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]

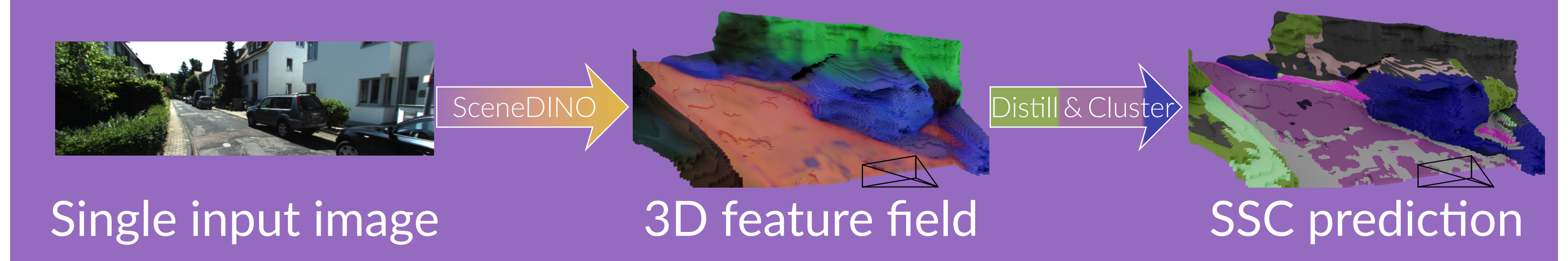
[1] TU Munich [2] TU Darmstadt [3] University of Oxford [4] MCML [5] ELIZA [6] hessian.AI [*] equal contribution

ICCV HONOLULU HAWAII OCT 19-23, 2025

Project Page

TL;DR

SceneDINO is unsupervised and infers 3D geometry and expressive features from a single image in a feed-forward manner, using multi-view self-supervision. Distilling and clustering features lead to unsupervised semantic scene completion predictions.



Introduction

Unsupervised Semantic Scene Completion aims to estimate the *dense 3D geometry* of a scene and partition the scene into *semantically meaningful regions* from a single image without any form of human supervision.

Motivation:

- Mitigate limitations of human-labeled 3D data (*e. g.*, high cost, inherent bias, *etc.*)
- Omit the need for costly and complex depth sensors (*e. g.*, LiDAR)
- Provide a foundation for approaching 3D scene understanding tasks using labels

Related work:

- Most existing approaches use significant geometric and semantic *supervision* [5]
- Some approaches only utilize 2D semantic supervision (*e. g.*, S4C [4])
- To the best of our knowledge, no existing fully *unsupervised* SSC approach
- No *feed-forward* approach for estimating general 3D features from a single image

Goal: Propose the *first fully unsupervised* semantic scene completion (SSC) approach.

References & Acknowledgments

[1] Mathilde Caron et al. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021.
[2] Stephanie Fu et al. FeatUp: A model-agnostic framework for features at any resolution. In *ICLR*, 2024.
[3] Mark Hamilton et al. Unsupervised semantic segmentation by distilling feature correspondences. In *ICLR*, 2022.
[4] Adrian Hayler et al. S4C: Self-supervised semantic scene completion with neural fields. In *3DV*, 2024.
[5] Yiming Li et al. SSCBench: A large-scale 3D semantic scene completion benchmark for autonomous driving. In *IROS*, 2024.
[6] Maxime Oquab et al. DINOv2: Learning robust visual features without supervision. *Trans. Mach. Learn. Res.*, 2024.
[7] Felix Wimbauer et al. Behind the scenes: Density fields for single view reconstruction. In *CVPR*, 2023.
[8] Yuanwen Yue et al. Improving 2D feature representations by 3D-aware fine-tuning. In *ECCV*, 2024.

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 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 (ELIZA). C. Rupprecht is supported by an Amazon Research Award. Finally, we acknowledge the support of ELLIS and thank Mateo de Mayo as well as Igor Cvišić for help with estimating camera poses.

Method

SceneDINO's unsupervised training comprises two stages:
(1) Learning a feed-forward *3D feature field* grounded in DINO [1] features
(2) Distilling and *clustering* the 3D feature field into unsupervised SSC predictions

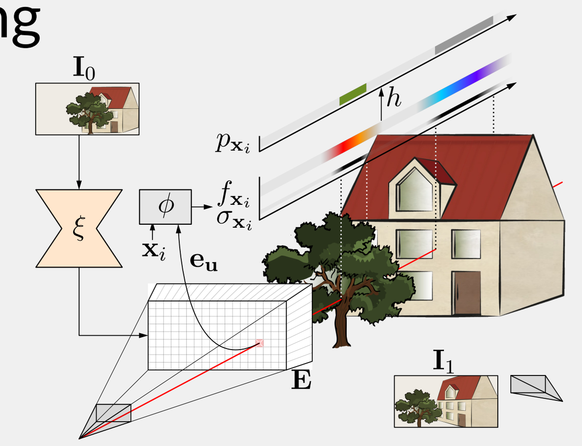
Stage 1: Feature-field training

SceneDINO 3D inference

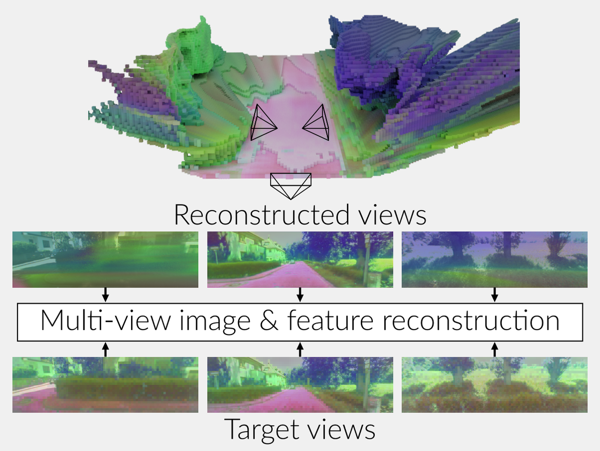
- 2D encoder-decoder ξ predicts dense embeddings $\mathbf{E}$ from image $\mathbf{I}_0$
- MLP ϕ estimates **density** $\sigma_{\mathbf{x}_i}$ and **features** $f_{\mathbf{x}_i}$ at 3D position $\mathbf{x}_i$ as

$$(\sigma_{\mathbf{x}_i}, f_{\mathbf{x}_i}) = \phi(\mathbf{e}_\mathbf{u}, \gamma(\mathbf{x}_i)),$$

with the interpolated embedding $\mathbf{e}_\mathbf{u}$ and the positional encoding γ



Multi-view self-supervision



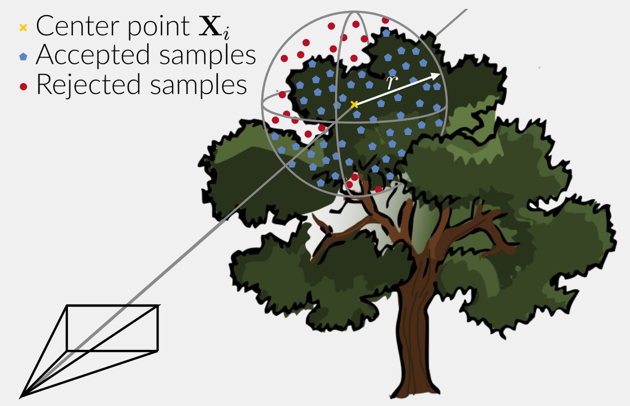
- SceneDINO is trained using multi-view images, (unsupervised) camera poses, and 2D DINO target features
- Single image fed into SceneDINO to predict a feature field
- Target views are reconstructed from the feature field using differentiable volume rendering and color sampling [7]
- Image/feature reconstruction and smoothness loss used
- Learned downsampler accounts for low target-feature resolution [2]

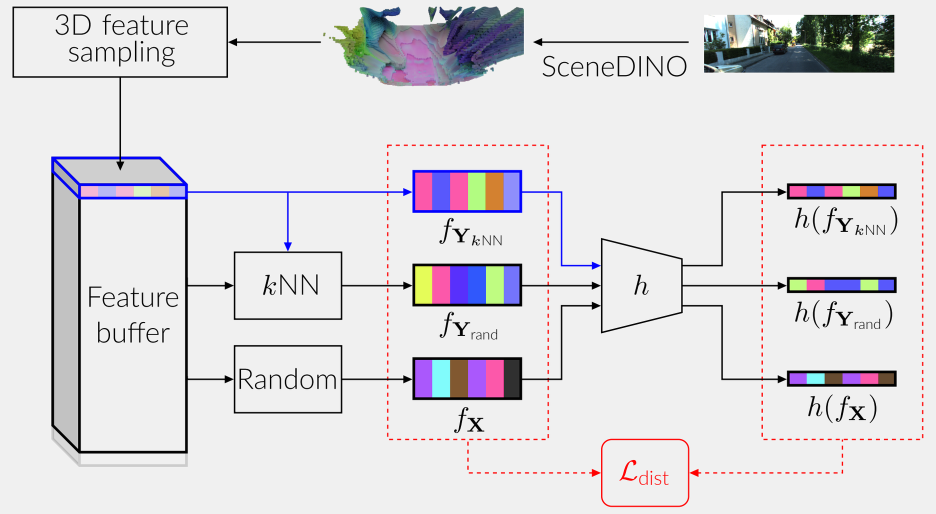
Stage 2: Unsupervised SSC

3D feature sampling

Intuition: Sample semantically rich 3D features and capture different semantic concepts.

- Sample center point $\mathbf{X}_i$ from all visible surface points
- Sample occupied points within the radius r around $\mathbf{X}_i$ to construct **feature batch** $f_\mathbf{X}$
- Repeat n-times: sample a new center point sufficiently far from existing center points





Feature distillation and clustering

- Given a feature batch $f_\mathbf{X}$, we sample random and kNN feature batches from the buffer
- **Contrastive loss** [3] amplifies feature dis/similarities and reduces feature space by
  - $f_\mathbf{X} \Leftrightarrow f_\mathbf{X}$ self-correlation
  - $f_\mathbf{X} \Leftrightarrow f_{\mathbf{Y}_{\text{rand}}}$ random correlation
  - $f_\mathbf{X} \Leftrightarrow f_{\mathbf{Y}_{\text{kNN}}}$ kNN correlation
- **Distilled features** $h(f_\mathbf{X})$ clustered using k-means

Results

Experiment setup: We train SceneDINO using on KITTI-360 (train). Next, we learn an unsupervised segmentation head by distilling and clustering SceneDINO's feature field. Hungarian matching is used to align pseudo semantics with the ground truth for validation.

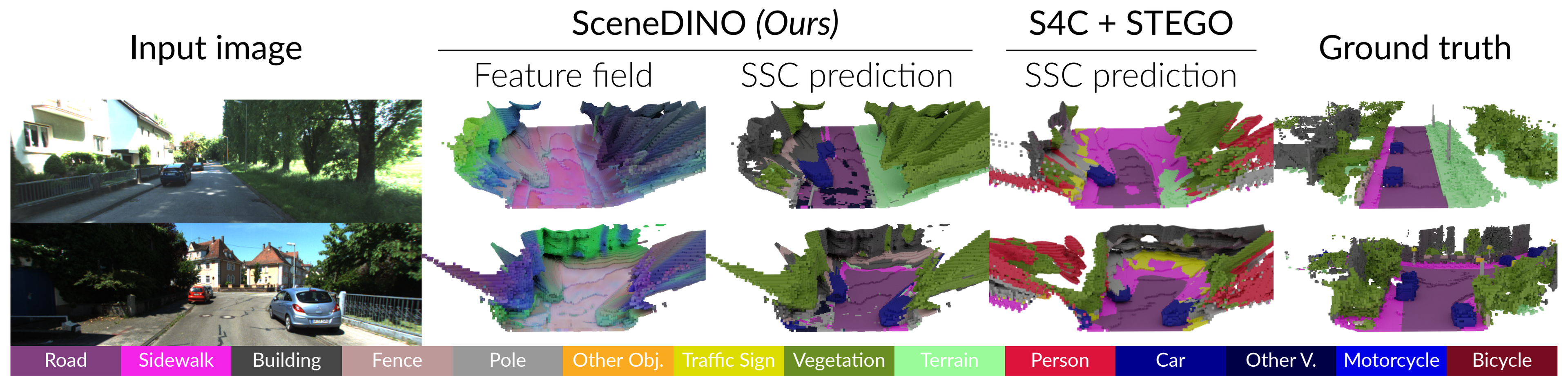


Table 1. **SSCBench-KITTI-360 results.** Semantic results using mIoU, and geometric results using IoU, Precision, and Recall (all in %, ↑) on SSCBench-KITTI-360 test.

Method	S4C [4] + STEGO [3]			SceneDINO (Ours)			S4C [4]		
Supervision	Unsupervised						2D supervision		
Range	12.8 m	25.6 m	51.2 m	12.8 m	25.6 m	51.2 m	12.8 m	25.6 m	51.2 m
Semantic validation									
mIoU	10.53	9.26	6.60	**10.76**	**10.01**	**8.00**	16.94	13.94	10.19
Geometric validation									
IoU	49.32	41.08	36.39	**49.54**	**42.27**	**37.60**	54.64	45.57	39.35
Precision	54.04	46.23	41.91	53.27	46.10	41.59	59.75	50.34	43.59
Recall	84.95	78.69	73.43	87.61	83.59	79.67	86.47	82.79	80.16

Table 2. **Linear probing** SceneDINO using different target features, mIoU (in %, ↑).

Probing approach	Target features	mIoU
Linear	DINO [1]	9.34
	DINOv2 [6]	**10.57**
S4C (full training)	n/a	10.19

Table 3. **Multi-view consistency results** on RE10K using L_1 (↓), L_2 (↓), and Cos-Sim (↑).

Method	L_1	L_2	Cos-Sim
DINOv2 [6]	14.20	0.66	0.75
FiT3D [8]	5.67	0.27	0.95
SceneDINO (w/ DINOv2)	**4.87**	**0.22**	**0.97**

Table 4. **SceneDINO analysis** on SSCBench-KITTI-360 test, using mIoU (in %, ↑) and 51.2 m range.

(a) **Training components ablation**

Δ mIoU	mIoU	Configuration
-1.18	6.82	No downsampler (bilinear up. + aug.)
-0.74	7.26	No pos. enc. decomposition
-0.12	7.88	w/ estimated ORB-SLAM3 poses
–	8.00	Full framework (SceneDINO)
+1.08	9.08	DINOv2 target features (*vs.* DINO)

(b) **Feature distillation analysis**

Δ mIoU	mIoU	Configuration
-1.61	6.39	No distillation
-1.35	6.65	No kNN-correlation loss ($\lambda_{\text{kNN}} = 0$)
-0.97	7.03	No neighborhood sampling
-0.47	7.53	5-crop sampling [3] (instead 3D sampling)
–	8.00	Full framework (SceneDINO)

Conclusion

- SceneDINO effectively estimates 3D geometry and lifts self-supervised DINO features using *multi-view self-supervision*
- Distilling and clustering SceneDINO's feature field in 3D leads to *state-of-the-art accuracy* in unsupervised semantic scene completion and 2D semantic segmentation
- SceneDINO offers *multi-view consistent features* and demonstrates *strong domain generalization*, *linear probing*, and *2D unsupervised semantic segmentation* results