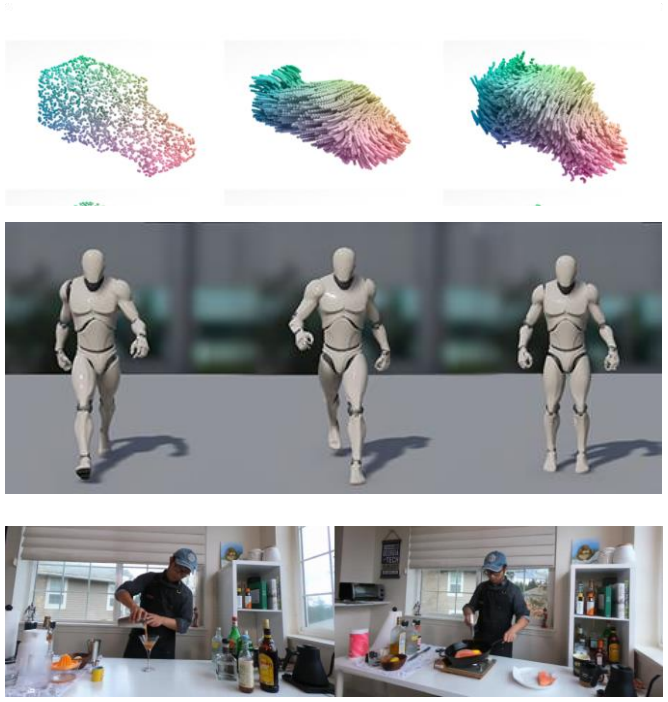
Current Progress in Multimodal Dynamic 3D Understanding





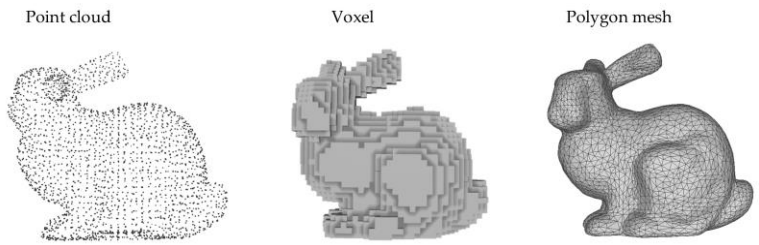
NeRF, Mildenhall et al. 2020

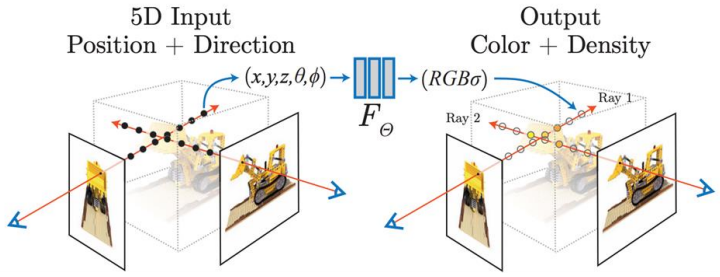
3D Representations



Dynamics

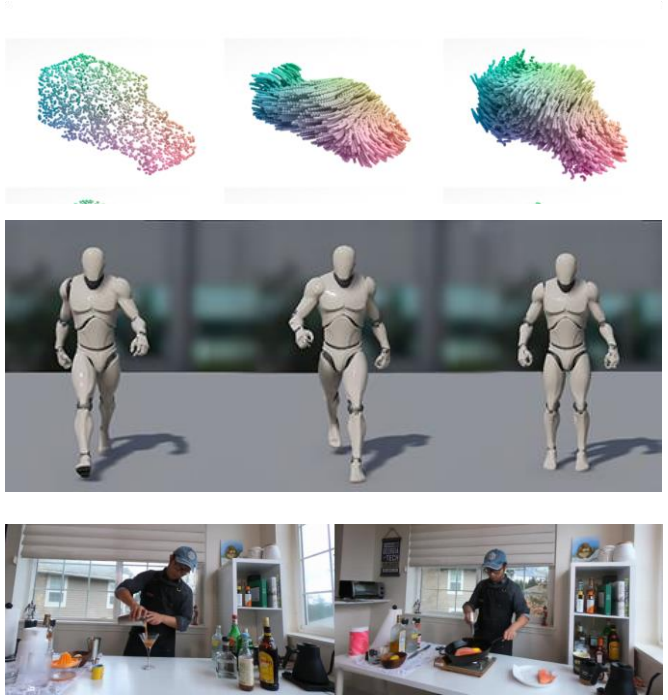
Current Progress in Multimodal Dynamic 3D Understanding



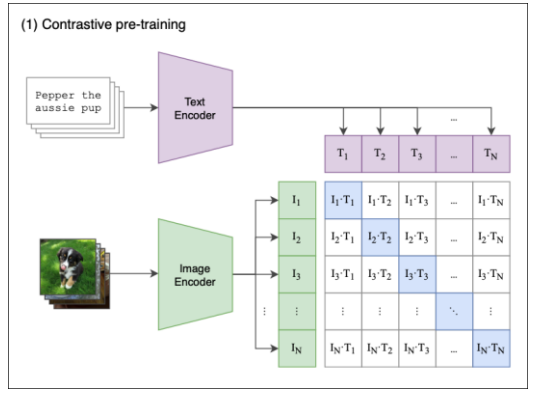


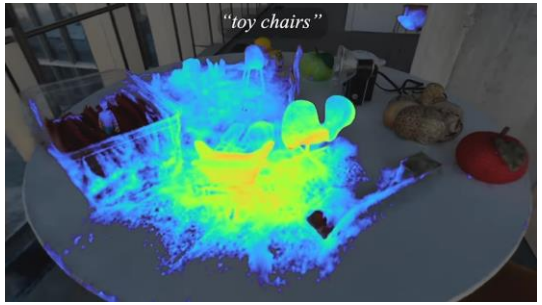
NeRF, Mildenhall et al. 2020

3D Representations



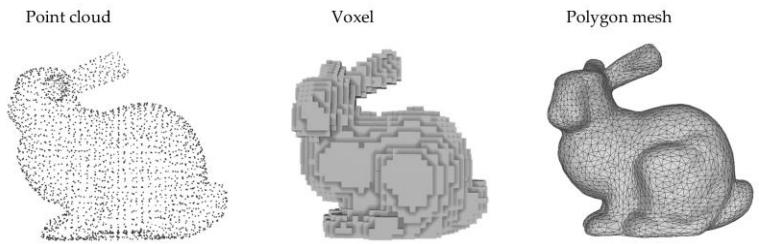
Dynamics

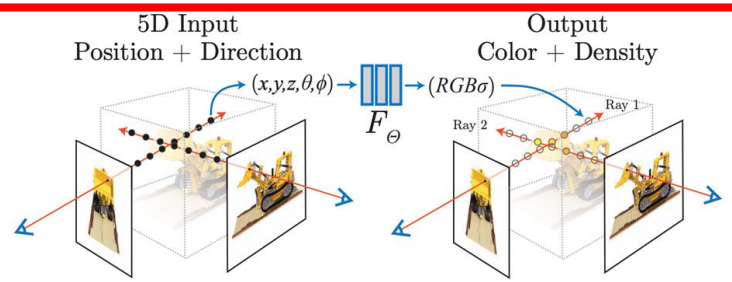




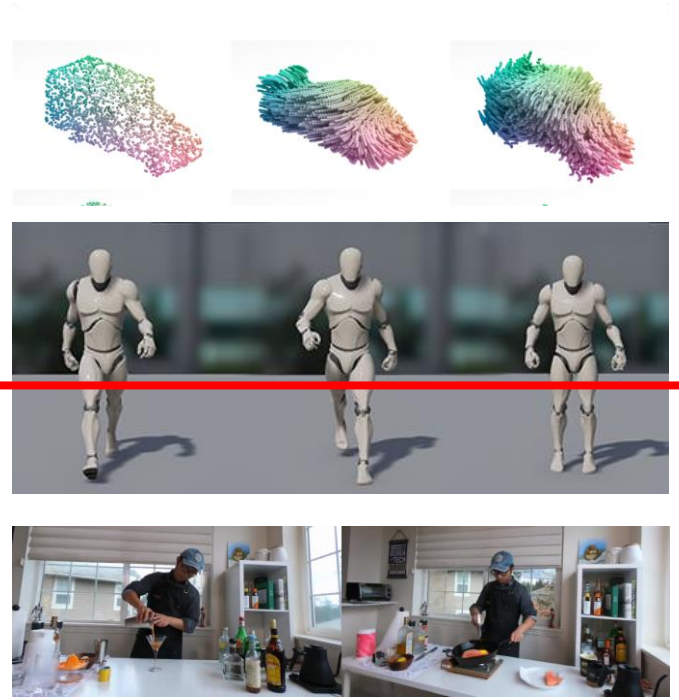
Multimodality

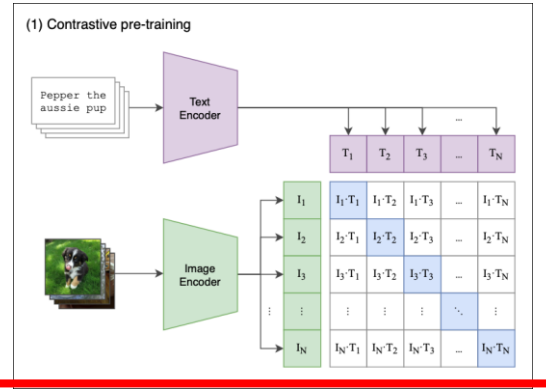
Current Progress in Multimodal Dynamic 3D Understanding

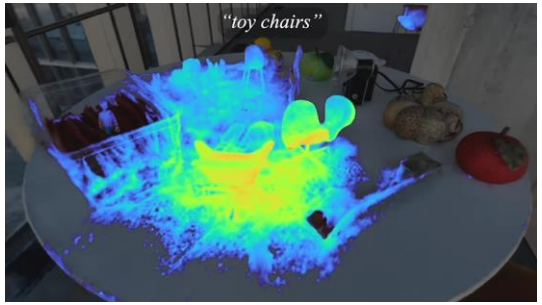




NeRF, Mildenhall et al. 2020





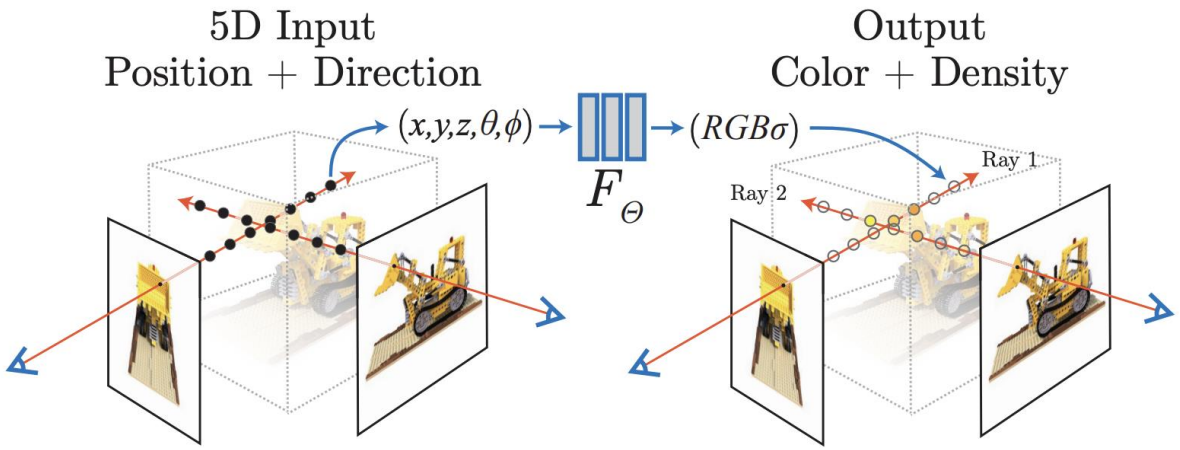


3D Representations

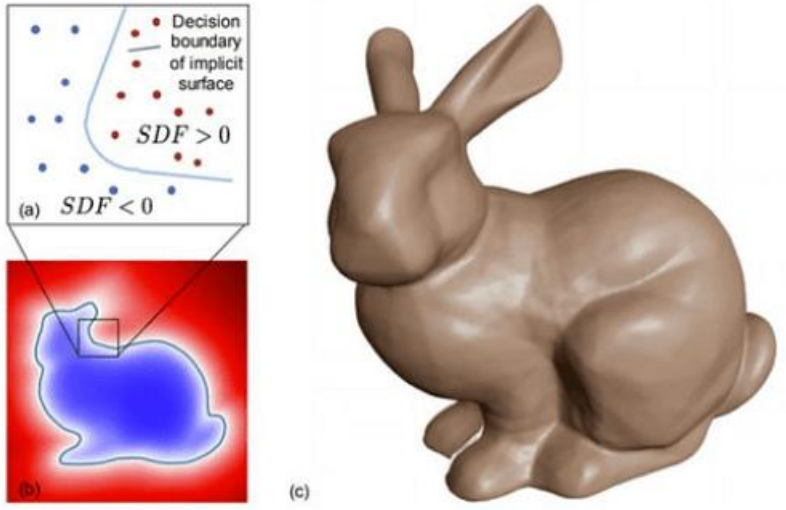
Dynamics

Multimodality

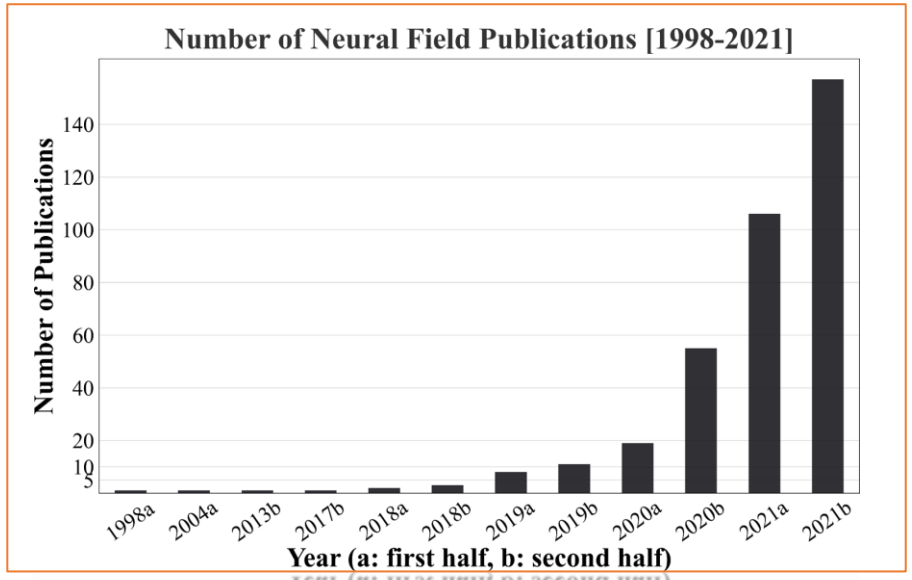
Neural Fields



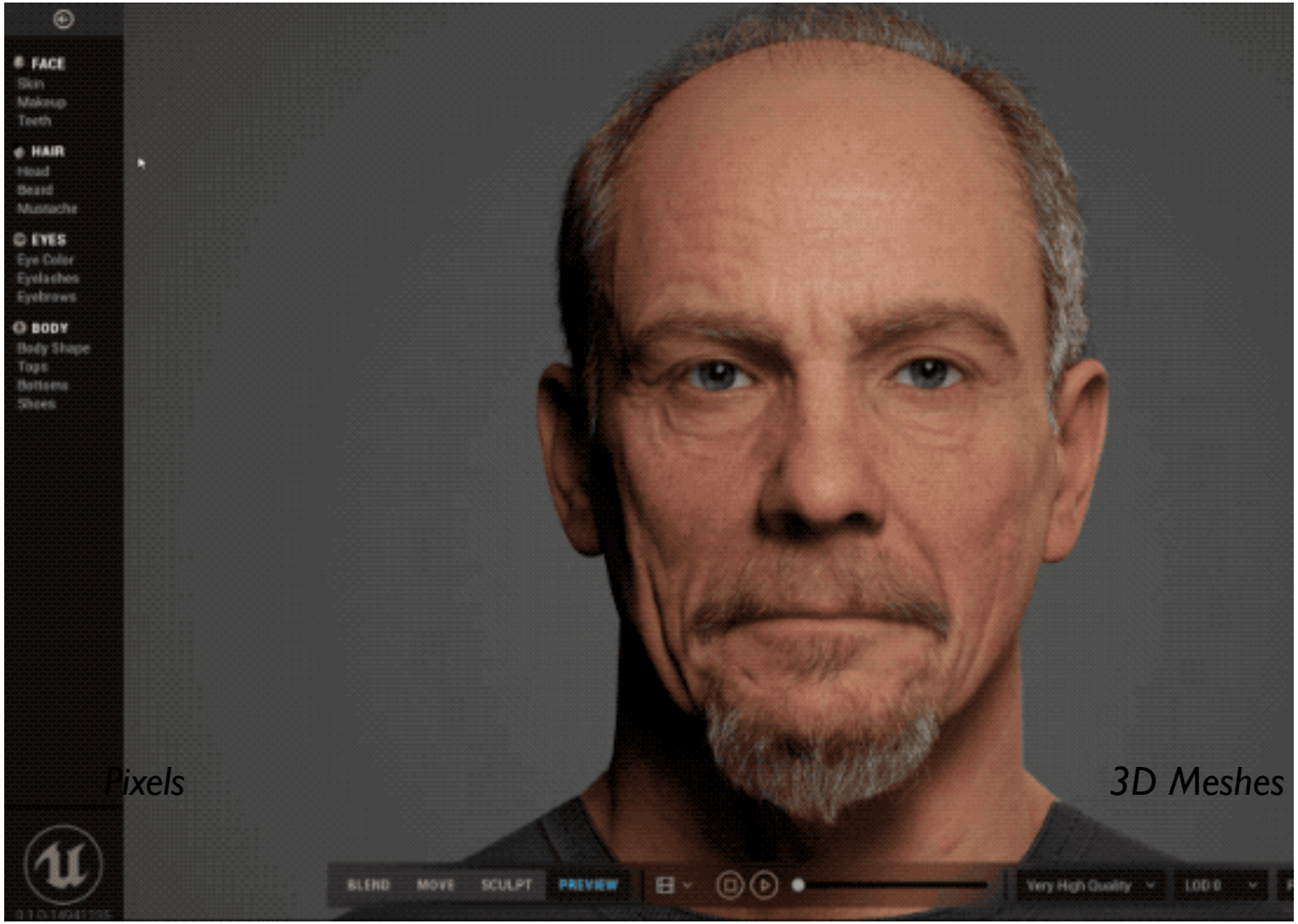
NeRF, Mildenhall et al. 2020



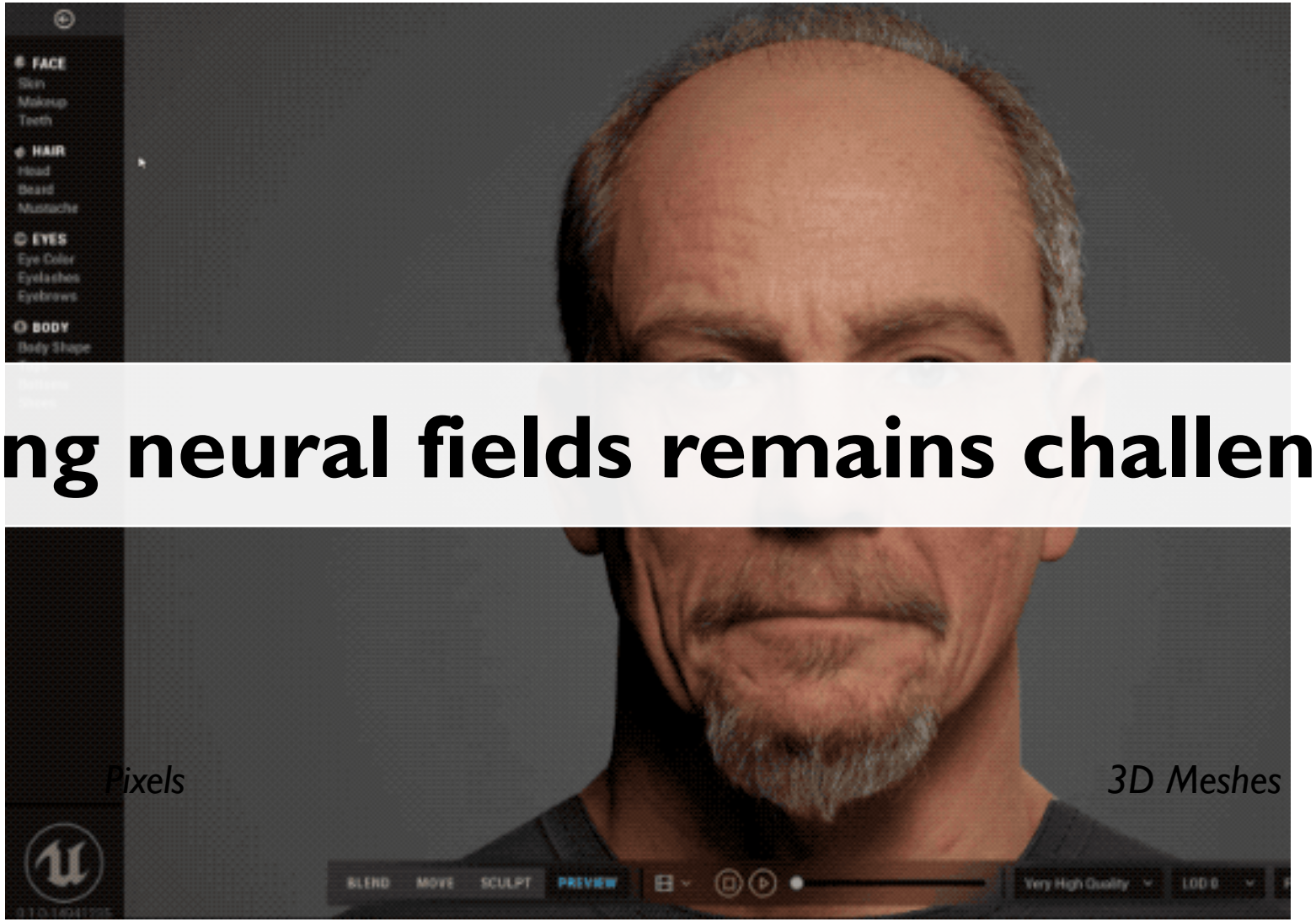
DeepSDF, Park et al. 2019



Everyday Tasks



Everyday Tasks



Editing neural fields remains challenging

Pixels

3D Meshes

Dynamic Neural Fields / NeRFs

Neural 3D Video Synthesis from Multi-view Video, Li et al., CVPR 2022

Dynamic Neural Fields / NeRFs

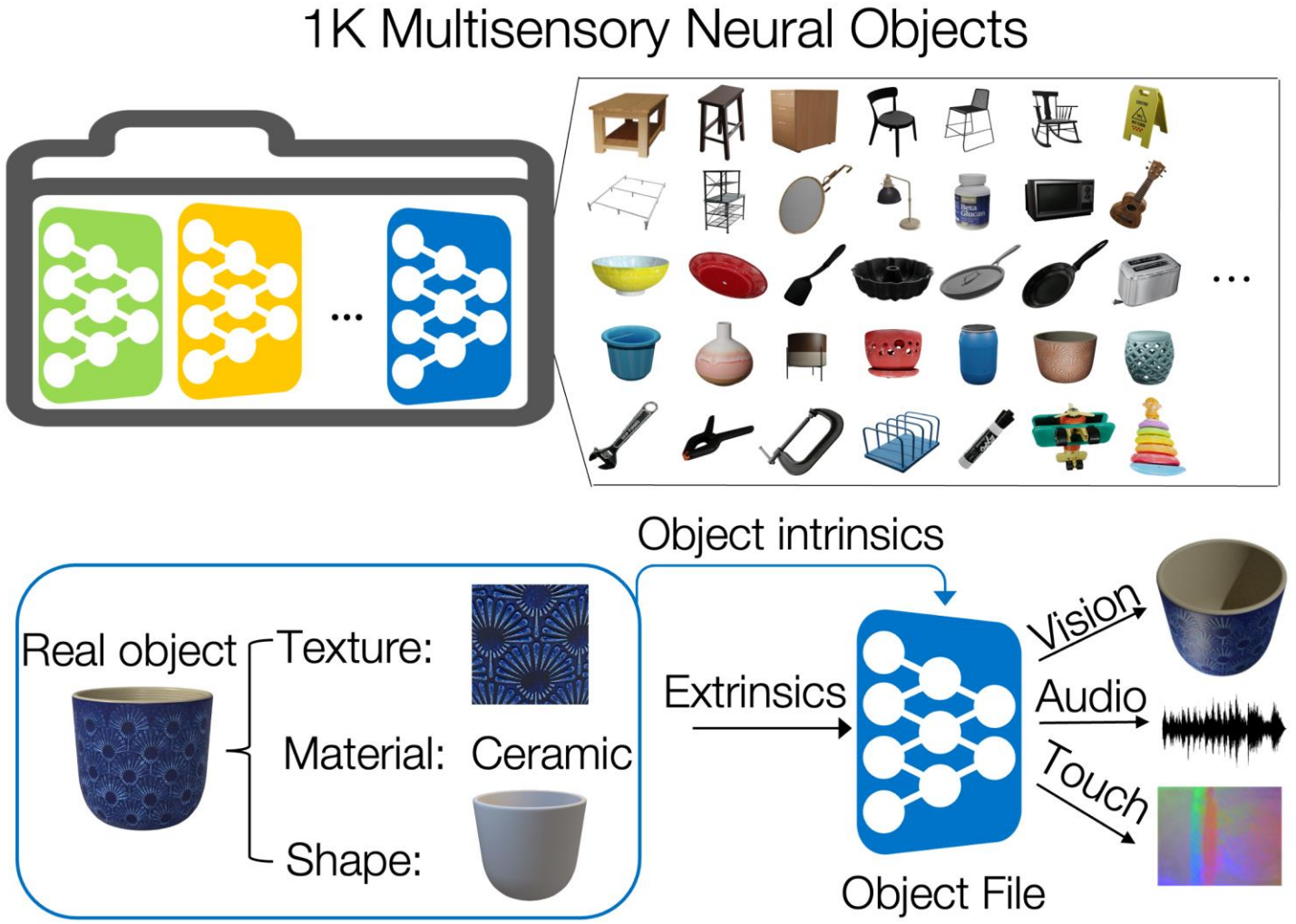
Neural 3D Video Synthesis from Multi-view Video, Li et al., CVPR 2022

Dynamic Neural Fields / NeRFs

Limited duration: 1-30 seconds

Neural 3D Video Synthesis from Multi-view Video, Li et al., CVPR 2022

Multimodal 3D Neural Fields



The ObjectFolder Benchmark: Multisensory Learning with Neural and Real Objects, Gao et al. 2023

Multimodal 3D Neural Fields

The ObjectFolder Benchmark: Multisensory Learning with Neural and Real Objects, Gao et al. 2023

Part of the Reason: Datasets

Type	Dataset	Real	360° view	Dynamic	Caption	Canonical	Audio
Scene	DTU[27]	✓	✓	✗	✗	—	✗
	BlendedMVS[80]	✗	✓	✗	✗	—	✗
	ScanNet[13]	✓	✗	✗	✗	—	✗
	LLFF[45]	✓	✗	✗	✗	—	✗
	Mip-NeRF 360[5]	✓	✓	✗	✗	—	✗
	Block-NeRF[65]	✓	✗	✓	✗	—	✗
	DyNeRF[35]	✓	✗	✓	✗	—	✗
	HyperNeRF[53]	✓	✗	✓	✗	—	✗
	NDSD[82]	✓	✗	✓	✗	—	✗
	ILFV[7]	✓	✗	✓	✗	—	✗
	Deep3DMV[39]	✓	✗	✓	✗	—	✗
Object	ShapeNet[9]	✗	✓	✗	✗	✓	✗
	NeRF[46]	✗	✓	✗	✗	✓	✗
	CO3D[55]	✓	✓	✗	✗	✗	✗
	ScanNeRF[14]	✓	✓	✗	✗	✓	✗
	OmniObject3D[75]	✓	✓	✗	✗	✓	✗
	ObjectFolder[23]	✗	✓	✗	✗	✓	✓
Hybrid	PeRFception[28]	✓	✓	✗	✗	✗	✗
	Objaverse[17]	✗	✓	✓	✓	✓	✗

Part of the Reason: Datasets

Type	Dataset	Real	360° view	Dynamic	Caption	Canonical	Audio
Scene	DTU[27]	✓	✓	✗	✗	—	✗
	BlendedMVS[80]	✗	✓	✗	✗	—	✗
	ScanNet[13]	✓	✗	✗	✗	—	✗
	LLFF[45]	✓	✗	✗	✗	—	✗
	Mip-NeRF 360[5]	✓	✓	✗	✗	—	✗
	Block-NeRF[65]	✓	✗	✓	✗	—	✗
	DyNeRF[35]	✓	✗	✓	✗	—	✗
	HyperNeRF[53]	✓	✗	✓	✗	—	✗
	NDSD[82]	✓	✗	✓	✗	—	✗
	ILFV[7]	✓	✗	✓	✗	—	✗
	Deep3DMV[39]	✓	✗	✓	✗	—	✗
Object	ShapeNet[9]	✗	✓	✗	✗	✓	✗
	NeRF[46]	✗	✓	✗	✗	✓	✗
	CO3D[55]	✓	✓	✗	✗	✗	✗
	ScanNeRF[14]	✓	✓	✗	✗	✓	✗
	OmniObject3D[75]	✓	✓	✗	✗	✓	✗
	ObjectFolder[23]	✗	✓	✗	✗	✓	✓
Hybrid	PeRFception[28]	✓	✓	✗	✗	✗	✗
	Objaverse[17]	✗	✓	✓	✓	✓	✗

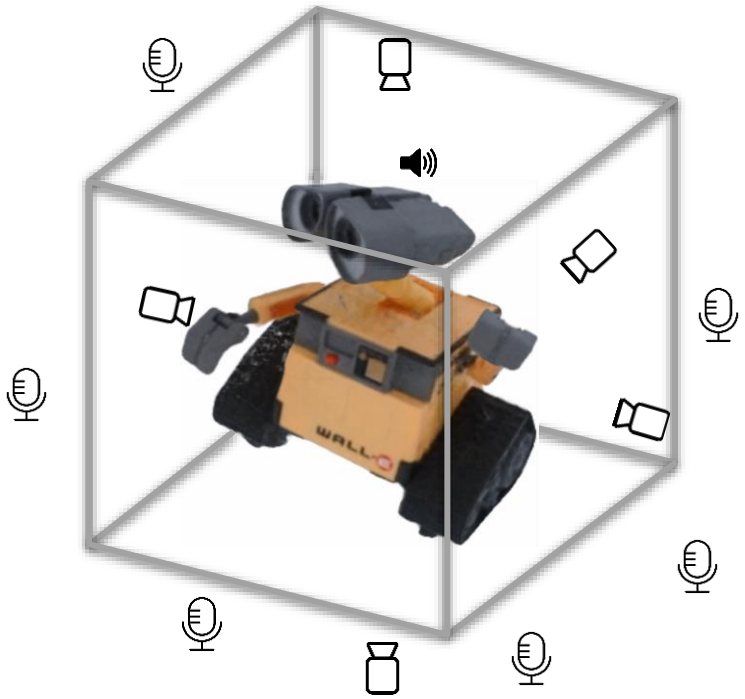
DiVA-360: The Dynamic Visuo-Audio Dataset

DiVA-360: The Dynamic Visuo-Audio Dataset

53× cam
1280x720 @ 120 FPS

6× mic

audio-visual scenes

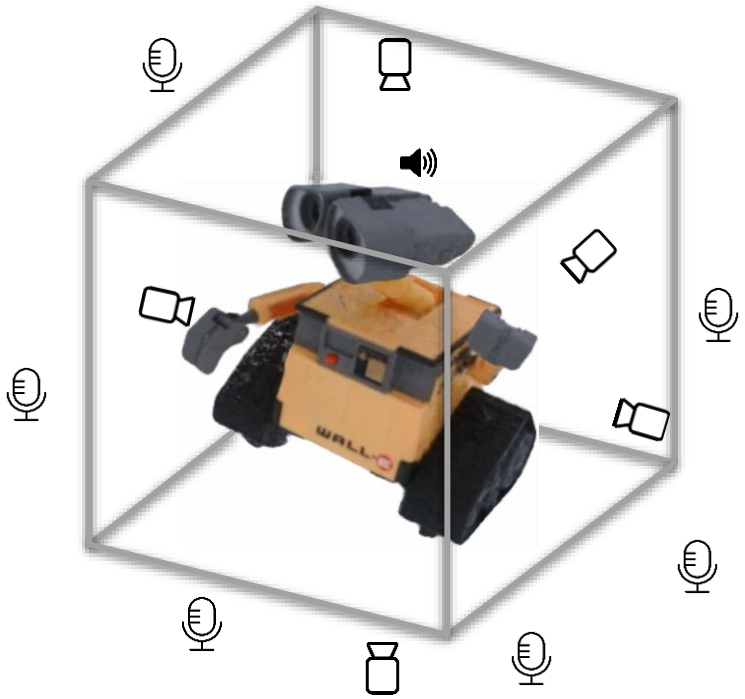


DiVA-360: The Dynamic Visuo-Audio Dataset

53× cam
1280x720 @ 120 FPS

6× mic

audio-visual scenes



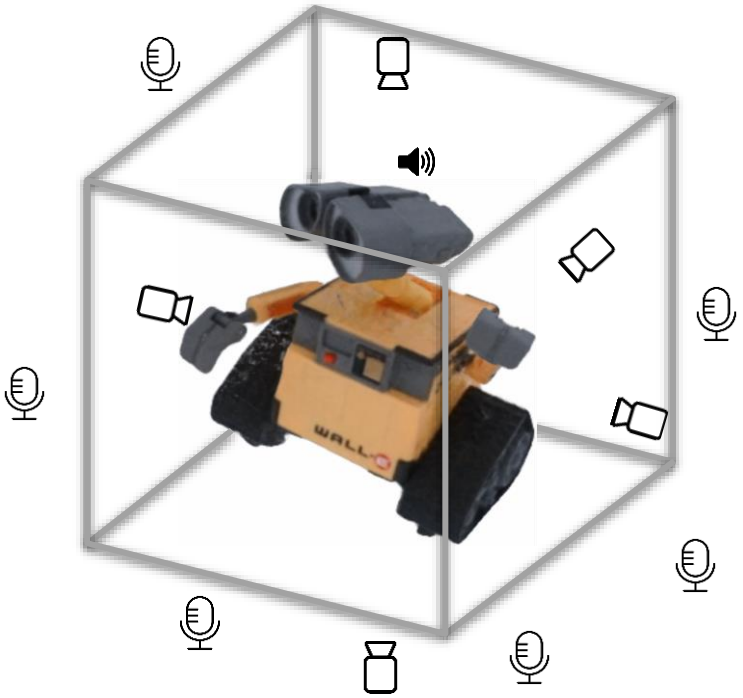
DiVA-360 Dynamic Dataset

DiVA-360: The Dynamic Visuo-Audio Dataset

53× cam
1280x720 @ 120 FPS

6× mic

audio-visual scenes



DiVA-360 Dynamic Dataset

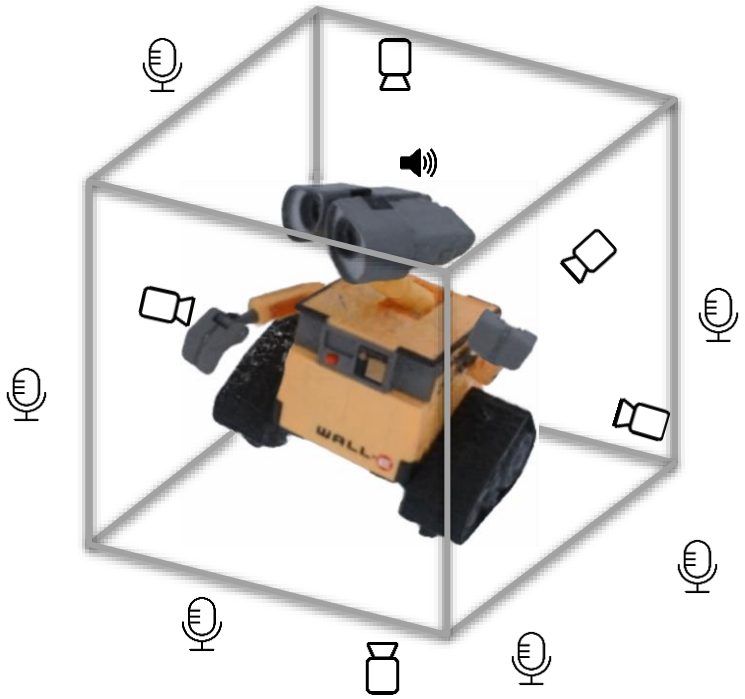
DiVA-360 Static Dataset

DiVA-360: The Dynamic Visuo-Audio Dataset

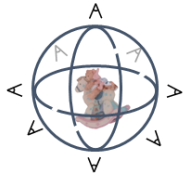
53× cam
1280x720 @ 120 FPS

6× mic

audio-visual scenes



360° View



DiVA-360 Dynamic Dataset

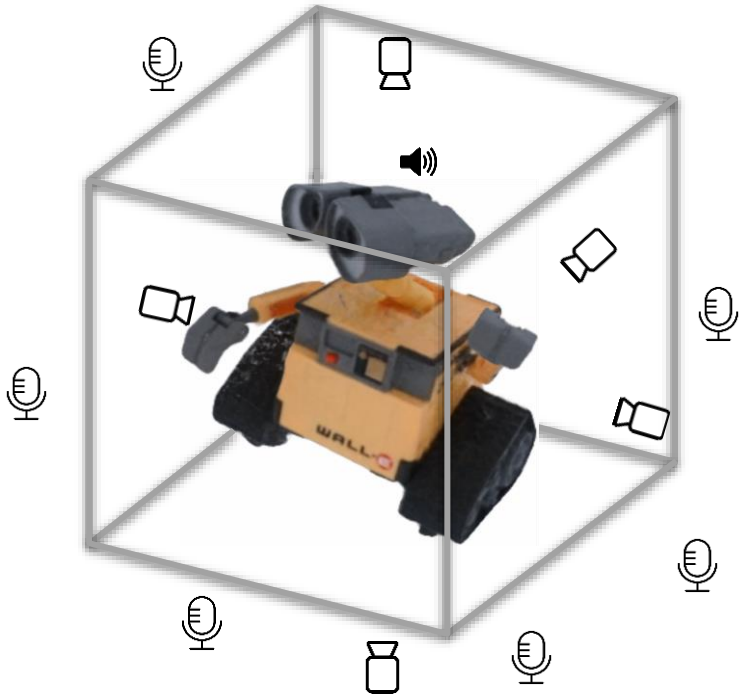
DiVA-360 Static Dataset

DiVA-360: The Dynamic Visuo-Audio Dataset

53× cam
1280x720 @ 120 FPS

6× mic

audio-visual scenes



DiVA-360 Dynamic Dataset

DiVA-360 Static Dataset

360° View

Text Description

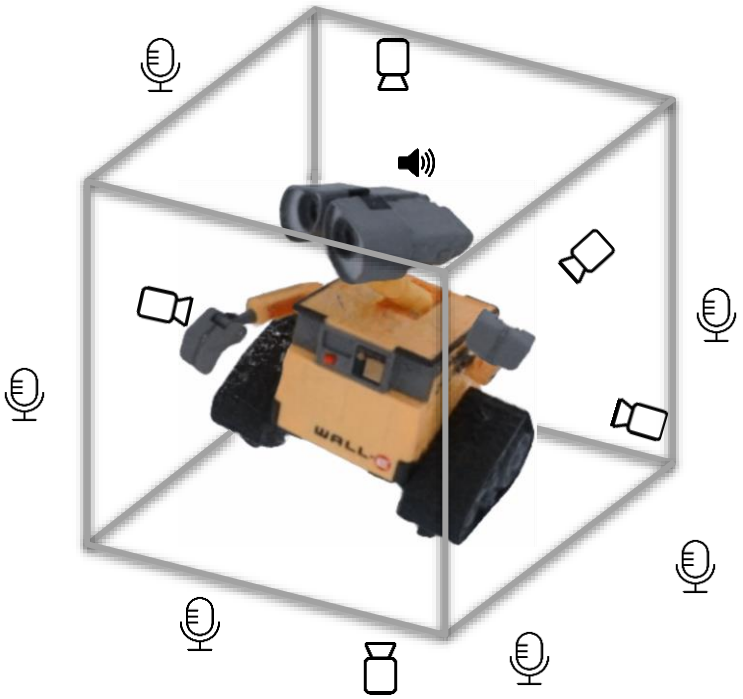
"a hypercar with a bright blue front and black back, two doors, white accent lines, a tall racing spoiler, and an engine visible behind the driver's seat."

DiVA-360: The Dynamic Visuo-Audio Dataset

53× cam
1280x720 @ 120 FPS

6× mic

audio-visual scenes



DiVA-360 Dynamic Dataset

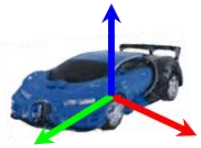
DiVA-360 Static Dataset

360° View

Text Description

"a hypercar with a bright blue front and black back, two doors, white accent lines, a tall racing spoiler, and an engine visible behind the driver's seat."

Canonicalized

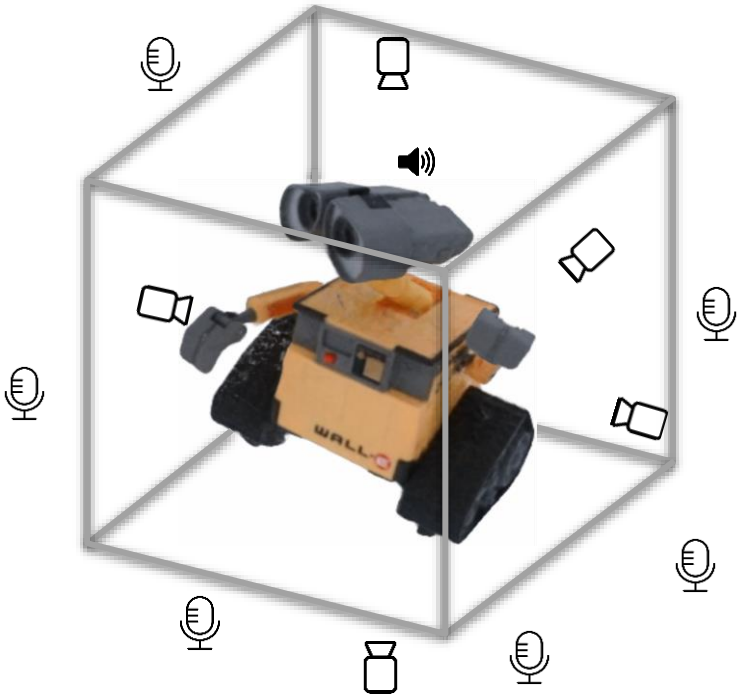


DiVA-360: The Dynamic Visuo-Audio Dataset

53× cam
1280x720 @ 120 FPS

6× mic

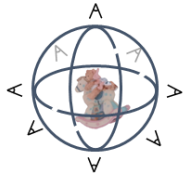
audio-visual scenes



DiVA-360 Dynamic Dataset

DiVA-360 Static Dataset

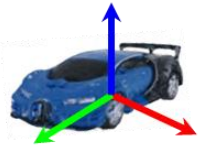
360° View



Text Description

"a hypercar with a bright blue front and black back, two doors, white accent lines, a tall racing spoiler, and an engine visible behind the driver's seat."

Canonicalized



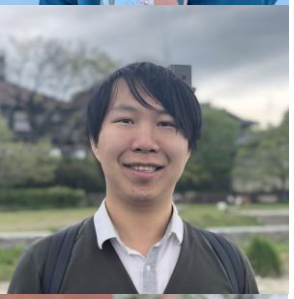
Spatial Audio

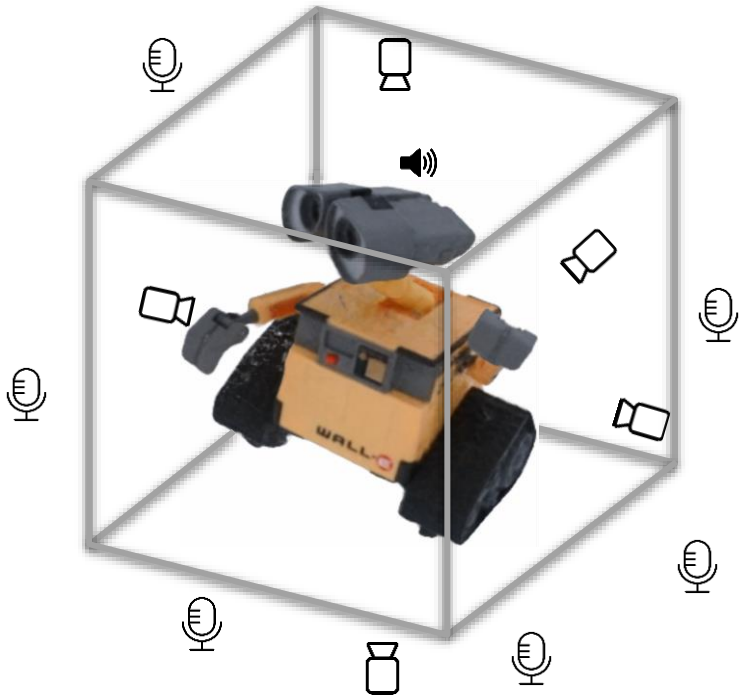
DiVA-360: The Dynamic Visuo-Audio Dataset

53× cam
1280x720 @ 120 FPS

6× mic

audio-visual scenes





DiVA-360 Dynamic Dataset

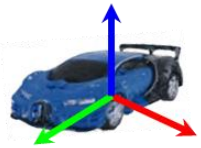
DiVA-360 Static Dataset

360° View

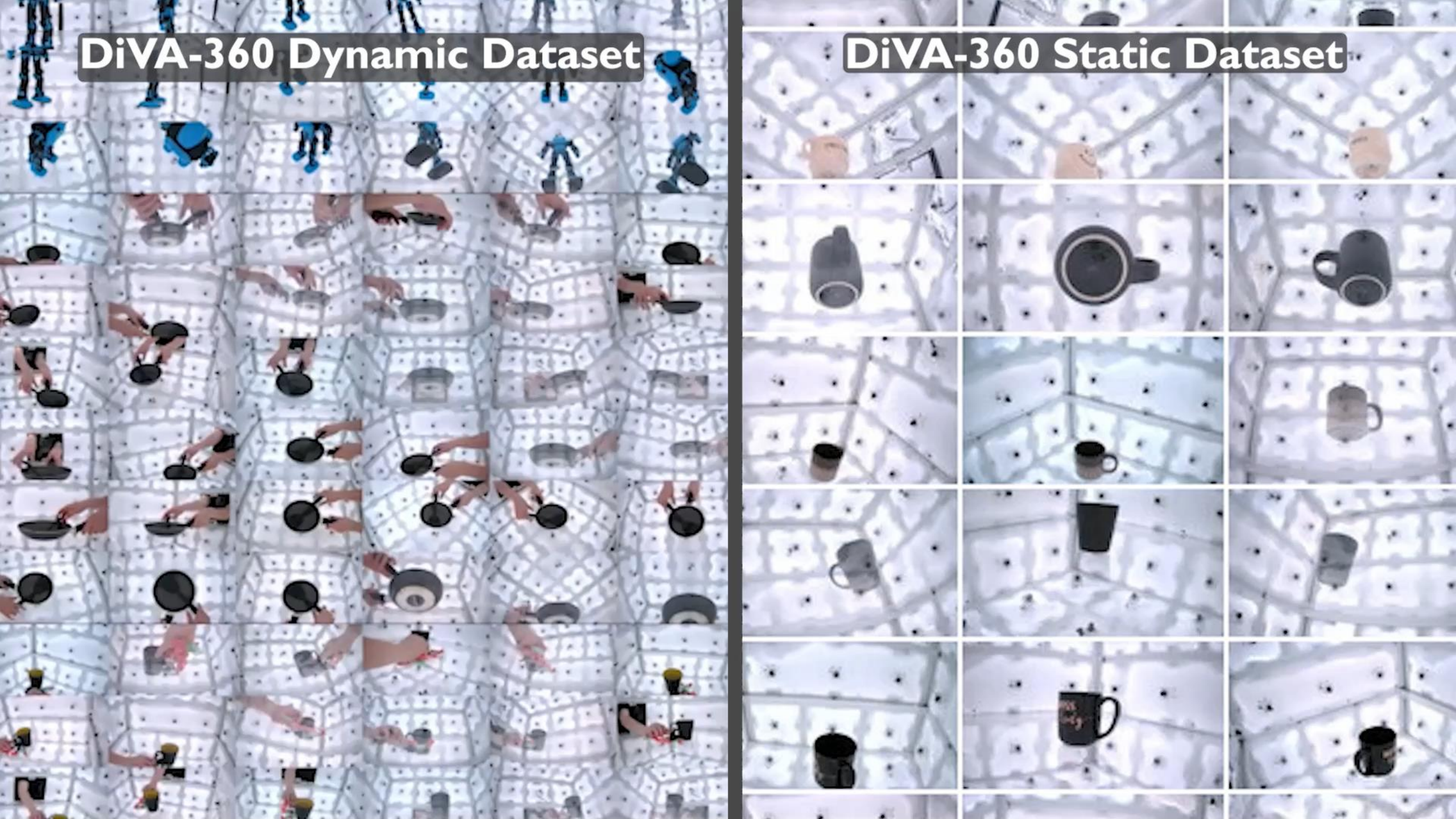
Text Description

"a hypercar with a bright blue front and black back, two doors, white accent lines, a tall racing spoiler, and an engine visible behind the driver's seat."

Canonicalized



Spatial Audio


DiVA-360 Dynamic Dataset
DiVA-360 Static Dataset

DiVA-360 Dynamic Dataset

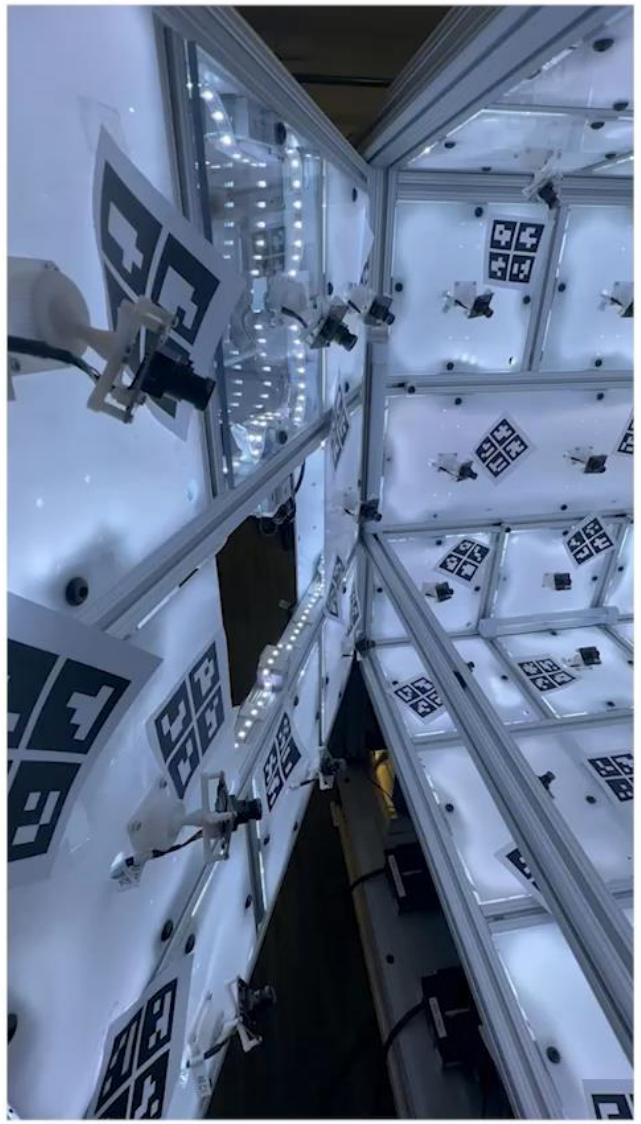
DiVA-360 Static Dataset

Challenges

Capture system (hardware+software)
Sensor synchronization
Calibration
Benchmarking Metrics

Capture Hardware & Software

Exterior Frame Assembly



Interior View

BRown Interaction Capture System (BRICS)

Capture Hardware & Software

Exterior Frame Assembly

Interior View

Table-scale scenes

Custom software for sensor synchronization

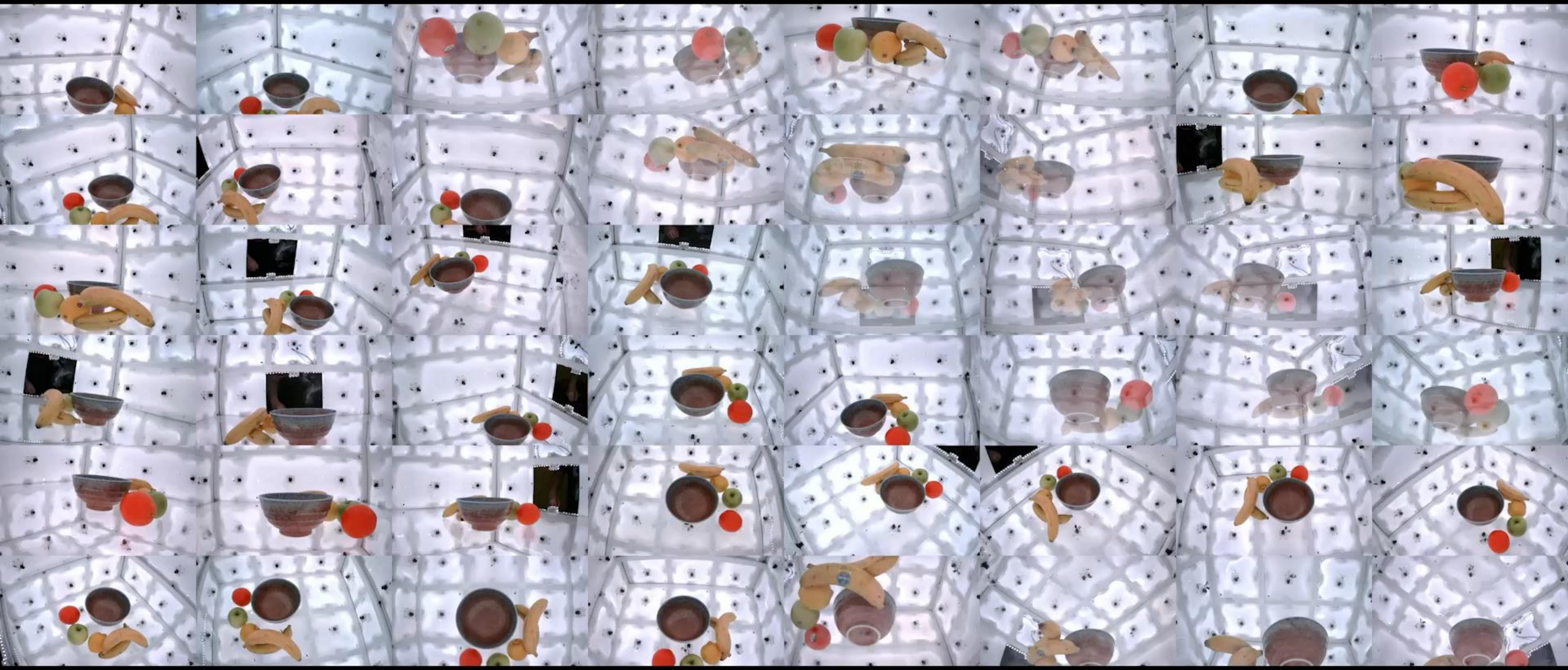
Calibration of cameras/lights

Manually labeled text descriptions

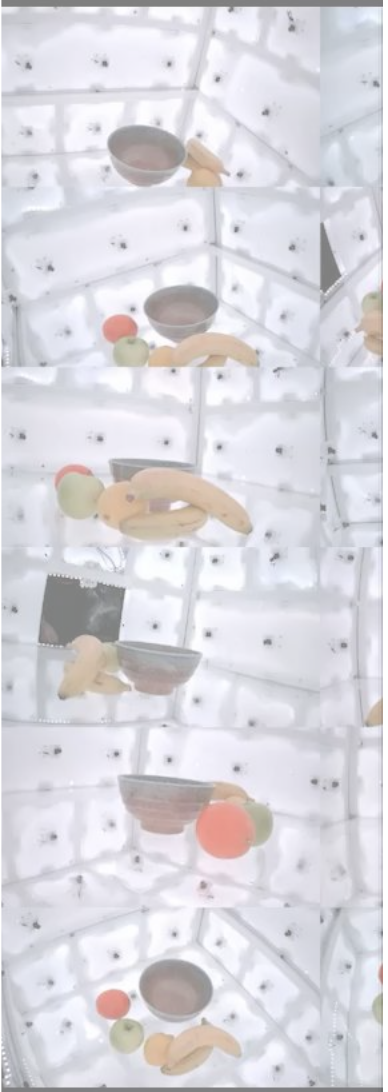
BRown Interaction Capture System (BRICS)

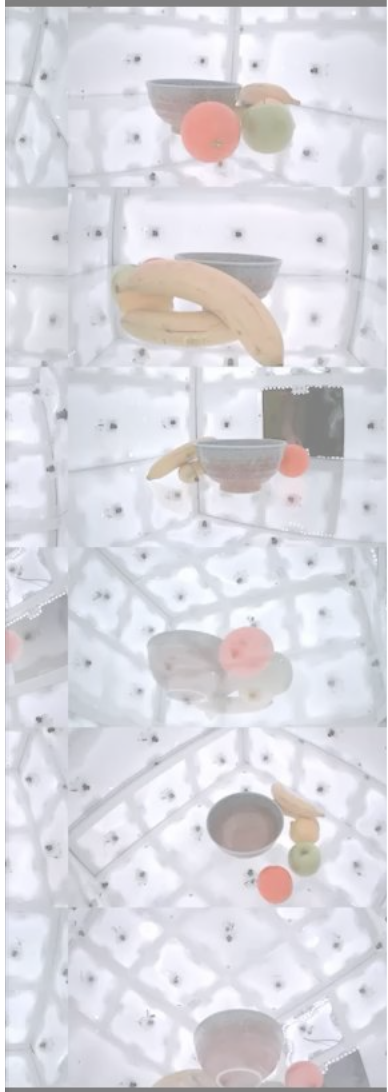
put_fruit

only 46/53 views shown
only 1/6 audio channels played
showing level 3 text descriptions



a hand picking up assorted fruit around a brown painted ceramic bowl and placing them into the bowl; the fruit includes an orange, a green apple, a lemon, and two bananas





REPORTS

Hearing Sounds, Understanding Actions: Action Representation in Mirror Neurons

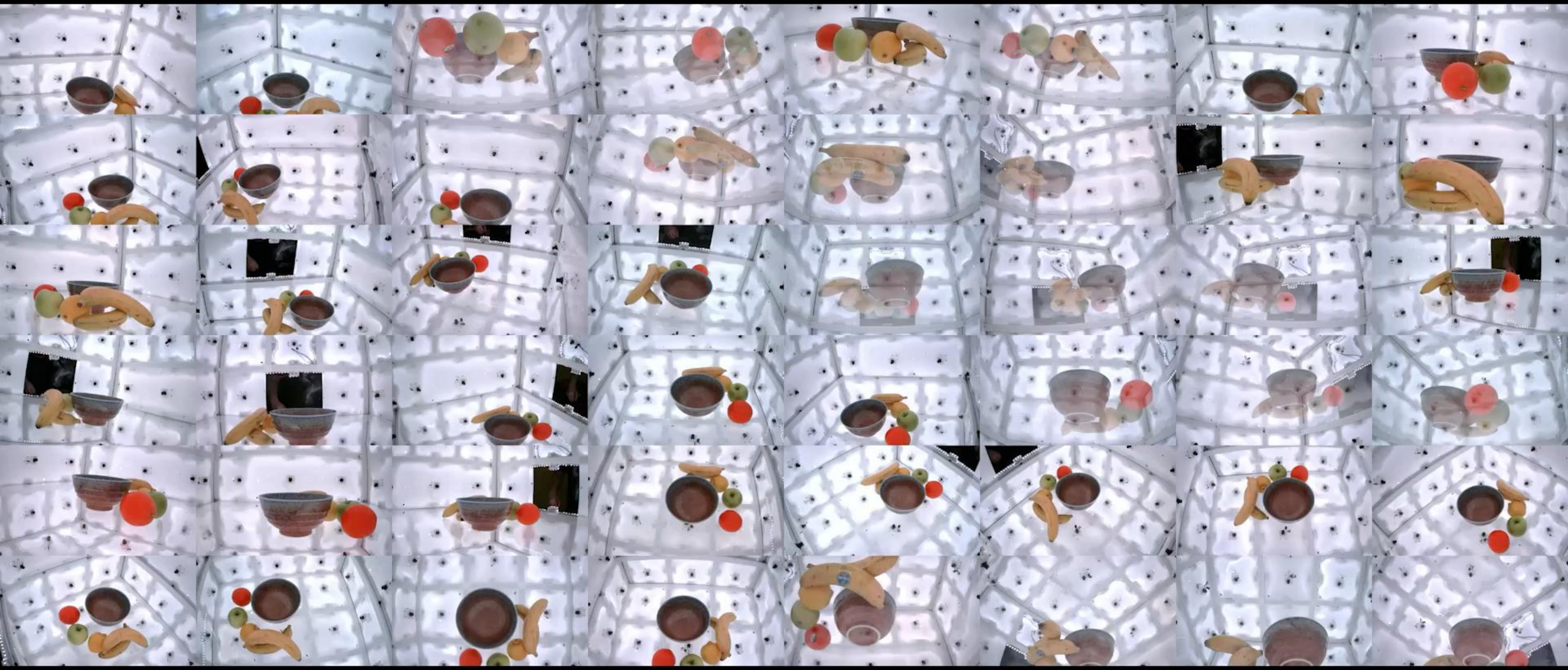
Evelyne Kohler,[1] Christian Keysers,[1] M. Alessandra Umiltà,[1] Leonardo Fogassi,[2] Vittorio Gallese,[1] Giacomo Rizzolatti[1]*

Many object-related actions can be recognized by their sound. We found neurons in monkey premotor cortex that discharge when the animal performs a specific action and when it hears the related sound. Most of the neurons also discharge when the monkey observes the same action. These audiovisual mirror neurons code actions independently of whether these actions are performed, heard, or seen. This discovery in the monkey homolog of Broca's area might shed light on the origin of language: audiovisual mirror neurons code abstract contents—the meaning of actions—and have the auditory access typical of human language to these contents.

the bowl; the fruit includes an orange, a green apple, a lemon, and two bananas

put_fruit

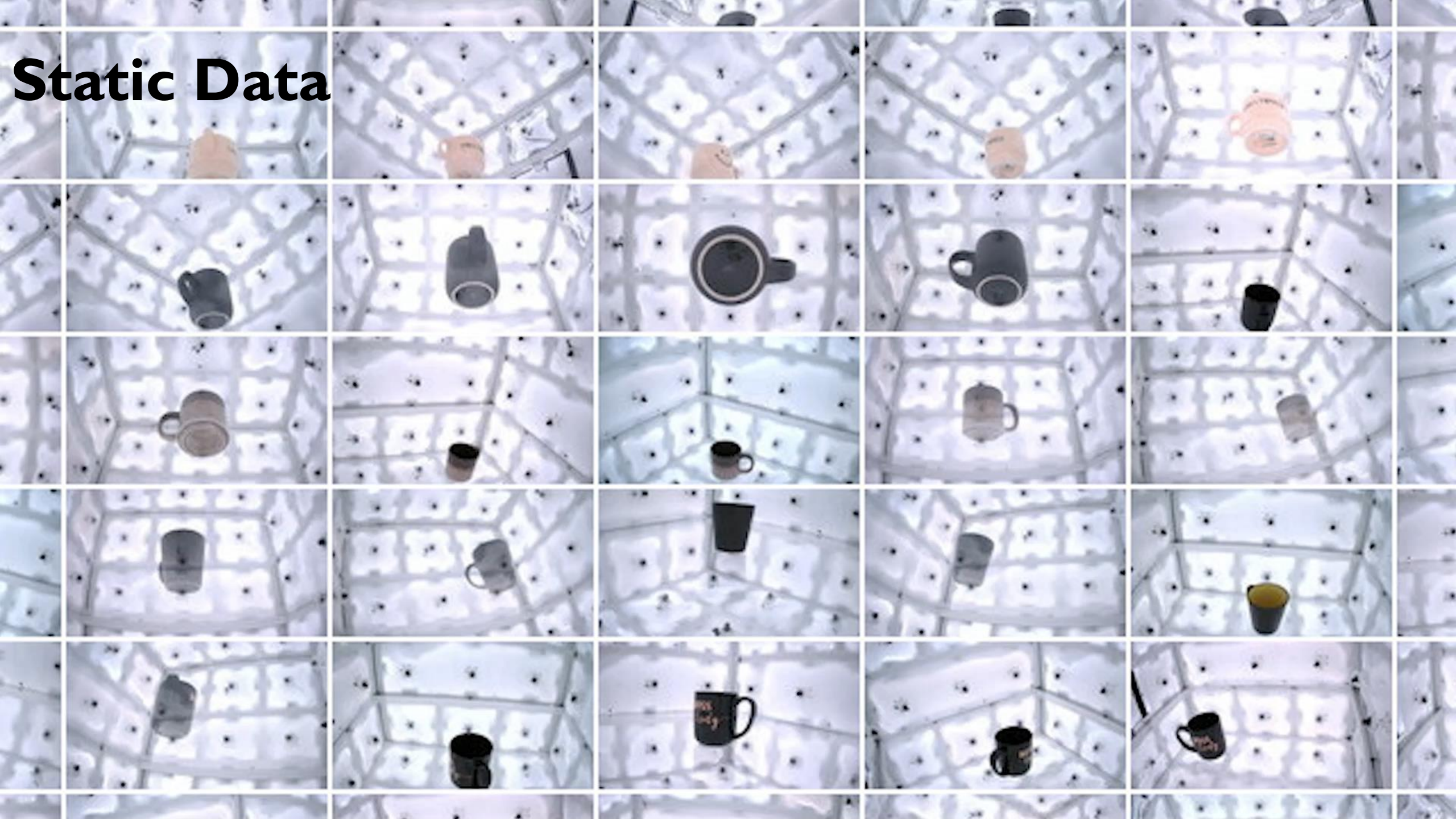
only 46/53 views shown
only 1/6 audio channels played
showing level 3 text descriptions



a hand picking up assorted fruit around a brown painted ceramic bowl and placing them into the bowl; the fruit includes an orange, a green apple, a lemon, and two bananas

Static Data

Static Data

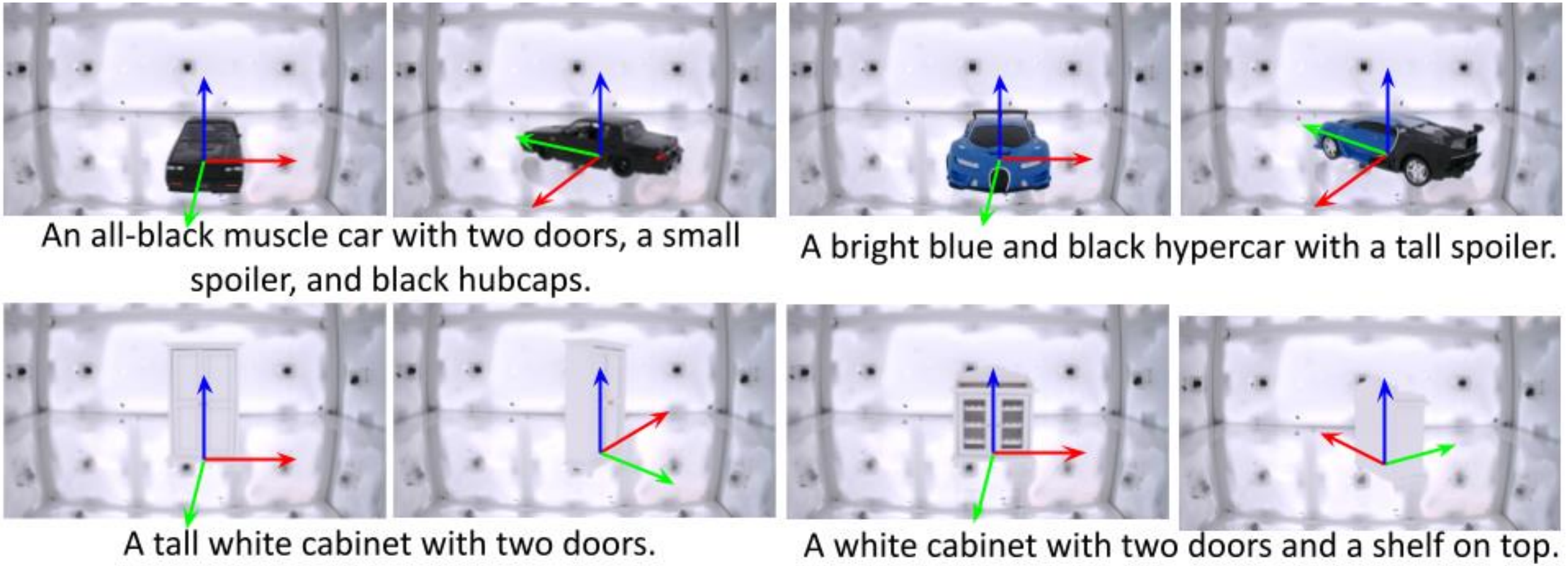


Additional Information

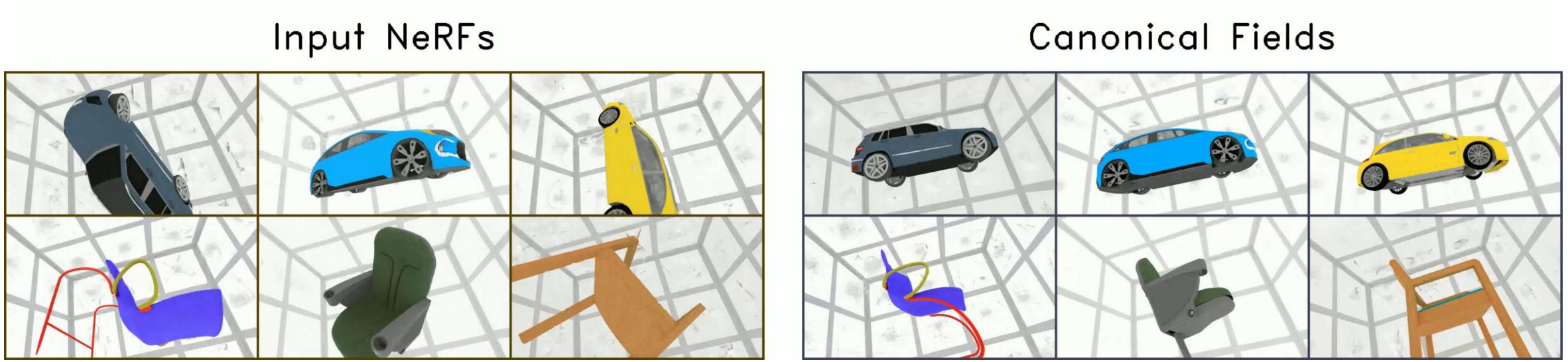
- FG-BG segmentation masks and 6 DoF pose

Additional Information

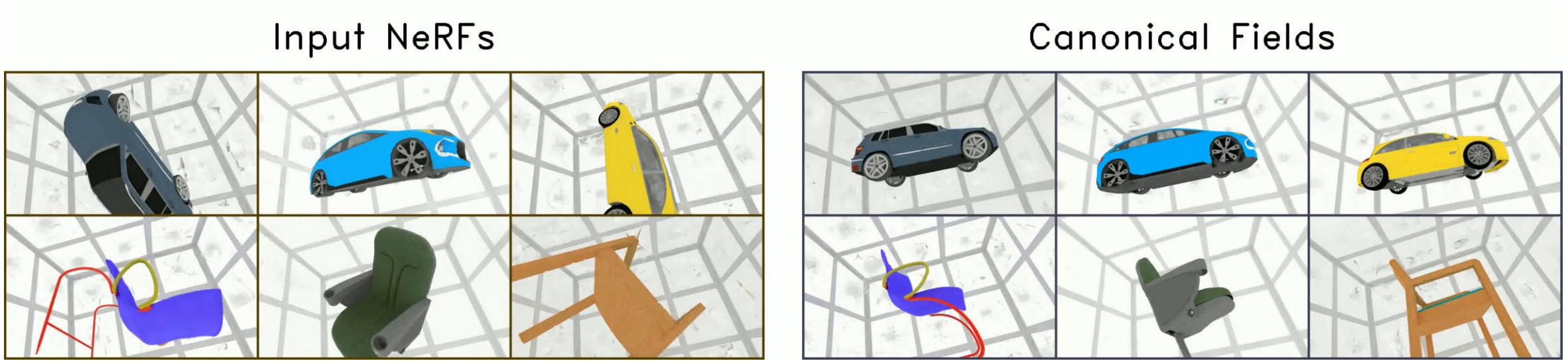
- FG-BG segmentation masks and 6 DoF pose



Detour: Self-Supervised Canonicalization of Fields



Detour: Self-Supervised Canonicalization of Fields



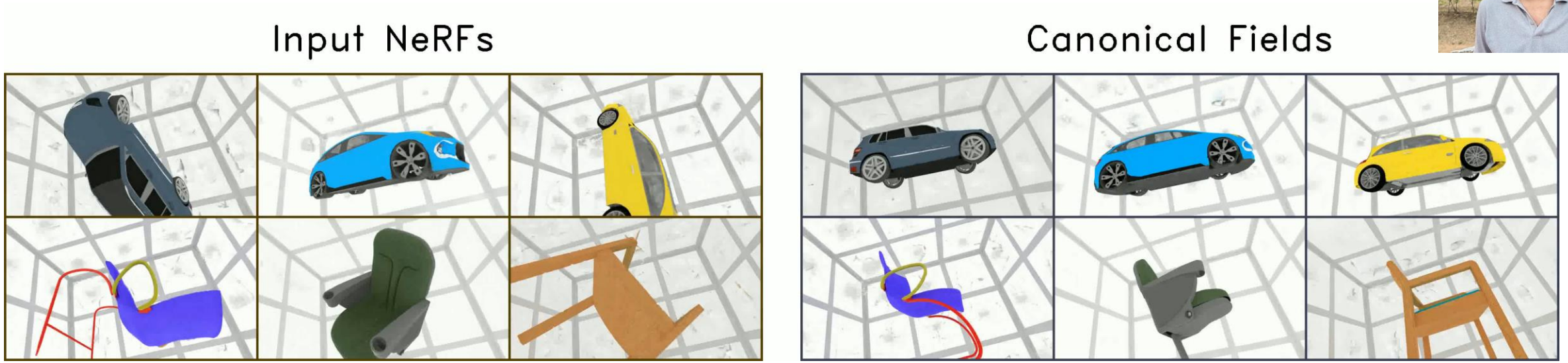
Highlight

Canonical Fields: Self-Supervised Learning of Pose-Canonicalized Neural Fields

Rohith Agaram, Shaurya Dewan, Rahul Sajnani, Adrien Poulenard, Madhava Krishna, Srinath Sridhar

CVPR 2023

Detour: Self-Supervised Canonicalization of Fields



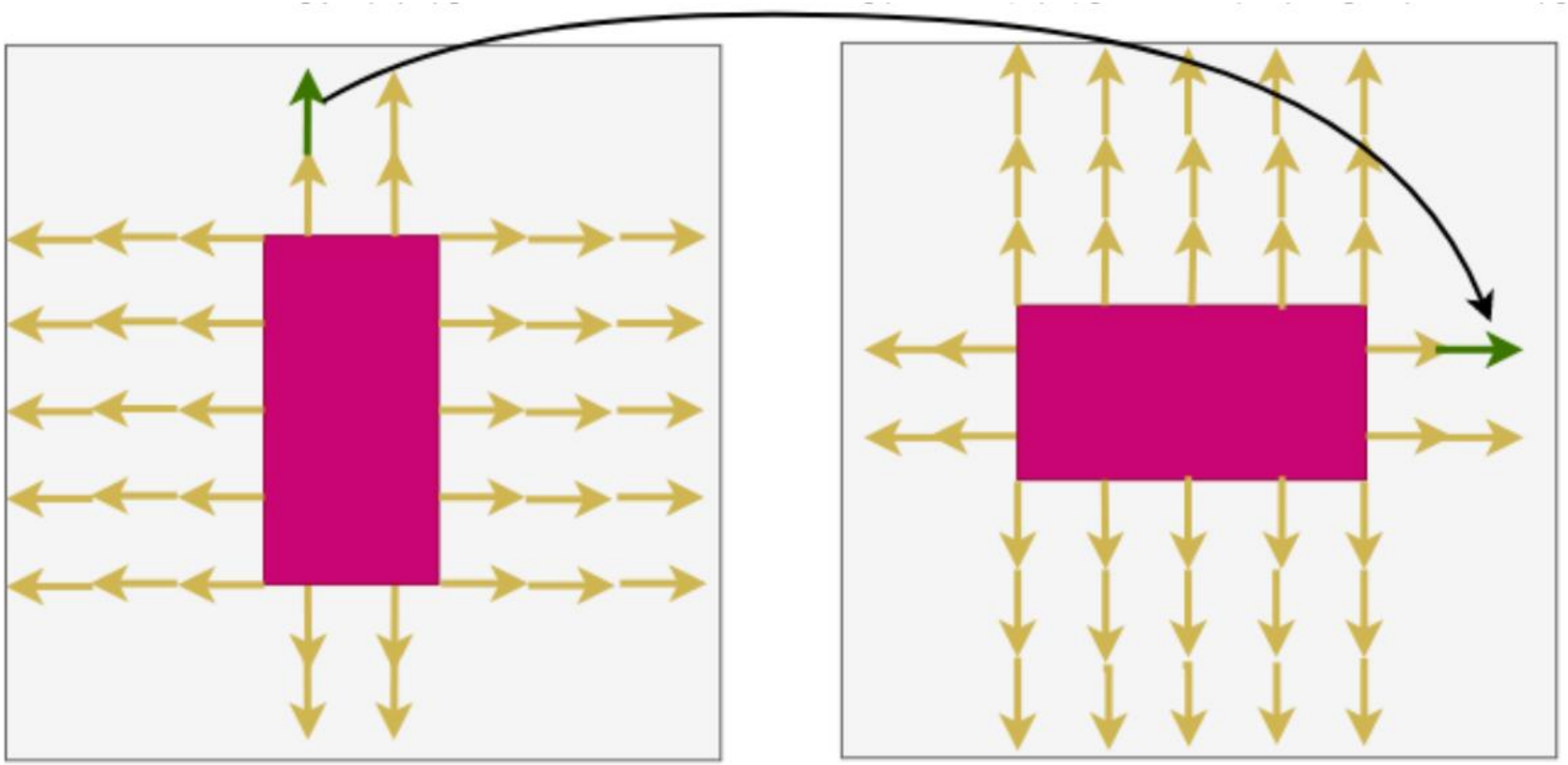
Canonical Fields: Self-Supervised Learning of Pose-Canonicalized Neural Fields

Rohith Agaram, Shaurya Dewan, Rahul Sajnani, Adrien Poulenard, Madhava Krishna, Srinath Sridhar

Highlight

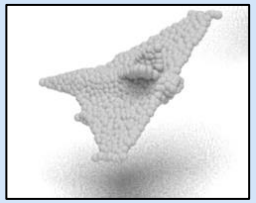
CVPR 2023

Rotation Equivariance in Vector Fields



Canonical Fields: Self-Supervised Learning of Pose-Canonicalized Neural Fields, Agaram et al., CVPR 2023 (**Highlight**)

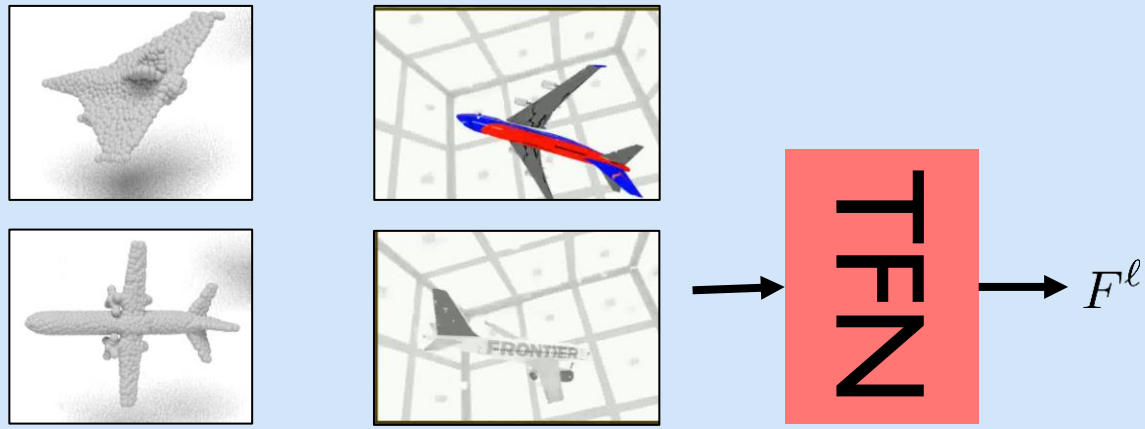
Rotation Canonicalization



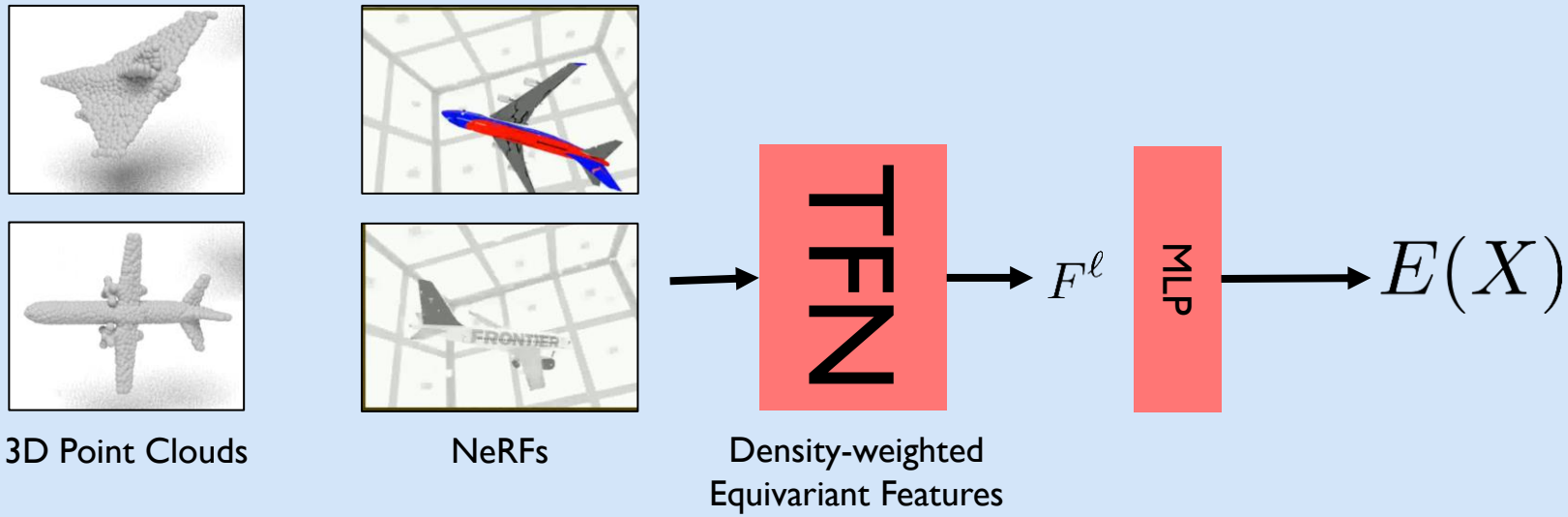
3D Point Clouds

NeRFs

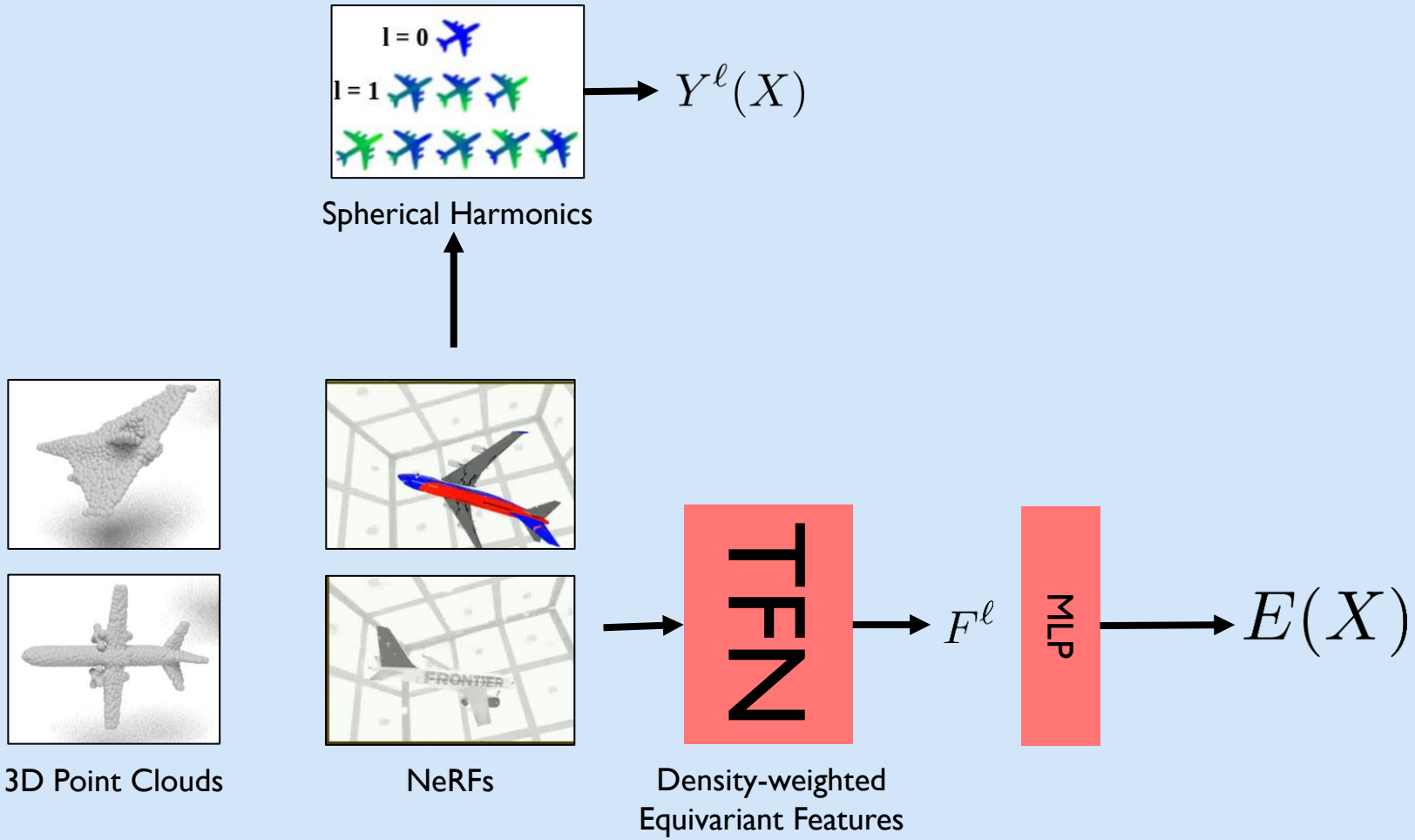
Rotation Canonicalization



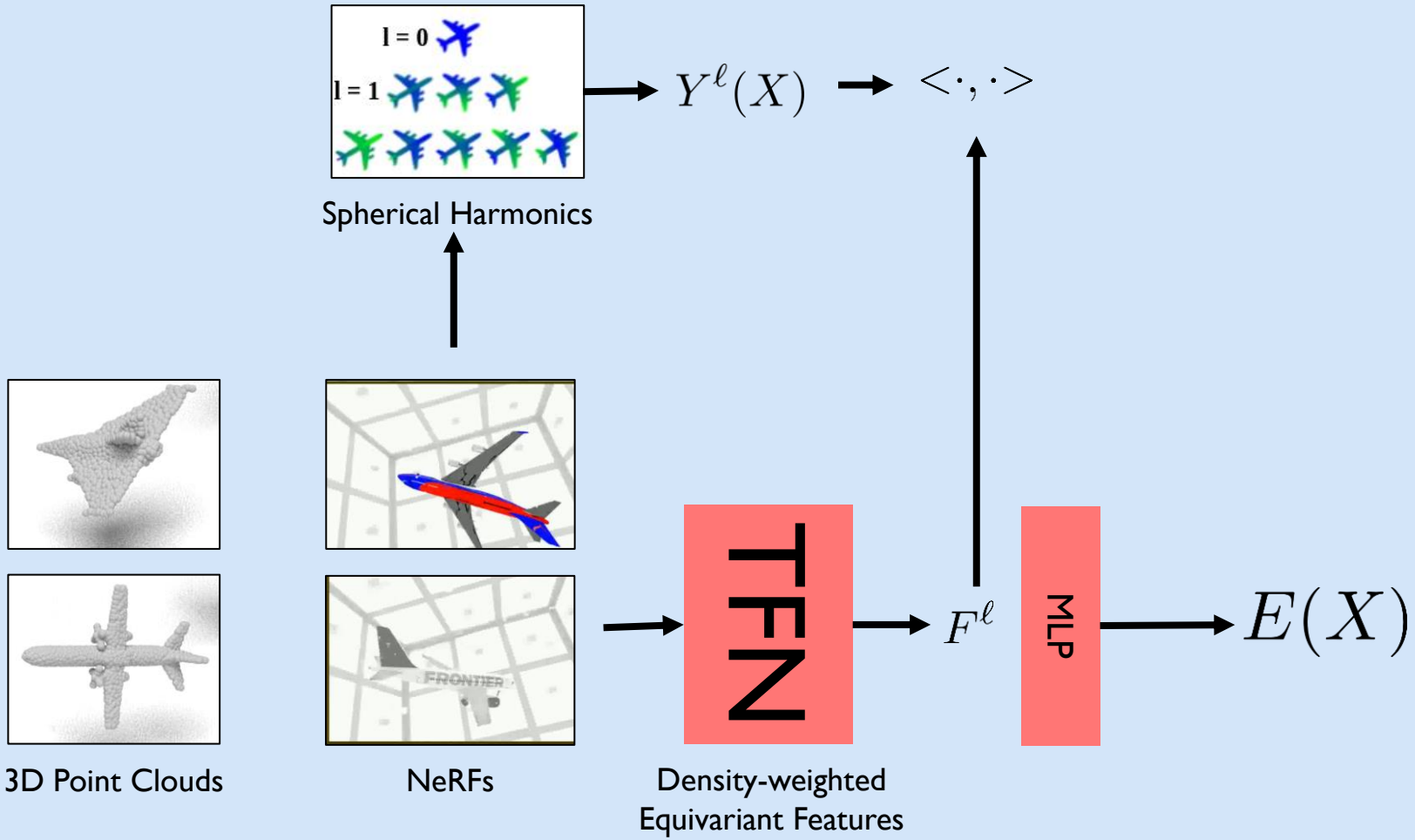
Rotation Canonicalization



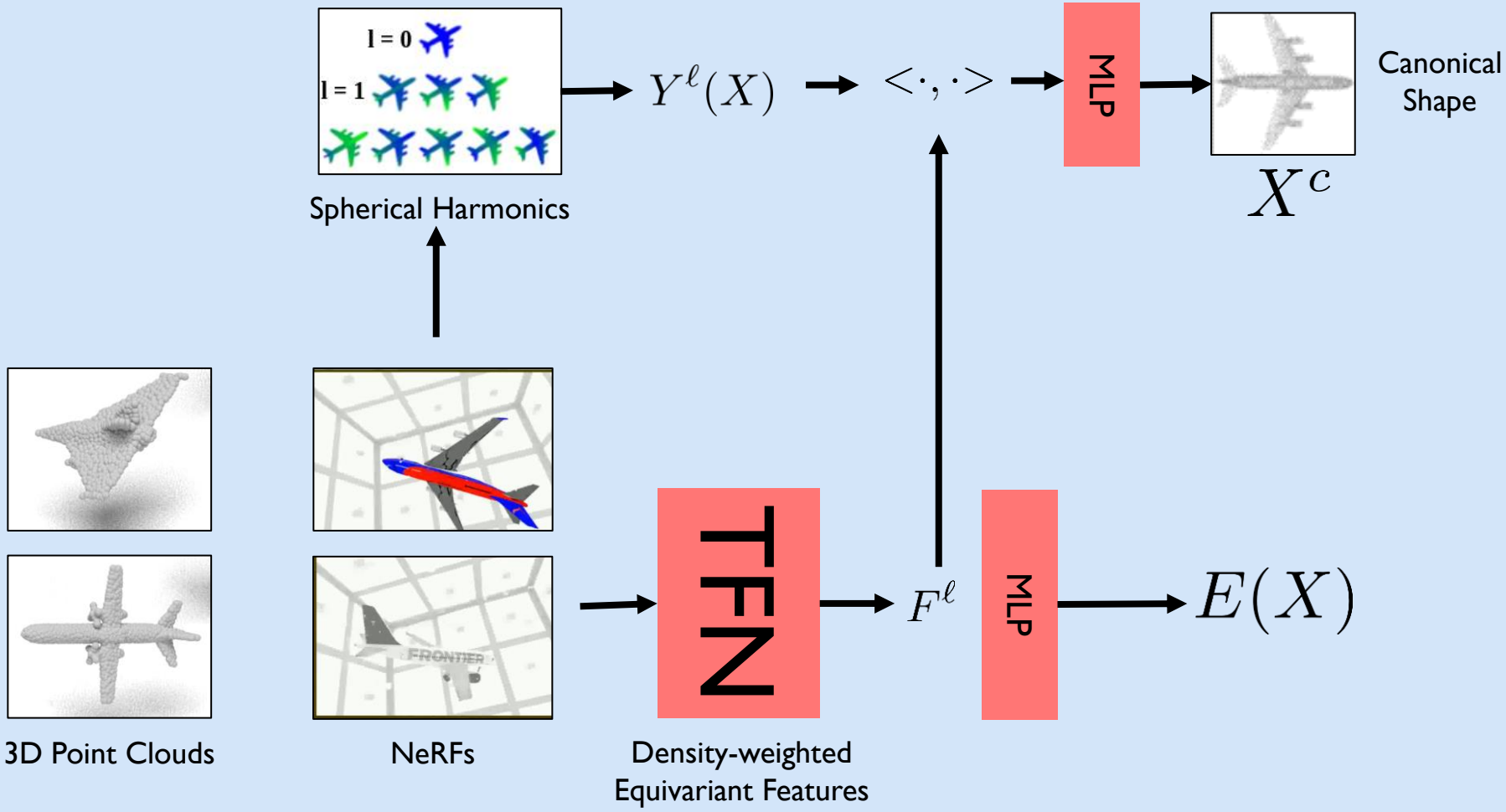
Rotation Canonicalization



Rotation Canonicalization

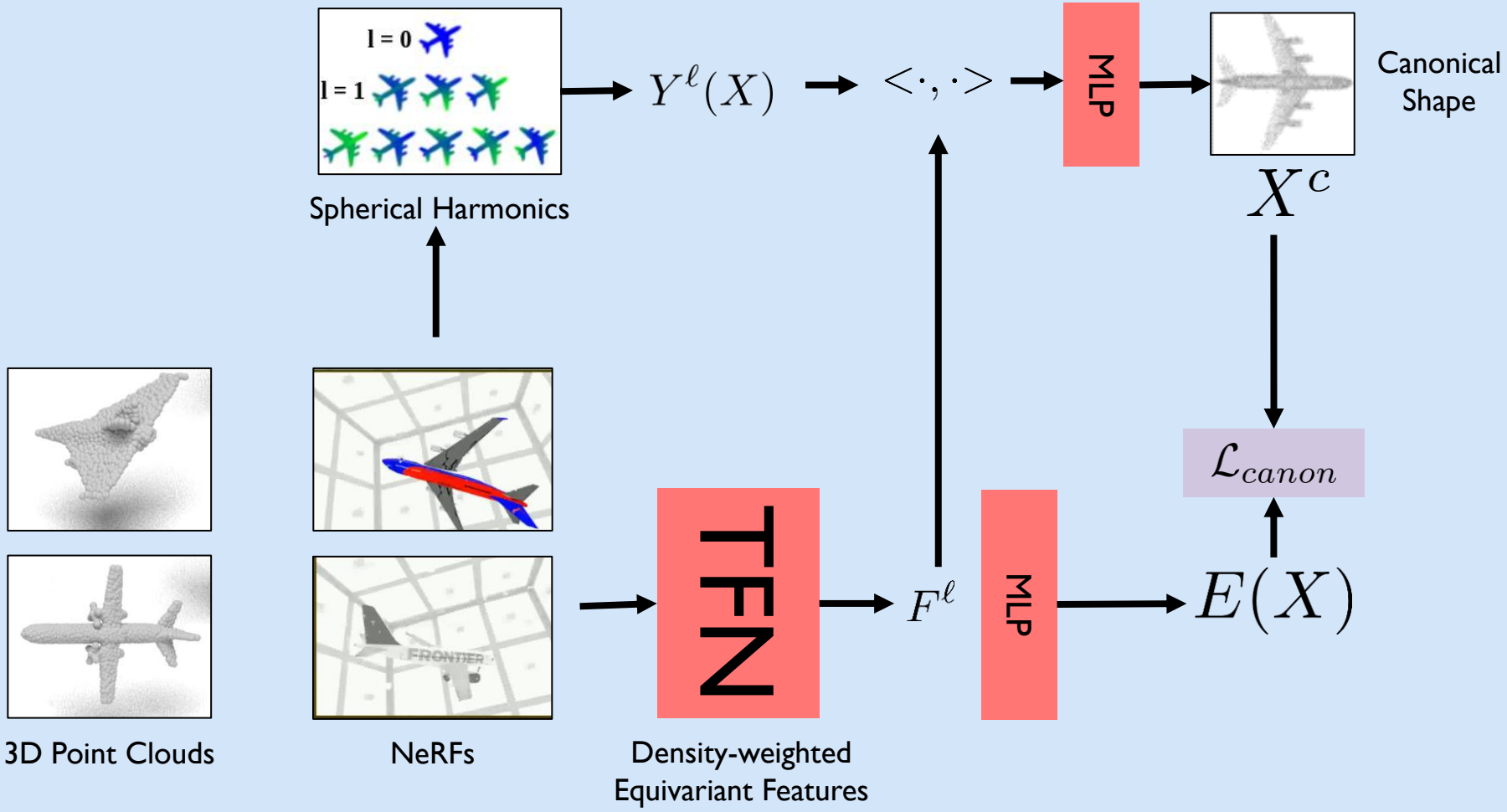


Rotation Canonicalization



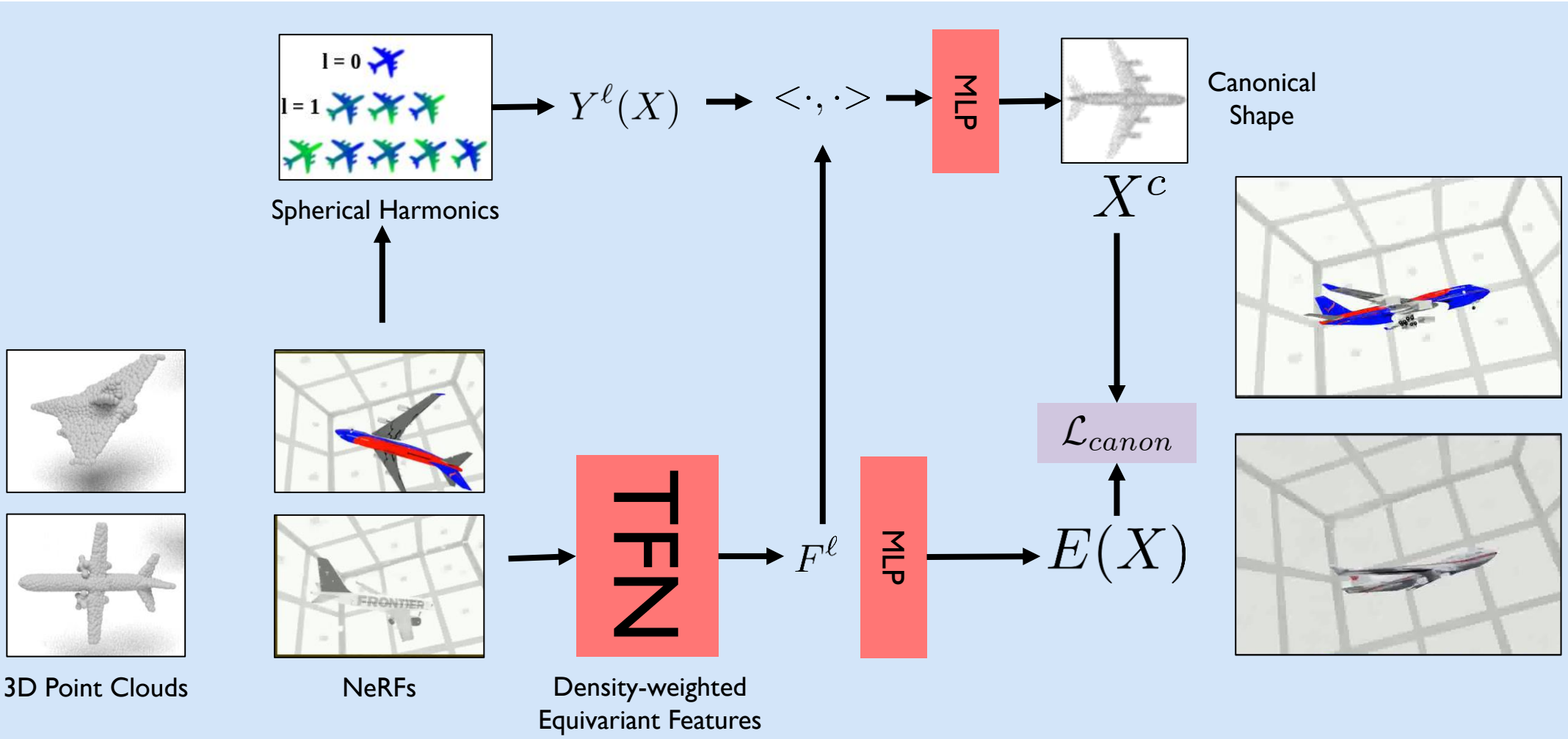
Rotation Canonicalization

$$\mathcal{L}_{canon} = \frac{1}{K} \sum_i \|EX_i^c - X_i\|_2.$$



Rotation Canonicalization

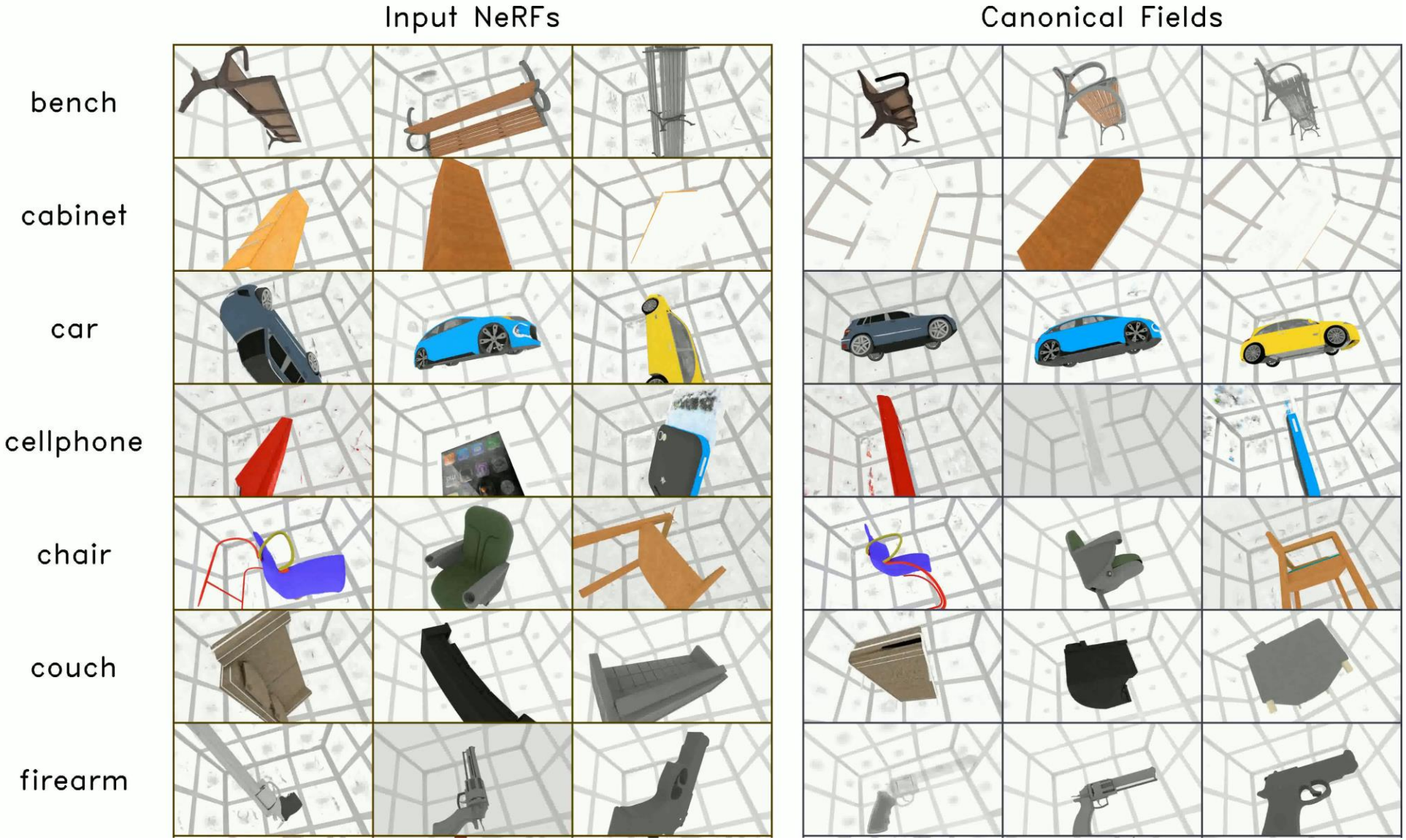
$$\mathcal{L}_{canon} = \frac{1}{K} \sum_i \|EX_i^c - X_i\|_2.$$



Results: Neural Radiance Fields

Canonical Fields: Self-Supervised Learning of Pose-Canonicalized Neural Fields, Agaram et al., CVPR 2023 (**Highlight**)

Results: Neural Radiance Fields



Canonical Fields: Self-Supervised Learning of Pose-Canonicalized Neural Fields, Agaram et al., CVPR 2023 (**Highlight**)

DiVA-360: The Dynamic Visuo-Audio Dataset

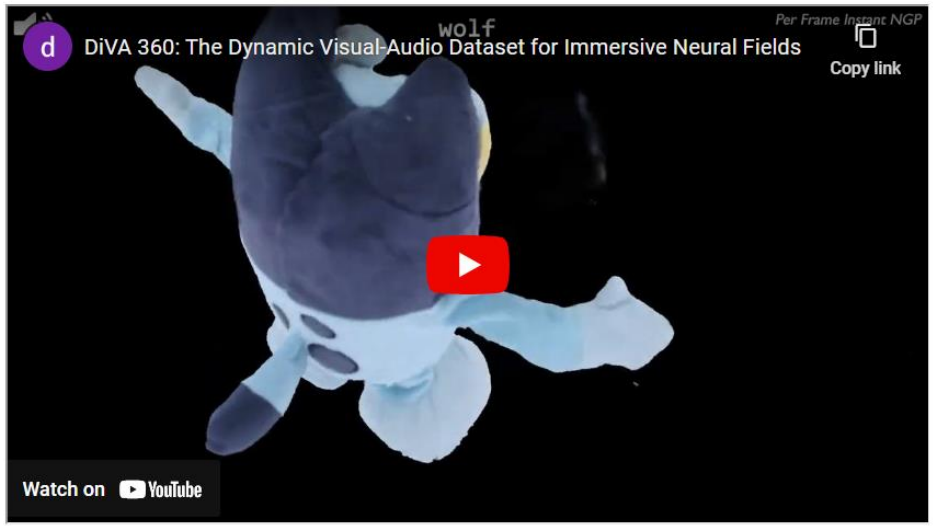
diva360.github.io

DiVA360: The Dynamic Visual-Audio Dataset for Immersive Neural Fields

Anonymous Authors

Paper

Video



Benchmarking Metrics

- Dynamic Data (per-frame on 6 held-out views)

Benchmarking Metrics

- Dynamic Data (per-frame on 6 held-out views)

Metrics	Methods
Peak Signal-to-Noise Ratio (PSNR) Structural Similarity Index Measure (SSIM) Learned Perceptual Image Patch Similarity (LPIPS) Just Objectionable Difference (JOD)	Instant-NGP (Mueller et al. 2022) MixVoxels (Wang et al. 2022)

Benchmarking Metrics

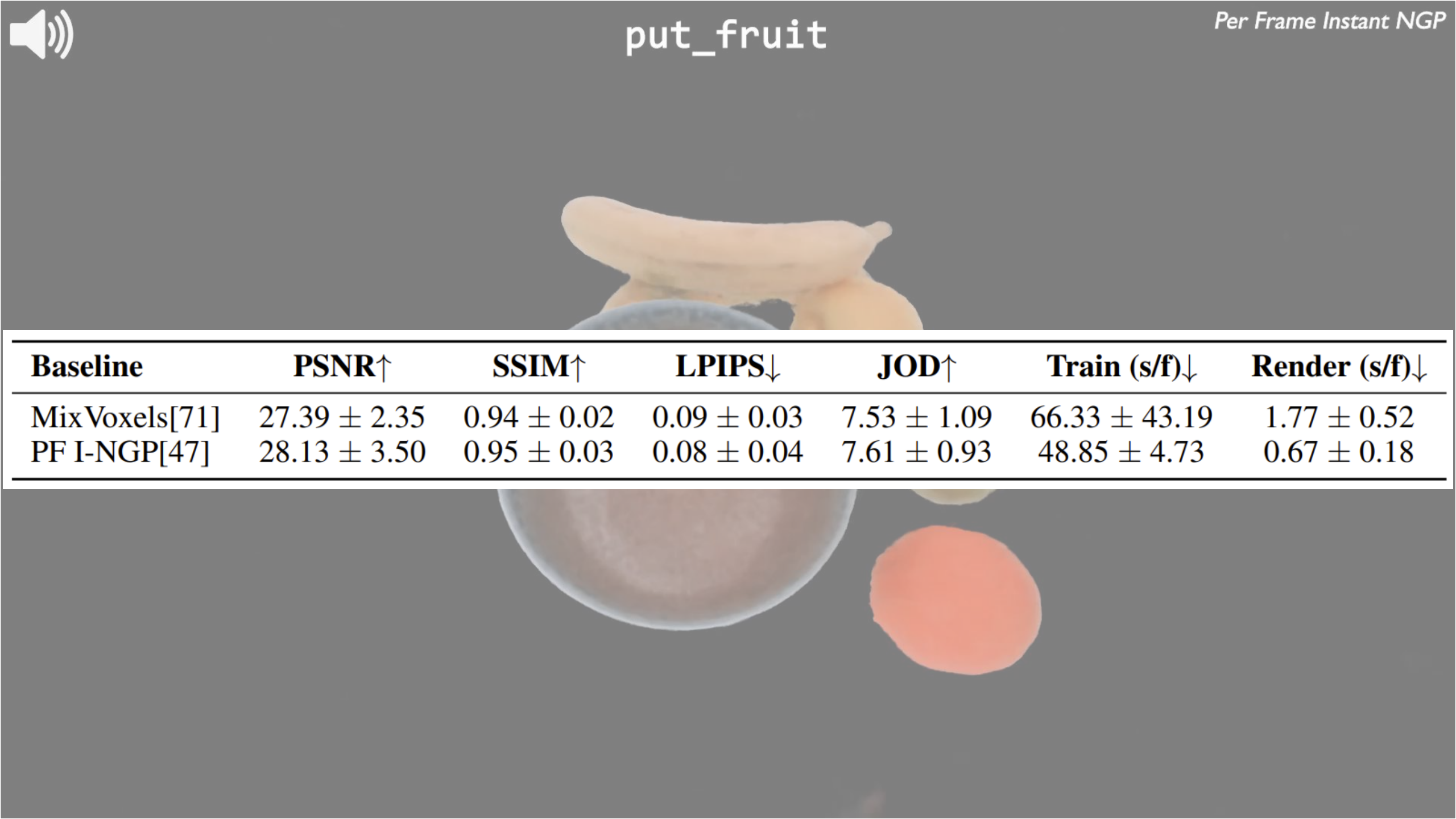
- Dynamic Data (per-frame on 6 held-out views)

Metrics	Methods
Peak Signal-to-Noise Ratio (PSNR) Structural Similarity Index Measure (SSIM) Learned Perceptual Image Patch Similarity (LPIPS) Just Objectionable Difference (JOD)	Instant-NGP (Mueller et al. 2022) MixVoxels (Wang et al. 2022)

- Static Data

Metrics	Methods
PSNR, SSIM, LPIPS, JOD 6DoF Pose Canonicalization (Sajnani et al. 2022)	Instant-NGP (Mueller et al. 2022) CaFi-Net (Agaram et al. 2023)


put_fruit
Per Frame Instant NGP



Baseline	PSNR↑	SSIM↑	LPIPS↓	JOD↑	Train (s/f)↓	Render (s/f)↓
MixVoxels[71]	27.39 ± 2.35	0.94 ± 0.02	0.09 ± 0.03	7.53 ± 1.09	66.33 ± 43.19	1.77 ± 0.52
PF I-NGP[47]	28.13 ± 3.50	0.95 ± 0.03	0.08 ± 0.04	7.61 ± 0.93	48.85 ± 4.73	0.67 ± 0.18


put_fruit
Per Frame Instant NGP

car00_random

Instant NGP

car00_random

Instant NGP

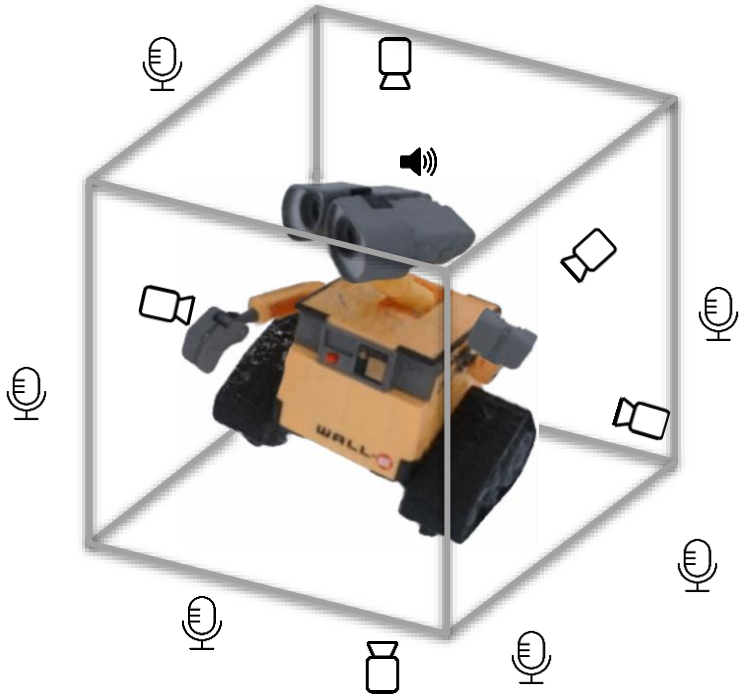
Category	PSNR ↑	SSIM* ↑	LPIPS* ↓	Category	PSNR ↑	SSIM* ↑	LPIPS* ↓
Chair	32.95±2.11	98±0.9	5±2.3	Keyboard	31.57±1.58	97±0.8	7±1.0
Table	38.11±1.88	99±0.3	2±0.6	Car	34.06±1.40	98±0.4	3±0.5
Cabinet	38.02±1.65	99±0.3	3±0.8	Couch	32.87±1.51	98±0.6	4±0.9
Mug	32.98±2.12	98±0.6	4±0.7	Plane	34.04±3.30	98±0.8	4±2.1
Fruit	34.91±2.36	99±0.3	3±0.8	Utensil	36.45±3.19	99±0.4	6±2.9
Mouse	35.03±2.40	99±0.2	3±0.4	Scenes	27.41±2.48	95±2.0	8±2.7

Categories	CC ↓	IC ↓	Categories	CC↓	IC ↓
chair	0.0411	0.0203	keyboard	0.1396	0.0719
table	0.0630	0.0292	car	0.0626	0.0224
cabinet	0.0736	0.0374	couch	0.0604	0.0343
mouse	0.0443	0.0494	plane	0.0538	0.0443
utensil	0.1536	0.1134			

DiVA-360: The Dynamic Visuo-Audio Dataset

53× cam 6× mic

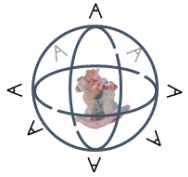
audio-visual scenes



DiVA-360 Dynamic Dataset

DiVA-360 Static Dataset

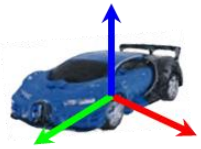
360° View



Text Description

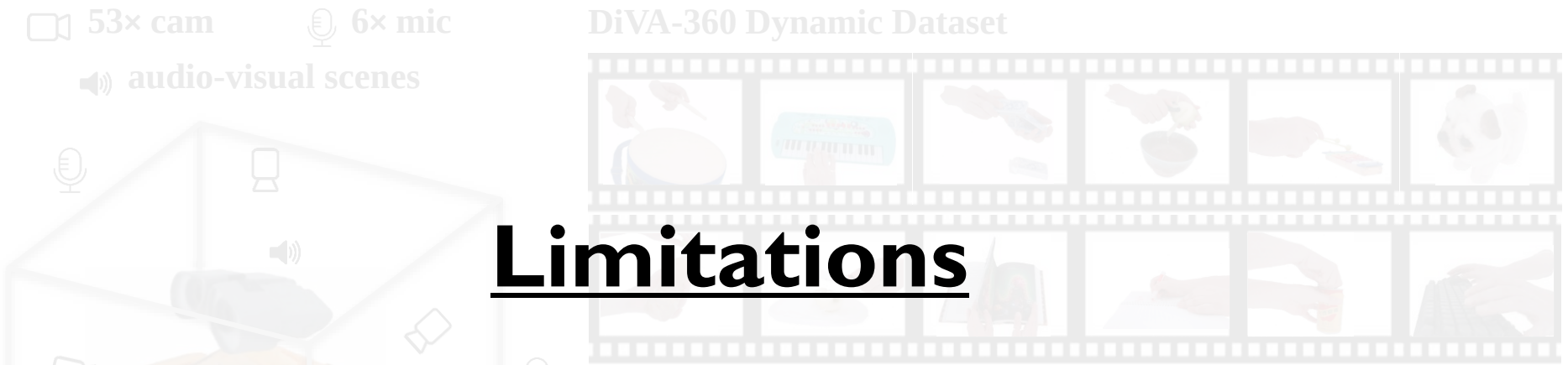
"a hypercar with a bright blue front and black back, two doors, white accent lines, a tall racing spoiler, and an engine visible behind the driver's seat."

Canonicalized



Spatial Audio

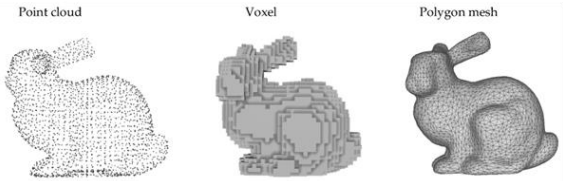
DiVA-360: The Dynamic Visuo-Audio Dataset

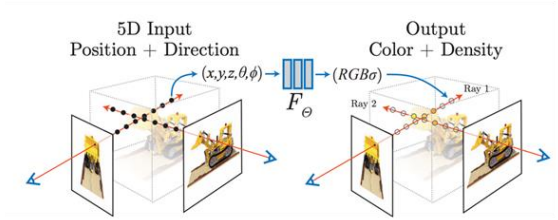


Limitations

Focus not on number of objects/categories
Benchmarking metrics limited to images
Not (yet) open world

Summary

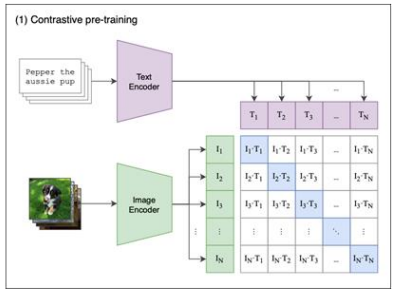


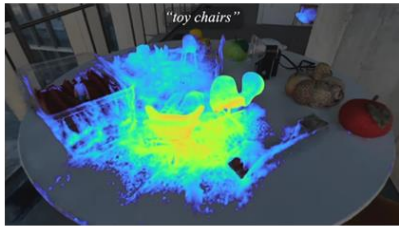


NeRF, Mildenhall et al. 2020

3D Representations

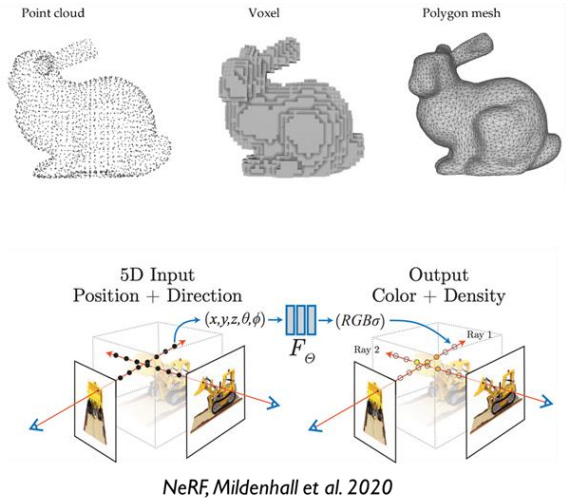
Dynamic 3D Understanding



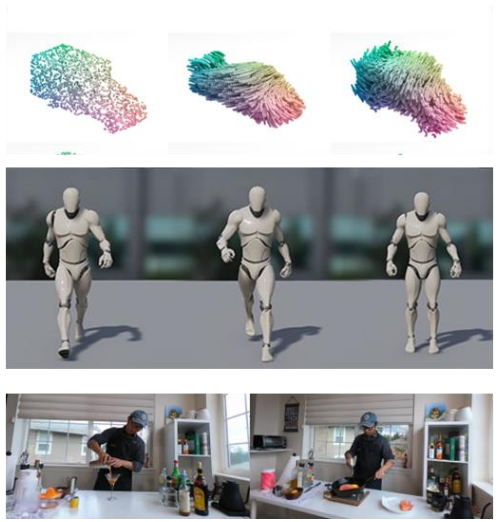


Multimodality

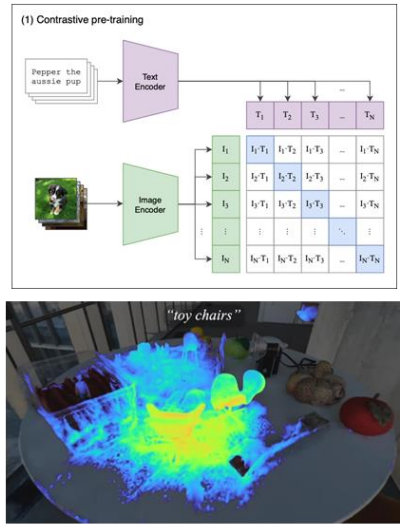
Summary



3D Representations

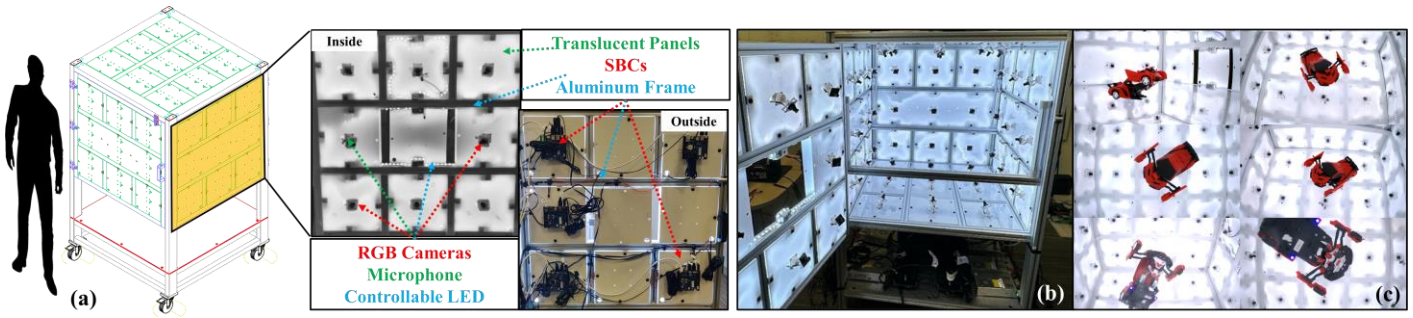


Dynamic 3D Understanding

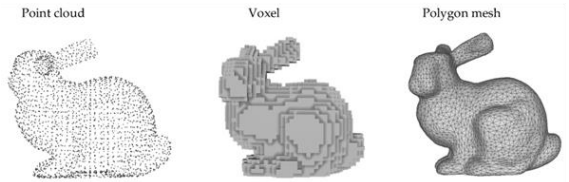


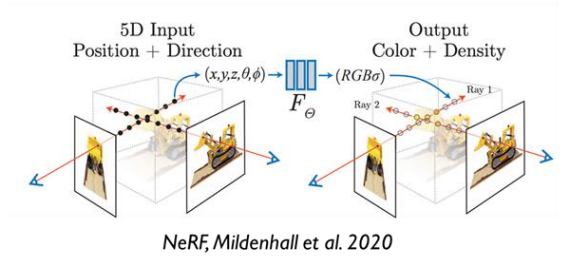
Multimodality

Building Hardware/Software for Multimodal Capture



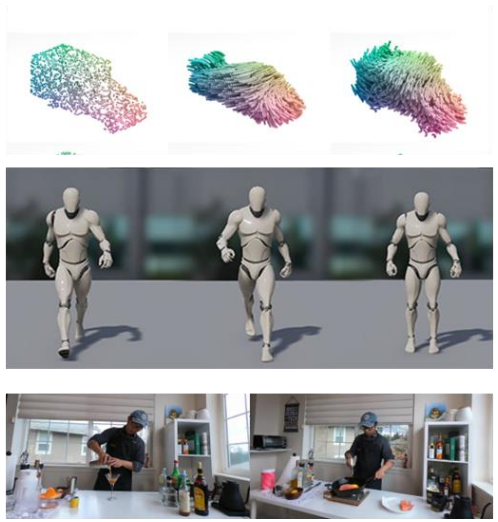
Summary



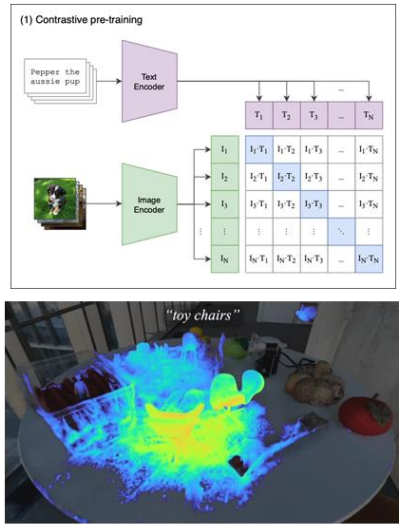


NeRF, Mildenhall et al. 2020

3D Representations

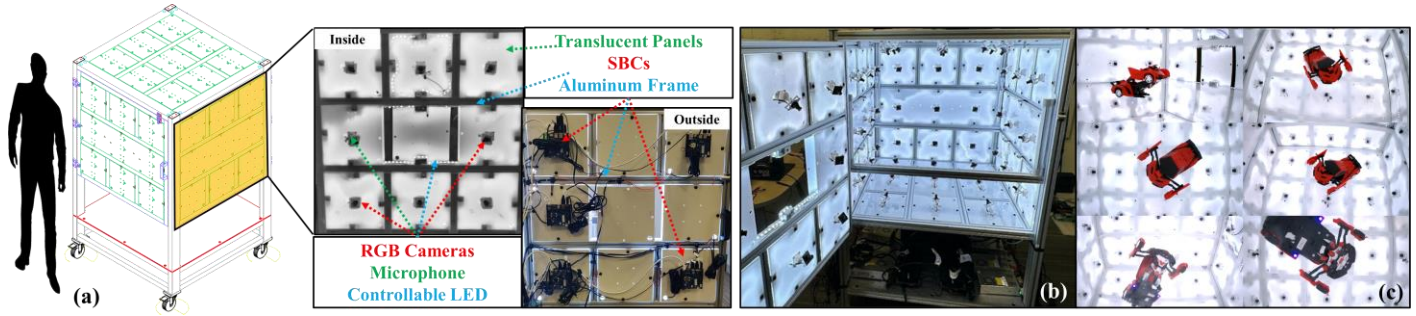


Dynamic 3D Understanding



Multimodality

Building Hardware/Software for Multimodal Capture



Future Work

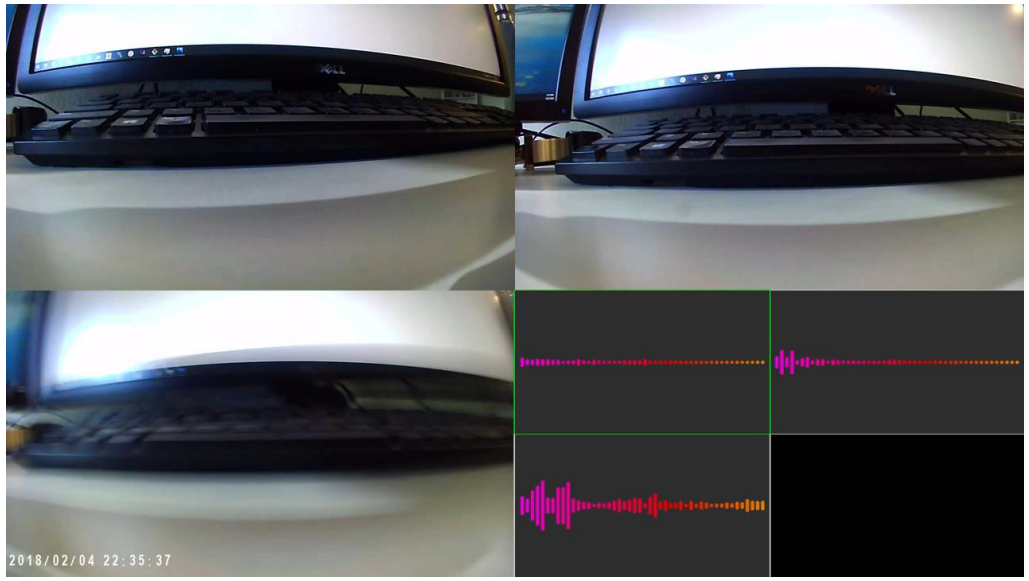
Future Work

Large-Scale Open-World Reasoning
Other Modalities: Touch, etc.

Tschernezki, Darkhalil, et al., EPIC Fields: Marrying 3D Geometry and Video Understanding

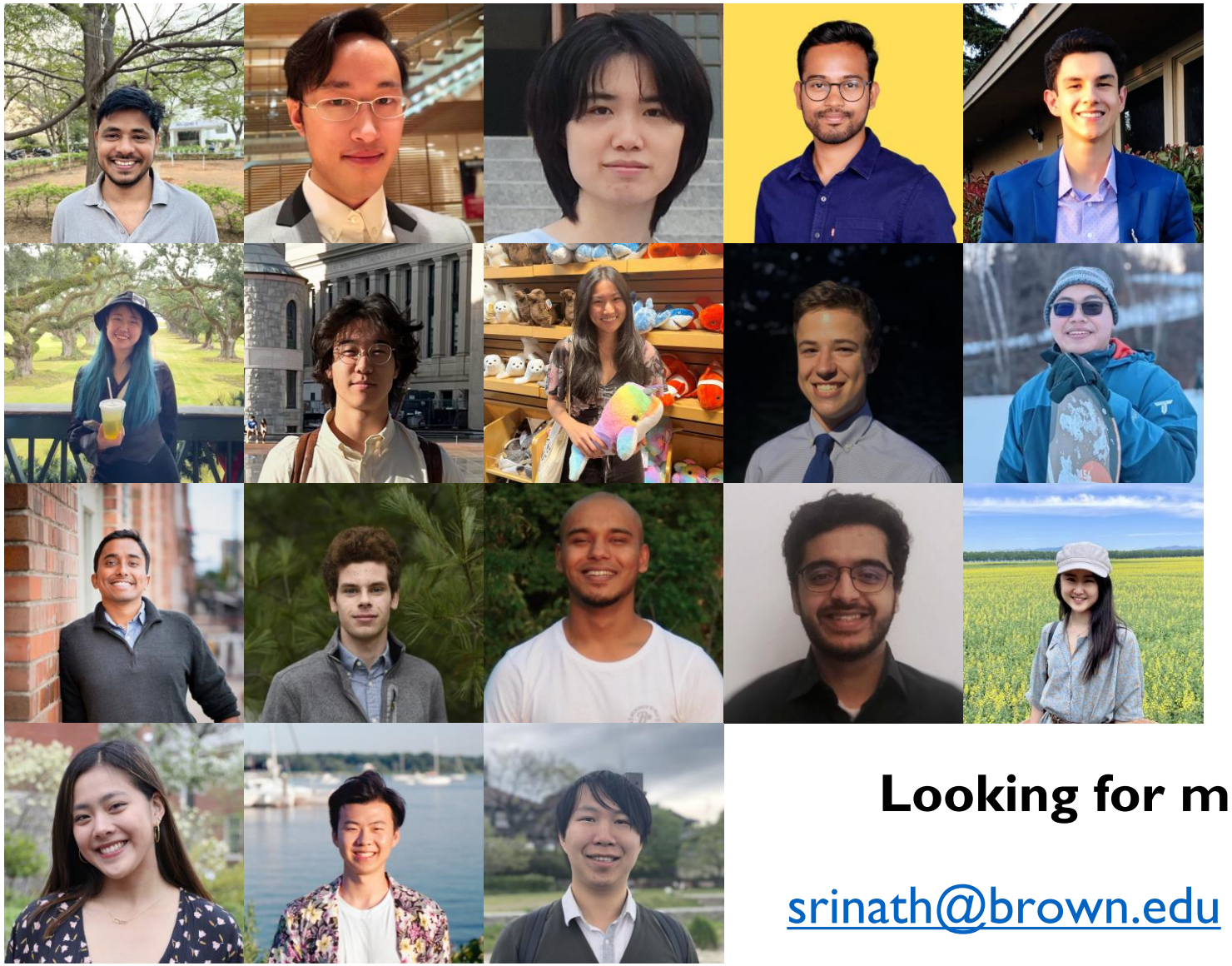
Future Work

Large-Scale Open-World Reasoning
Other Modalities: Touch, etc.



Tschernezki, Darkhalil, et al., EPIC Fields: Marrying 3D Geometry and Video Understanding

Thank you!



Rohith Agaram | Kefan Chen | Yiwen Chen | Arnab Dey | Jacob Frausto | Rao Fu | Dylan Hu | Iris Huang | Filip Kierzenka | Cheng-You Lu | Rugved Mavidipalli | Theo McArn | Chandradeep Pokhariya | Rahul Sajnani | Qiuhong Wei | Angela Xing | Xiao Zhan | Peisen Zhou

Looking for motivated students / postdocs!

srinath@brown.edu | @drsrinathsridha | ivl.cs.brown.edu