ICCV HONOLULU HAWAII OCT 19-23, 2025

Feed-Forward SceneDINO for Unsupervised Semantic Scene Completion


Aleksandar Jevtić* [1]


Christoph Reich* [1,2,4,5]


Felix Wimbauer[1,4]


Oliver Hahn[2]


Christian Rupprecht[3]

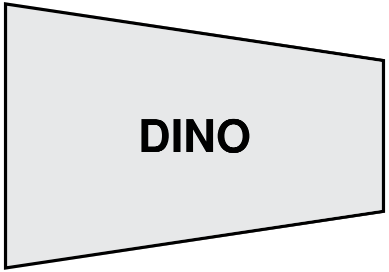

Stefan Roth[2,5,6]


Daniel Cremers[1,4,5]

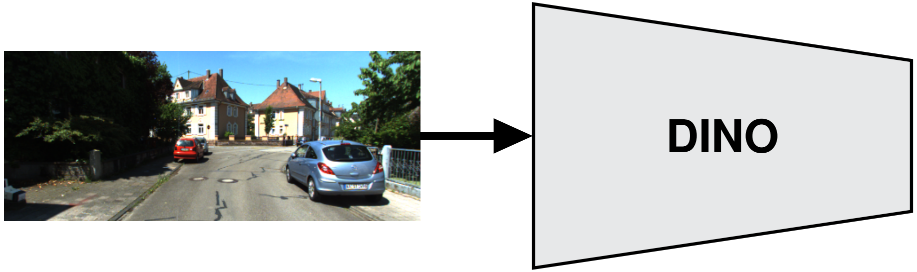
*equal contribution

1 TUM
2 TECHNISCHE UNIVERSITÄT DARMSTADT
3 UNIVERSITY OF OXFORD
4 mcml Munich Center for Machine Learning
5 ZUSE SCHOOL ELIZA
6 hessian.AI

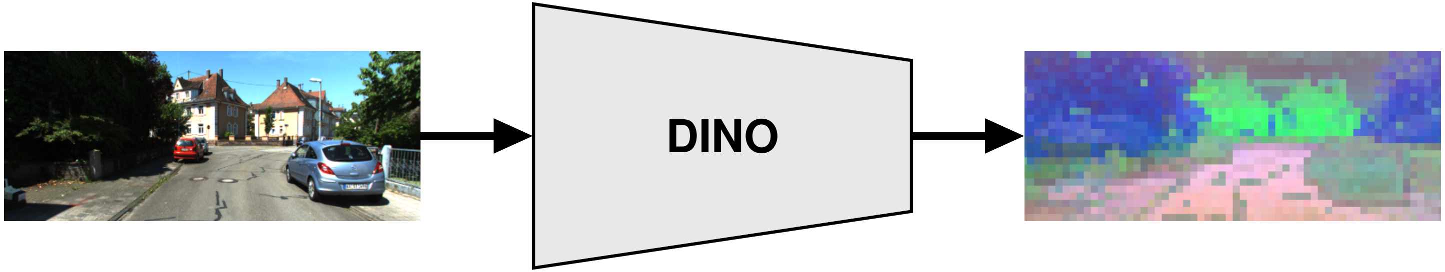
Motivation



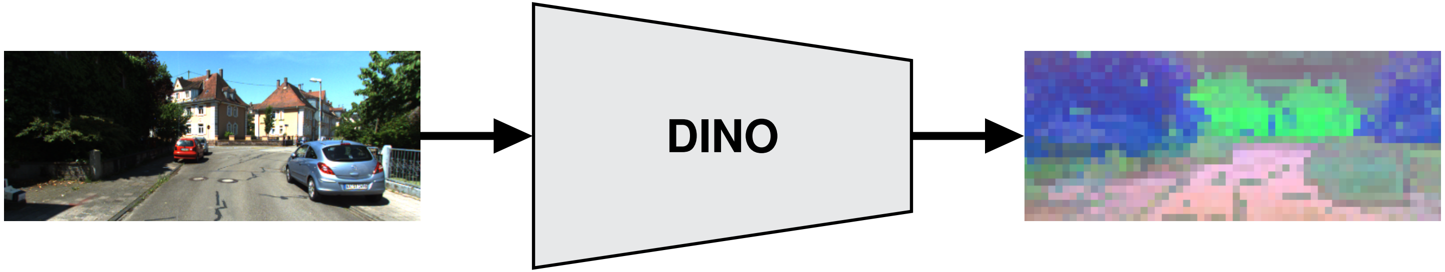
Motivation



Motivation

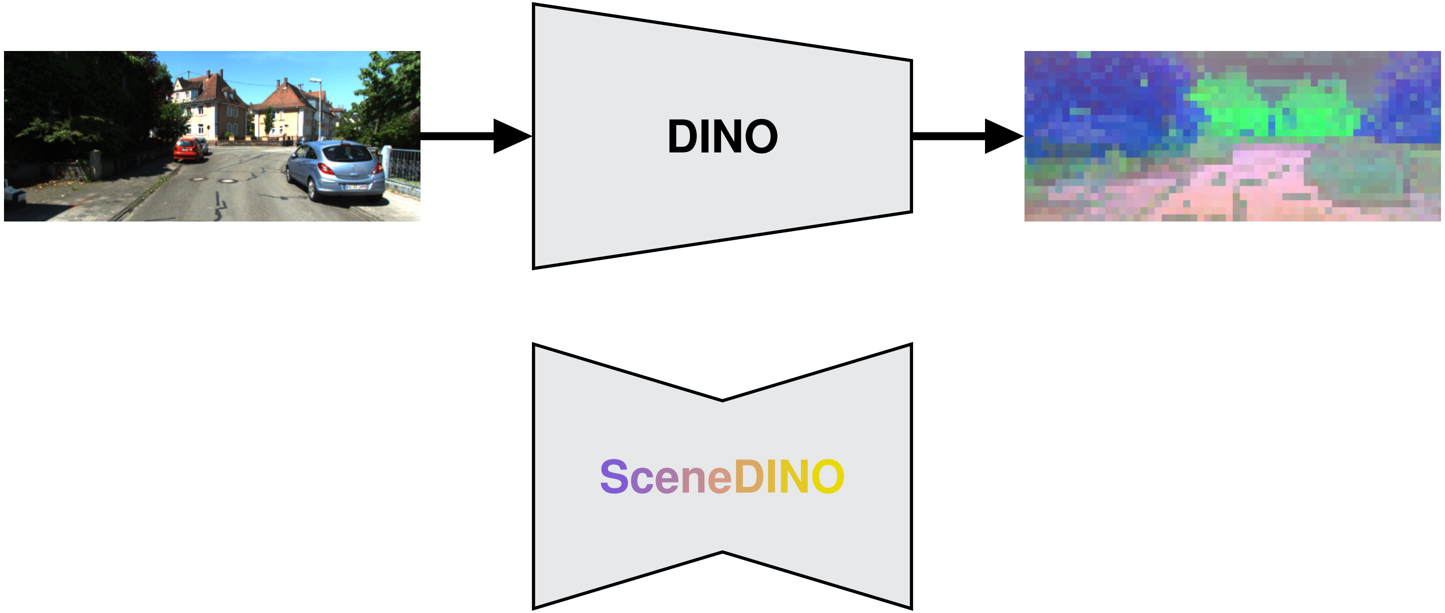


Motivation



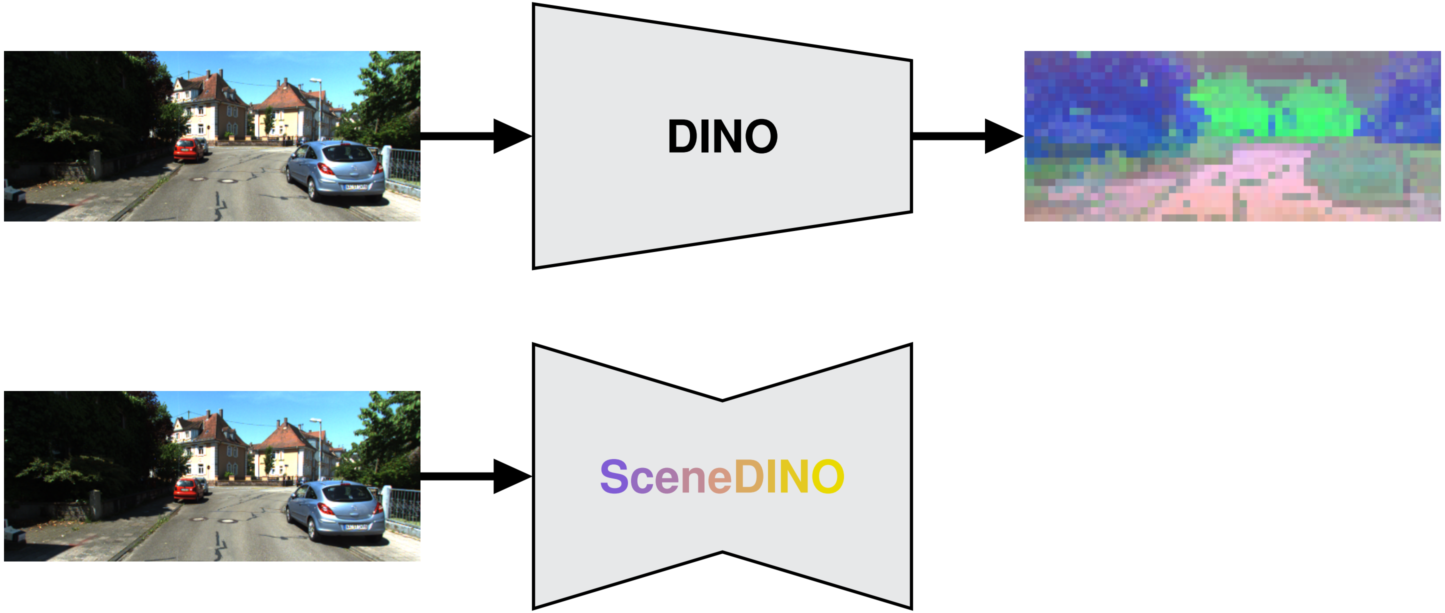
Bring DINO to 3D 🚀

Motivation



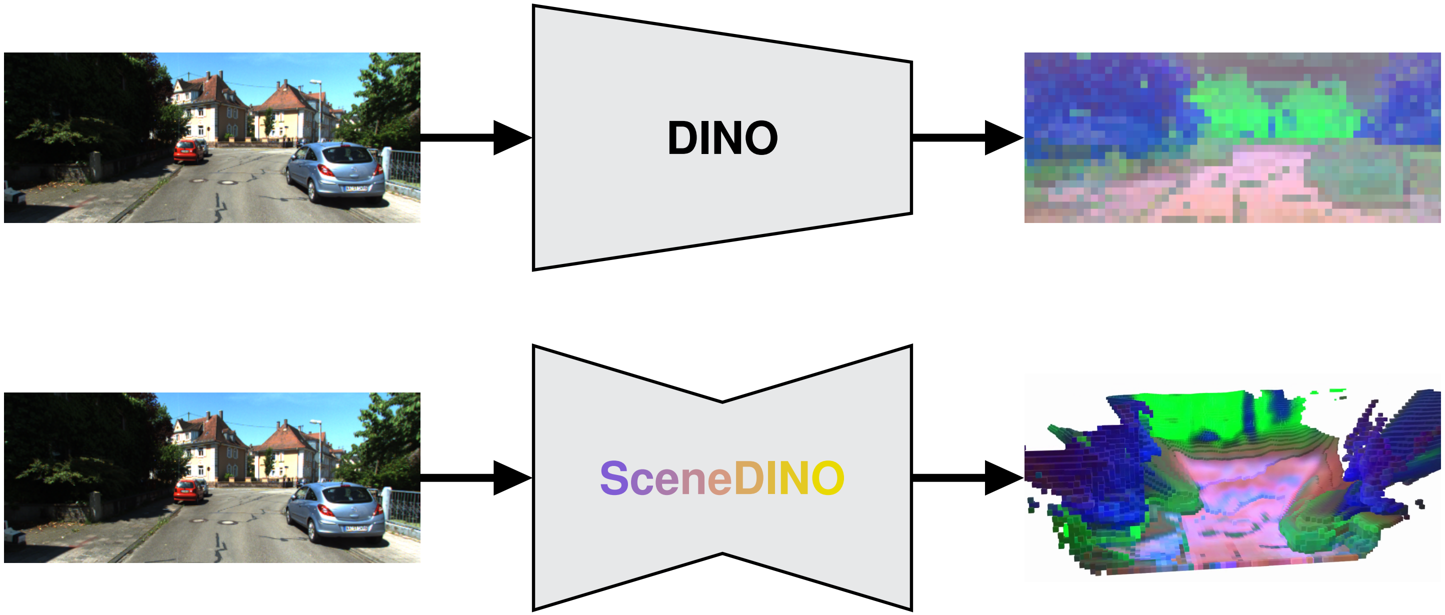
Bring DINO to 3D 🚀

Motivation



Bring DINO to 3D 🚀

Motivation



Bring DINO to 3D 🚀

Semantic Scene Completion (SSC)

a.k.a. Semantic Occupancy Prediction

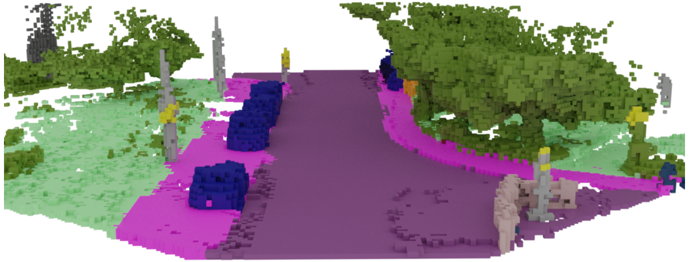
n **input images**

Semantic Scene Completion (SSC)

a.k.a. Semantic Occupancy Prediction

n **input images**

Dense 3D geometry & semantics



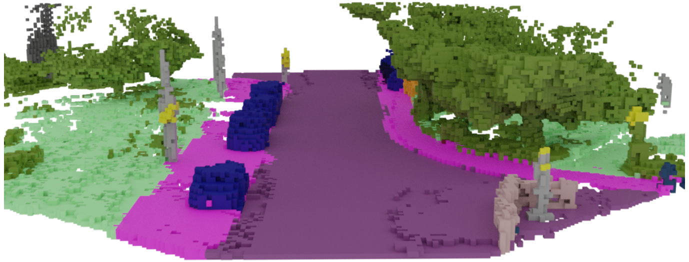
Semantic Scene Completion (SSC)

a.k.a. Semantic Occupancy Prediction

ICCV OCT 19-23, 2025 HONOLULU HAWAII

Single input image

Dense 3D geometry & semantics

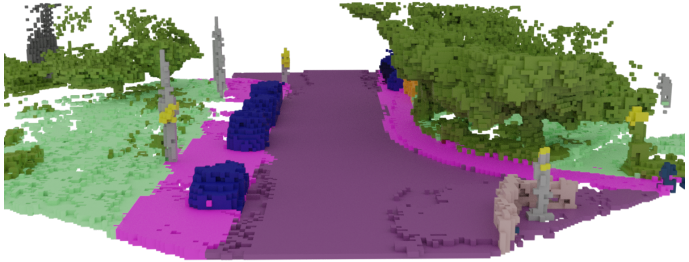


Semantic Scene Completion (SSC)

a.k.a. Semantic Occupancy Prediction

Single input image

Dense 3D geometry & semantics



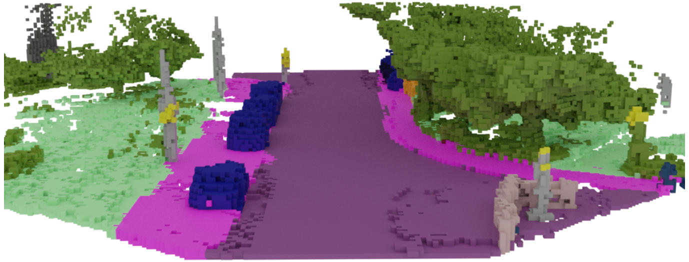
✓ Comprehensive 3D scene understanding task

Semantic Scene Completion (SSC)

a.k.a. Semantic Occupancy Prediction

Single input image

Dense 3D geometry & semantics

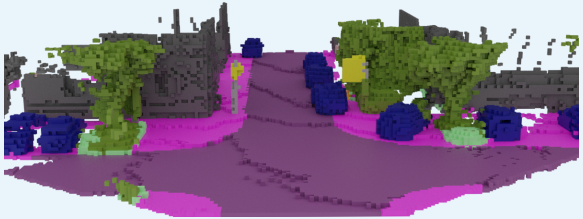


✓ Comprehensive 3D scene understanding task

✓ Applications in robotics, autonomous driving, medical image analysis, and civil engineering

Related Work: SSC

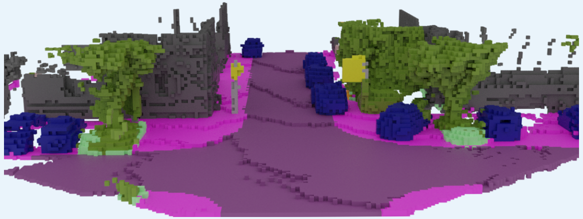
Geometric & 3D semantic supervision (*e.g.*, [1])



[1] S. Song *et al.*, "Semantic scene completion from a single depth image," in *CVPR*, 2017, pp. 190–198.
[2] Y. Huang *et al.*, "SelfOcc: Self-supervised vision-based 3D occupancy prediction," in *CVPR*, 2024, pp. 19 946–19 956.

Related Work: SSC

Geometric & 3D semantic supervision (*e.g.*, [1])

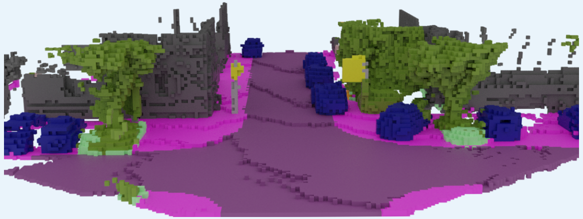


– Ground truth very expensive
– Special hardware needed
– Infeasible to scale

[1] S. Song *et al.*, "Semantic scene completion from a single depth image," in *CVPR*, 2017, pp. 190–198.
[2] Y. Huang *et al.*, "SelfOcc: Self-supervised vision-based 3D occupancy prediction," in *CVPR*, 2024, pp. 19 946–19 956.

Related Work: SSC

Geometric & 3D semantic supervision (*e.g.*, [1])



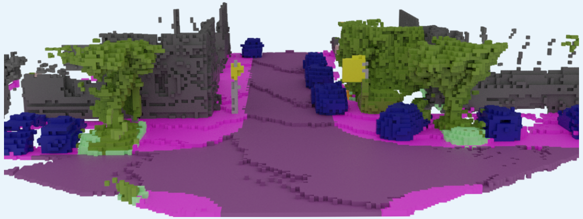
– Ground truth very expensive
– Special hardware needed
– Infeasible to scale

2D supervision (*e.g.*, [2])

[1] S. Song *et al.*, “Semantic scene completion from a single depth image,” in *CVPR*, 2017, pp. 190–198.
[2] Y. Huang *et al.*, “SelfOcc: Self-supervised vision-based 3D occupancy prediction,” in *CVPR*, 2024, pp. 19 946–19 956.

Related Work: SSC

Geometric & 3D semantic supervision (*e.g.*, [1])



– Ground truth very expensive
– Special hardware needed
– Infeasible to scale

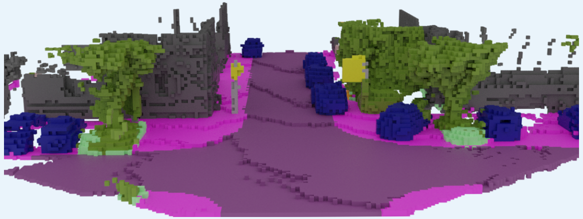
2D supervision (*e.g.*, [2])

– Still, expensive to obtain
– Limited generalization

[1] S. Song *et al.*, "Semantic scene completion from a single depth image," in *CVPR*, 2017, pp. 190–198.
[2] Y. Huang *et al.*, "SelfOcc: Self-supervised vision-based 3D occupancy prediction," in *CVPR*, 2024, pp. 19 946–19 956.

Related Work: SSC

Geometric & 3D semantic supervision (*e.g.*, [1])



– Ground truth very expensive
– Infeasible to scale
– Special hardware needed

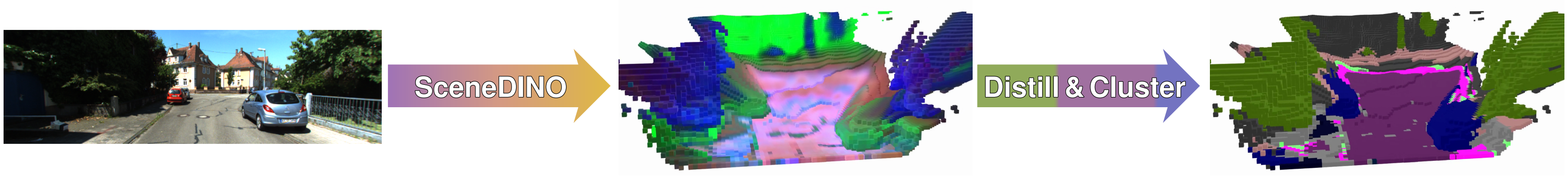
2D supervision (*e.g.*, [2])

– Still, expensive to obtain
– Limited generalization

Large-scale SSC annotations infeasible → unsupervised SSC

[1] S. Song *et al.*, "Semantic scene completion from a single depth image," in *CVPR*, 2017, pp. 190–198.
[2] Y. Huang *et al.*, "SelfOcc: Self-supervised vision-based 3D occupancy prediction," in *CVPR*, 2024, pp. 19 946–19 956.

SceneDINO

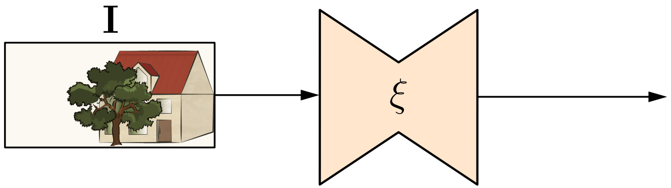


Single Input Image | 3D Feature Field | SSC Prediction

✓ **Fully unsupervised** ✓ **Multi-view self-supervision** ✓ **Feed-forward inference**

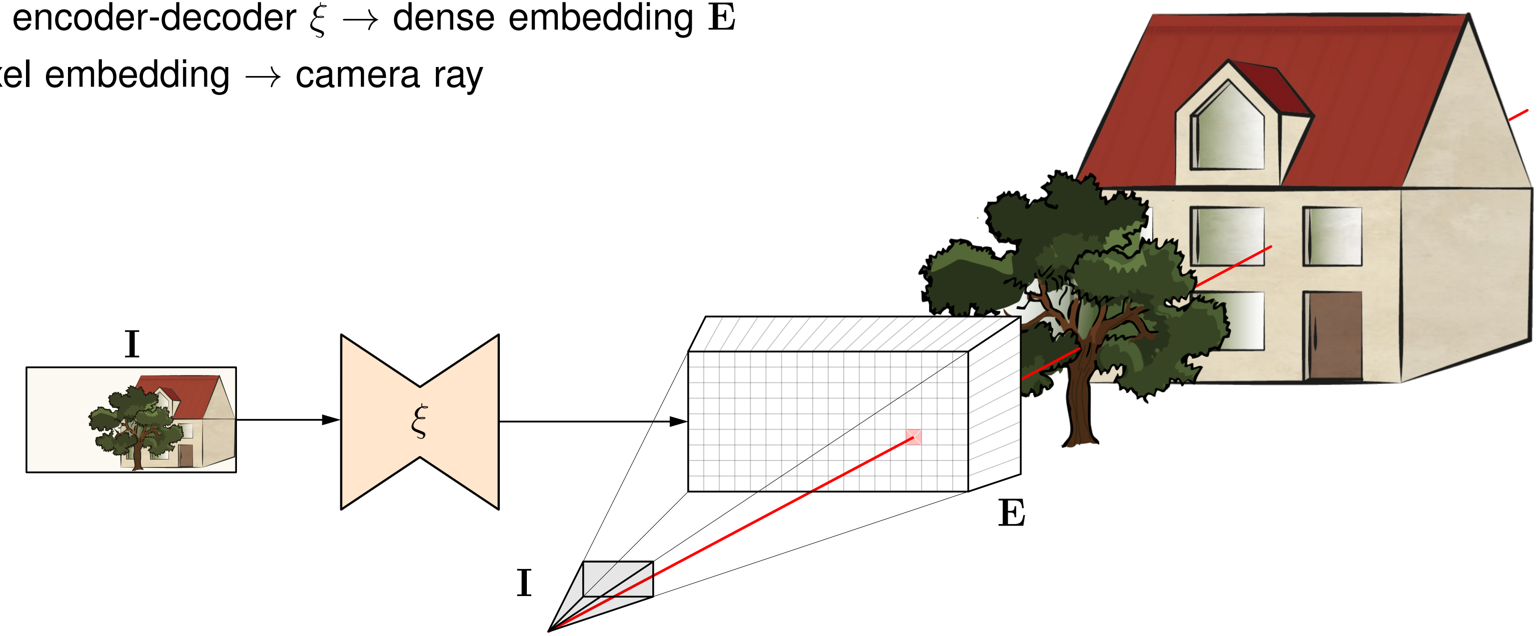
Model Architecture

- Single input image $\mathbf{I}$
- 2D encoder-decoder ξ $\rightarrow$ dense embedding $\mathbf{E}$



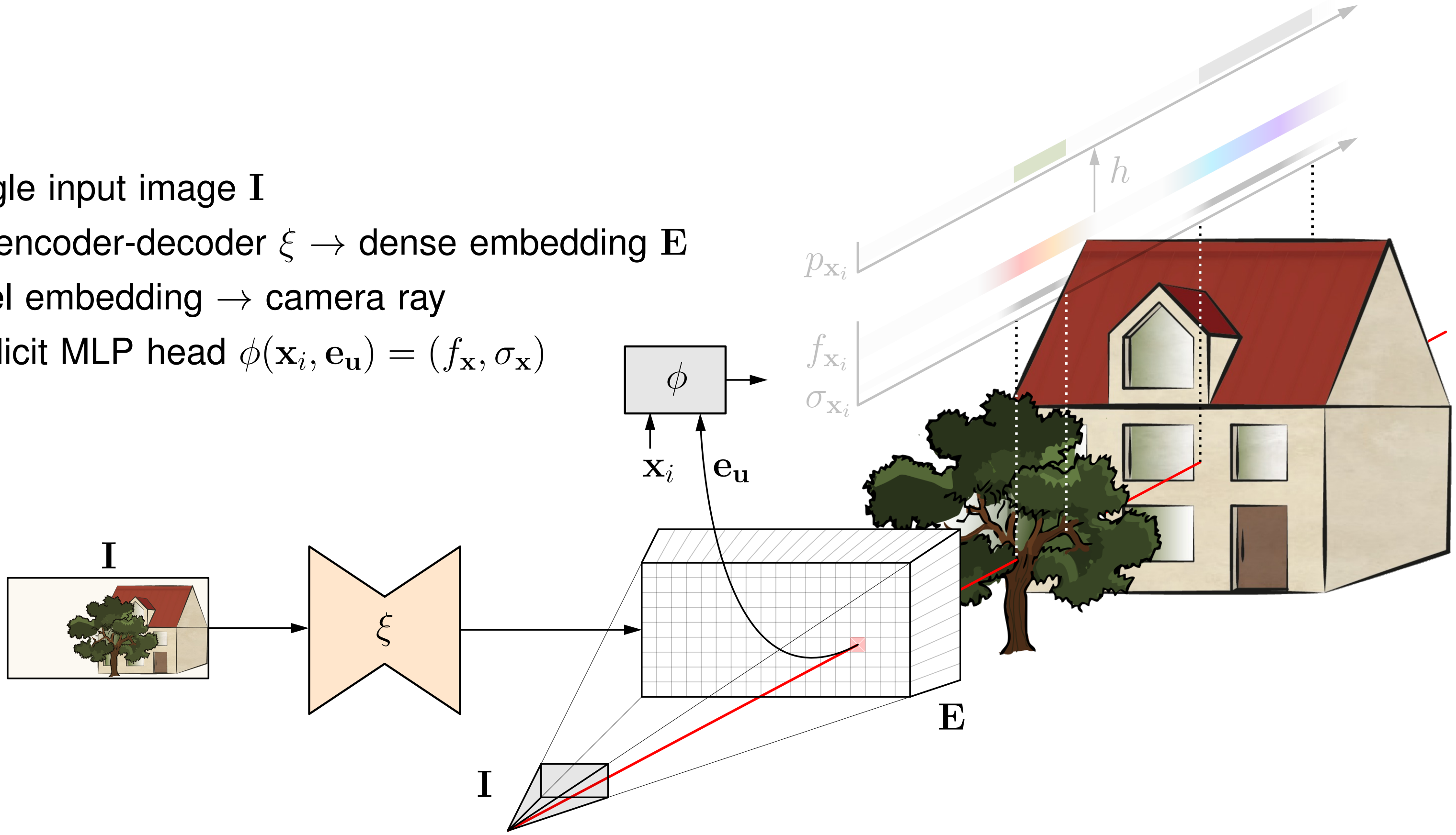
Model Architecture

- Single input image $\mathbf{I}$
- 2D encoder-decoder $\xi \rightarrow$ dense embedding $\mathbf{E}$
- Pixel embedding $\rightarrow$ camera ray



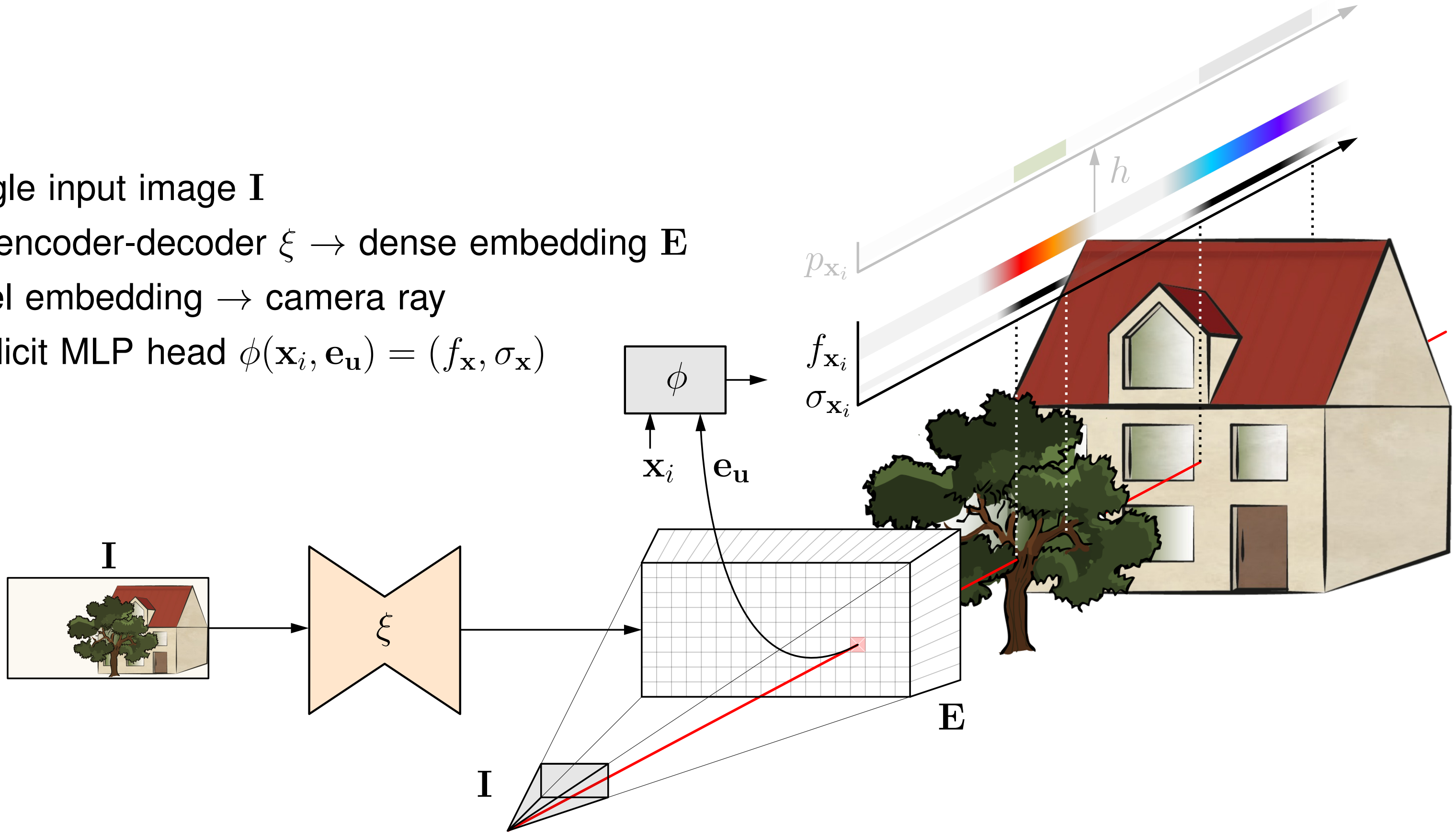
Model Architecture

- Single input image $\mathbf{I}$
- 2D encoder-decoder $\xi \rightarrow$ dense embedding $\mathbf{E}$
- Pixel embedding $\rightarrow$ camera ray
- Implicit MLP head $\phi(\mathbf{x}_i, \mathbf{e}_\mathbf{u}) = (f_\mathbf{x}, \sigma_\mathbf{x})$



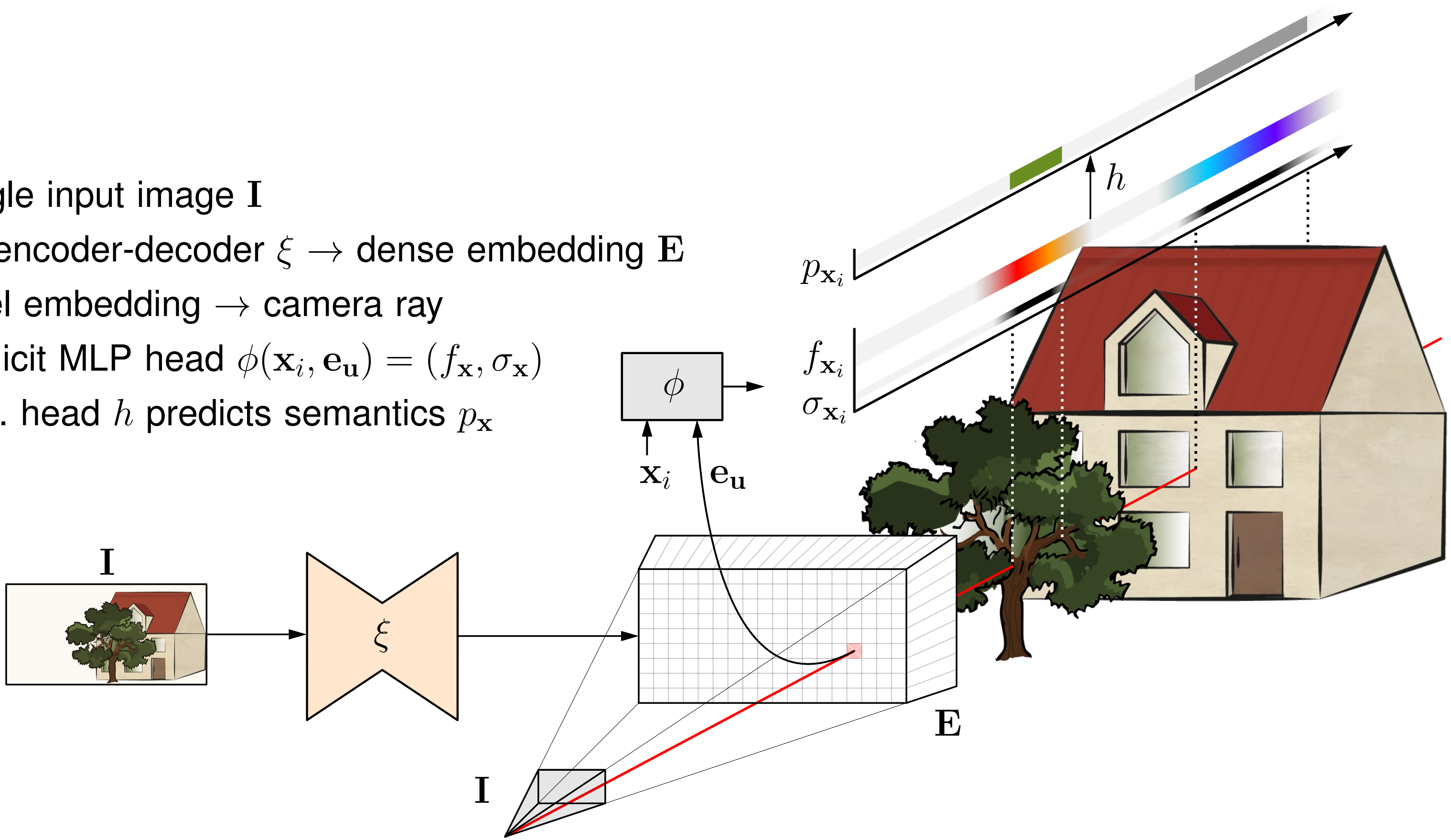
Model Architecture

- Single input image $\mathbf{I}$
- 2D encoder-decoder $\xi \rightarrow$ dense embedding $\mathbf{E}$
- Pixel embedding $\rightarrow$ camera ray
- Implicit MLP head $\phi(\mathbf{x}_i, \mathbf{e_u}) = (f_{\mathbf{x}}, \sigma_{\mathbf{x}})$

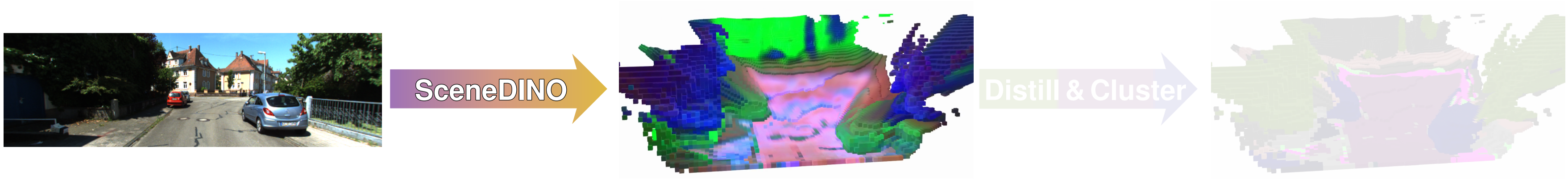


Model Architecture

- Single input image $\mathbf{I}$
- 2D encoder-decoder $\xi \rightarrow$ dense embedding $\mathbf{E}$
- Pixel embedding $\rightarrow$ camera ray
- Implicit MLP head $\phi(\mathbf{x}_i, \mathbf{e_u}) = (f_\mathbf{x}, \sigma_\mathbf{x})$
- Seg. head h predicts semantics $p_\mathbf{x}$

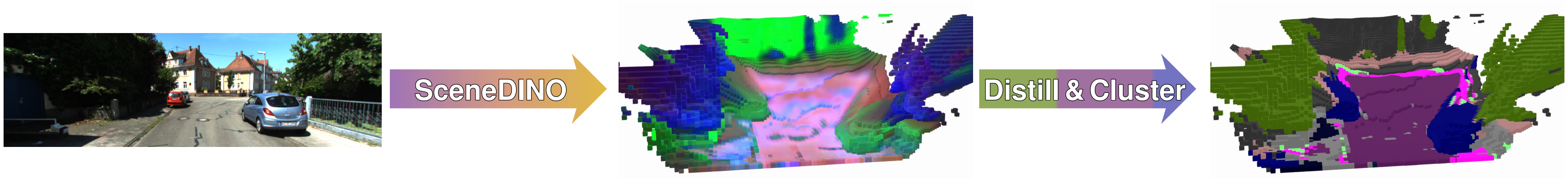


SceneDINO Training



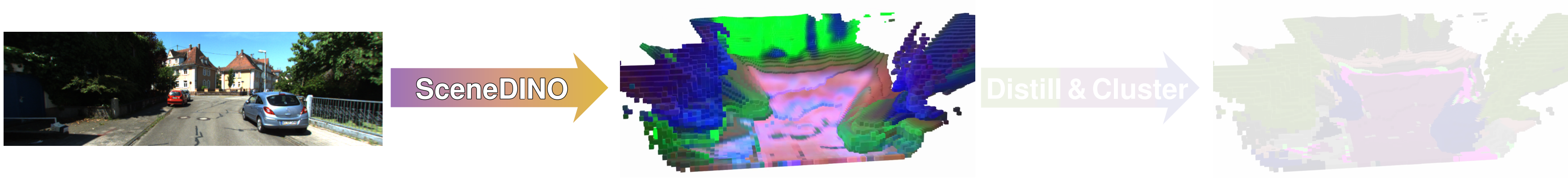
Single Input Image **3D Feature Field** **SSC Prediction**

SceneDINO Training



Single Input Image 3D Feature Field SSC Prediction

SceneDINO Training

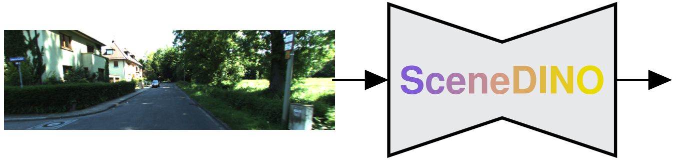


Single Input Image

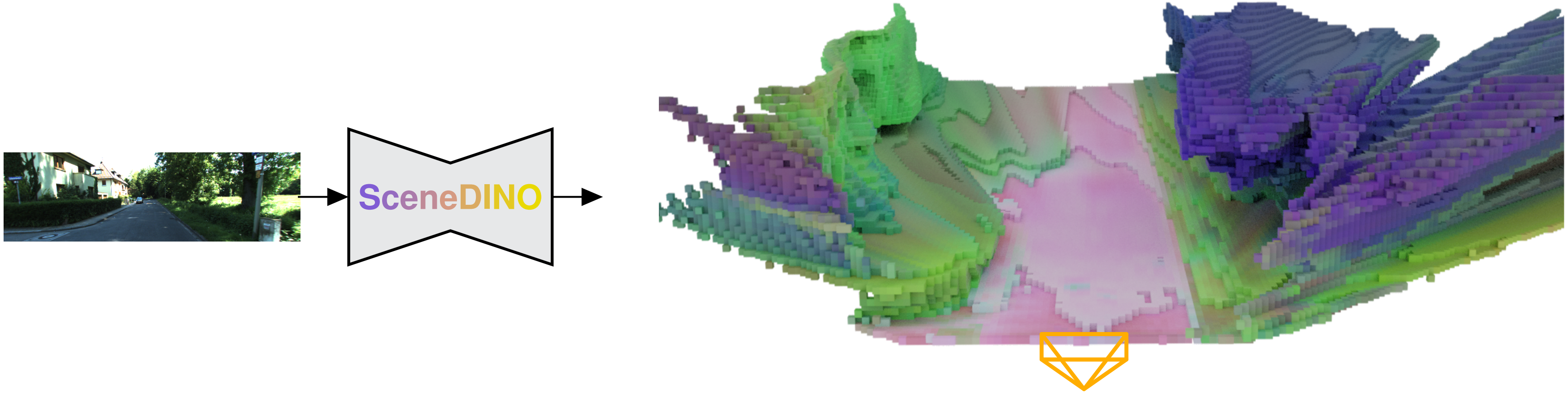
3D Feature Field

SSC Prediction

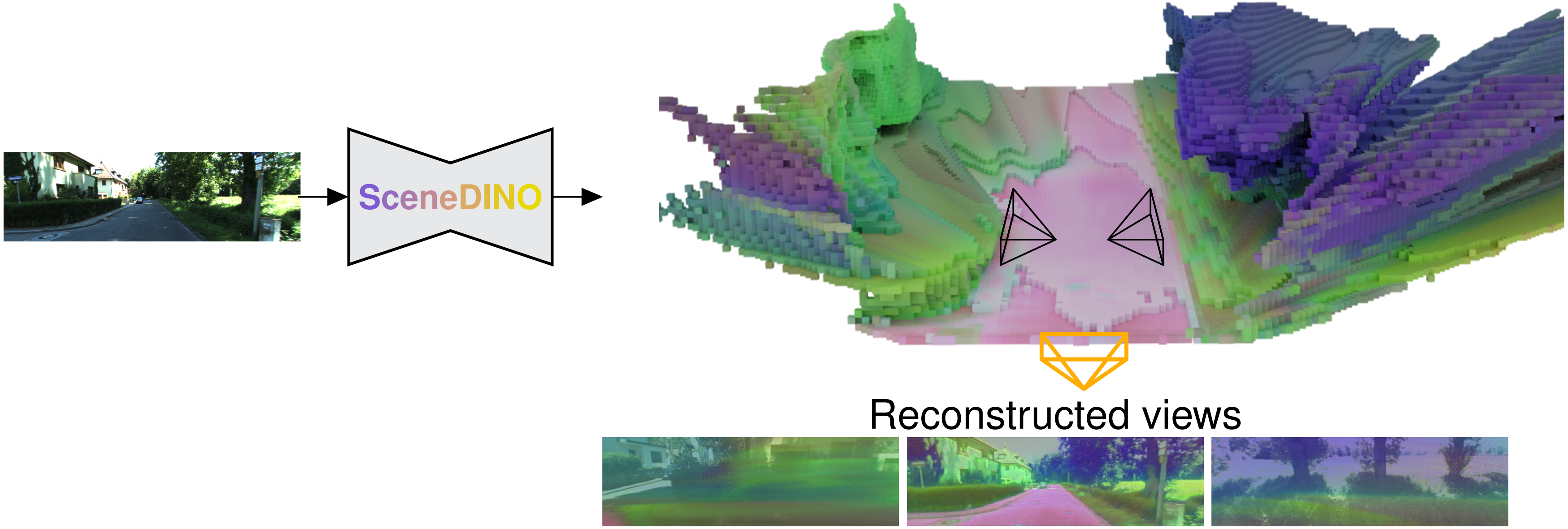
Multi-View Self-Supervision



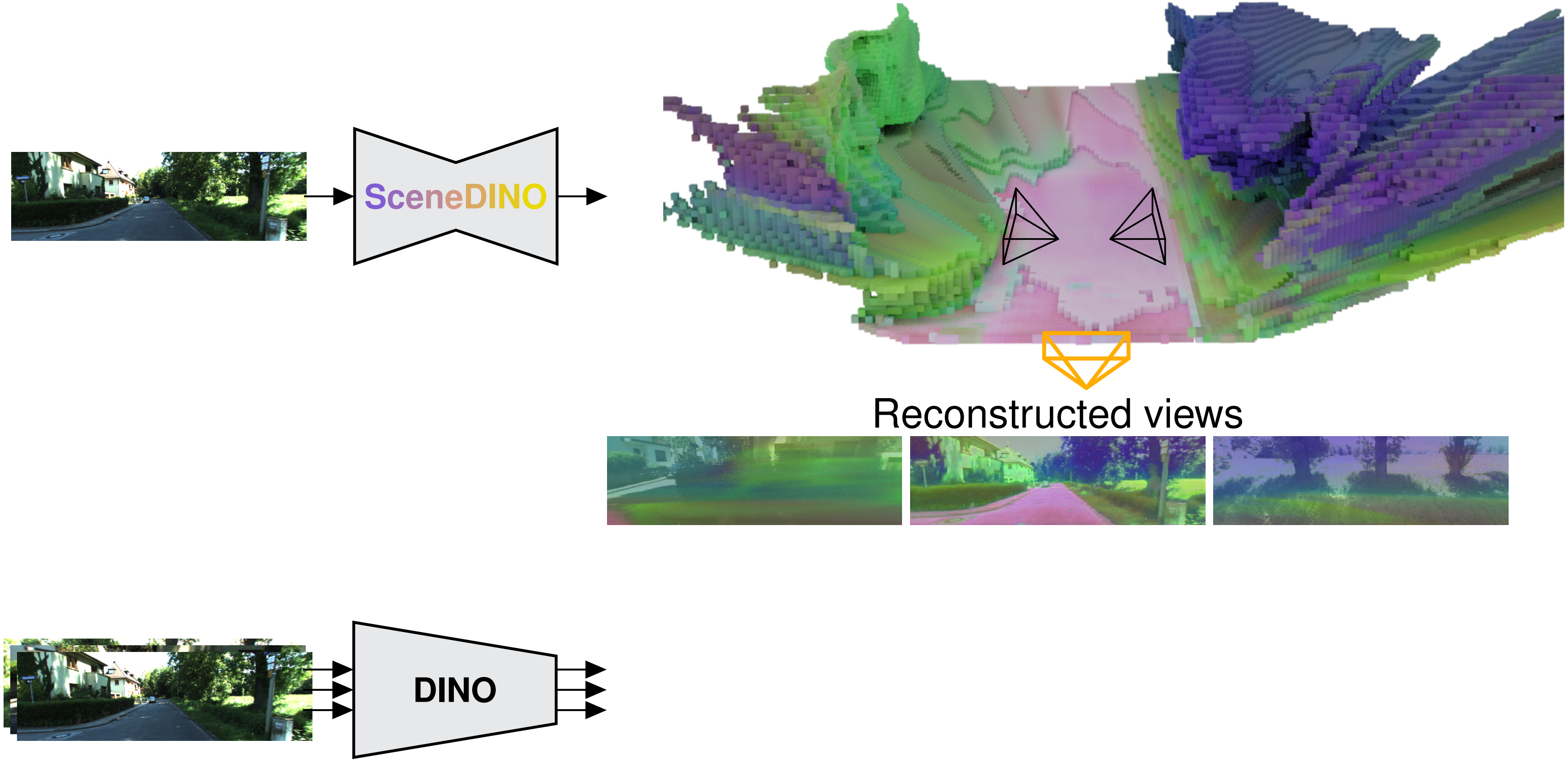
Multi-View Self-Supervision



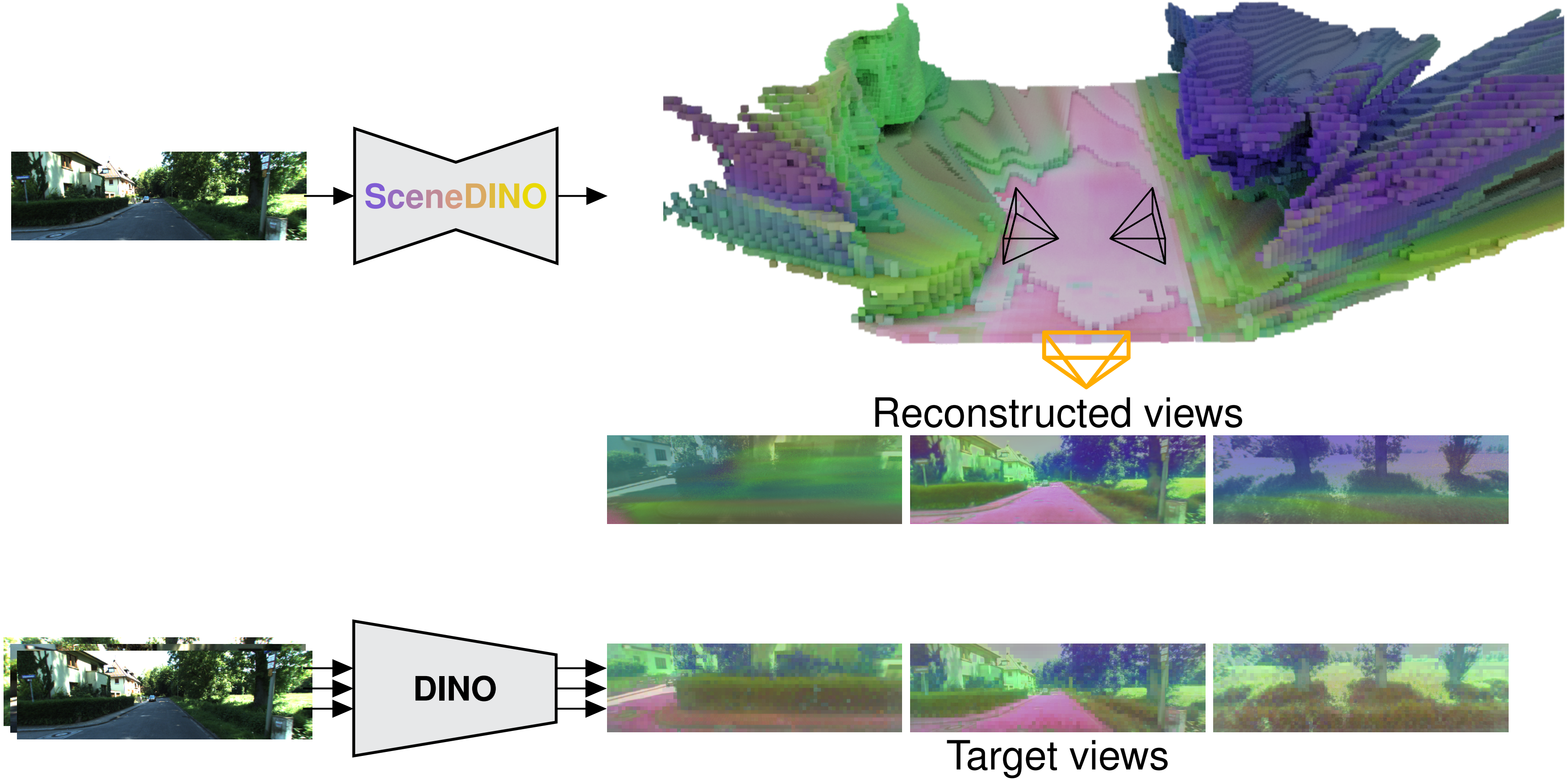
Multi-View Self-Supervision



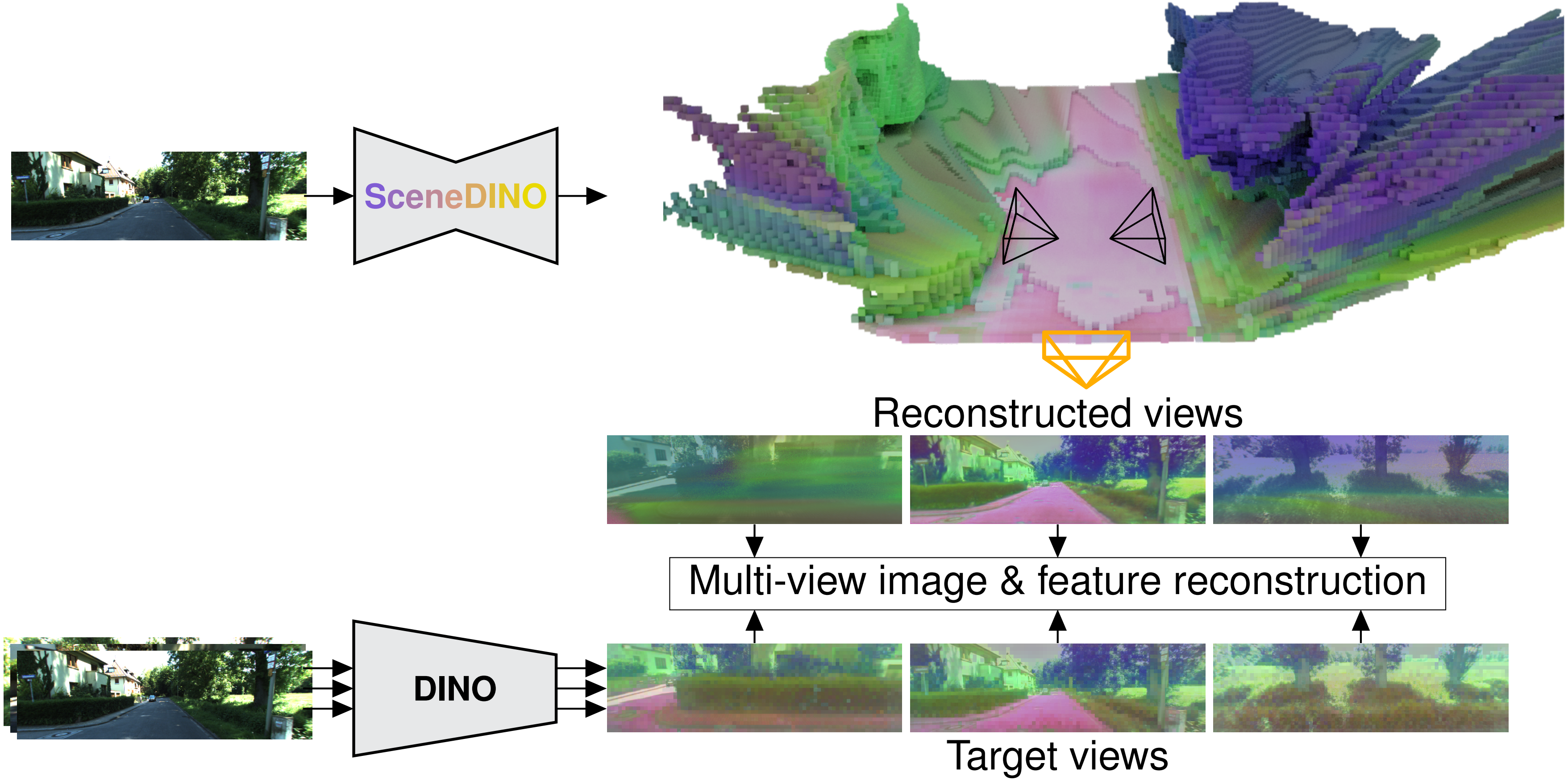
Multi-View Self-Supervision



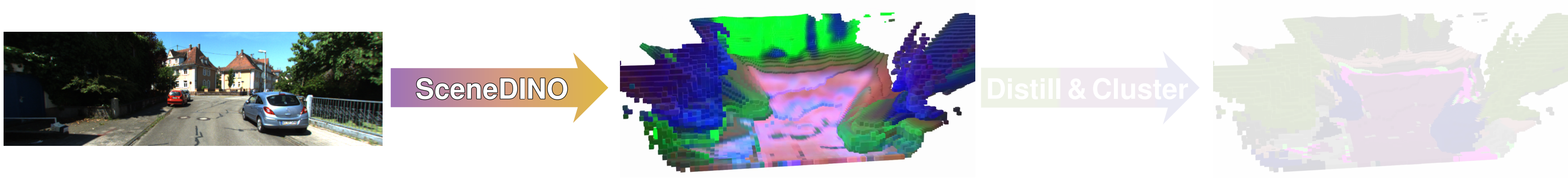
Multi-View Self-Supervision



Multi-View Self-Supervision



SceneDINO Training

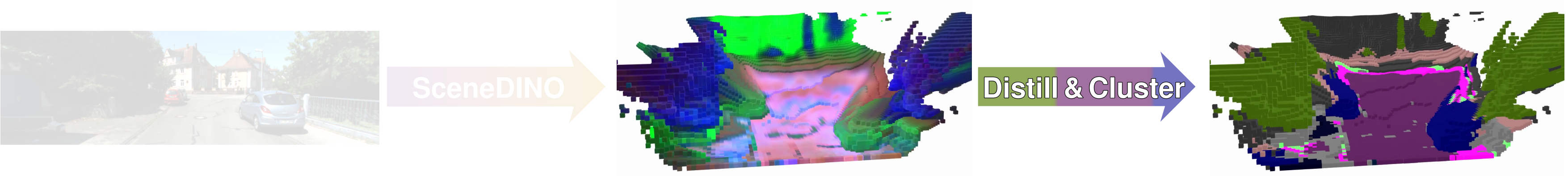


Single Input Image

3D Feature Field

SSC Prediction

SceneDINO Training



3D Feature Field **SSC Prediction**

Unsupervised Segmentation

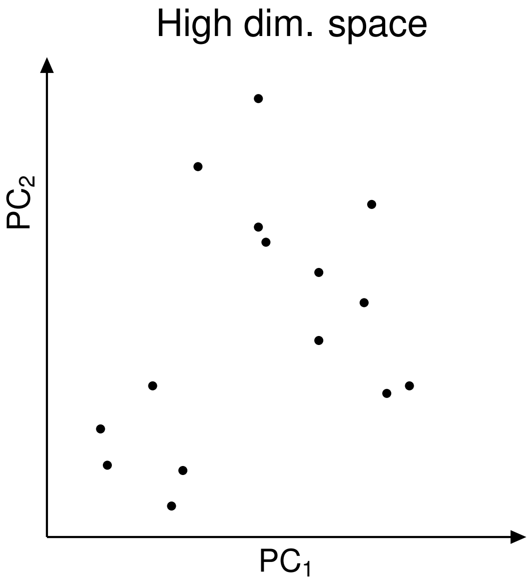
- **Goal:** Learn unsupervised segmentation head

Unsupervised Segmentation

- **Goal:** Learn unsupervised segmentation head
- **Idea:** Magnify semantic correspondence & cluster features

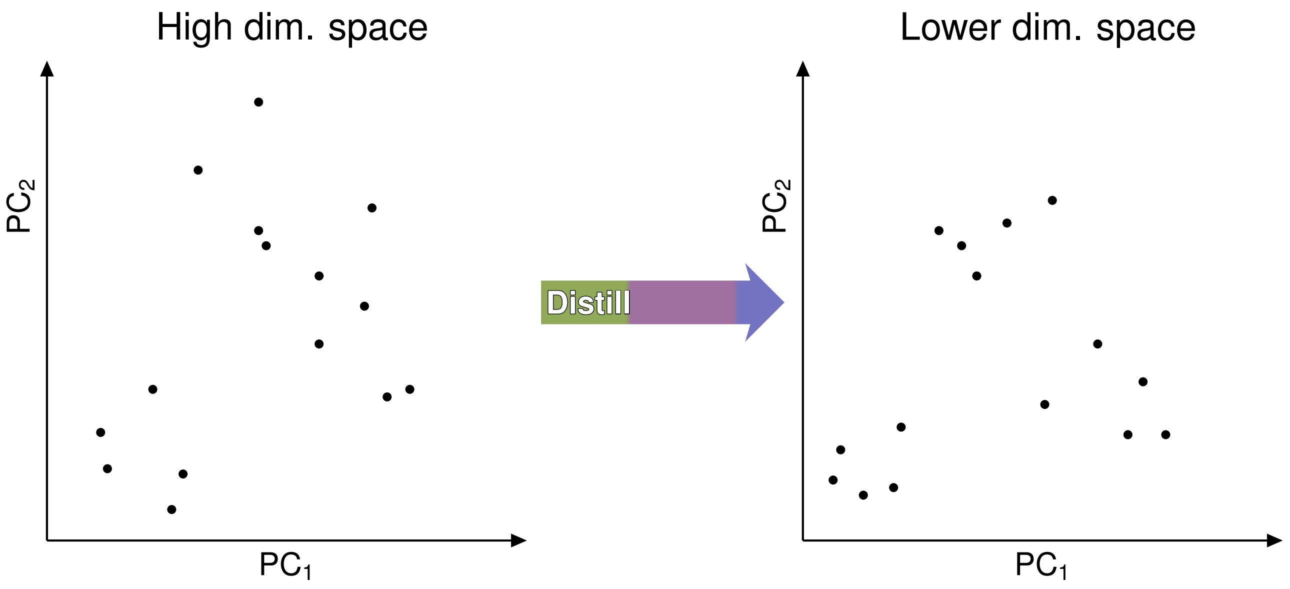
Unsupervised Segmentation

- **Goal:** Learn unsupervised segmentation head
- **Idea:** Magnify semantic correspondence & cluster features



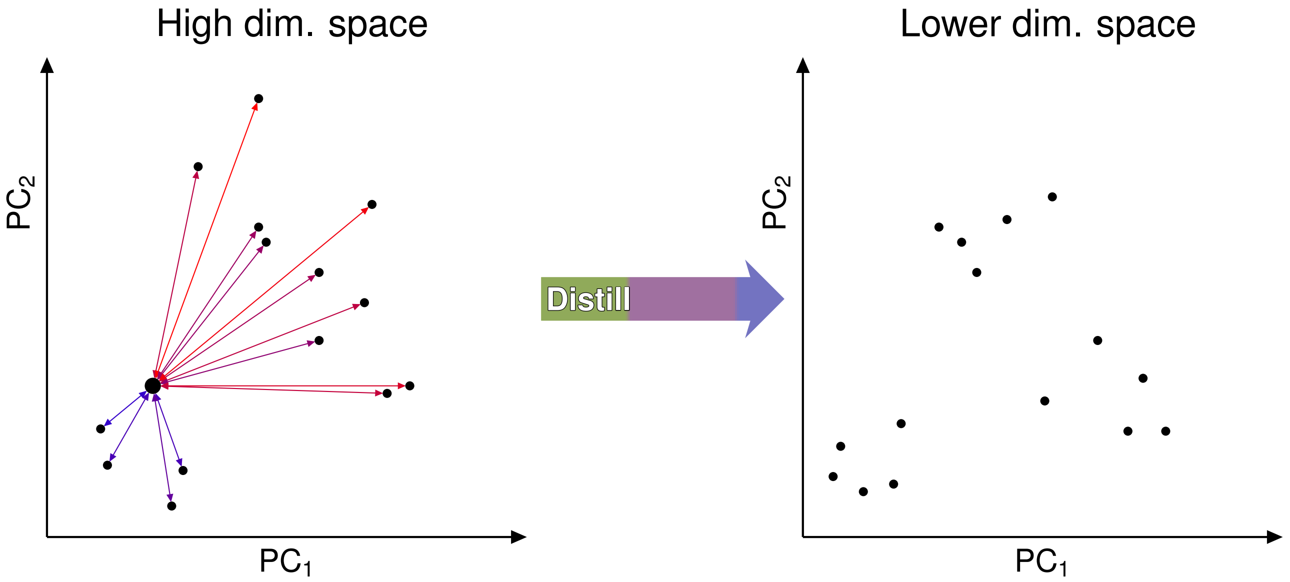
Unsupervised Segmentation

- **Goal:** Learn unsupervised segmentation head
- **Idea:** Magnify semantic correspondence & cluster features



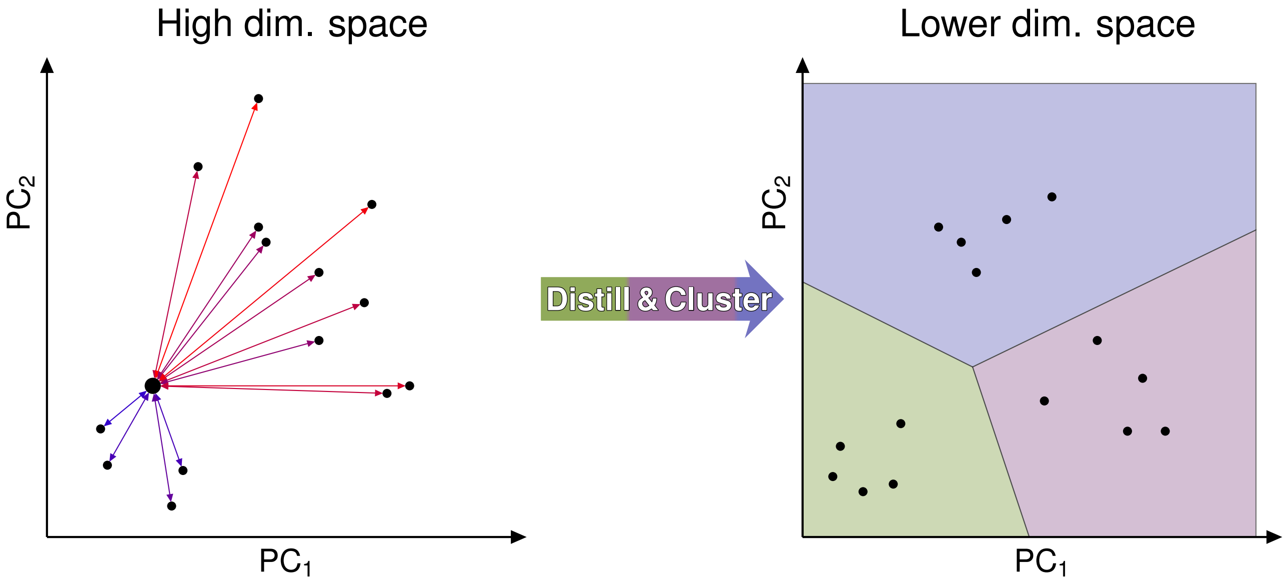
Unsupervised Segmentation

- **Goal:** Learn unsupervised segmentation head
- **Idea:** Magnify semantic correspondence & cluster features



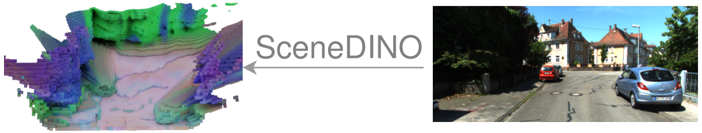
Unsupervised Segmentation

- **Goal:** Learn unsupervised segmentation head
- **Idea:** Magnify semantic correspondence & cluster features



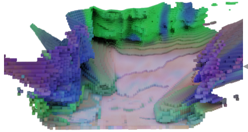
Unsupervised Segmentation

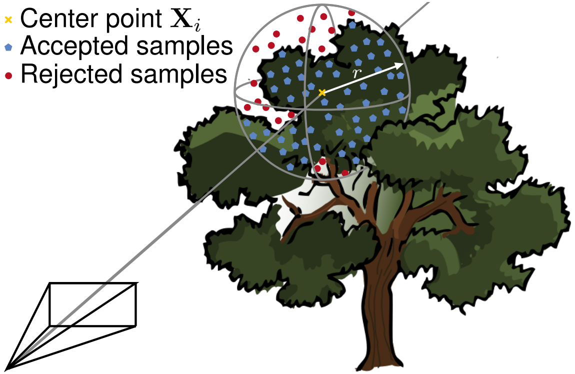
- **Goal:** Learn unsupervised segmentation head
- **Idea:** Magnify semantic correspondence & cluster features



Unsupervised Segmentation

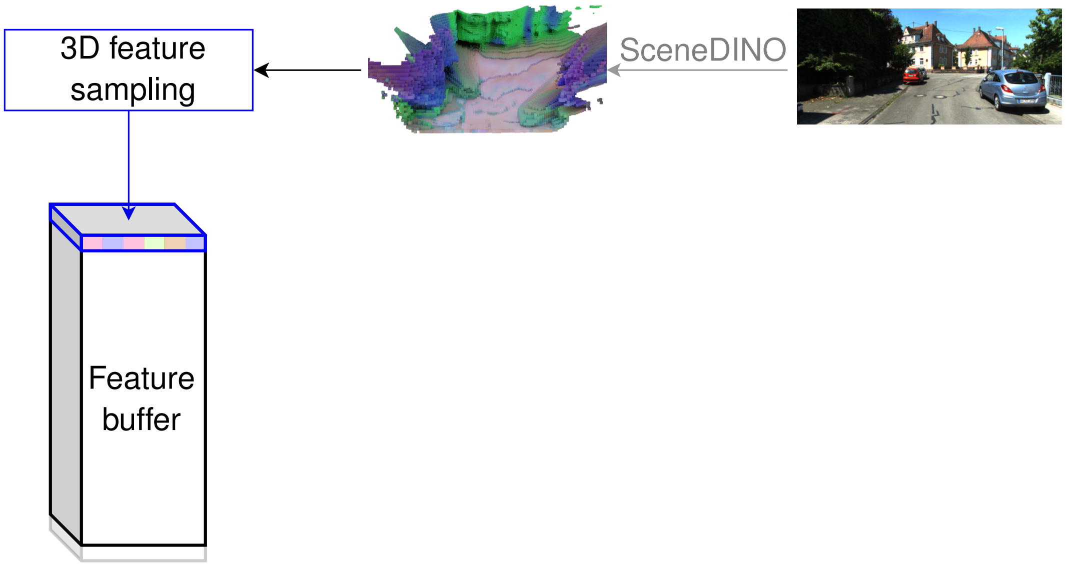
- **Goal:** Learn unsupervised segmentation head
- **Idea:** Magnify semantic correspondence & cluster features

3D feature sampling ←  ← SceneDINO ← 



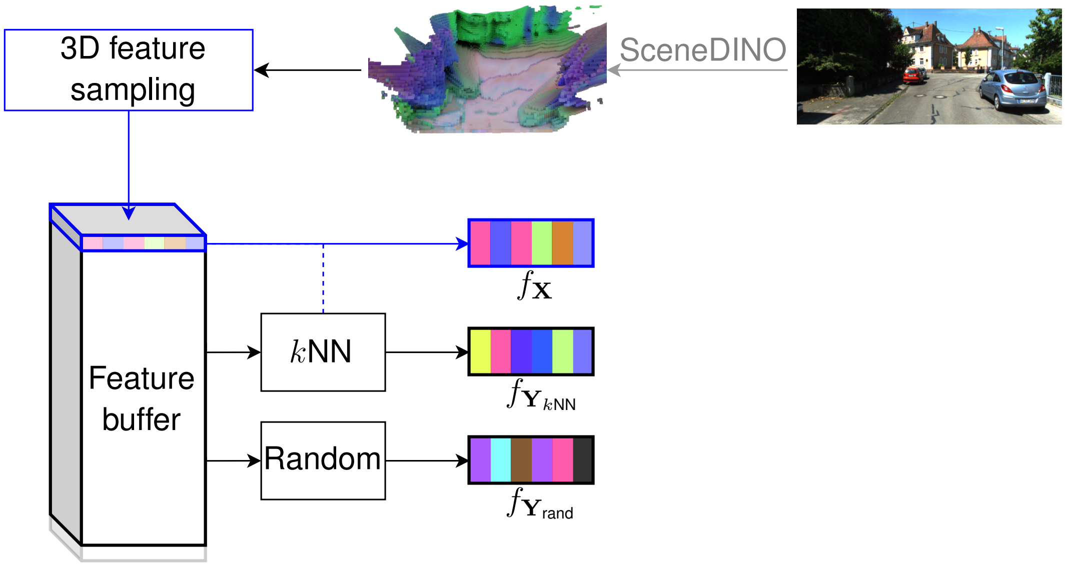
Unsupervised Segmentation

- **Goal:** Learn unsupervised segmentation head
- **Idea:** Magnify semantic correspondence & cluster features



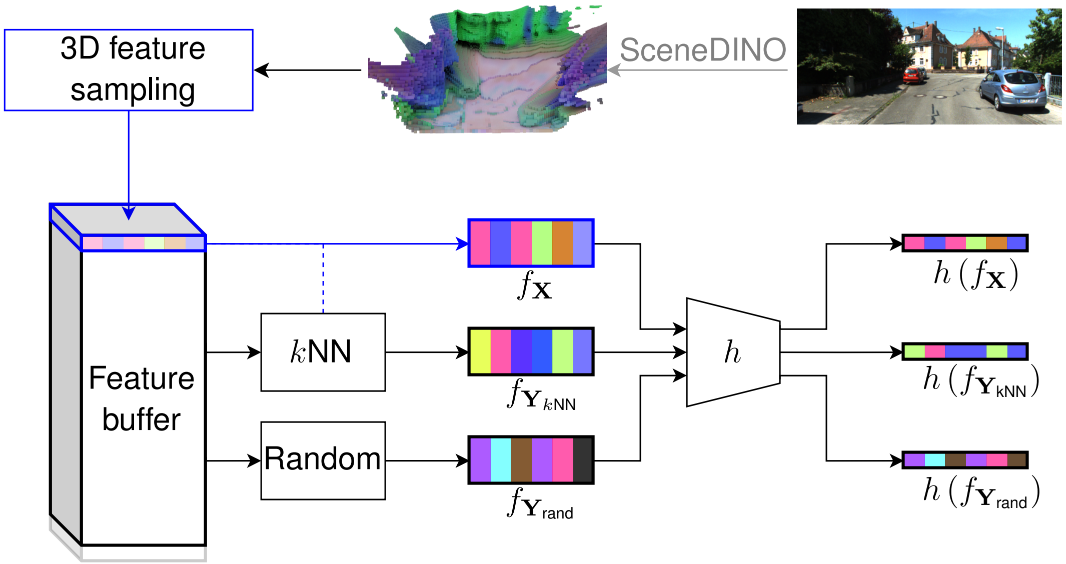
Unsupervised Segmentation

- **Goal:** Learn unsupervised segmentation head
- **Idea:** Magnify semantic correspondence & cluster features



Unsupervised Segmentation

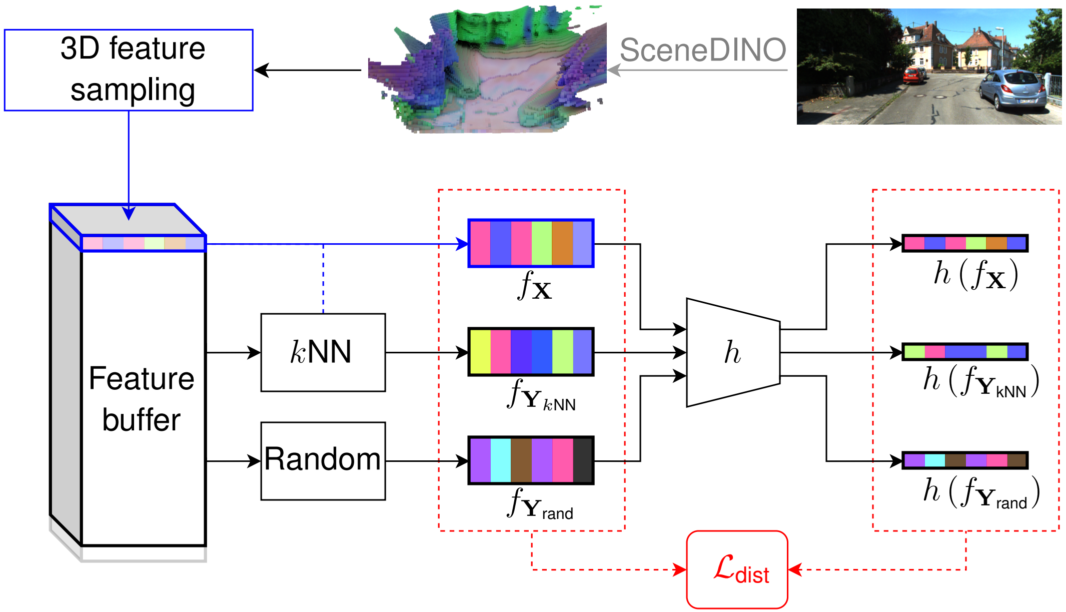
- **Goal:** Learn unsupervised segmentation head
- **Idea:** Magnify semantic correspondence & cluster features



- Head projects features down

Unsupervised Segmentation

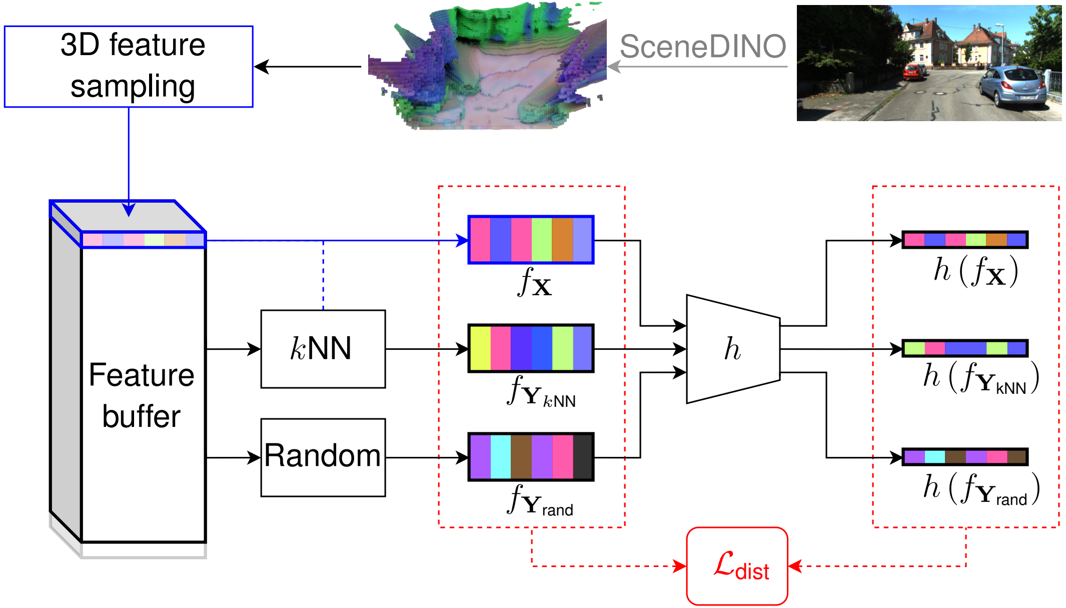
- **Goal:** Learn unsupervised segmentation head
- **Idea:** Magnify semantic correspondence & cluster features



- Head projects features down
- $\mathcal{L}_{\mathrm{dist}}$ aligns correspondences

Unsupervised Segmentation

- **Goal:** Learn unsupervised segmentation head
- **Idea:** Magnify semantic correspondence & cluster features



- Head projects features down
- $\mathcal{L}_{\text{dist}}$ aligns correspondences

$$f_{\mathbf{X}} \longleftrightarrow f_{\mathbf{X}}$$
$$f_{\mathbf{X}} \longleftrightarrow f_{\mathbf{Y}_{k\text{NN}}}$$
$$f_{\mathbf{X}} \longleftrightarrow f_{\mathbf{Y}_{\text{rand}}}$$

- k-means cluster distilled features

Results: SceneDINO

Results: Unsupervised SSC

- KITTI-360-SSCBench experiments (full range 51.2 m validation)

Method	Unsupervised	Target features	mIoU (in %, ↑)
S4C [3] (2D supervised)	✗	n/a	10.19
S4C [3] + STEGO [4]	✓	DINO	6.60

[3] A. Hayler *et al.*, “S4C: Self-supervised semantic scene completion with neural fields,” in *3DV*, 2024.
[4] M. Hamilton *et al.*, “Unsupervised semantic segmentation by distilling feature correspondences,” in *ICLR*, 2022.

Results: Unsupervised SSC

- KITTI-360-SSCBench experiments (full range 51.2 m validation)

Method	**Unsupervised**	**Target features**	**mIoU** (in %, ↑)
S4C [3] (2D supervised)	✗	n/a	10.19
S4C [3] + STEGO [4]	✓	DINO	6.60
SceneDINO (Ours)	✓	DINO	**8.00**

[3] A. Hayler *et al.*, "S4C: Self-supervised semantic scene completion with neural fields," in *3DV*, 2024.
[4] M. Hamilton *et al.*, "Unsupervised semantic segmentation by distilling feature correspondences," in *ICLR*, 2022.

Results: Unsupervised SSC

- KITTI-360-SSCBench experiments (full range 51.2 m validation)

Method	**Unsupervised**	**Target features**	**mIoU** (in %, ↑)
S4C [3] (2D supervised)	✗	n/a	10.19
S4C [3] + STEGO [4]	✓	DINO	6.60
SceneDINO (Ours)	✓	DINO	8.00
SceneDINO (Ours)	✓	DINOv2	**9.08**

[3] A. Hayler *et al.*, "S4C: Self-supervised semantic scene completion with neural fields," in *3DV*, 2024.
[4] M. Hamilton *et al.*, "Unsupervised semantic segmentation by distilling feature correspondences," in *ICLR*, 2022.

Results: Unsupervised SSC

- KITTI-360-SSCBench experiments (full range 51.2 m validation)

Method	**Unsupervised**	**Target features**	**mIoU** (in %, ↑)
S4C [3] (2D supervised)	✗	n/a	10.19
S4C [3] + STEGO [4]	✓	DINO	6.60
SceneDINO (Ours)	✓	DINO	8.00
SceneDINO (Ours)	✓	DINOv2	**9.08**

State-of-the-art unsupervised semantic scene completion accuracy

[3] A. Hayler *et al.*, "S4C: Self-supervised semantic scene completion with neural fields," in *3DV*, 2024.
[4] M. Hamilton *et al.*, "Unsupervised semantic segmentation by distilling feature correspondences," in *ICLR*, 2022.

Results: Unsupervised SSC

- KITTI-360-SSCBench experiments (full range 51.2 m validation)

Method	**Unsupervised**	**Target features**	**mIoU** (in %, ↑)
S4C [3] (2D supervised)	✗	n/a	10.19
S4C [3] + STEGO [4]	✓	DINO	6.60
SceneDINO (Ours)	✓	DINO	8.00
SceneDINO (Ours)	✓	DINOv2	9.08
SceneDINO (Ours)	✗ (linear)	DINOv2	10.57

[3] A. Hayler *et al.*, “S4C: Self-supervised semantic scene completion with neural fields,” in *3DV*, 2024.
[4] M. Hamilton *et al.*, “Unsupervised semantic segmentation by distilling feature correspondences,” in *ICLR*, 2022.

Results: Unsupervised SSC

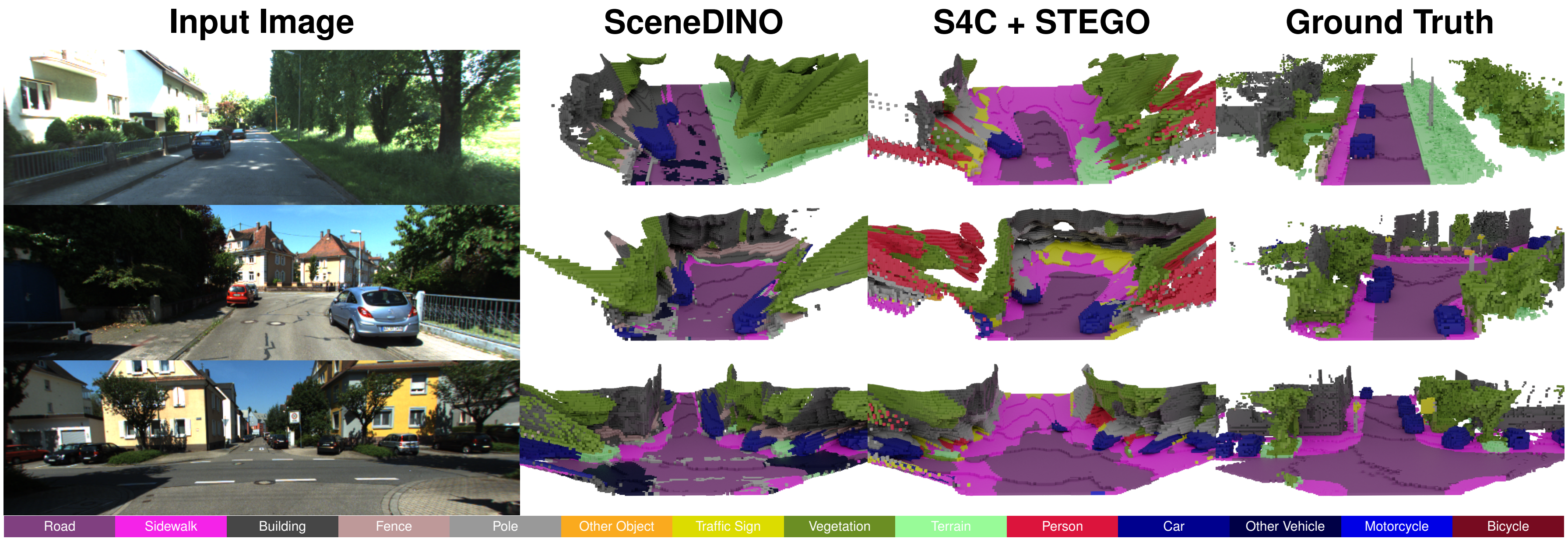
- KITTI-360-SSCBench experiments (full range 51.2 m validation)

Method	**Unsupervised**	**Target features**	**mIoU** (in %, ↑)
S4C [3] (2D supervised)	✗	n/a	10.19
S4C [3] + STEGO [4]	✓	DINO	6.60
SceneDINO (Ours)	✓	DINO	8.00
SceneDINO (Ours)	✓	DINOv2	9.08
SceneDINO (Ours)	✗ (linear)	DINOv2	10.57

Linear probing outperforms 2D supervised S4C

[3] A. Hayler *et al.*, "S4C: Self-supervised semantic scene completion with neural fields," in *3DV*, 2024.
[4] M. Hamilton *et al.*, "Unsupervised semantic segmentation by distilling feature correspondences," in *ICLR*, 2022.

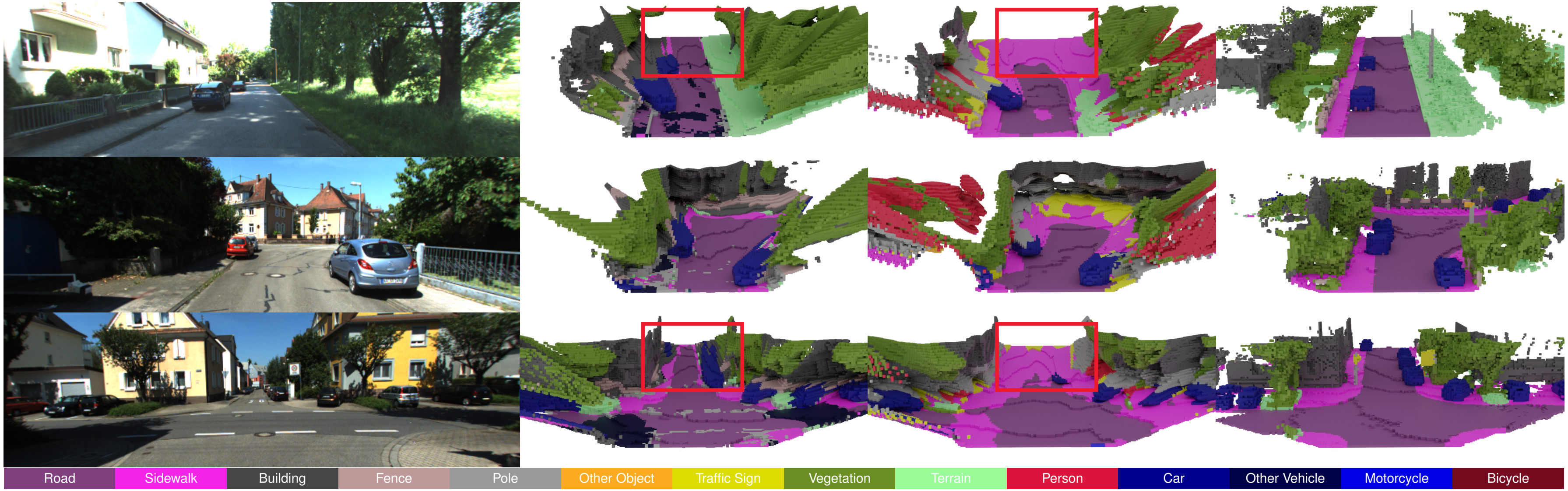
Results: Unsupervised SSC



Results: Unsupervised SSC

ICCV HONOLULU HAWAII OCT 19-23, 2025

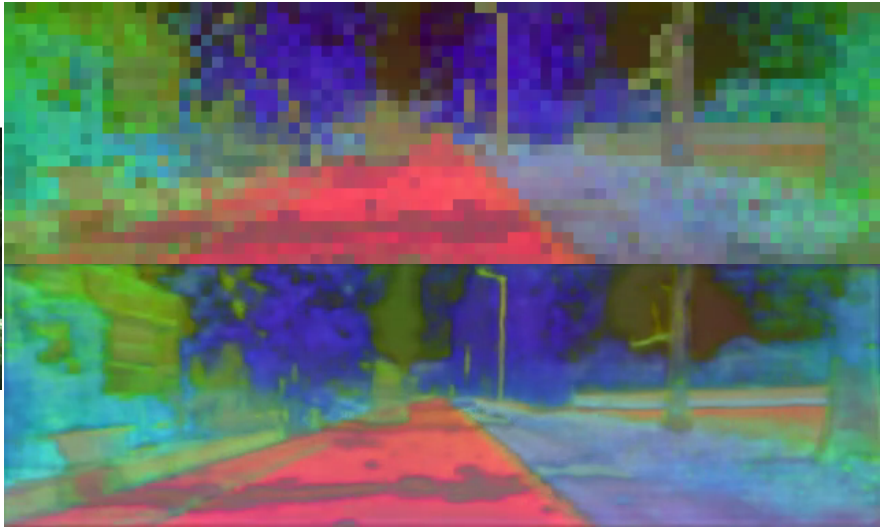
Input Image | SceneDINO | S4C + STEGO | Ground Truth



Results: SceneDINO in 2D

Input Image

DINO



SceneDINO

Results: Multi-View Feature Consistency

ICCV OCT 19-23, 2025 HONOLULU HAWAII

- Multi-view consistency results using optical flow alignment

Method	**KITTI-360**		**RealEstate10K**	
	L_1 (↓)	**cos-sim (↑)**	**L_1 (↓)**	**cos-sim (↑)**
DINO [5]	16.06	0.70	14.41	0.75
SceneDINO (w/ DINO)	**6.45**	**0.93**	**5.87**	**0.95**
DINOv2 [6]	15.83	0.70	14.20	0.75
FiT3D [7]	7.02	0.93	5.67	0.95
SceneDINO (w/ DINOv2)	**5.24**	**0.96**	**4.87**	**0.97**

[5] M. Caron *et al.*, “Emerging properties in self-supervised vision transformers,” in *ICCV*, 2021.
[6] M. Oquab *et al.*, “DINOv2: Learning robust visual features without supervision,” *TMLR*, 2024.
[7] Y. Yue *et al.*, “Improving 2D feature representations by 3D-aware fine-tuning,” in *ECCV*, 2024.

Results: Multi-View Feature Consistency

- Multi-view consistency results using optical flow alignment

Method	KITTI-360		RealEstate10K	
	L_1 (↓)	cos-sim (↑)	L_1 (↓)	cos-sim (↑)
DINO [5]	16.06	0.70	14.41	0.75
SceneDINO (w/ DINO)	**6.45**	**0.93**	**5.87**	**0.95**
DINOv2 [6]	15.83	0.70	14.20	0.75
FiT3D [7]	7.02	0.93	5.67	0.95
SceneDINO (w/ DINOv2)	**5.24**	**0.96**	**4.87**	**0.97**

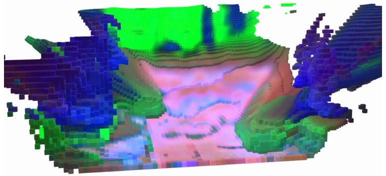
SceneDINO's features are significantly more multi-view consistent

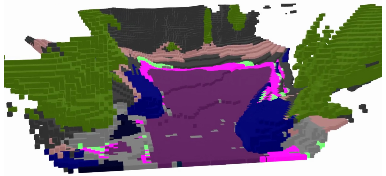
[5] M. Caron *et al.*, "Emerging properties in self-supervised vision transformers," in *ICCV*, 2021.
[6] M. Oquab *et al.*, "DINOv2: Learning robust visual features without supervision," *TMLR*, 2024.
[7] Y. Yue *et al.*, "Improving 2D feature representations by 3D-aware fine-tuning," in *ECCV*, 2024.

Conclusion

We presented SceneDINO for unsupervised SSC

- **Multi-view self-supervision** effective for 3D scene understanding
- Single image $\rightarrow$ **3D geometry** & **expressive features**
- Distilling & clustering leads to **SoTA accuracy** in unsupervised SSC
- Strong **linear probing**, **multi-view consistency**, and **domain generalization**





ICCV OCT 19-23, 2025 HONOLULU HAWAII

Paper | **Project Page** | **Code & Weights**

https://visinf.github.io/scenedino/

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 866008), the ERC Advanced Grant SIMULACRON, the DFG project CR 250/26-1 "4D-YouTube", the GNI Project "AICC", and the State of Hesse within the LOEWE emergenCITY center. This work was partially supported by the Deutsche Forschungsgemeinschaft (German Research Foundation, DFG) under Germany's Excellence Strategy (EXC 3066/1 "The Adaptive Mind", Project No. 533717223). C. Reich is supported by the Konrad Zuse School of Excellence in Learning and Intelligent Systems. C. Rupprecht is supported by an Amazon Research Award.

erc European Research Council Established by the European Commission

emergenCITY

ZUSE SCHOOL ELIZA

The adaptive Mind

Results: SceneDINO Analysis

- Analysing camera poses and target features

Δ mIoU	mIoU	Configuration
-0.12	7.88	w/ estimated ORB-SLAM3 poses
—	8.00	Full framework (SceneDINO)
+1.08	9.08	DINOv2 target features (*vs.* DINO)

Results: SceneDINO Analysis

- Analysing camera poses and target features

△ mIoU	mIoU	Configuration
-0.12	7.88	w/ estimated ORB-SLAM3 poses
—	8.00	Full framework (SceneDINO)
+1.08	9.08	DINOv2 target features (*vs.* DINO)

SceneDINO can benefit from better target features

Results: SceneDINO Analysis

ICCV OCT 19-23, 2025 HONOLULU HAWAII

- Analysing camera poses and target features

△ mIoU	mIoU	Configuration
-0.12	7.88	w/ estimated ORB-SLAM3 poses
—	8.00	Full framework (SceneDINO)
+1.08	9.08	DINOv2 target features (*vs.* DINO)

SceneDINO can benefit from better target features

- Linear probing features (w/ 2D sem. GT)

Probing approach	Target features	mIoU
Linear	DINO	9.34
	DINOv2	**10.57**
S4C (full training)	n/a	10.19

Results: SceneDINO Analysis

ICCV OCT 19-23, 2025 HONOLULU HAWAII

- Analysing camera poses and target features

△ mIoU	mIoU	Configuration
-0.12	7.88	w/ estimated ORB-SLAM3 poses
—	8.00	Full framework (SceneDINO)
+1.08	9.08	DINOv2 target features (*vs.* DINO)

SceneDINO can benefit from better target features

- Linear probing features (w/ 2D sem. GT)

Probing approach	Target features	mIoU
Linear	DINO	9.34
	DINOv2	**10.57**
S4C (full training)	n/a	10.19

Linear probing outperforms 2D supervised S4C