ICCV OCT 19-23, 2025 HONOLULU HAWAII

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

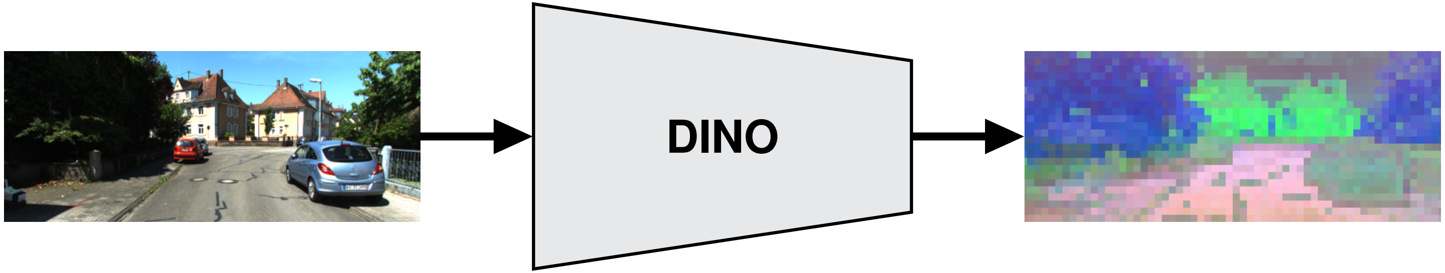
3 UNIVERSITY OF OXFORD

4 mcml Munich Center for Machine Learning

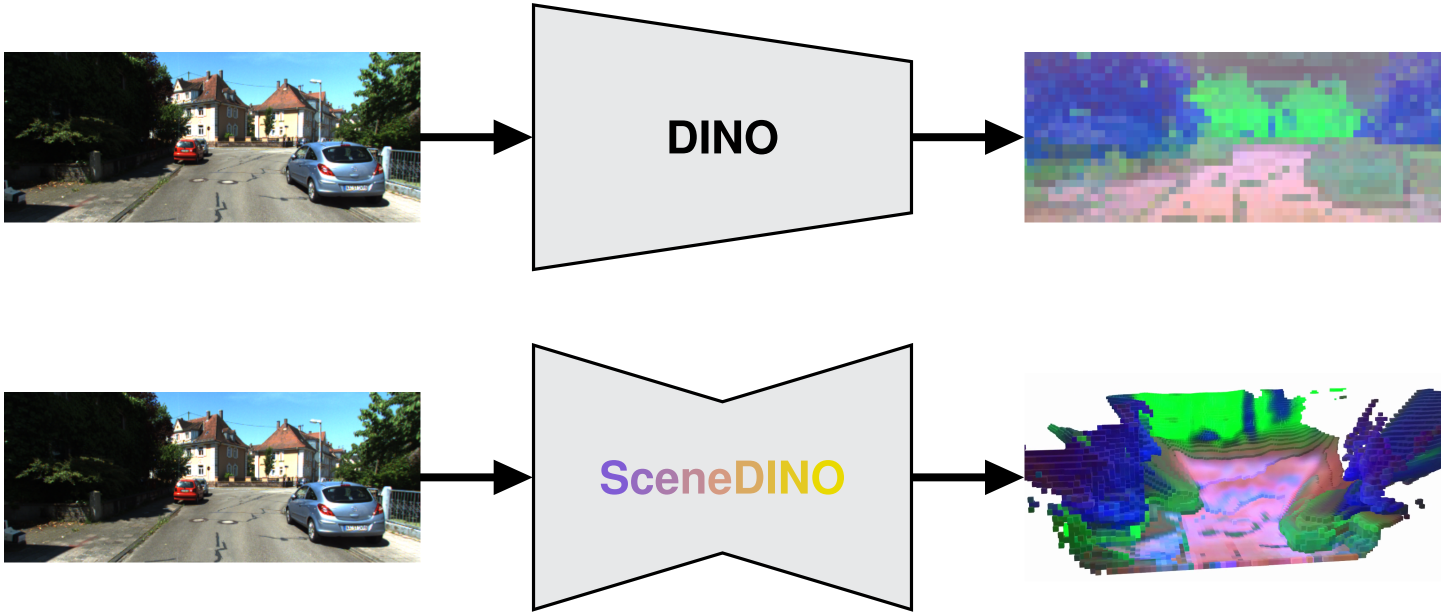
5 ZUSE SCHOOL ELIZA

6 hessian.AI

Motivation



Motivation



Bring DINO to 3D 🚀

Semantic Scene Completion (SSC)

a.k.a. Semantic Occupancy Prediction

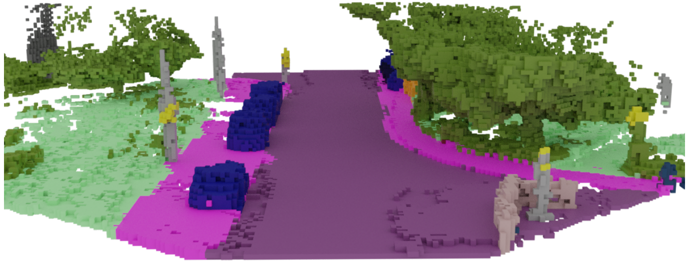
Single input image

Semantic Scene Completion (SSC)

a.k.a. Semantic Occupancy Prediction

Single input image

Dense 3D geometry & semantics

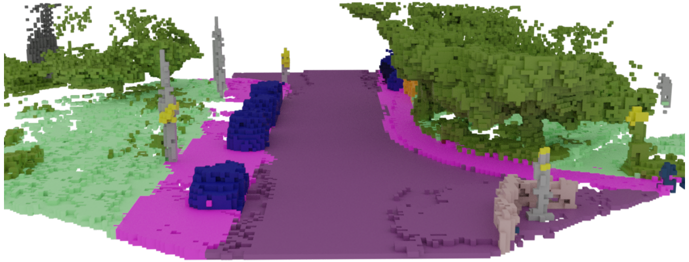


Semantic Scene Completion (SSC)

a.k.a. Semantic Occupancy Prediction

Single input image

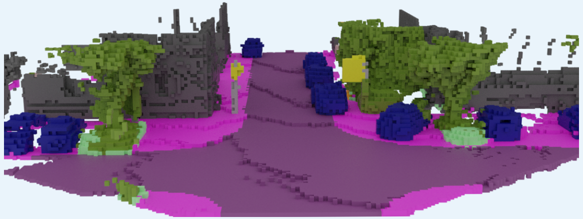
Dense 3D geometry & semantics



- ✓ Comprehensive 3D scene understanding task
- ✓ Applications in robotics, autonomous driving, medical image analysis, and civil engineering

Related Work: SSC

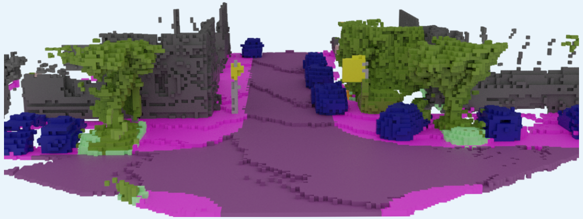
Geometric & 3D semantic supervision (*e.g.*, [1])



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

Related Work: SSC

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

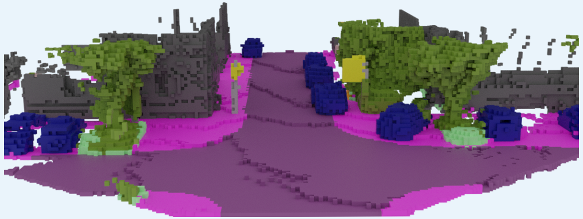


- – Ground truth very expensive
- – Special hardware needed
- – Infeasible to scale

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

Related Work: SSC

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



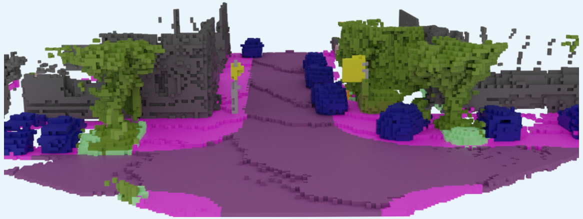
- Ground truth very expensive
- Special hardware needed
- Infeasible to scale

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

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

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])

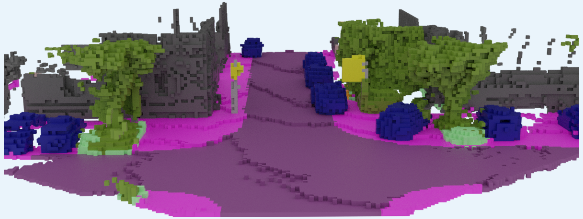
- Still, expensive to obtain
- Limited generalization

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

Related Work: SSC

ICCV OCT 19-23, 2025 HONOLULU HAWAII

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



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

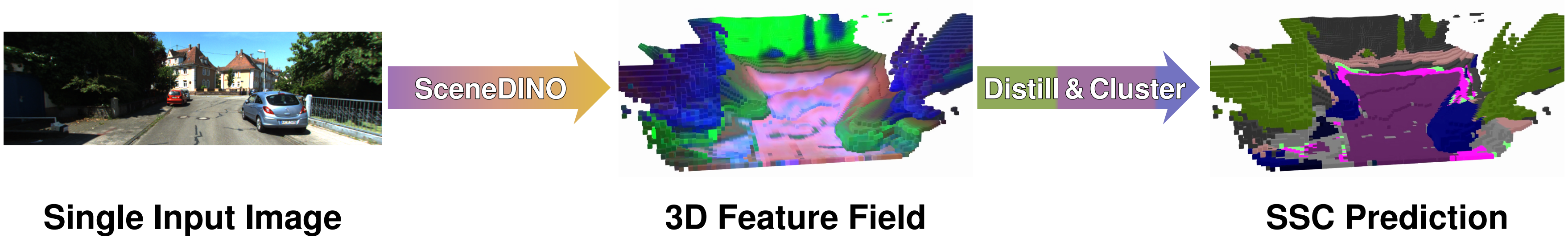
2D supervision (*e.g.*, [2])

- 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.
[2] Y. Huang *et al.*, "SelfOcc: Self-supervised vision-based 3D occupancy prediction," in *CVPR*, 2024.

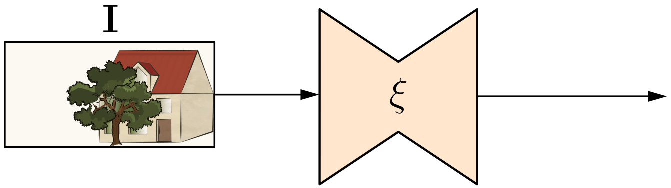
SceneDINO



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

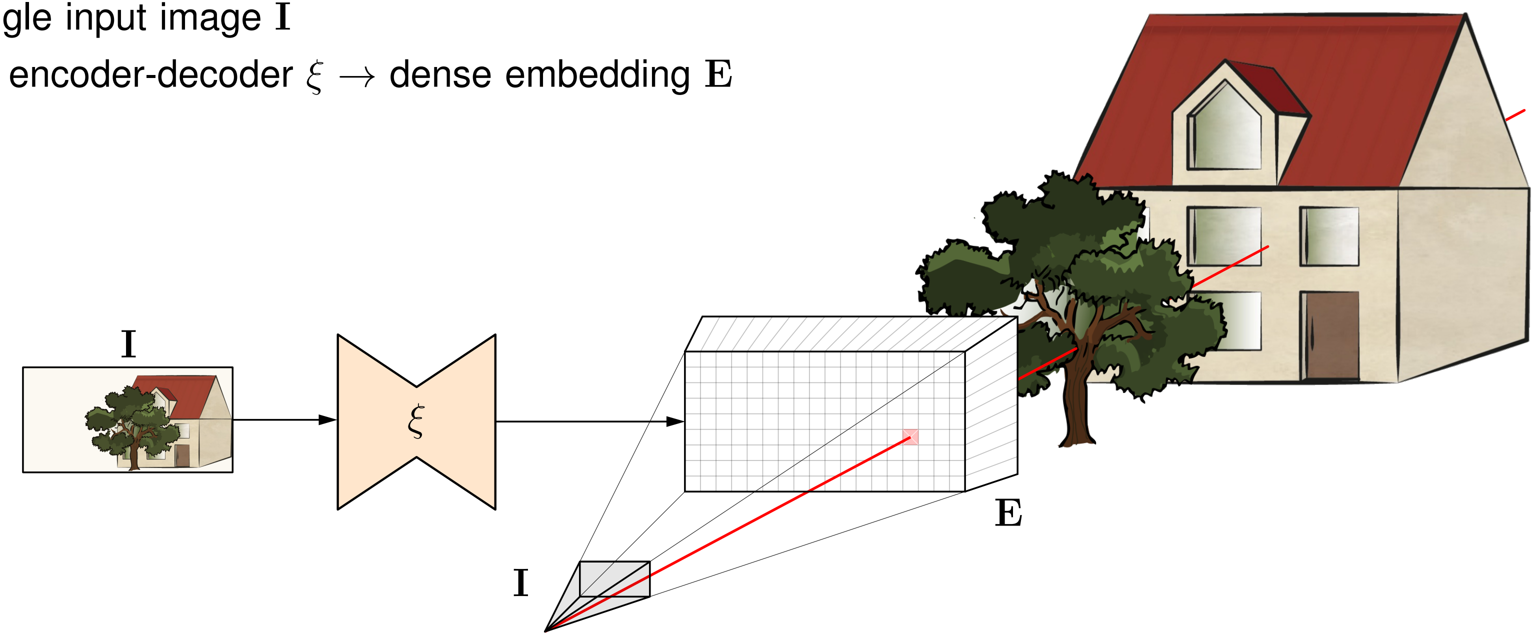
Model Architecture

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



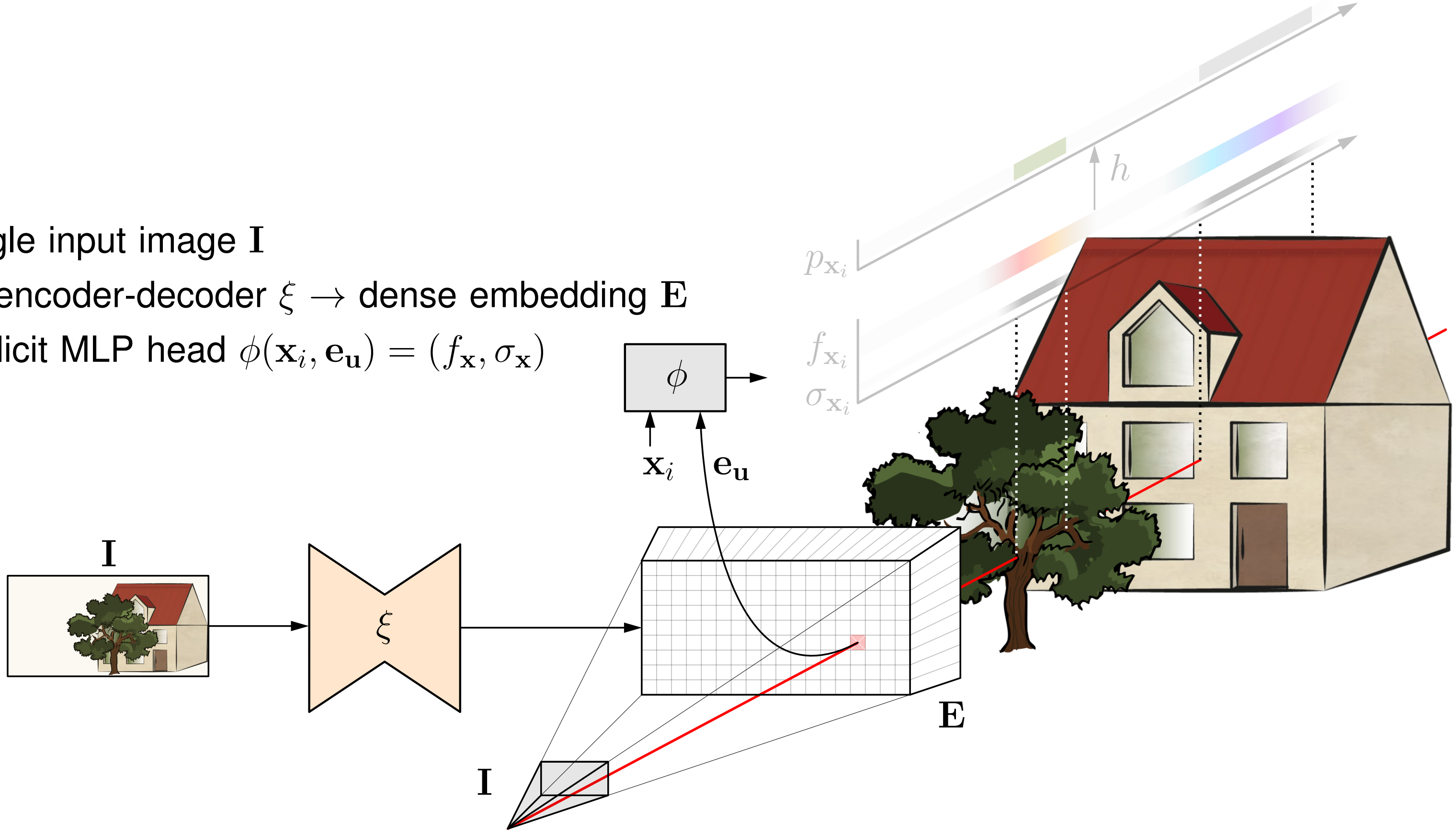
Model Architecture

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



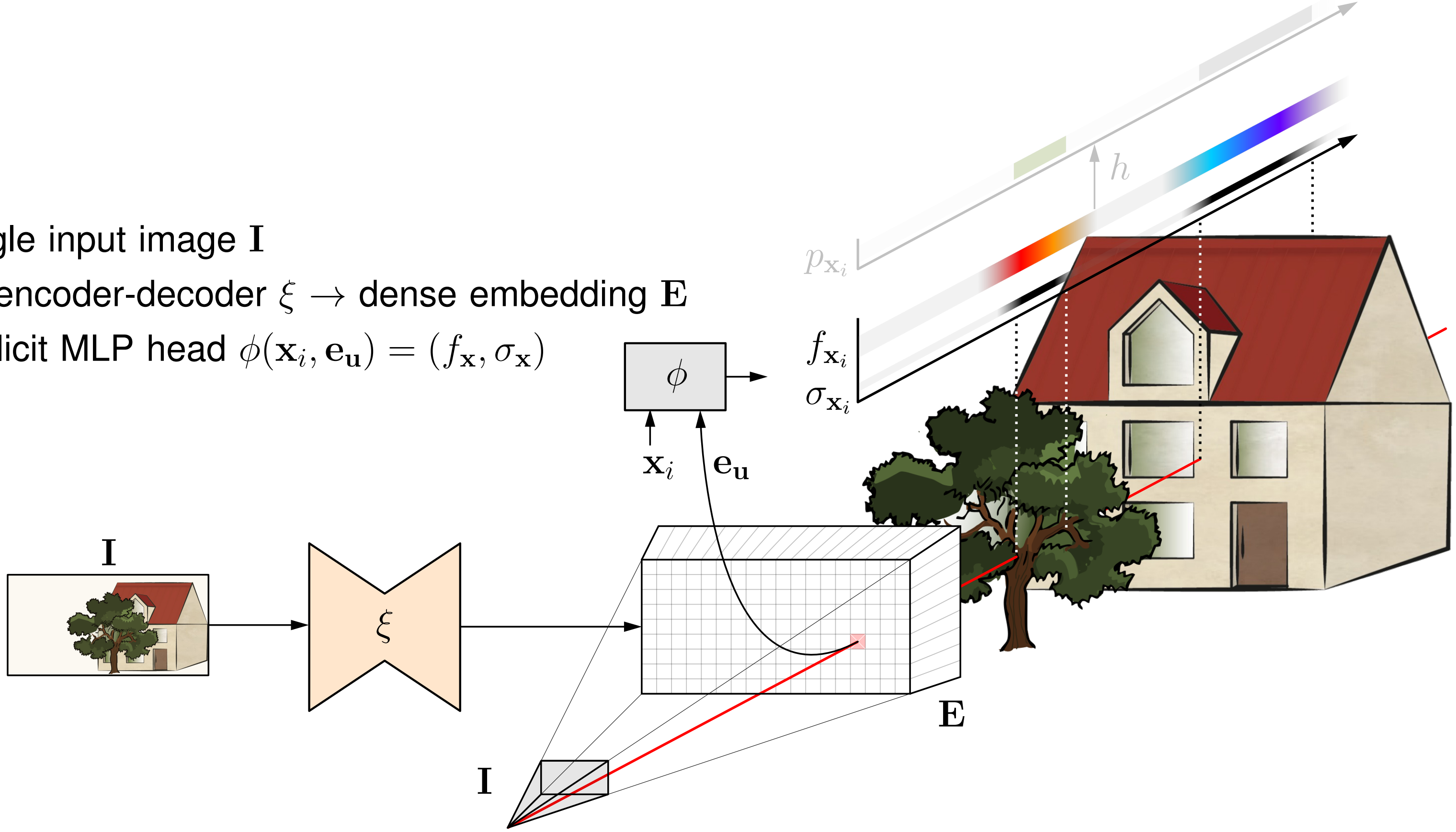
Model Architecture

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



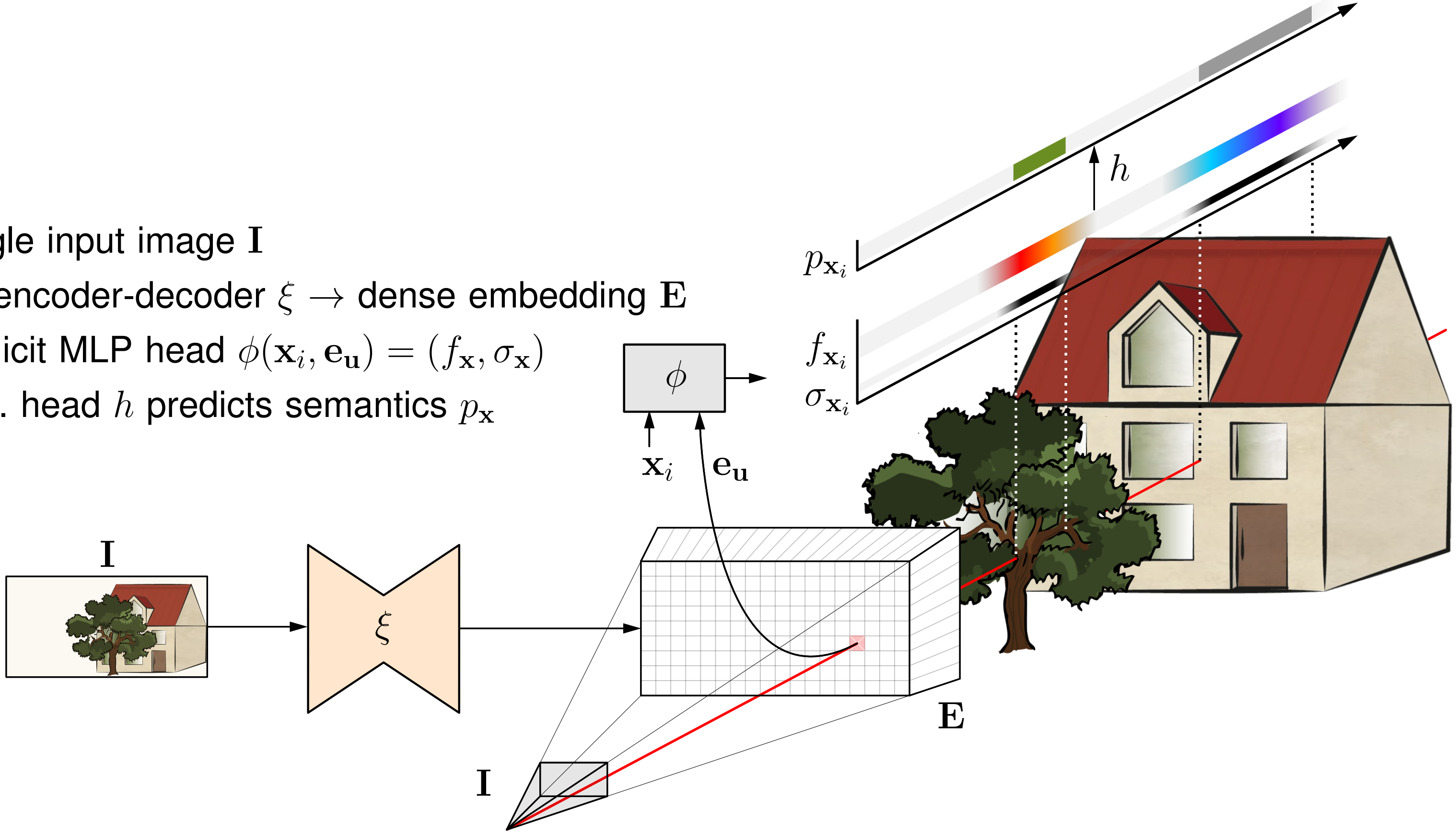
Model Architecture

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

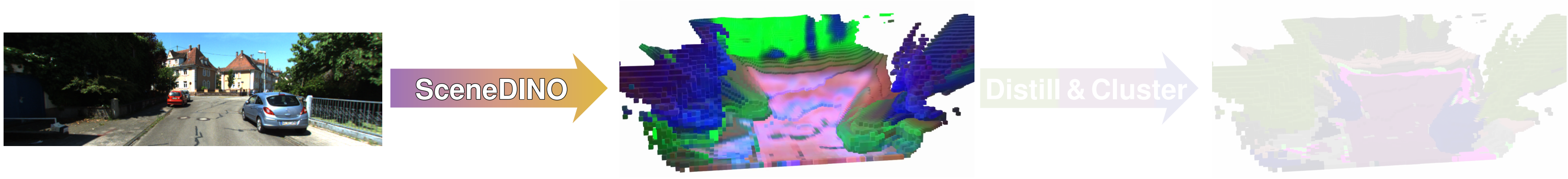


Model Architecture

- Single input image $\mathbf{I}$
- 2D encoder-decoder ξ → dense embedding $\mathbf{E}$
- Implicit MLP head $\phi(\mathbf{x}_i, \mathbf{e}_\mathbf{u}) = (f_\mathbf{x}, \sigma_\mathbf{x})$
- Seg. head h predicts semantics $p_\mathbf{x}$



SceneDINO Training



SceneDINO

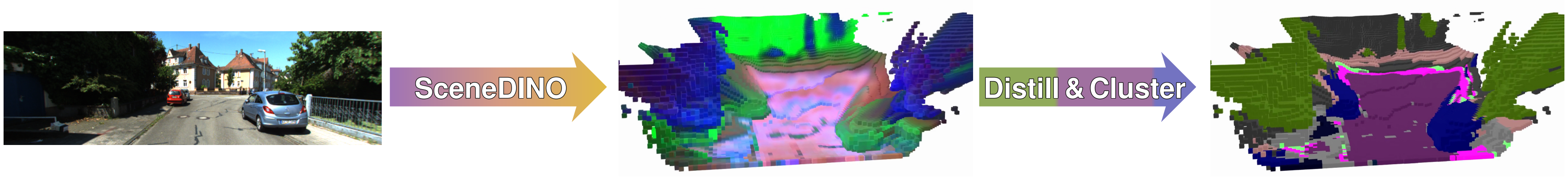
Distill & Cluster

Single Input Image

3D Feature Field

SSC Prediction

SceneDINO Training



SceneDINO

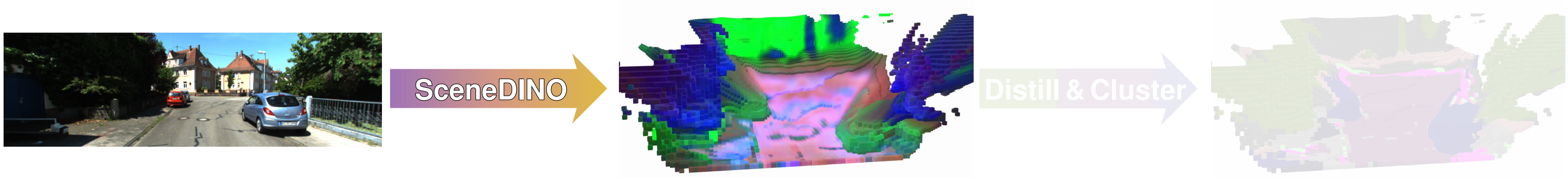
Distill & Cluster

Single Input Image

3D Feature Field

SSC Prediction

SceneDINO Training

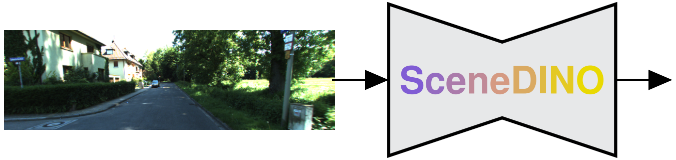


Single Input Image

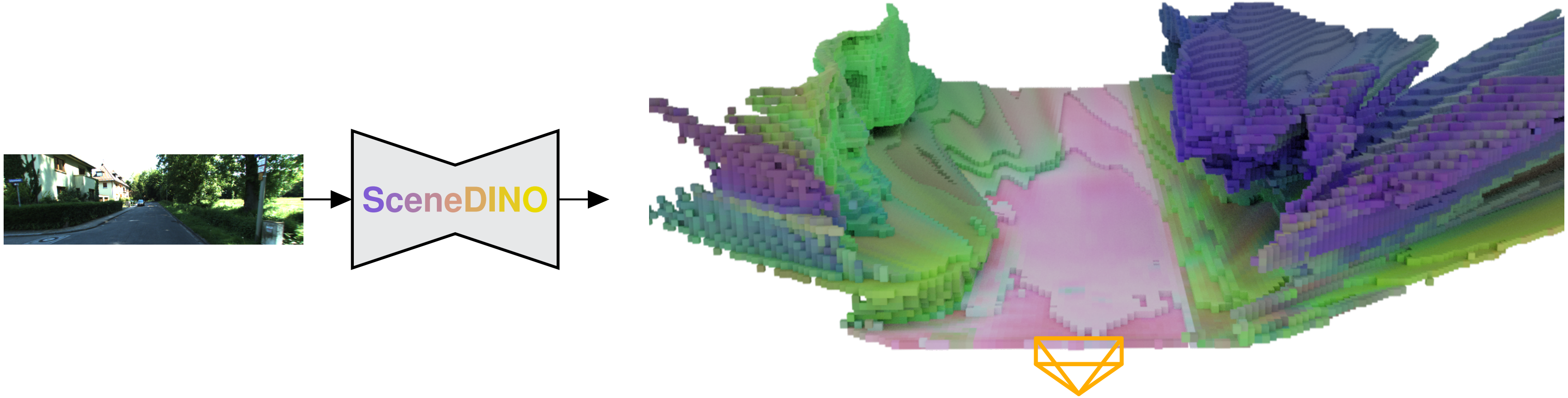
3D Feature Field

SSC Prediction

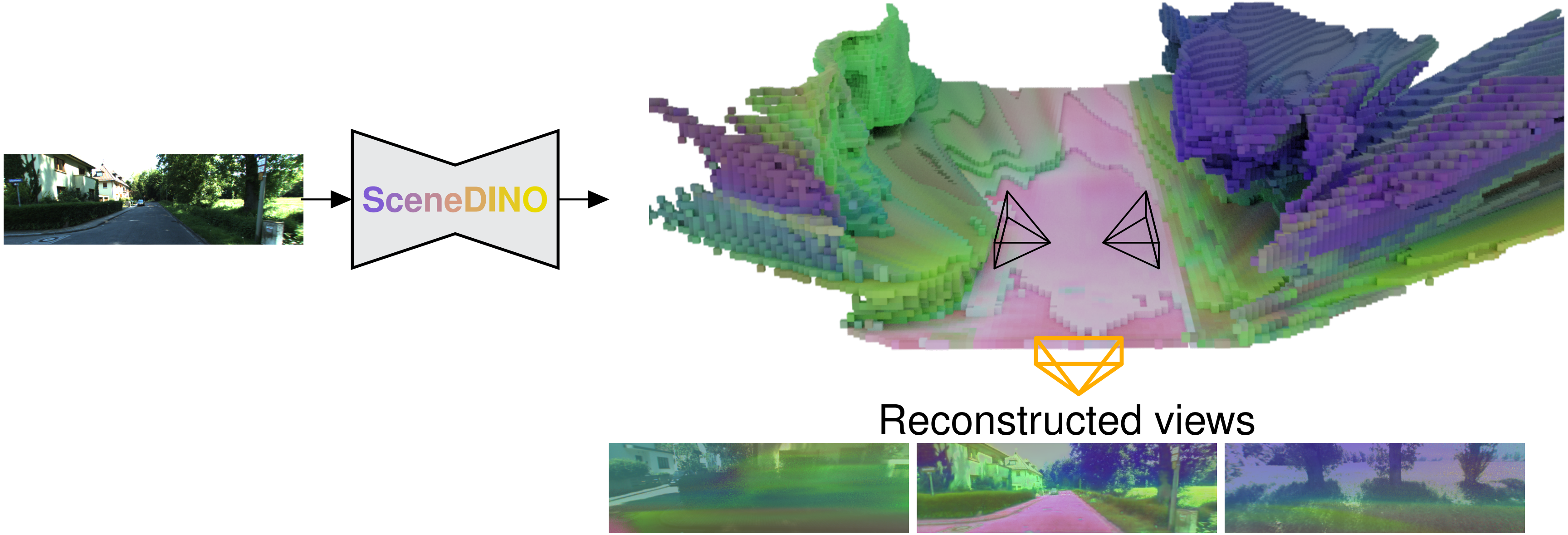
Multi-View Self-Supervision



Multi-View Self-Supervision

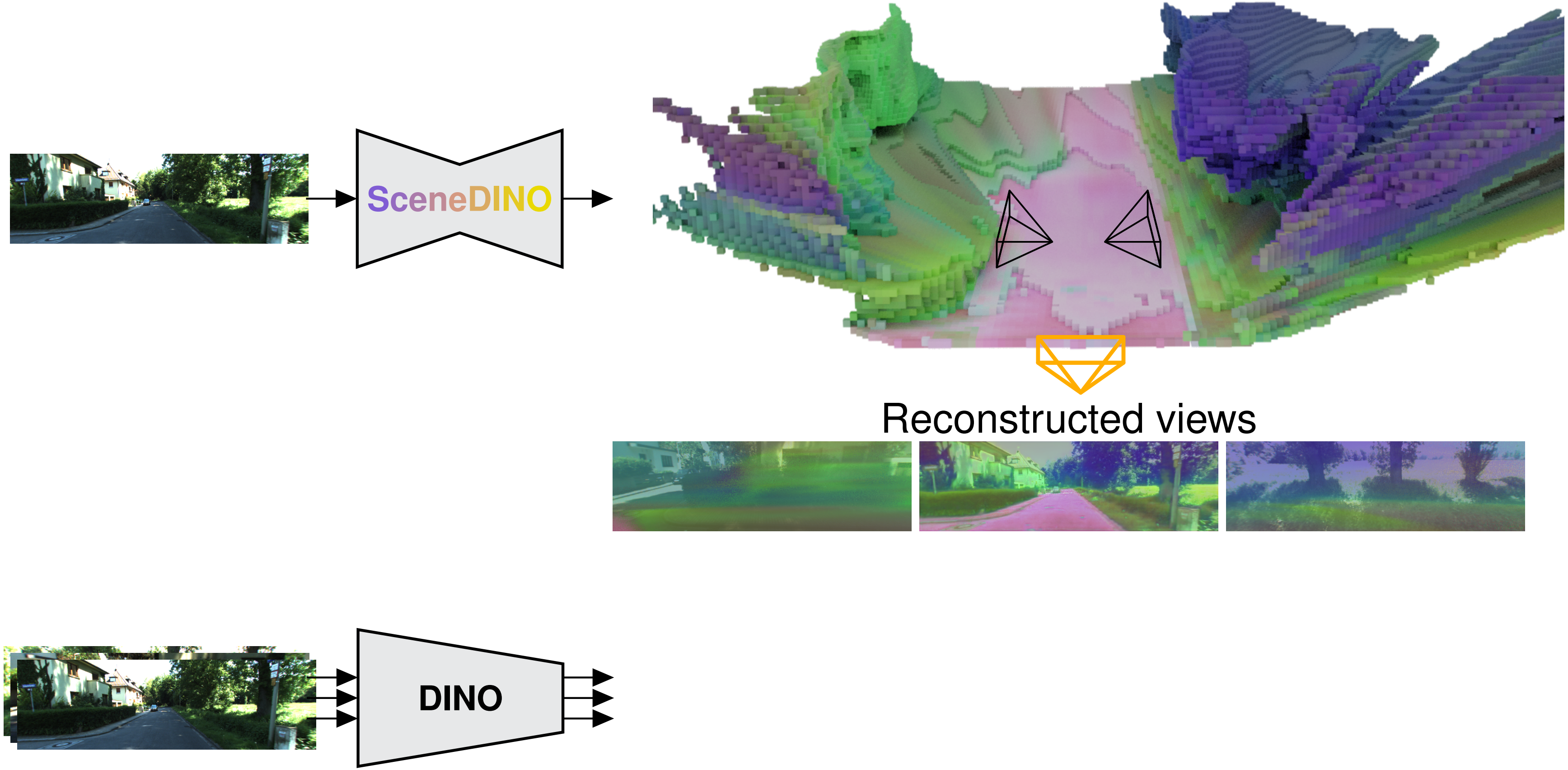


Multi-View Self-Supervision



Reconstructed views

Multi-View Self-Supervision

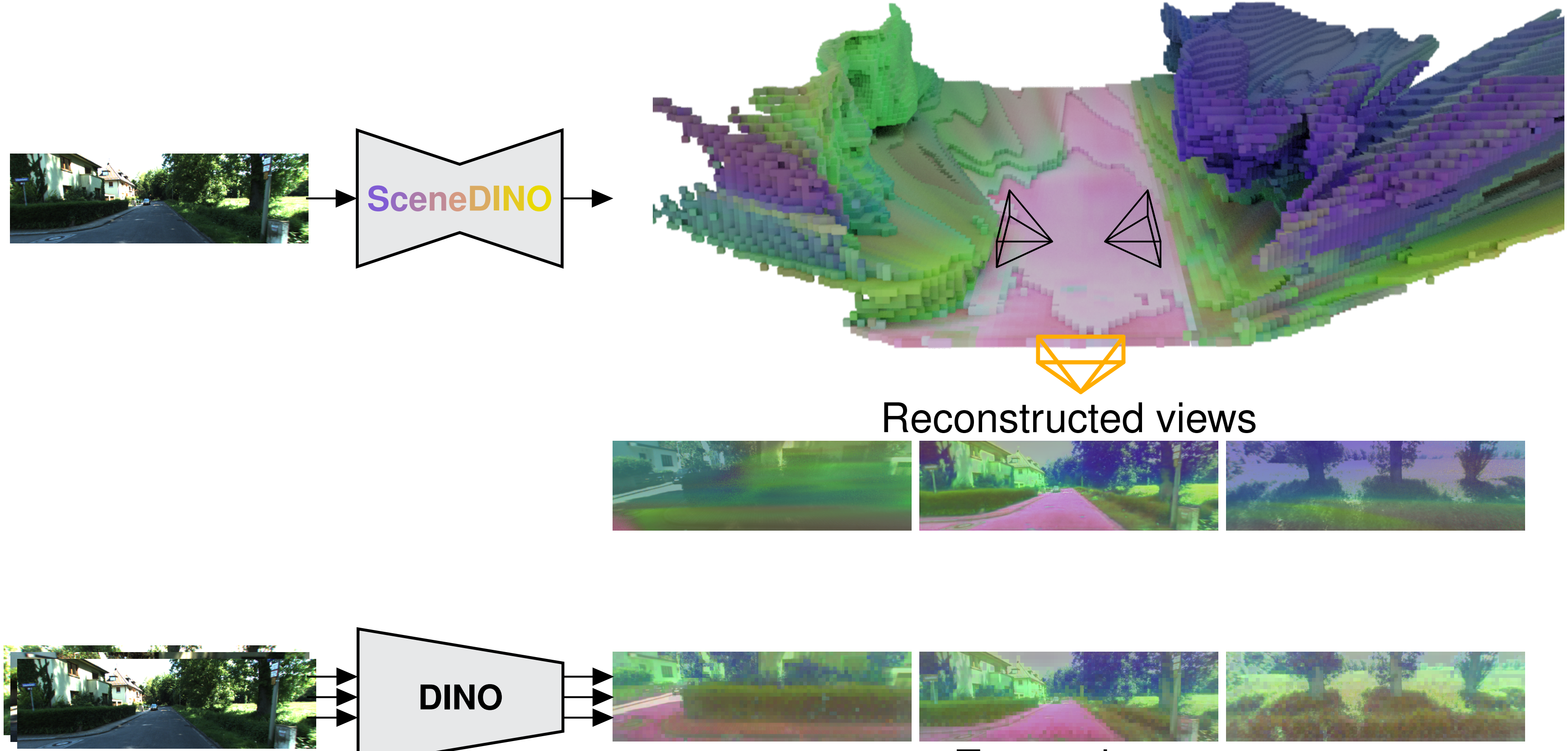


SceneDINO

Reconstructed views

DINO

Multi-View Self-Supervision



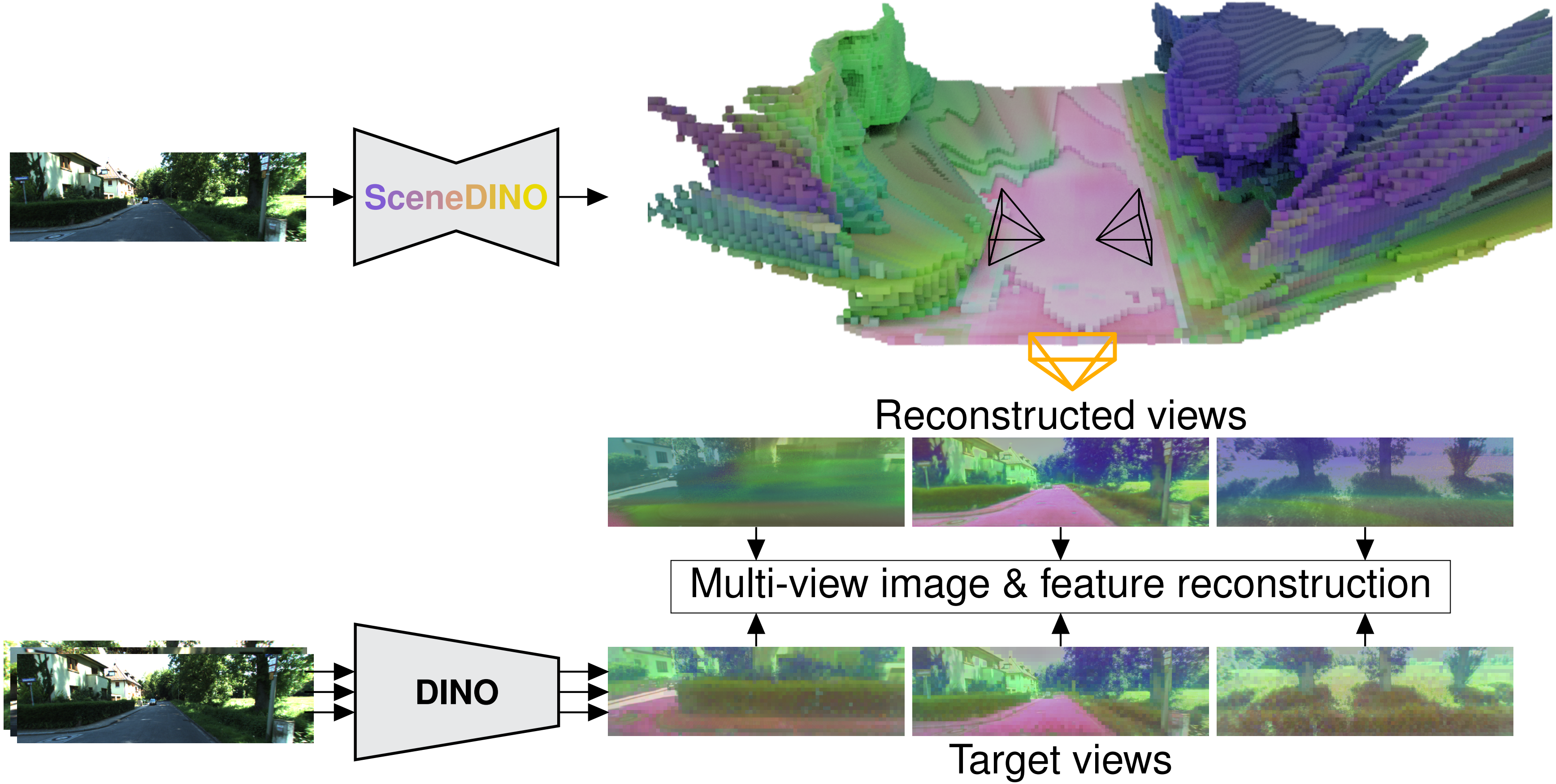
SceneDINO

DINO

Reconstructed views

Target views

Multi-View Self-Supervision



SceneDINO

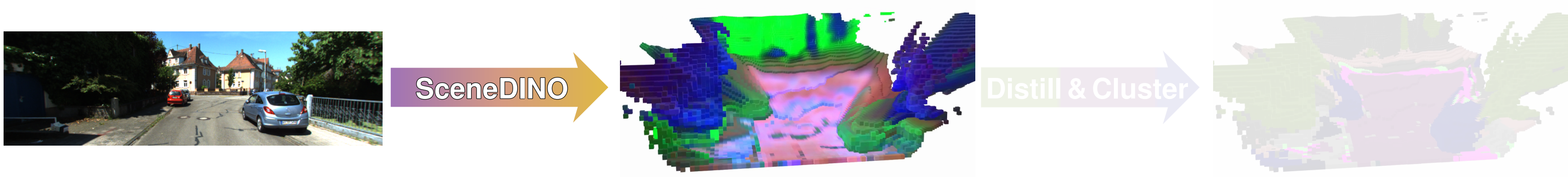
Reconstructed views

Multi-view image & feature reconstruction

DINO

Target views

SceneDINO Training



SceneDINO

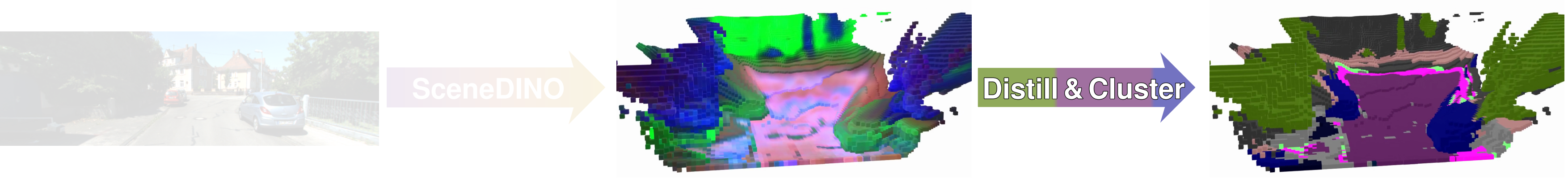
Distill & Cluster

Single Input Image

3D Feature Field

SSC Prediction

SceneDINO Training



SceneDINO

Distill & Cluster

Single Input Image

3D Feature Field

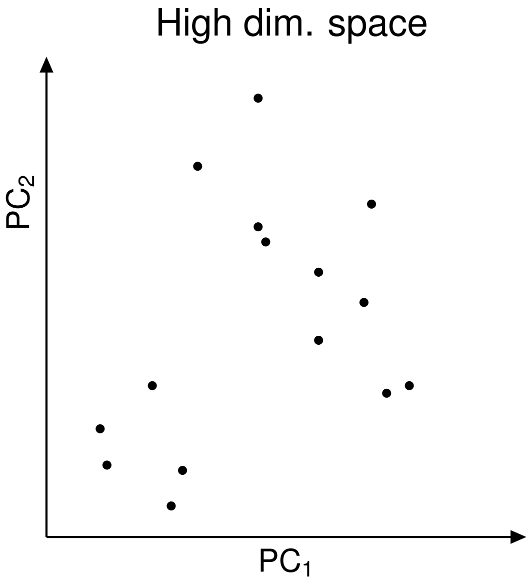
SSC Prediction

Unsupervised Segmentation

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

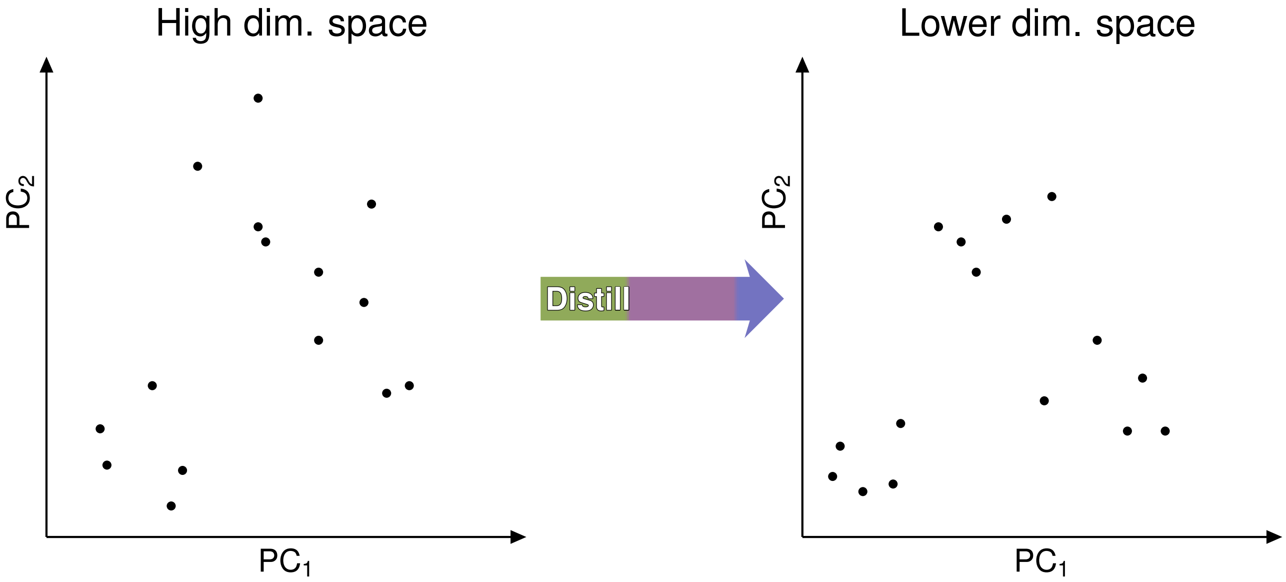
Unsupervised Segmentation

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



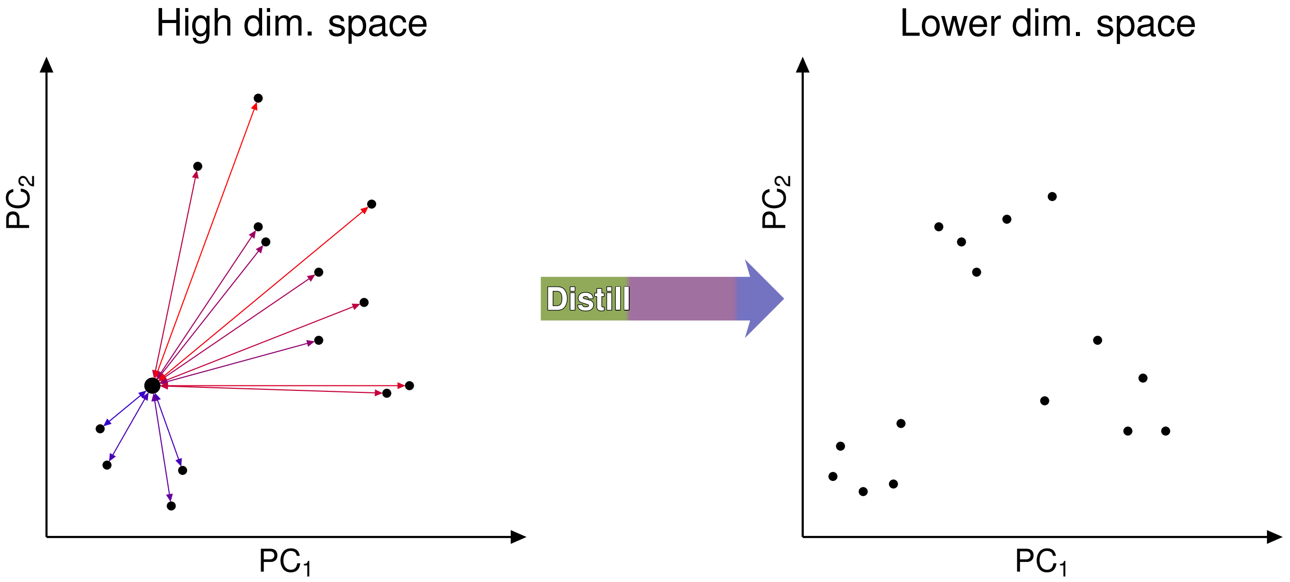
Unsupervised Segmentation

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



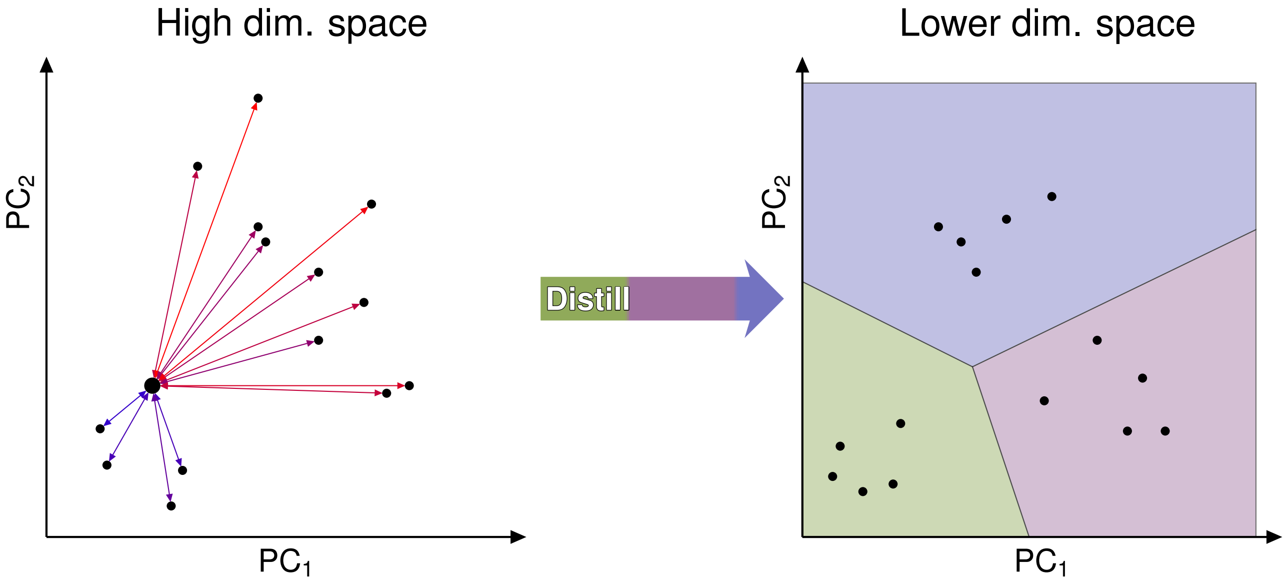
Unsupervised Segmentation

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



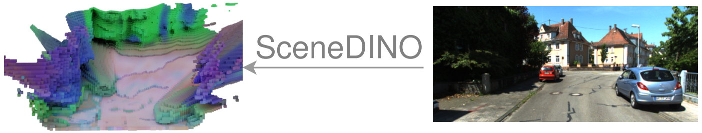
Unsupervised Segmentation

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



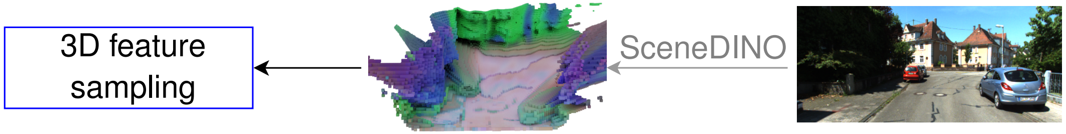
Unsupervised Segmentation

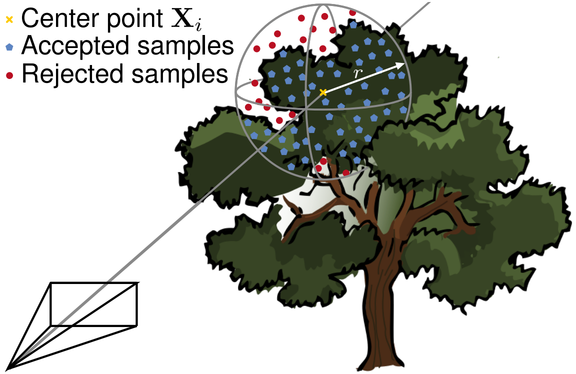
- **Goal:** Learn unsupervised segmentation head
- **Idea:** Magnify semantic correspondence & cluster features



Unsupervised Segmentation

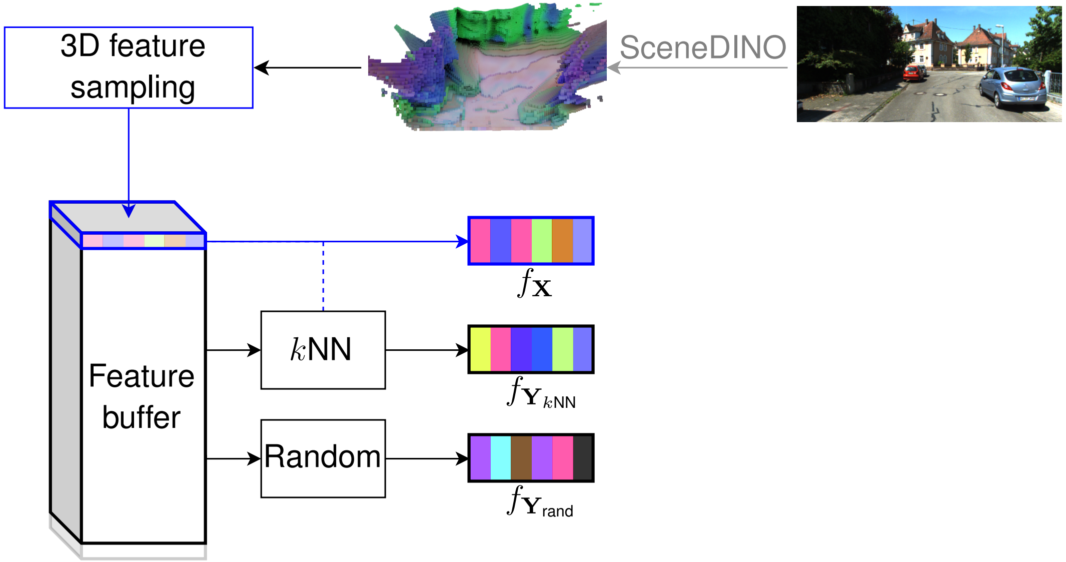
- **Goal:** Learn unsupervised segmentation head
- **Idea:** Magnify semantic correspondence & cluster features





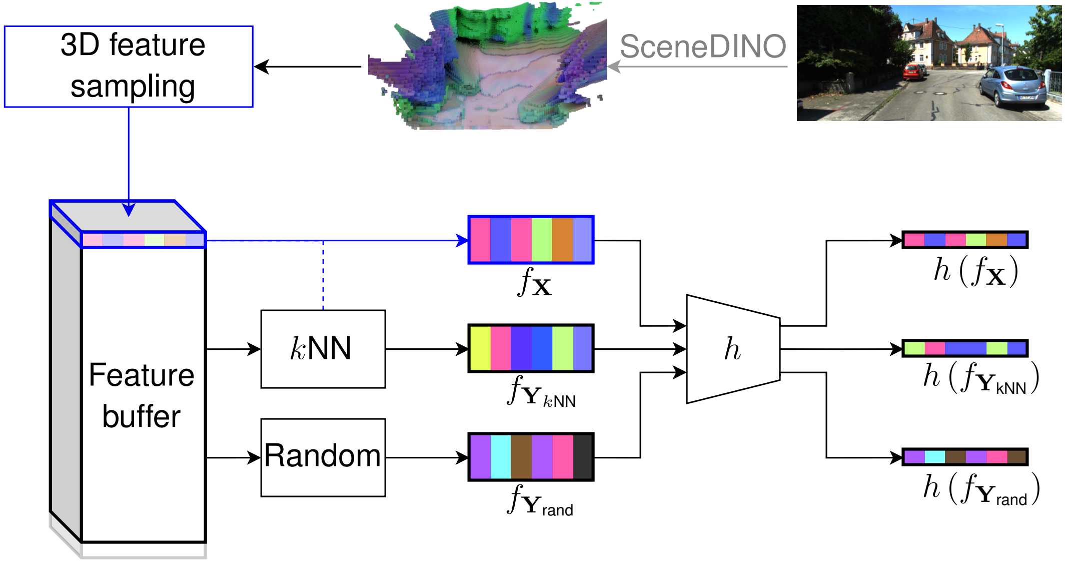
Unsupervised Segmentation

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



Unsupervised Segmentation

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

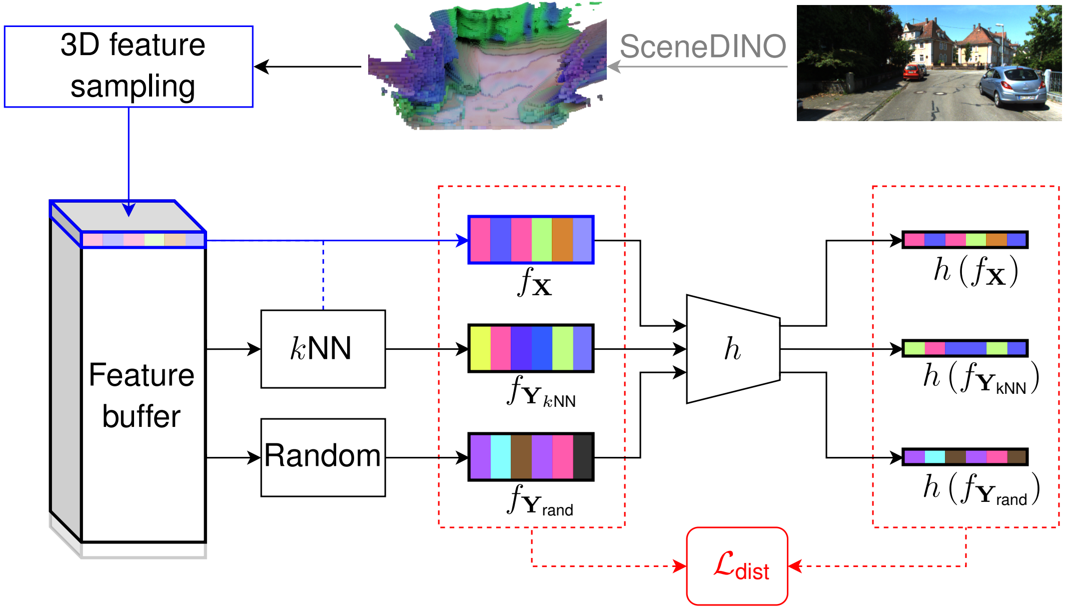


3D feature sampling
SceneDINO
Feature buffer
kNN
Random
$f_{\mathbf{X}}$
$f_{\mathbf{Y}_{k\text{NN}}}$
$f_{\mathbf{Y}_{\text{rand}}}$
h
$h\left(f_{\mathbf{X}}\right)$
$h\left(f_{\mathbf{Y}_{\text{kNN}}}\right)$
$h\left(f_{\mathbf{Y}_{\text{rand}}}\right)$

- Head projects features down

Unsupervised Segmentation

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

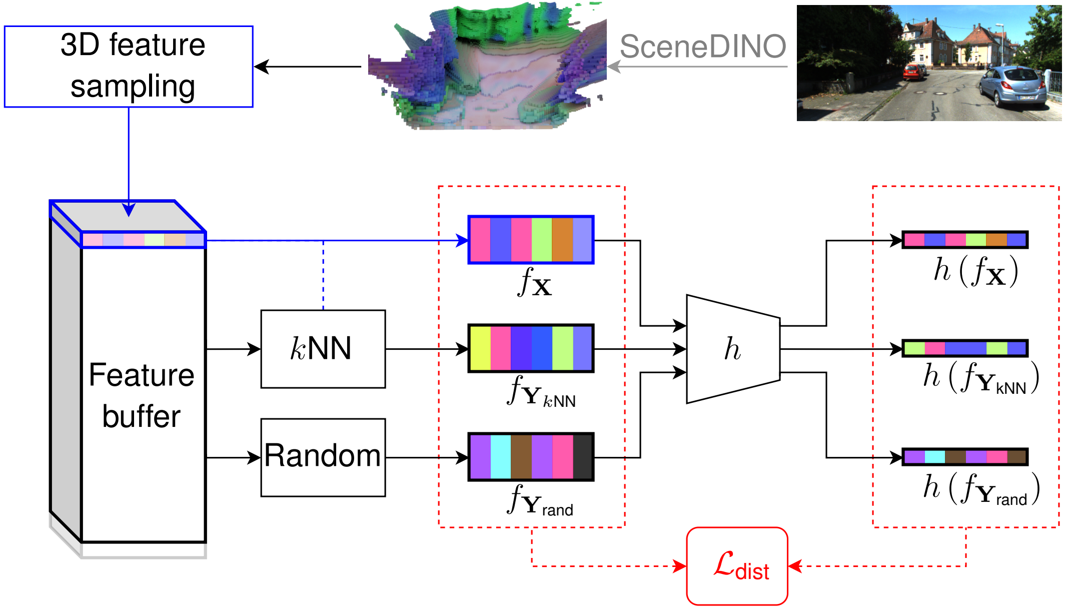

3D feature sampling
SceneDINO
Feature buffer
kNN
Random
$f_{\mathbf{X}}$
$f_{\mathbf{Y}_{k\text{NN}}}$
$f_{\mathbf{Y}_{\text{rand}}}$
h
$h\left(f_{\mathbf{X}}\right)$
$h\left(f_{\mathbf{Y}_{\text{kNN}}}\right)$
$h\left(f_{\mathbf{Y}_{\text{rand}}}\right)$
$\mathcal{L}_{\text{dist}}$

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

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

Unsupervised Segmentation

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



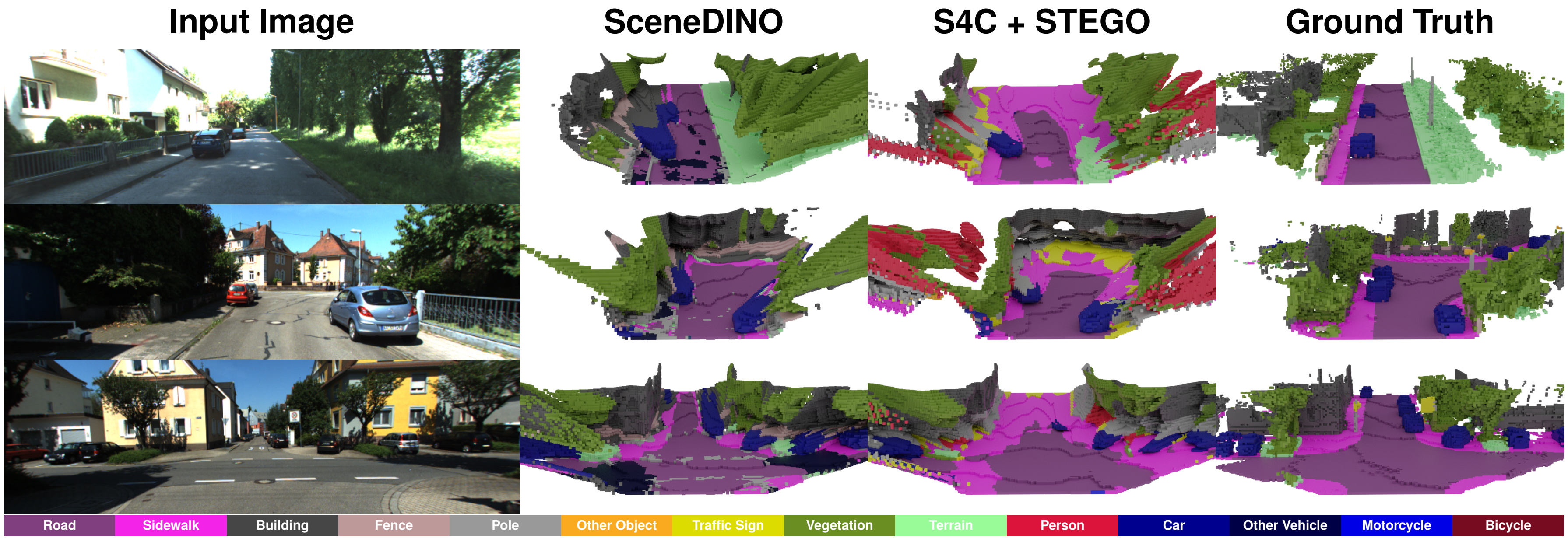
3D feature sampling
SceneDINO
Feature buffer
kNN
Random
$f_{\mathbf{X}}$
$f_{\mathbf{Y}_{k\text{NN}}}$
$f_{\mathbf{Y}_{\text{rand}}}$
h
$h\left(f_{\mathbf{X}}\right)$
$h\left(f_{\mathbf{Y}_{\text{kNN}}}\right)$
$h\left(f_{\mathbf{Y}_{\text{rand}}}\right)$
$\mathcal{L}_{\text{dist}}$

- 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: Unsupervised SSC



Input Image

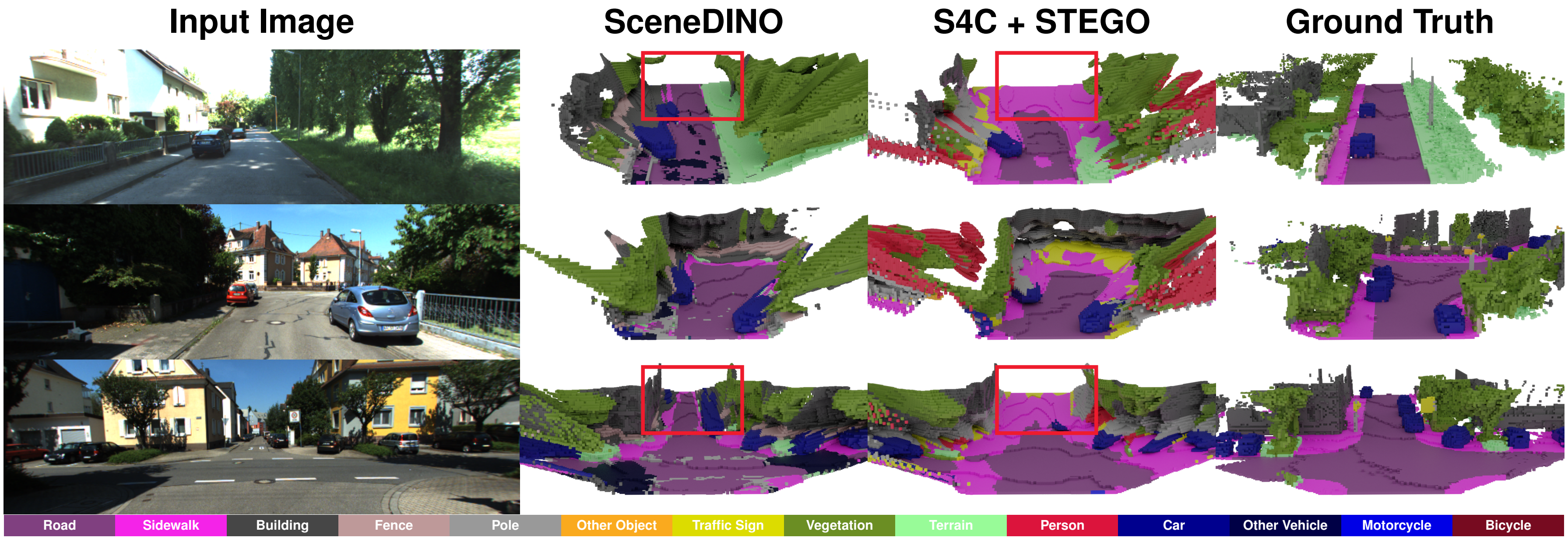
SceneDINO

S4C + STEGO

Ground Truth

Results: Unsupervised SSC

ICCV OCT 19-23, 2025 HONOLULU HAWAII



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

ICCV HONOLULU HAWAII OCT 19-23, 2025

- 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

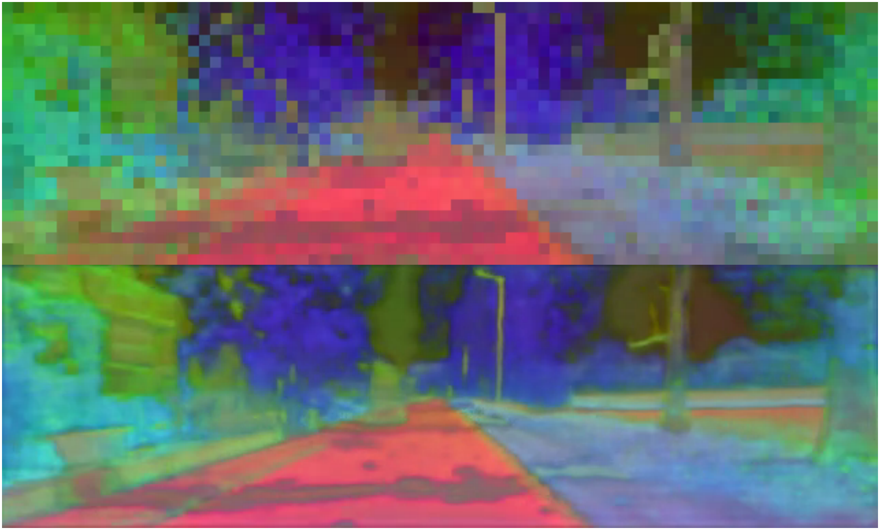
Linear probing outperforms 2D supervised S4C

[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: SceneDINO in 2D

Input Image

DINO [5]



SceneDINO

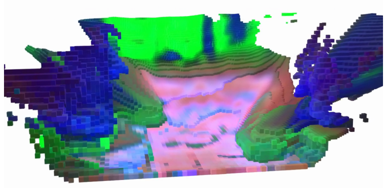
SceneDINO’s features are significantly more multi-view consistent

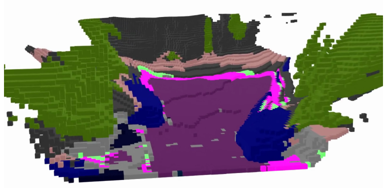
[5] M. Caron *et al.*, “Emerging properties in self-supervised vision transformers,” in *ICCV*, 2021.

Conclusion

We presented SceneDINO for unsupervised semantic scene completion

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





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