Retreival Augmentation for 3D Vision in the Long Tail

Sagie Benaim

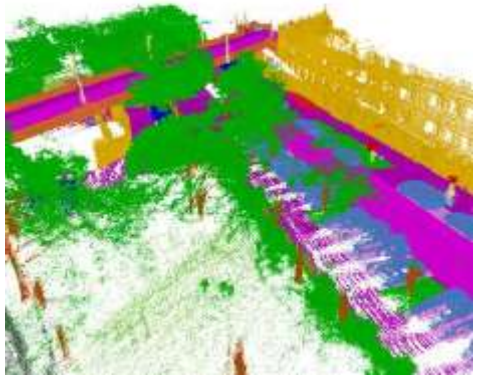
Hebrew University of Jerusalem, Amazon

Work done at the Hebrew University of Jerusalem

3D Vision

3D Tracking

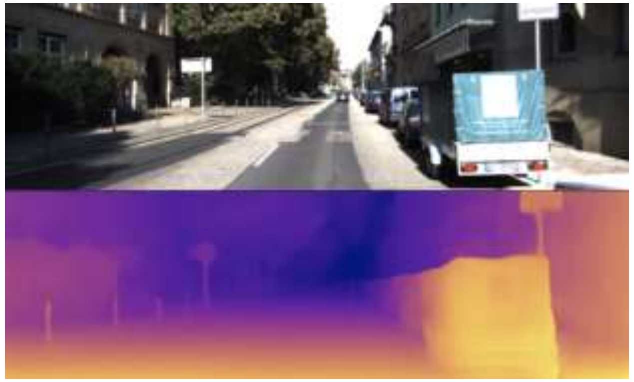
3D Segmentation



3D Reconstruction

3D Generation

Depth Estimation



Current Success is Driven By **Large Datasets**

3D Datasets

2D Datasets

Not All Data Can Be Collected

Generate
Rare 3D Objects

Estimate Depth of
Unexpected Scenarios

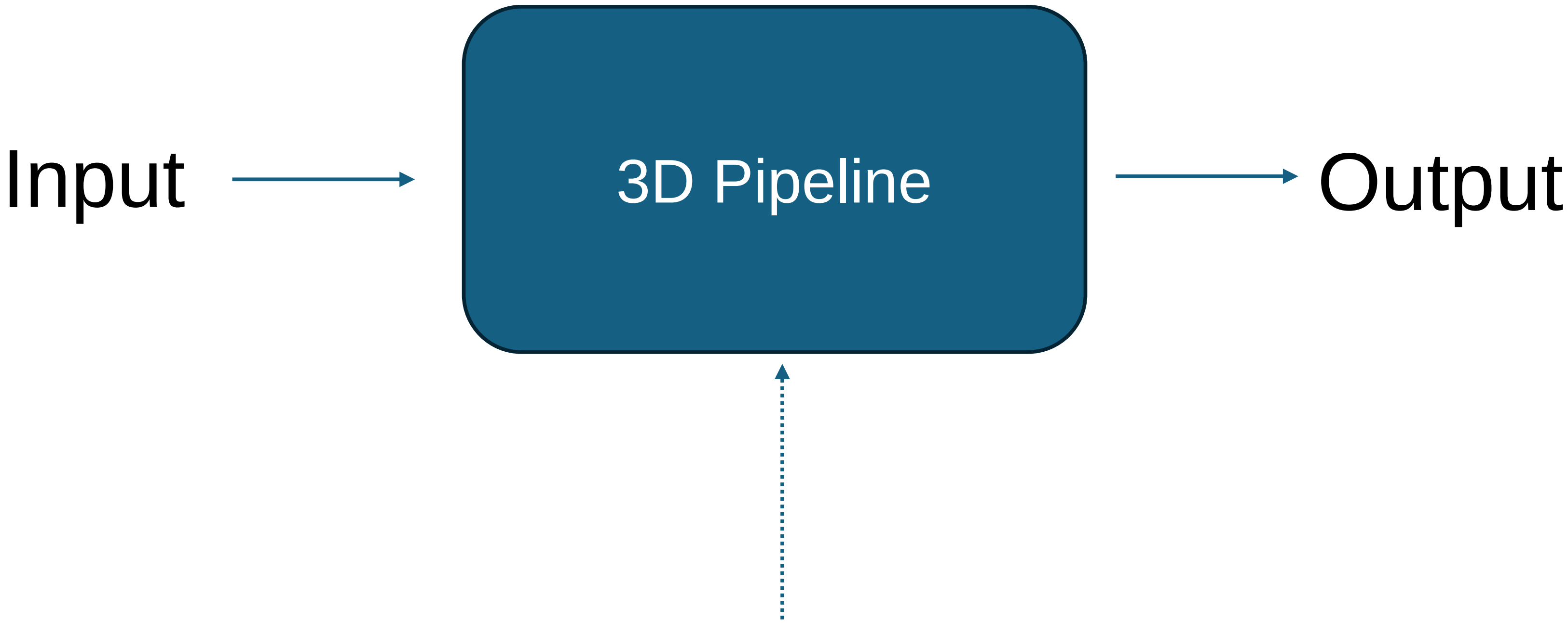
Reconstruct
Personal Objects

Long Tail of Data Distribution

The picture can't be displayed.

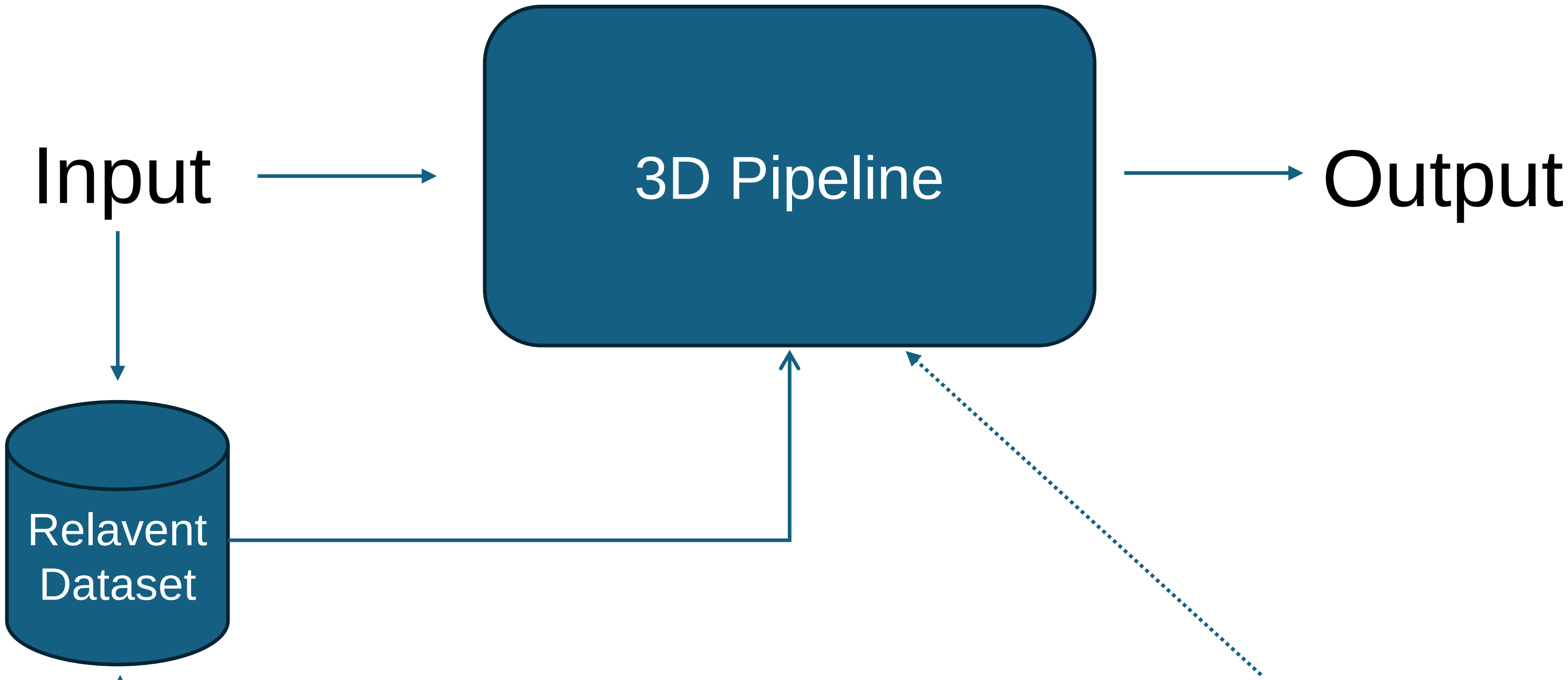
How Can We Tackle Such Cases?

Current Pipelines



Struggles to memorize long tail data

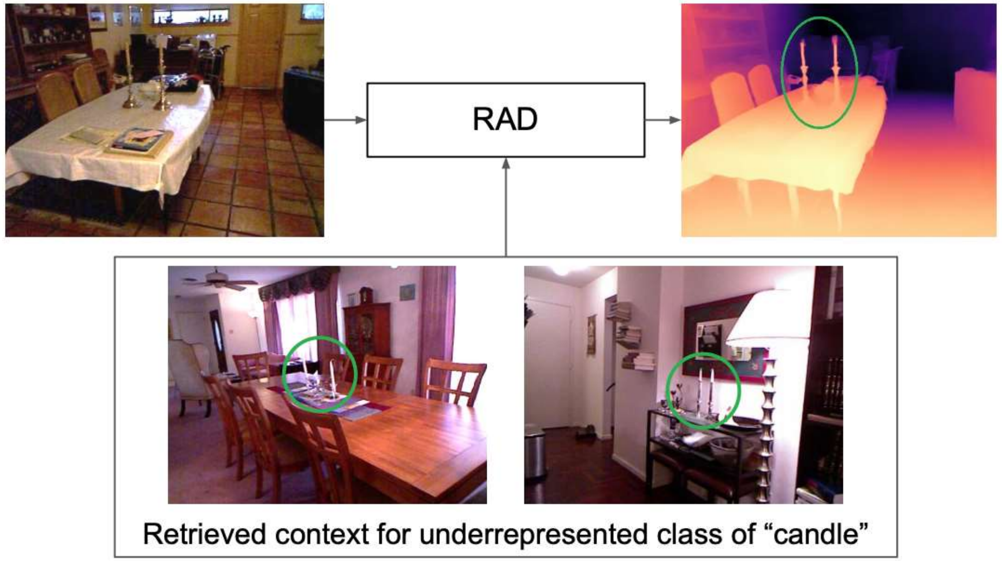
Main Idea: Retrival Augmentation



Focus on leveraging retrieved data.
No need to memorize long tail data!

Retrieval Augmentation For 3D

Depth Estimation



3D Generation

RAD: Retrieval-Augmented Monocular Metric Depth Estimation for Underrepresented Classes

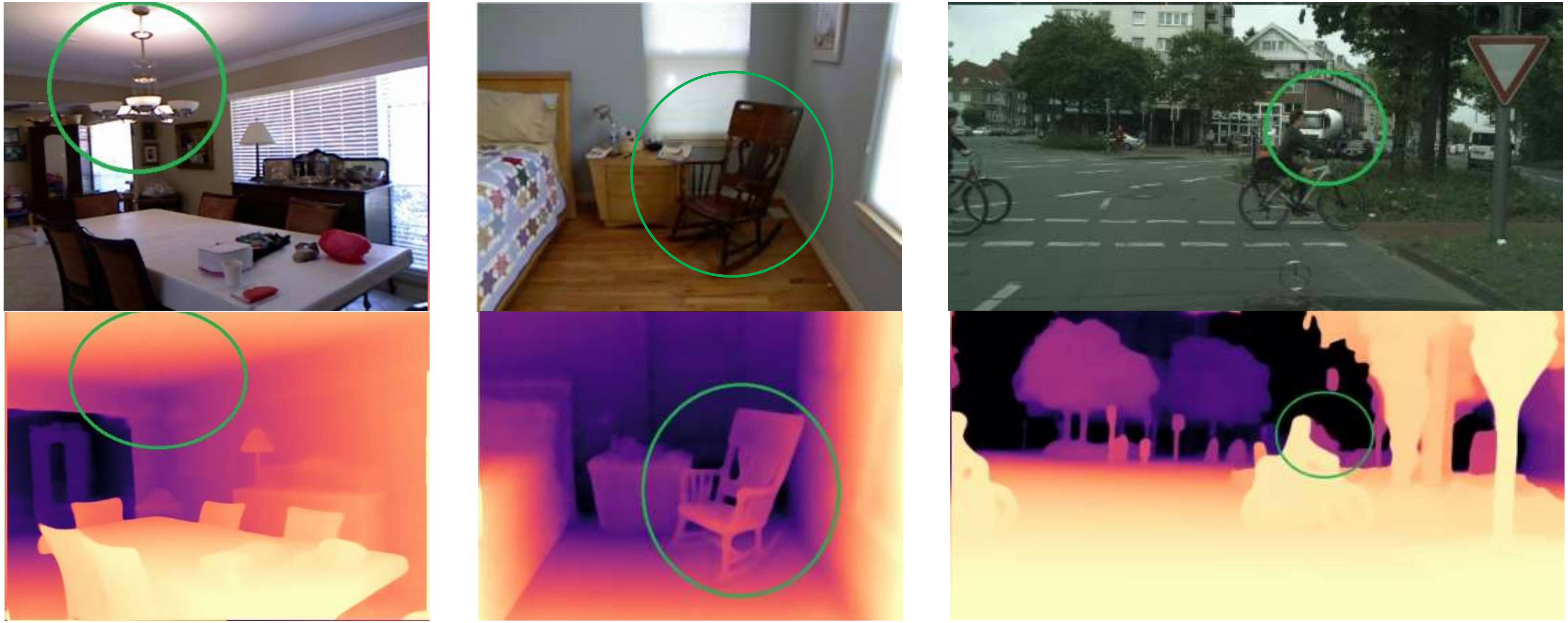
Michael Baltaxe[1, 2], Dan Levi[2], Sagie Benaim[1]

[1]The Hebrew University of Jerusalem, [2]General Motors

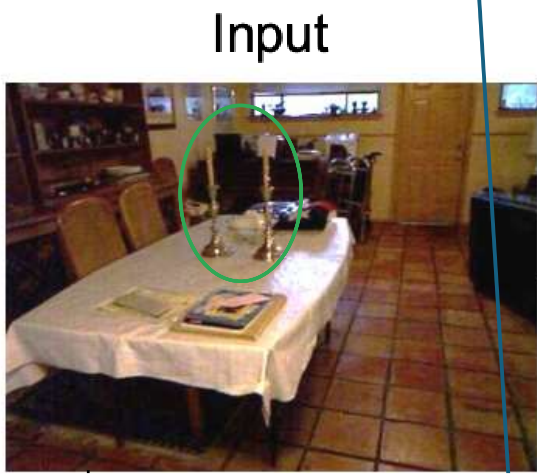
CVPR Findings 2026

“Long Tail” in Depth Estimation

- SOTA networks fail on **underrepresented classes** due to lack of strong priors.



Strong cues about the object geometry



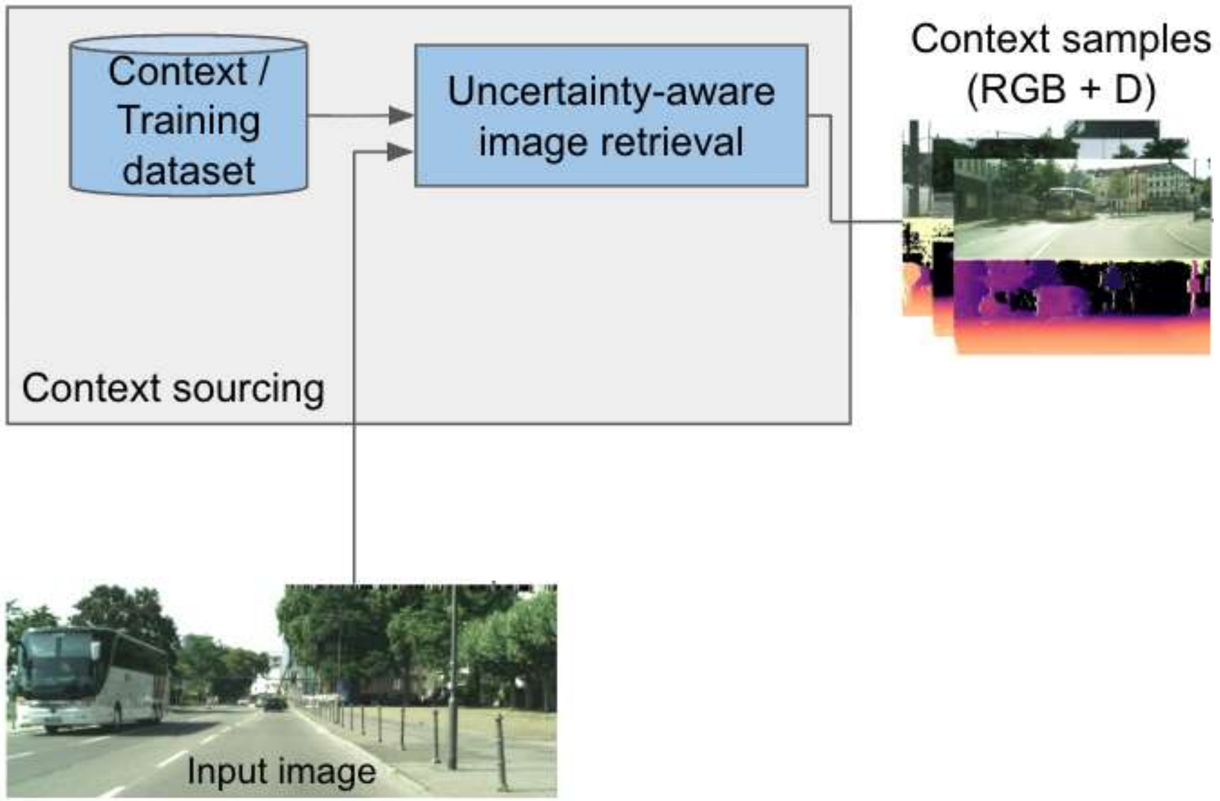
How Can We Build a Retrieval-Augmented Depth (RAD) Estimation Pipeline?

1. Retrieve Context Samples

Uncertainty-Aware Context Retrieval

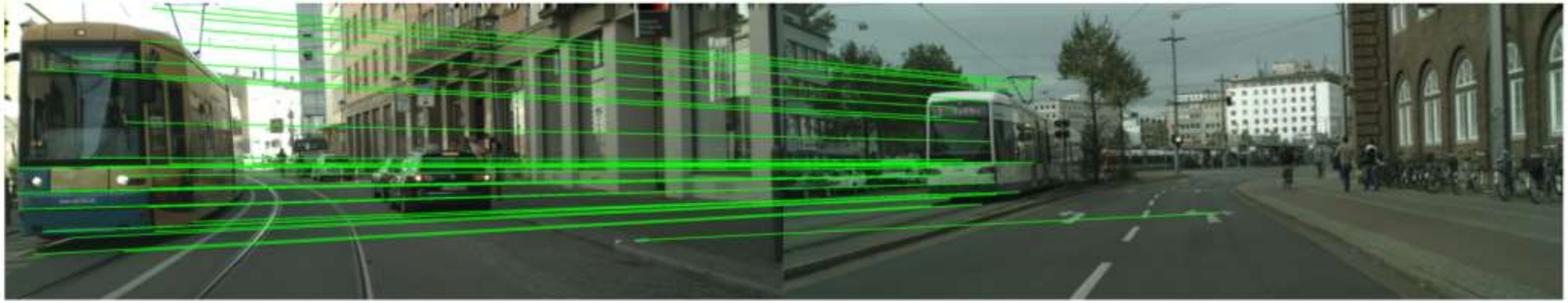
Input image

2. Find Corresponding Geometry

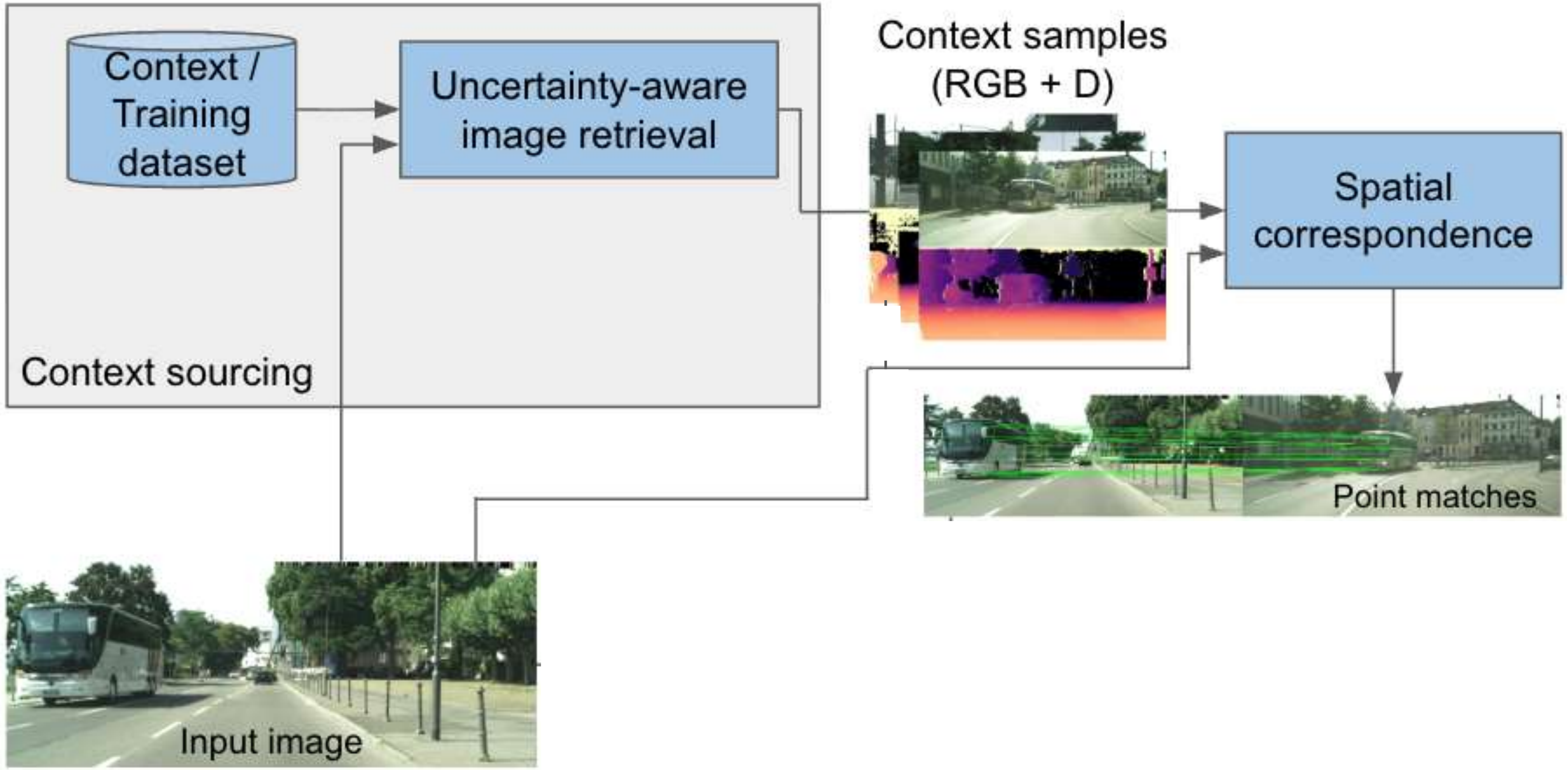


Spatial Correspondence

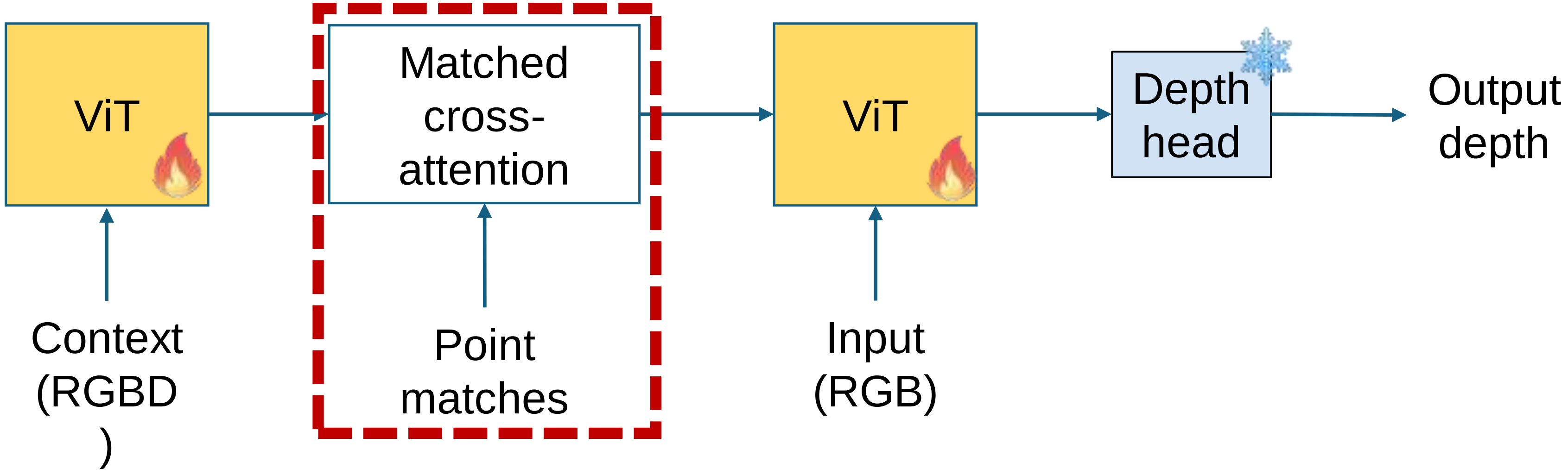
- **Purpose:** Precise spatial links between Input and Retrievals.
- **Method:** Off-the-shelf point matching algorithm.
- **Goal:** These correspondences guide the information transfer.



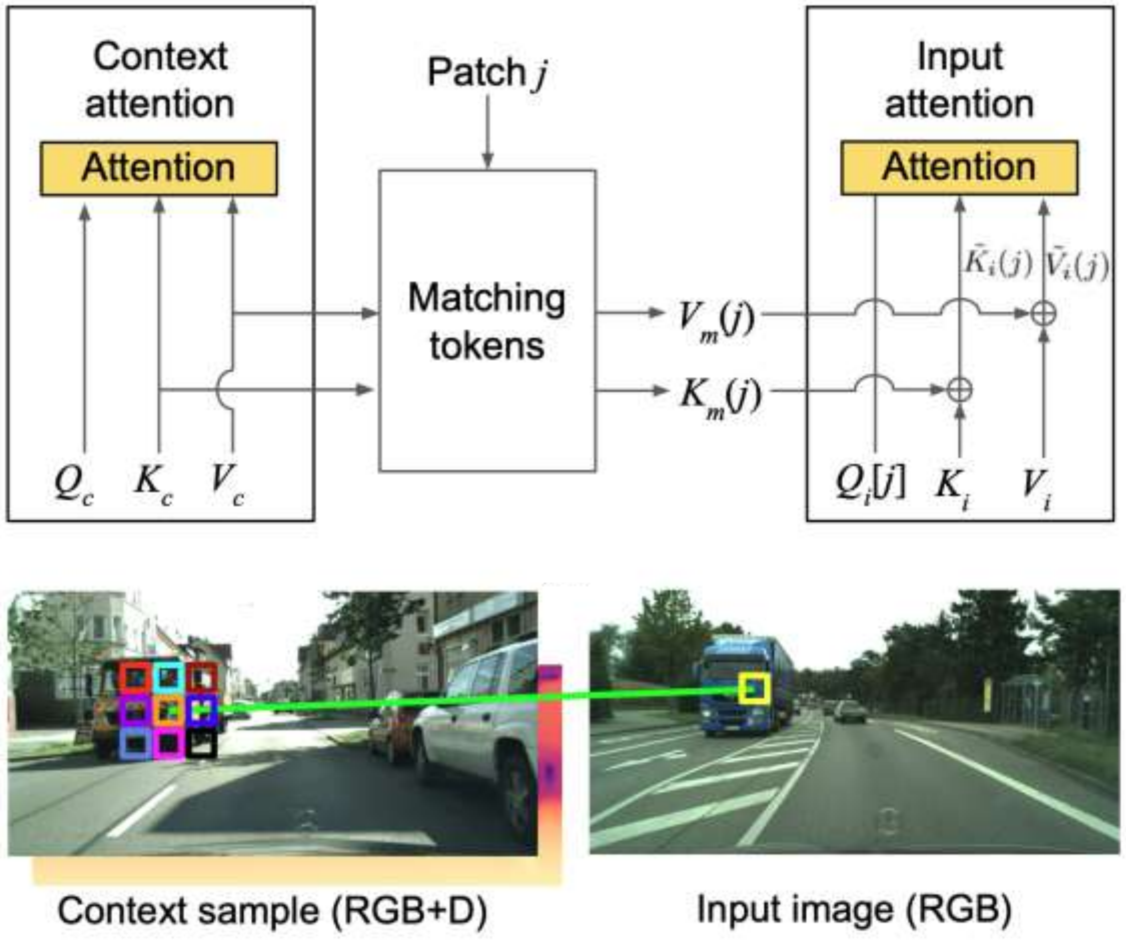
3. Integrate the Retrieved Information: Dual-Stream Depth Estimation Network



Dual Stream Depth Estimation Network

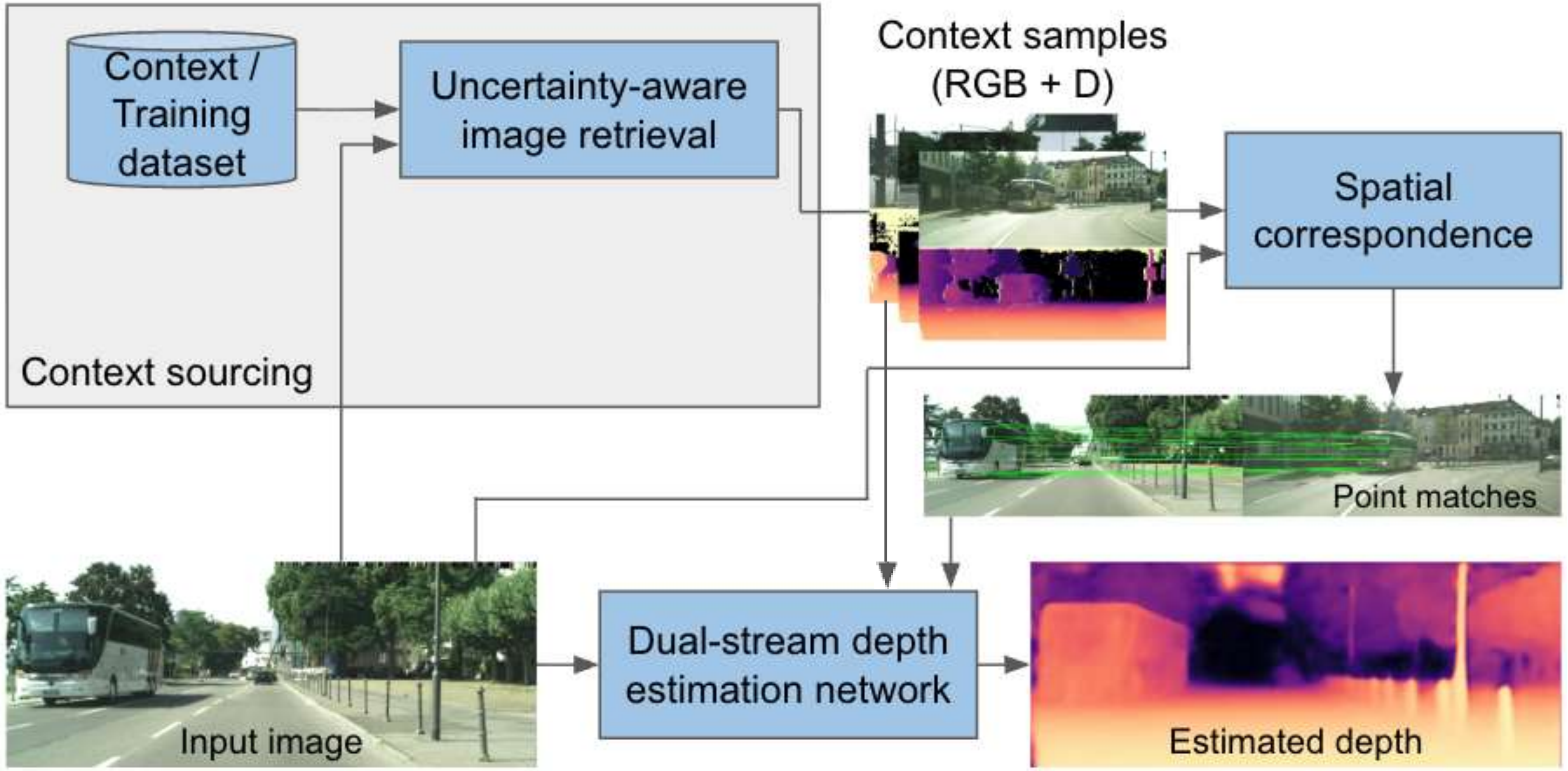


Matched Cross-Attention

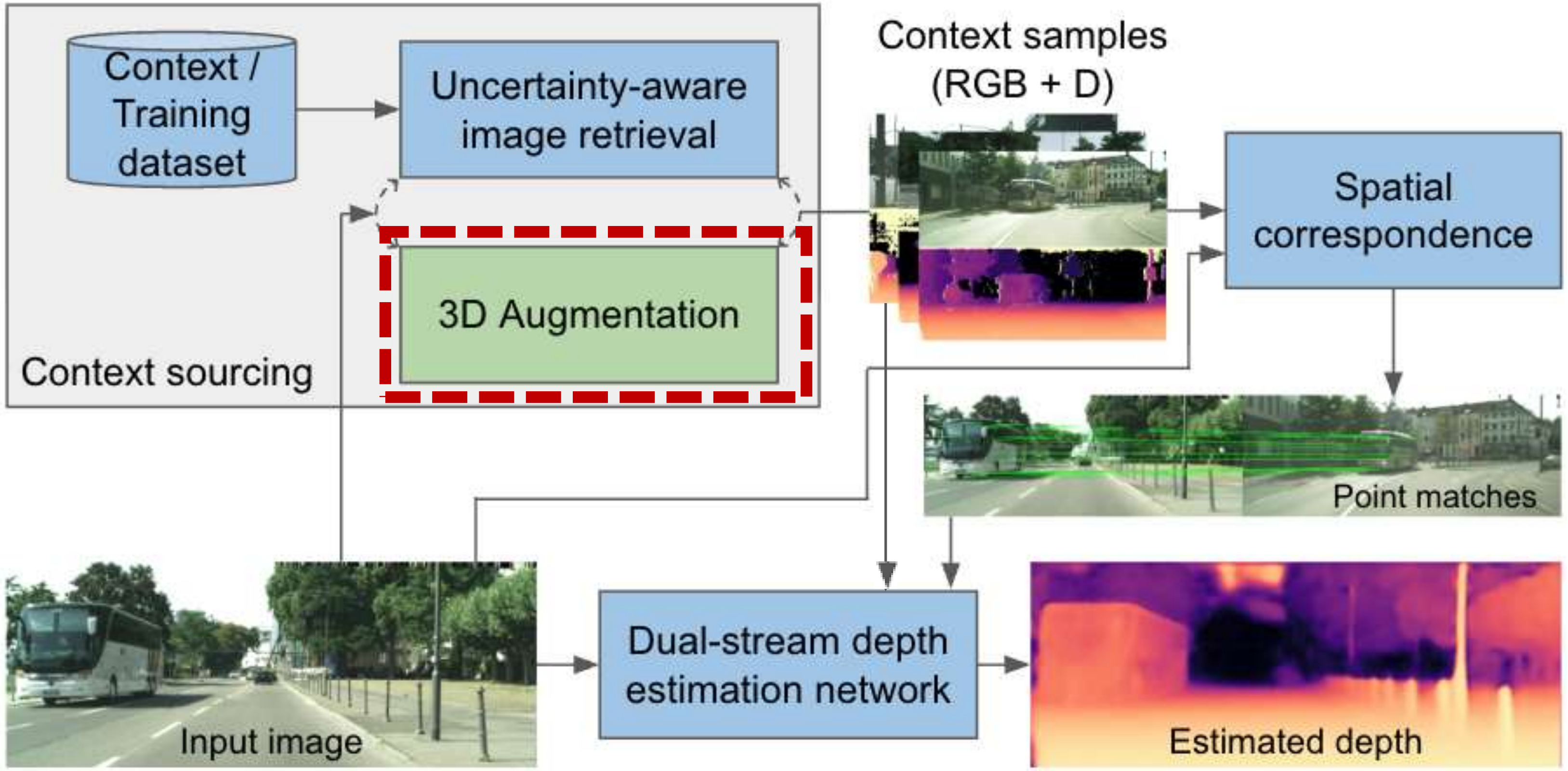


$\oplus$ Concatenation

Full Pipeline (Inference)



Full Pipeline (Training)

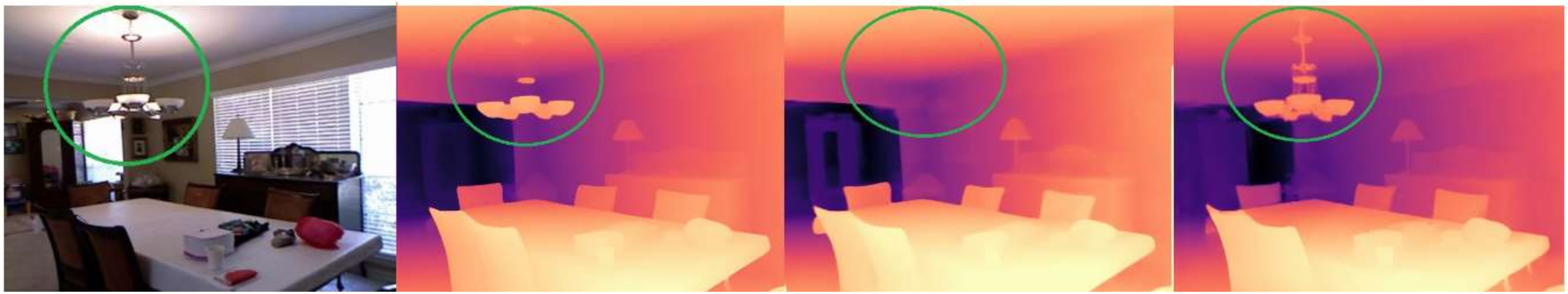


Quantitative Results: **Underrepresented Classes**

Method	NYU						KITTI						Cityscapes					
	↑			↓			↑			↓			↑			↓		
	δ_1	δ_2	δ_3	AbsRel	RMS	Log_{10}	δ_1	δ_2	δ_3	AbsRel	RMS	$\text{RMS}_{\log}$	δ_1	δ_2	δ_3	AbsRel	RMS	$\text{RMS}_{\log}$
AdaBins [2]	81.0	88.3	93.7	0.173	0.522	0.053	87.2	95.3	98.7	0.143	4.567	0.193	—	—	—	—	—	—
NewCRF [57]	80.3	89.7	95.4	0.169	0.497	0.049	89.1	96.7	99.1	0.121	4.247	0.166	—	—	—	—	—	—
DepthAnyV2 [53]	91.4	98.2	99.8	0.089	0.327	0.049	92.3	97.8	99.1	0.083	2.893	0.086	89.2	94.7	99.1	0.110	7.804	0.166
ZoeDepth† [4]	90.9	98.1	99.6	0.098	0.334	0.043	88.6	96.9	98.4	0.116	3.908	0.140	40.9	66.7	79.3	0.314	17.051	0.496
DepthPro [5]	89.6	97.6	99.1	0.106	0.375	0.048	77.9	97.7	99.3	0.160	2.698	0.170	47.5	87.8	98.0	0.224	11.973	0.285
Metric3Dv2 [56]	92.4	97.6	99.4	0.095	0.331	**0.039**	89.8	97.7	**99.4**	0.121	2.577	0.141	92.1	**97.1**	98.5	0.097	7.063	0.152
UniDepthV2 [39]	92.8	98.1	99.3	0.092	0.329	**0.039**	90.5	98.2	**99.4**	0.111	4.059	0.130	88.3	96.4	98.2	0.136	7.777	0.171
RAD-Small	95.3	98.5	99.1	0.084	0.321	0.057	92.2	95.9	98.0	0.143	3.943	0.097	86.7	96.6	98.5	0.133	5.573	0.168
RAD-Base	96.7	99.4	99.7	0.072	0.299	0.053	95.1	97.6	98.8	0.084	3.107	0.088	88.5	96.9	99.0	0.127	5.191	0.158
RAD-Large	**97.5**	**99.5**	**99.9**	**0.063**	**0.288**	0.040	**96.6**	**98.6**	**99.4**	**0.072**	**2.498**	**0.052**	**93.5**	97.0	**99.1**	**0.090**	**5.083**	**0.150**

Quantitative Results: **Underrepresented Classes**

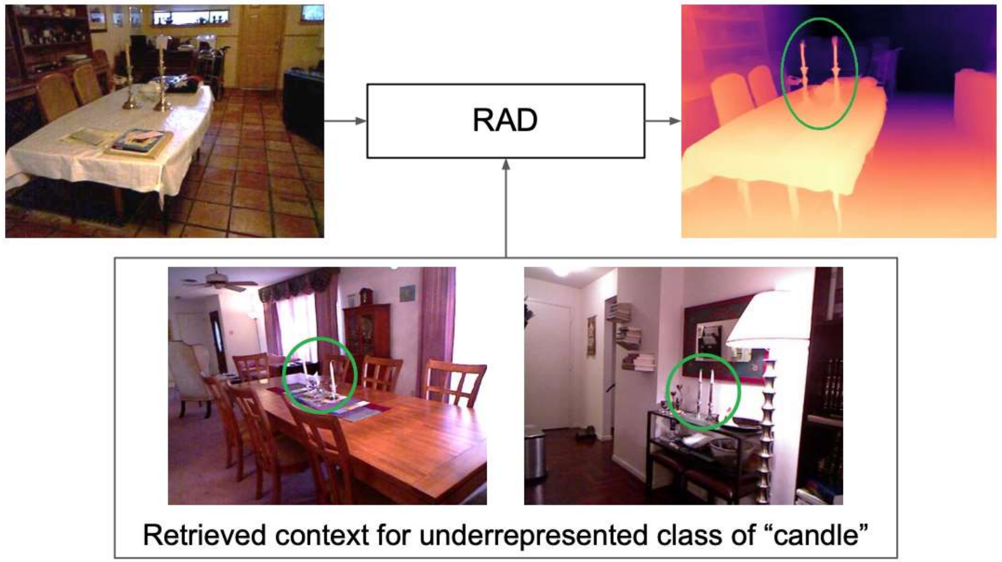
- **NYU Depth v2:** 29.2% reduction in Relative Absolute Error.
- **KITTI:** 13.3% reduction in Relative Absolute Error.
- **Cityscapes:** 7.2% reduction in Relative Absolute Error.
- **Conclusion:** RAD **significantly outperforms SoTA** on underrepresented classes.



Input UniDepth v2 Metric3D v2 **Ours**

Retrieval Augmentation For 3D

Depth Estimation



3D Generation

MV-RAG: Retrieval Augmented Multiview Diffusion

Yosef Dayani, Omer Benishu, Sagie Benaim

“Bolognese dog”

Pipeline

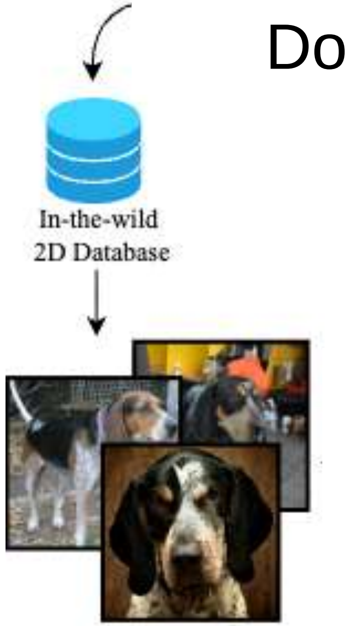
Coonhound
Dog

Pipeline

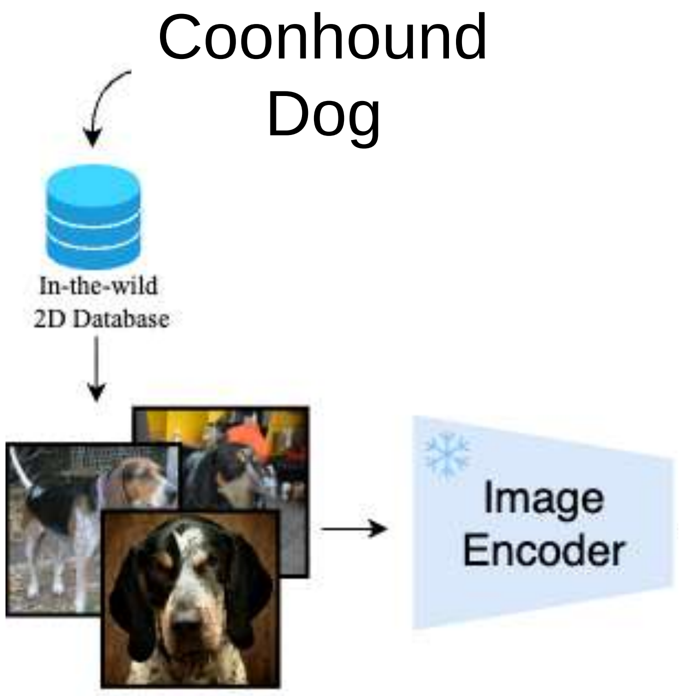
Coonhound
Dog

Pipeline

Coonhound
Dog



Pipeline



Pipeline

Coonhound
Dog

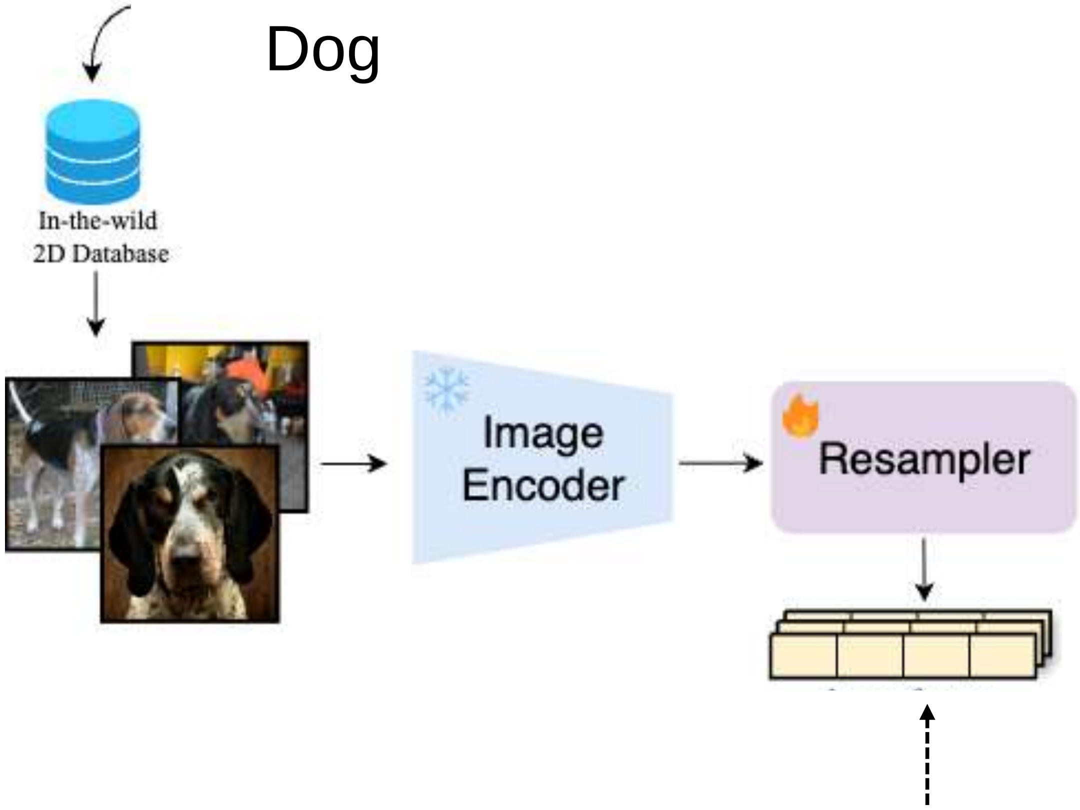
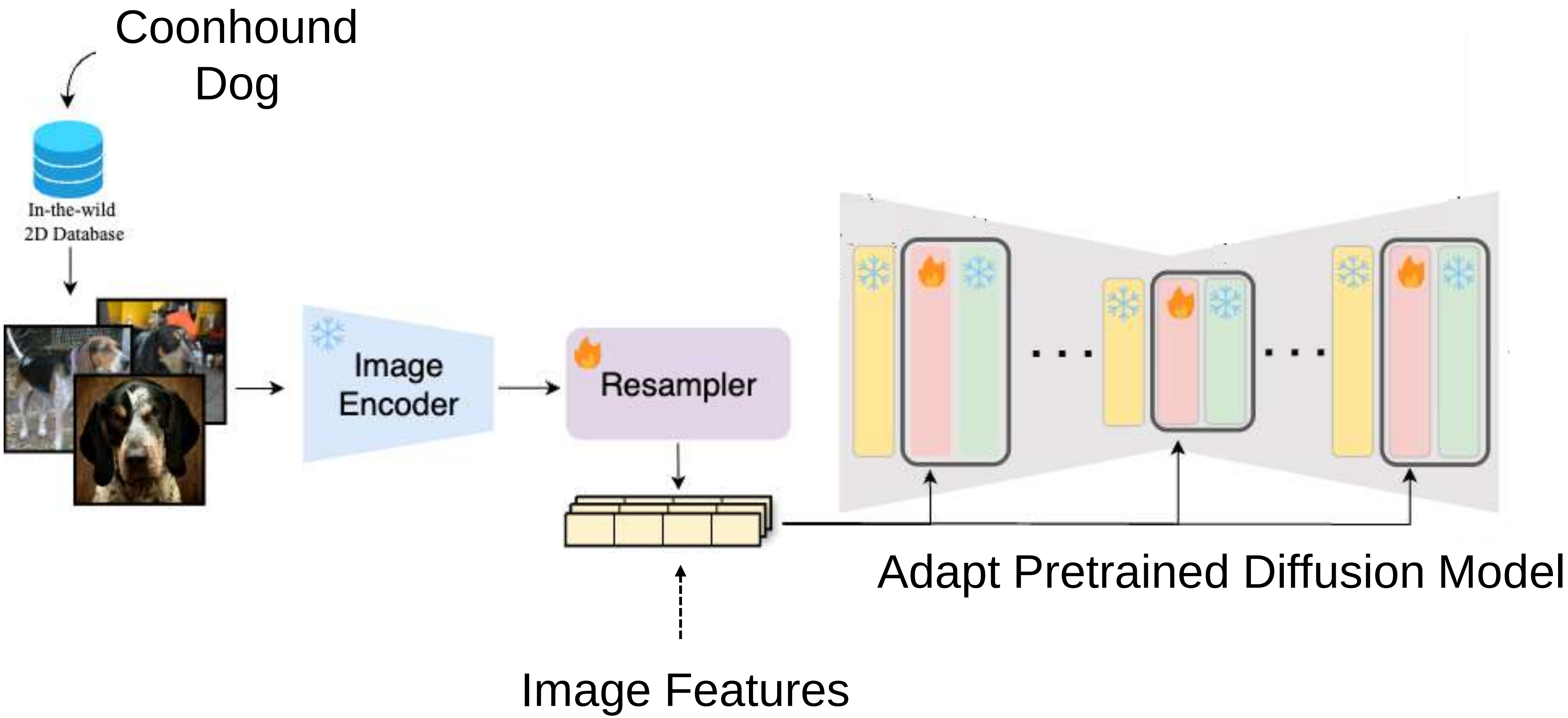
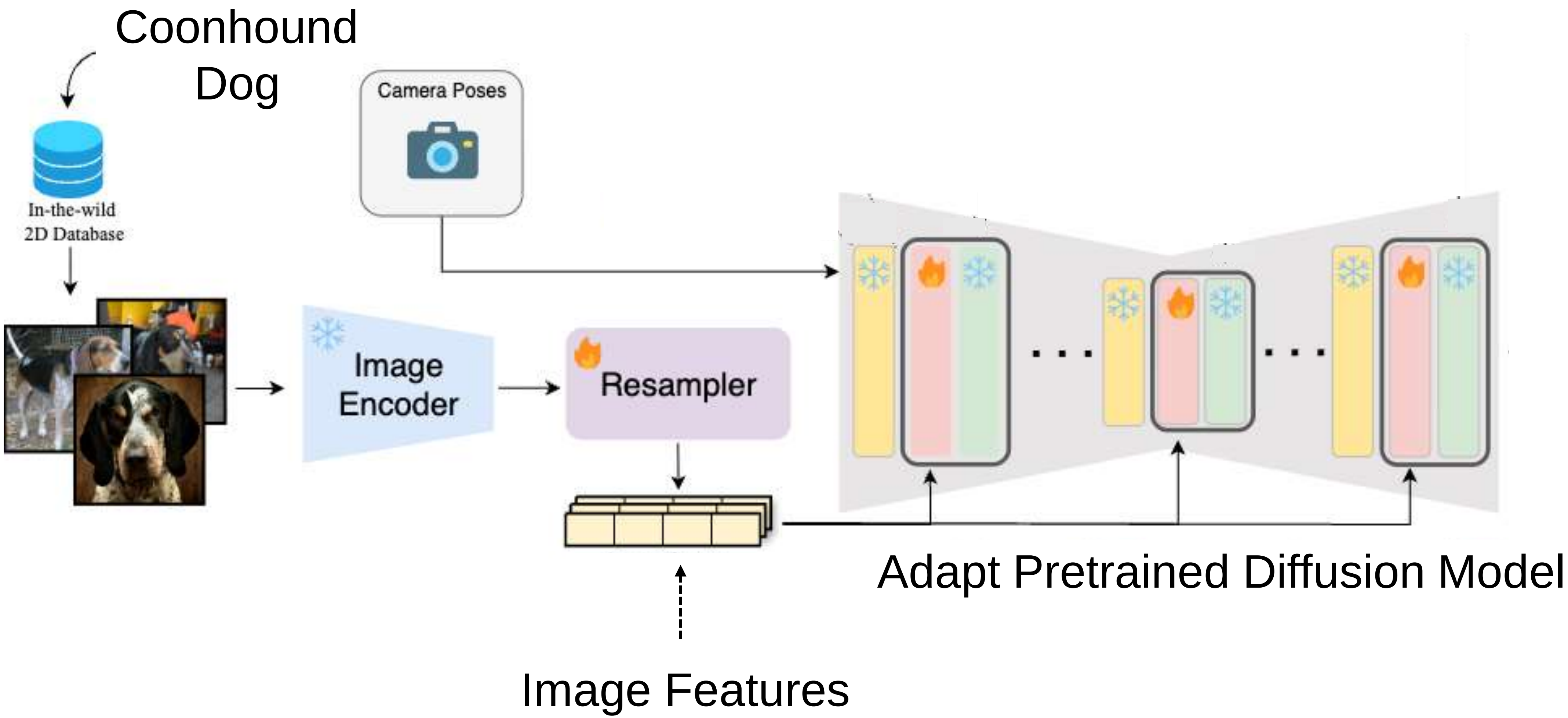


Image Features

Pipeline

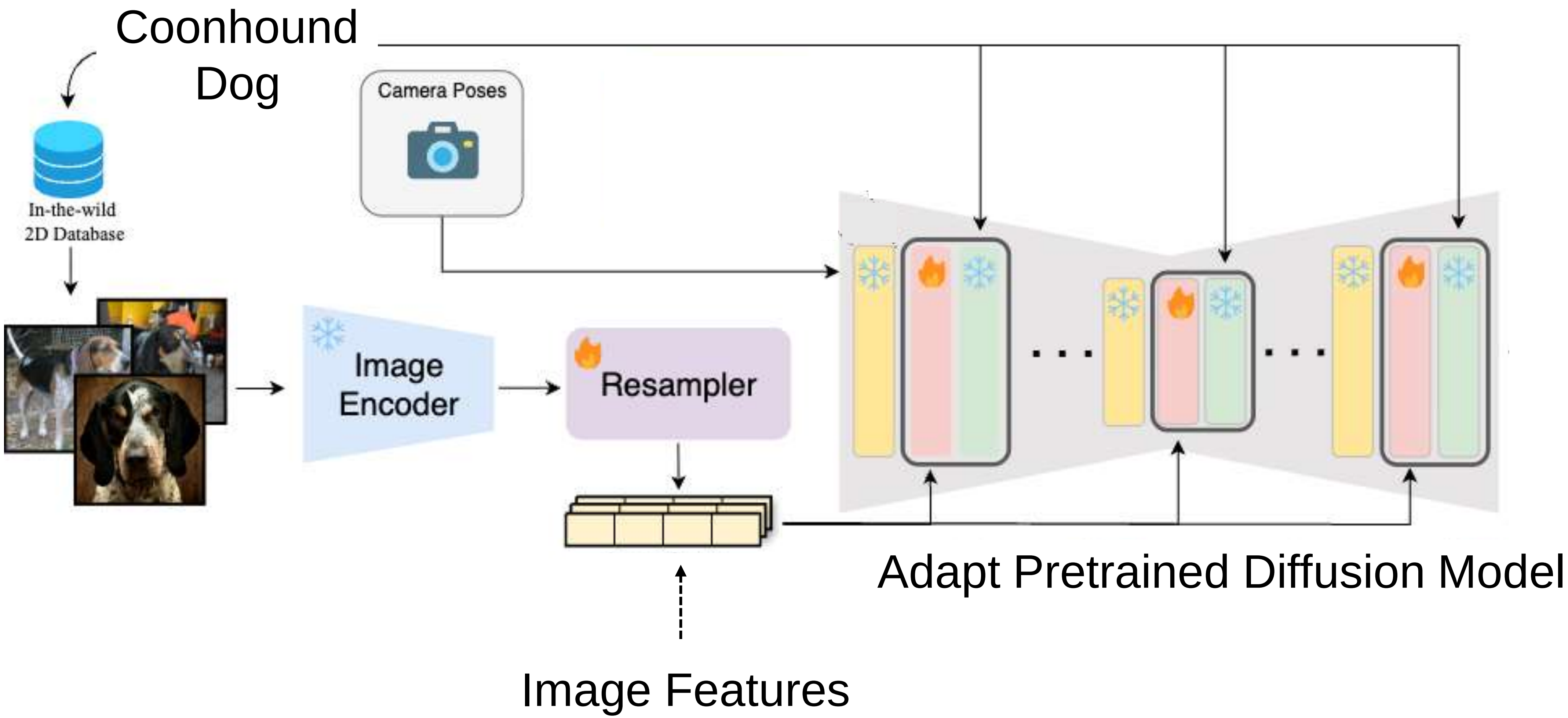

Coonhound
Dog
In-the-wild
2D Database
Image
Encoder
Resampler
Adapt Pretrained Diffusion Model
Image Features

Pipeline

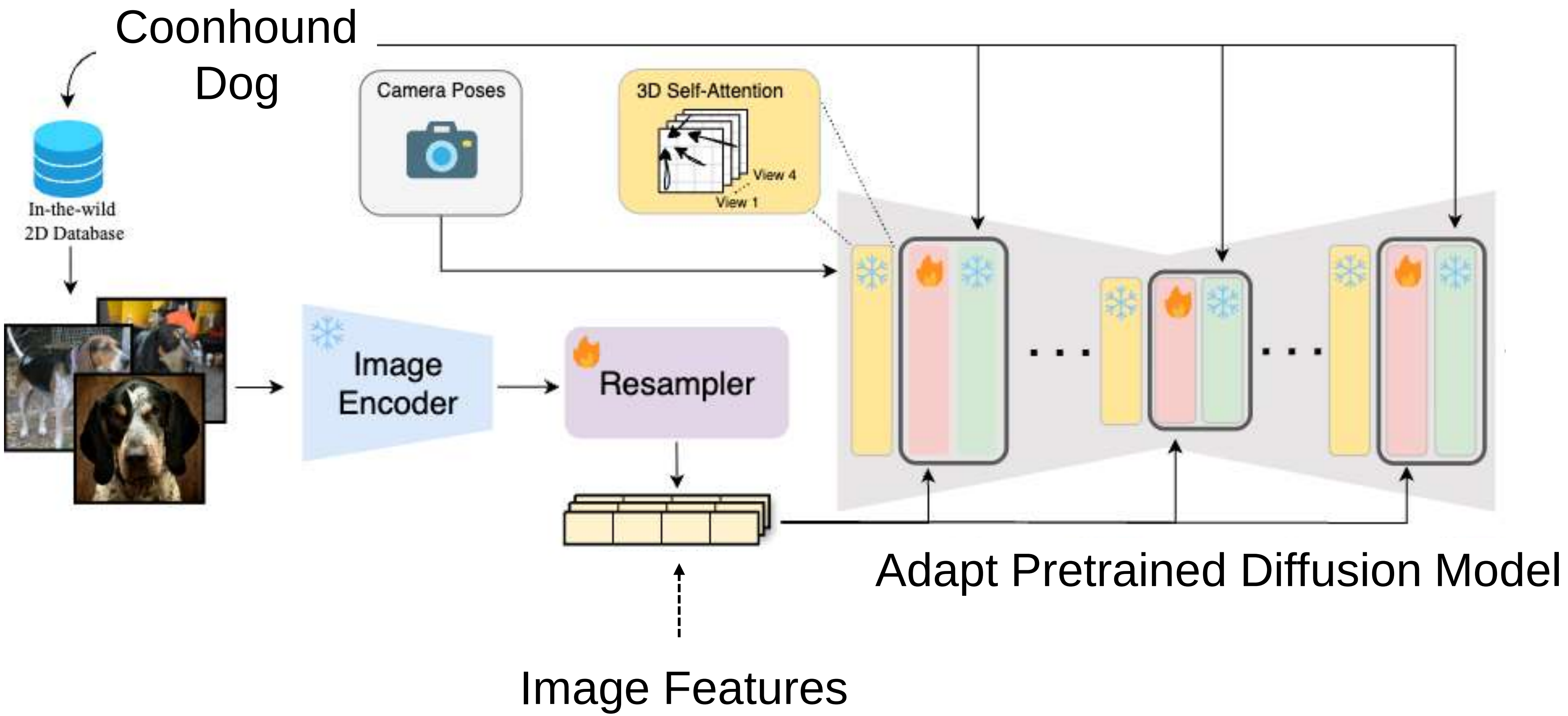


Coonhound
Dog
In-the-wild
2D Database
Camera Poses
Image
Encoder
Resampler
Adapt Pretrained Diffusion Model
Image Features

Pipeline

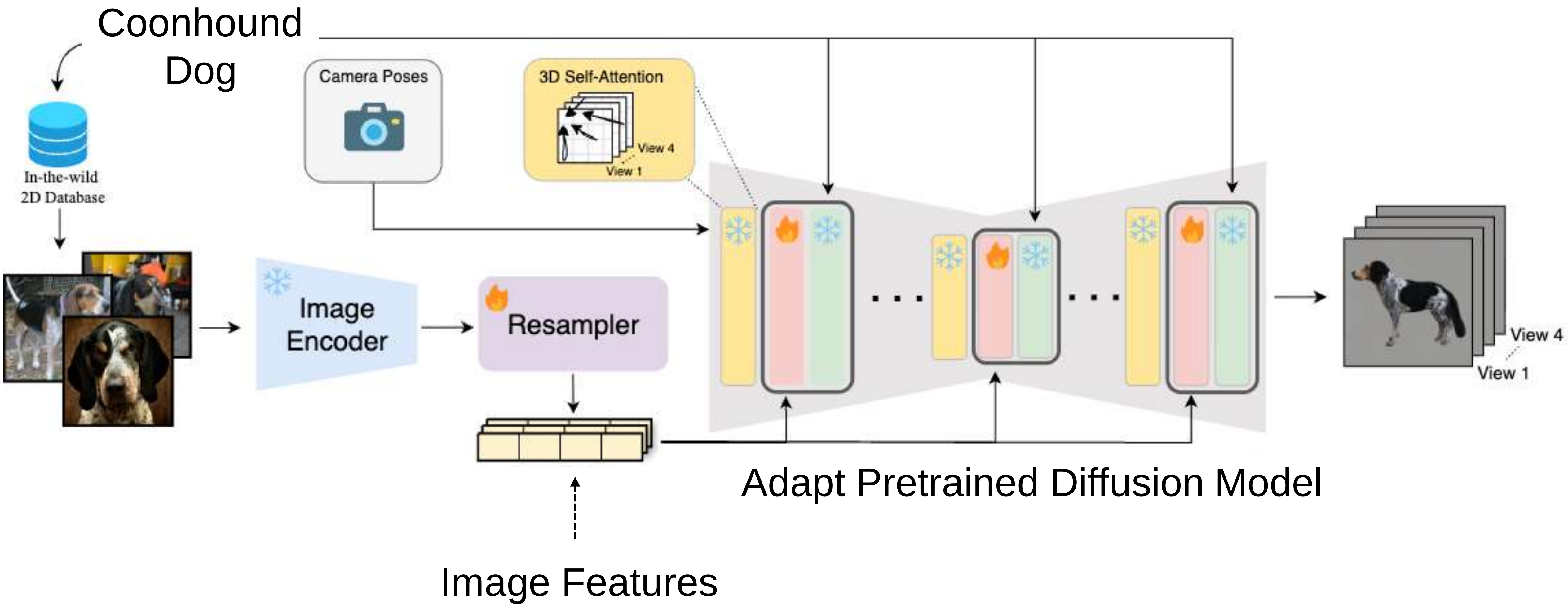

Coonhound
Dog
In-the-wild
2D Database
Camera Poses
Image
Encoder
Resampler
Image Features
Adapt Pretrained Diffusion Model

Pipeline



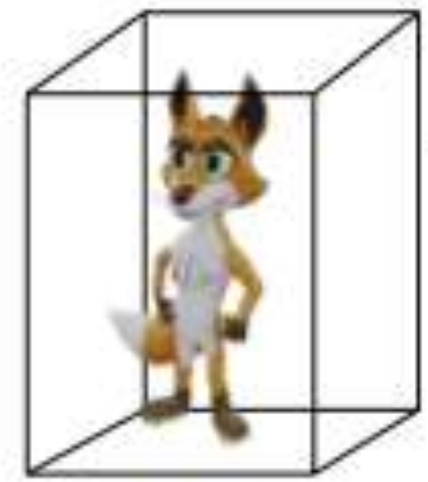
Coonhound Dog
In-the-wild 2D Database
Camera Poses
3D Self-Attention
View 4
View 1
Image Encoder
Resampler
Adapt Pretrained Diffusion Model
Image Features

Pipeline



Coonhound
Dog
In-the-wild
2D Database
Camera Poses
3D Self-Attention
View 4
View 1
Image
Encoder
Resampler
View 4
View 1
Adapt Pretrained Diffusion Model
Image Features

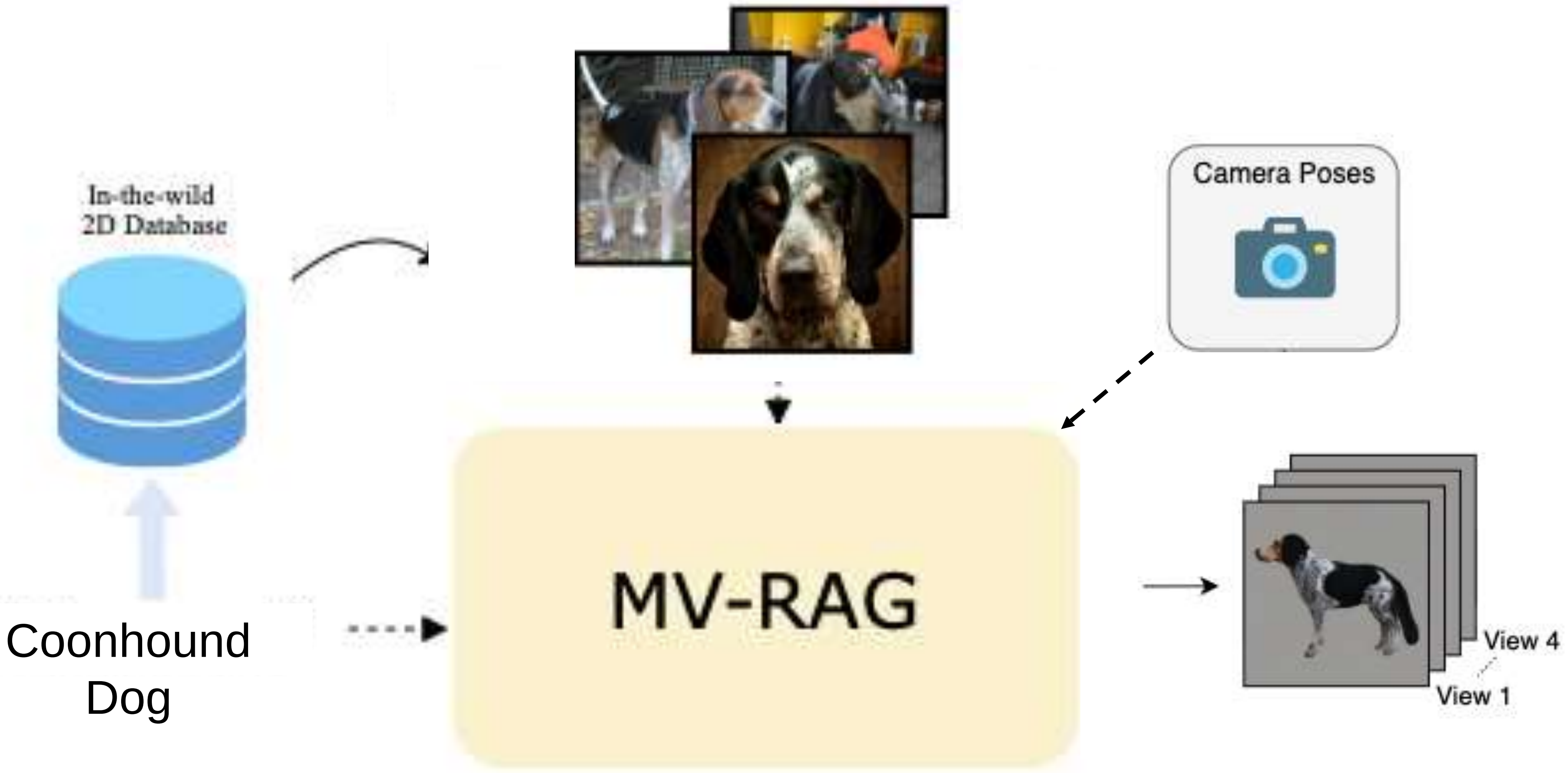
3D Training Mode



"Cartoon Fox"

2D Training Mode

Inference


In-the-wild
2D Database
Camera Poses
MV-RAG
Coonhound
Dog
View 4
View 1

"Peugeot 202"

Collaborators On These Projects

Michael Baltaxe
Hebrew University, General Motors

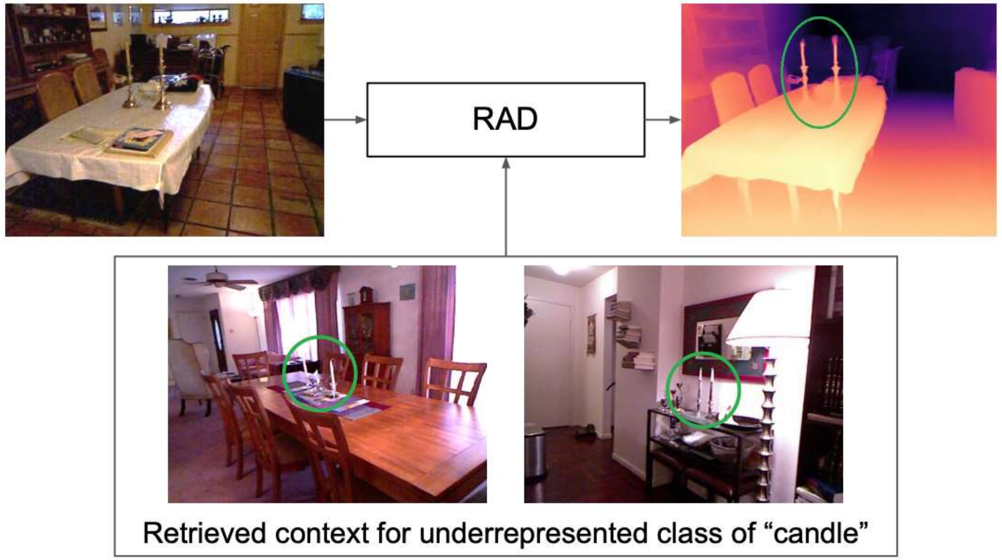
Dan Levi
General Motors

Yosef Dayani
Hebrew University, Mobileye

Omer Beinshu
Hebrew University, Nintex

Retrieval Augmentation For 3D Long Tail

Depth Estimation



3D Generation

Questions?

- General Framework.
- Other 3D Tasks? 3D Tracking, 3D Reconstruction, World Models …