

Feed-Forward SceneDINO for Unsupervised Semantic Scene Completion

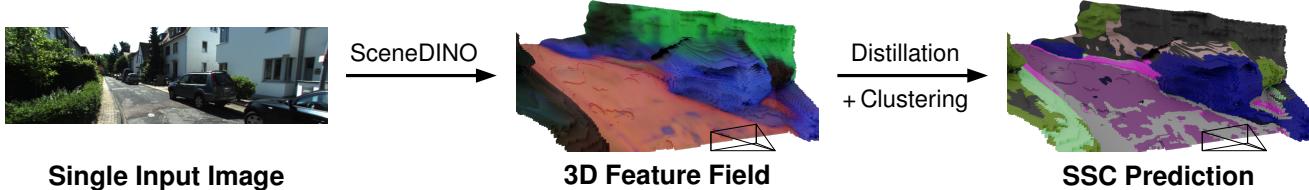
Aleksandar Jevtić^{* 1}Christoph Reich^{* 1,2,4,5}Felix Wimbauer^{1,4}Oliver Hahn²Christian Rupprecht³Stefan Roth^{2,5,6}Daniel Cremers^{1,4,5}¹TU Munich²TU Darmstadt³University of Oxford⁴MCML⁵ELIZA⁶hessian.AI^{*}equal contribution<https://visinf.github.io/scenedino>

Figure 1. **SceneDINO overview.** Given a single input image (*left*), SceneDINO predicts both 3D scene geometry and 3D features in the form of a feature field (*middle*) in a feed-forward manner, capturing the structure and semantics of the scene. Unsupervised distillation and clustering of SceneDINO’s feature space leads to unsupervised semantic scene completion predictions (*right*).

Abstract

Semantic scene completion (SSC) aims to infer both the 3D geometry and semantics of a scene from single images. In contrast to prior work on SSC that heavily relies on expensive ground-truth annotations, we approach SSC in an unsupervised setting. Our novel method, SceneDINO, adapts techniques from self-supervised representation learning and 2D unsupervised scene understanding to SSC. Our training exclusively utilizes multi-view consistency self-supervision without any form of semantic or geometric ground truth. Given a single input image, SceneDINO infers the 3D geometry and expressive 3D DINO features in a feed-forward manner. Through a novel 3D feature distillation approach, we obtain unsupervised 3D semantics. In both 3D and 2D unsupervised scene understanding, SceneDINO reaches state-of-the-art segmentation accuracy. Linear probing our 3D features matches the segmentation accuracy of a current supervised SSC approach. Additionally, we showcase the domain generalization and multi-view consistency of SceneDINO, taking the first steps towards a strong foundation for single image 3D scene understanding.

1. Introduction

Understanding the geometry and semantics of 3D scenes from image observations is a fundamental computer vision task with broad applications in robotics [26], autonomous driving [46, 65], medical image analysis [18, 112], and civil engineering [69]. The Semantic Scene Completion (SSC)

task unifies 3D geometry and semantic prediction from limited image observations [63, 88, 95]. Recent progress in SSC has been primarily driven by utilizing supervised learning [37, 87, 95]. However, acquiring large-scale 3D annotations is highly labor-intensive [165]. While significant resources have been invested in collecting human annotations for 2D tasks [52, 84], annotating similar amounts of data in 3D remains unapproached. This motivates approaching SSC without the need for manually annotated data.

Existing SSC approaches rely on ground-truth semantic annotations and frequently utilize additional supervision from LiDAR scans [37, 45, 73, 95]. In contrast, we are the first to approach SSC in a *fully unsupervised* setting, *i.e.* without task supervision or other supervised components. In particular, we aim to approach SSC from a *single image* without relying on any human annotations, only learning from unlabeled multi-view images using self-supervision. This setting is extremely challenging for two reasons: *first*, the human-defined nature of semantic taxonomies is ambiguous, and *second*, a single image only provides a partial observation of the scene with many invisible areas. We take inspiration from recent advances in self-supervised learning (SSL) of 2D representations and 3D reconstruction. 2D SSL representations, such as from DINO [11], have been shown effective for 2D unsupervised scene understanding [32, 103]. 3D reconstruction approaches successfully leveraged SSL from multi-view data to infer dense 3D geometry from a single image [33, 107].

In this paper, we present *SceneDINO*, to the best of our knowledge, the first approach for unsupervised semantic scene completion. Trained using 2D SSL features

from DINO [11] and multi-view self-supervision [107], SceneDINO predicts both 3D geometry and 3D features from a single image during inference in a feed-forward manner. Our general 3D feature representations enable us to approach unsupervised 3D scene understanding. Harnessing our expressive 3D features, we propose a novel 3D feature distillation approach for obtaining unsupervised semantic predictions in 3D. While we focus on the task of unsupervised SSC, SceneDINO’s features are general, offering a foundation for different 3D scene-understanding tasks by building on our 3D feature field.

Specifically, we make the following contributions: *(i)* We introduce SceneDINO, the first approach predicting dense 3D geometry *and* expressive 3D features in a *feed-forward manner* from a *single image*. *(ii)* We effectively distill SceneDINO’s feature field representation in 3D, obtaining unsupervised semantic predictions. *(iii)* We demonstrate the first fully unsupervised SSC results. We build a simple yet competitive unsupervised SSC baseline, lifting unsupervised 2D semantic predictions. Our SceneDINO approach outperforms this SSC baseline in unsupervised SSC as well as established 2D approaches in 2D semantic segmentation. *(iv)* Finally, we also showcase the domain generalization ability and multi-view consistency of SceneDINO.

2. Related Work

Single-image scene reconstruction. Estimating 3D geometry from image observations is a fundamental task in computer vision and has been studied for decades [36]. Traditional approaches, such as structure from motion [89], as well as recent neural radiance fields (NeRFs) [74], perform scene reconstruction using multiple images, as reviewed by multiple surveys [34, 108, 119]. Recently, estimating dense 3D geometry from a single image have been approached [8, 33, 80, 85, 96, 102, 107, 113]. Unlike monocular depth estimation [75], these approaches predict the depth for visible and occluded regions, reconstructing a complete scene. Behind the Scenes (BTS) [107] introduced an approach for unsupervised single-image *scene* reconstruction using multi-view self-supervision, which infers dense 3D geometry in a feed-forward manner. Our approach extends BTS by additionally lifting self-supervised features into 3D for unsupervised 3D scene understanding.

Semantic scene completion (SSC), also known as 3D semantic occupancy prediction, aims to jointly estimate the 3D geometry and semantics of a scene [62, 63, 95, 117]. Initial approaches used 3D semantic and geometric annotations and addressed indoor scenes [6, 13, 57–59, 67, 116], outdoor scenes with LiDAR [16, 61, 86, 87, 109], or both domains [8, 73]. Using birds-eye views has been proven effective for SSC [44, 64, 99]. To overcome the need for 3D annotations, approaches for using 2D annotations have

been proposed [37, 45, 81]. While SelfOcc [45] and RenderOcc [81] use multiple inference views, S4C [37] performs single-image SSC. In particular, S4C [37] employs a supervised 2D model and lifts 2D multi-view semantic predictions into 3D. In contrast to using 2D annotations, GaussTR [48] uses 2D foundation models for SSC and multiple views during inference. However, GaussTR relies on heavily supervised foundation models, including SAM [52] and Metric3Dv2 [42], and uses weak supervision from image/text pairs. To the best of our knowledge, there is no method for approaching SSC without the need for any ground-truth annotations. Our work presents the first unsupervised SSC approach, utilizing lifted SSL features and a single RGB input image for inference.

Self-supervised representation learning (SSL) aims to extract general features from data without annotations, facilitating various downstream tasks such as segmentation [24]. Recent SSL methods, often based on Vision Transformers (ViTs) [23], leverage clustering [2, 9, 10, 47, 60], masked modeling [20, 29, 39, 76, 106], contrastive learning [3, 12, 14, 38, 40, 41], or negative-free [4, 5, 11, 28, 79] pretext tasks [22, 78] for large-scale training. State-of-the-art models, *e.g.*, DINO [11], produce semantically rich, dense features, driving recent advances in 2D unsupervised scene understanding [32, 103]. We here aim to bring expressive features from DINO [11, 79] to 3D for SSC.

2D-to-3D feature lifting. The expressiveness of 2D visual representations has motivated lifting 2D features into 3D [93, 110]. Existing approaches utilize multi-view 2D features for 3D feature lifting [30, 43, 49, 53, 72, 82, 92, 93, 97, 100, 101, 105, 110, 115]. Lifting 2D features is effective in various tasks, including few-shot semantic occupancy prediction [110], and refining 2D representations [115]. However, existing feature-lifting approaches fit to a single scene [49, 53, 92, 93, 100, 101, 110, 115], require RGB-D inputs [30, 43, 72, 97, 105], or work on 3D point cloud inputs [82]. The only feed-forward approaches that use RGB inputs and lift 2D features, which we are aware of, are GaussTR [48]; MVSplat360 [15]. However, both approaches utilize multiple input images during inference, and MVSplat360 [15] only predicts low-dimensional feature representations, which are not suitable for unsupervised scene understanding. In contrast, we propose the first feed-forward approach for inferring lifted high-dimensional and rich 3D features using a single input image.

2D unsupervised semantic segmentation partitions images automatically into semantically meaningful regions without any form of human annotations. Early deep learning-based methods [17, 35, 47] approach the problem via representation learning. Leveraging SSL features from DINO as an inductive prior, STEGO [32] distills the feature representation into a lower-dimensional space for unsupervised probing. Building up on STEGO, subsequent methods

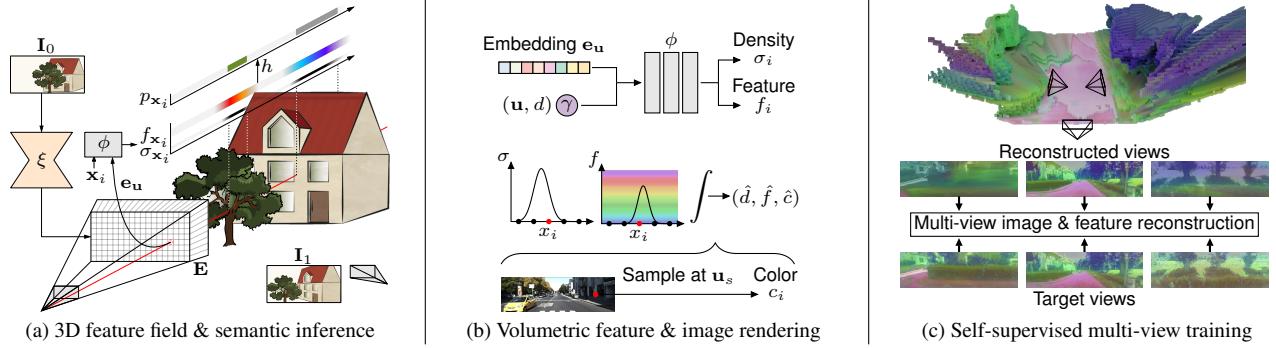


Figure 2. **SceneDINO architecture, rendering, and training.** (a) Inference: Given a single input image I_0 during inference, a 2D encoder-decoder ξ produces the embedding E from which the local embedding e_u is interpolated. The MLP encoder ϕ takes in e_u and 3D position x_i , and predicts both the density σ_{x_i} and the 3D feature f_{x_i} . Using a lightweight unsupervised segmentation head h , we can obtain semantic predictions p_{x_i} using f_{x_i} . (b) Rendering: Our feature field allows for volume rendering by shooting rays through it, yielding depth \hat{d} and \hat{f} in 2D. Color c_i is sampled from another view (e.g., I_1) using u_s and rendered to obtain the reconstructed color \hat{c} . (c) Multi-view training: We render 2D views (features & images) from our feature field and reconstruct the training views.

[31, 50, 91, 94] propose enhancements to the distillation. Our approach follows the idea of STEGO [32], extending it to 3D and integrating feature distillation using our 3D feature field to build the first unsupervised SSC approach.

3. Unsupervised Semantic Scene Completion

We approach semantic scene completion (SSC) without any form of manual supervision. To this end, we first describe SceneDINO, predicting *3D geometry* and expressive *3D features* from a *single image in a feed-forward manner* (Sec. 3.1), and SceneDINO’s multi-view training (Sec. 3.2). Next, we present our 3D feature distillation approach to obtain *unsupervised 3D semantic* predictions (Sec. 3.3). An overview of our full pipeline, including inference, rendering, and multi-view self-supervision, is provided in Fig. 2.

Notation. Let $I_0 \in [0, 1]^{3 \times H \times W}$ be a single RGB input image (for both training & inference) with corresponding pose $T_0 \in \mathbb{R}^{4 \times 4}$ and projection matrix $K_0 \in \mathbb{R}^{3 \times 4}$. For training, let (I_v, T_v, K_v) with $v \in \{1, 2, \dots, n\}$, be n additional views for multi-view self-supervision. Assuming a pinhole camera model, any 3D point $x \in \mathbb{R}^3$ in world coordinates can be projected onto the image plane of view v and the input view $v = 0$ with the perspective projection $\pi_v(x)$.

3.1. SceneDINO

Given a single input image I_0 , SceneDINO represents the dense geometric structure and features of a scene as a continuous mapping from world coordinates $x \in \mathbb{R}^3$ to a volumetric density $\sigma_x \in \mathbb{R}_+$ and a feature $f_x \in \mathbb{R}^D$. This continuous output representation is often called a *feature field*. While SceneDINO could represent any feature space, we aim for expressive SSL features from DINO [11, 79].

Architecture & feature field inference. Our SceneDINO architecture comprises two main parts: a 2D encoder-

decoder ξ and an MLP decoder (*cf.* Fig. 2a), following BTS [107]. ξ takes in I_0 and produces a per-pixel embedding $E \in \mathbb{R}^{D_E \times H \times W}$ with D_E dimensions. Intuitively, every spatial element of E represents a camera ray through a pixel, capturing both local geometry and features.

To infer the feature at a 3D position x , we employ a two-layer MLP decoder ϕ (*cf.* Fig. 2a). Given a position x within the camera frustum, we project x into the camera plane, obtaining the pixel location $u = \pi_0(x)$. We query E at the position u using bilinear interpolation, obtaining the local embedding e_u . Based on the embedding e_u , the pixel position u , and the distance $d_x \in \mathbb{R}_+$ of x to the camera, we obtain the density σ_x and feature prediction f_x as

$$(\sigma_x, f_x) = \phi(e_u, \gamma(u, d_x)), \quad (1)$$

where γ denotes a positional encoding [74].

Feature, depth & color volume rendering. SceneDINO predicts a continuous feature field from a single image. This representation can be used to render features and depth in 2D from an arbitrary viewpoint (*cf.* Fig. 2b), following the discretization strategy of Max *et al.* [71]. Given a viewpoint (T_r, K_r) , we sample L points x_i along the ray through pixel u_r , with distance δ_i between x_i and x_{i+1} . Based on the volumetric densities σ_{x_i} (*cf.* Eq. 1), we can compute the probabilities α_i of the ray ending between x_i and x_{i+1} , and accumulate these into V_i , the probability of x_i being visible:

$$V_i = \prod_{j=1}^{i-1} (1 - \alpha_j), \quad \text{with } \alpha_i = 1 - \exp(-\sigma_{x_i} \delta_i). \quad (2)$$

Using V_i and α_i , we render depth \tilde{d}_{u_r} and feature \tilde{f}_{u_r} from the estimated features f_{x_i} from Eq. (1) and distances d_{x_i} to x_i onto the image plane at position u_r as

$$\tilde{f}_{u_r} = \sum_{i=1}^L V_i \alpha_i f_{x_i} \quad \tilde{d}_{u_r} = \sum_{i=1}^L V_i \alpha_i d_{x_i}. \quad (3)$$

The differentiability of this rendering process enables us to self-supervise SceneDINO using multi-view images and their 2D feature representations (*e.g.*, from DINO [11]). SceneDINO predicts 3D geometry and features, but does not predict color as we focus on semantic downstream tasks. To obtain color for image reconstruction during training, we follow the color sampling approach of BTS [107].

3.2. 3D feature field training

We train SceneDINO using *multi-view self-supervision* (*cf.* Fig. 2c), aiming to obtain an expressive and view-consistent feature field without the need for any form of manual annotations. For self-supervision, we sample $n + 1$ views \mathbf{I}_v with camera parameters¹ T_v, K_v from the data and obtain dense 2D features \mathbf{F}_v from a self-supervised ViT (*e.g.*, DINO [11]). Note that the 2D features entail a resolution of $\mathbf{F}_v \in \mathbb{R}^{D \times \frac{H}{p} \times \frac{W}{p}}$, due to the ViT patch size p . The set of training views and features $\mathbb{V} = \{(\mathbf{I}_v, T_v, K_v, \mathbf{F}_v) \mid v = 0, \dots, n\}$ is randomly partitioned into two subsets $\mathbb{V}_{\text{source}}$ and $\mathbb{V}_{\text{target}}$. Training reconstructs the views $\mathbb{V}_{\text{target}}$ using the views of $\mathbb{V}_{\text{source}}$. In practice, we use a randomly sampled set of image patches that align with the ViT patches instead of the full image. In the following, we still refer to images for the sake of brevity.

Image reconstruction. We aim to learn the geometry of our feature field via multi-view photometric consistency. In particular, for every image $\mathbf{I}_t \in \mathbb{V}_{\text{target}}$ we derive a reconstructed image $\hat{\mathbf{I}}_{t,s}$ from every view s in $\mathbb{V}_{\text{source}}$ using differentiable rendering and color sampling. Equipped with both the reconstructed image $\hat{\mathbf{I}}_{t,s}$ and the target image \mathbf{I}_t , we compute our photometric loss per patch as

$$\mathcal{L}_p = \min_{\mathbf{I}_s \in \mathbb{V}_{\text{source}}} \left(\lambda_1 \mathcal{L}_1(\mathbf{I}_t, \hat{\mathbf{I}}_{t,s}) + \lambda_{\text{SSIM}} \mathcal{L}_{\text{SSIM}}(\mathbf{I}_t, \hat{\mathbf{I}}_{t,s}) \right). \quad (4)$$

We only consider the minimum per-patch loss across the different views in $\mathbb{V}_{\text{source}}$. The scalars λ_1 and λ_{SSIM} weight the absolute error \mathcal{L}_1 and the SSIM loss $\mathcal{L}_{\text{SSIM}}$ [104].

To regularize the 3D geometry prediction, we impose smoothness using an edge-aware smoothness loss [27]. Based on the estimated depth $\tilde{d}_{\mathbf{u}_t}$ (*cf.* Eq. 3), we obtain the inverse and mean-normalized depth $\tilde{d}_{\mathbf{u}_t}^*$. Using $\tilde{d}_{\mathbf{u}_t}^*$, we compute the edge-aware smoothness \mathcal{L}_s for pixel \mathbf{u}_t as

$$\mathcal{L}_s = |\nabla_x \tilde{d}_{\mathbf{u}_t}^*| e^{-|\nabla_x \mathbf{I}_t|} + |\nabla_y \tilde{d}_{\mathbf{u}_t}^*| e^{-|\nabla_y \mathbf{I}_t|}, \quad (5)$$

using the first spatial derivatives ∇_x and ∇_y at \mathbf{u}_t .

Feature reconstruction. We learn a multi-view consistent and expressive 3D feature field using the 2D features \mathbf{F}_t from $\mathbb{V}_{\text{target}}$. As we aim to learn a high-resolution (continuous) feature field, we render 2D features using Eq. 3 at the full image resolution $\mathbf{F}_t \in \mathbb{R}^{D \times H \times W}$. To compensate for

¹Note, camera poses can be obtained using unsupervised visual SLAM [7], strictly adhering to the fully unsupervised setting.

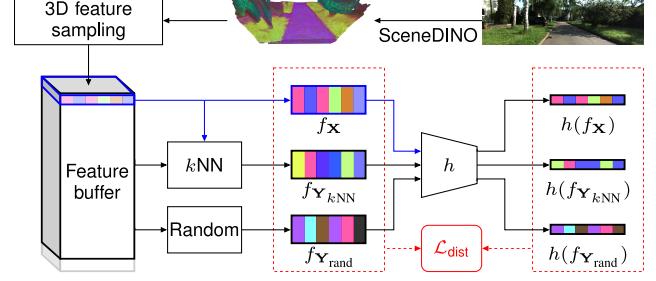


Figure 3. **3D feature distillation.** Given an input image, SceneDINO predicts a 3D feature field. 3D features f_X are sampled from the feature field. For f_X , we obtain $f_{Y_{kNN}}$ and $f_{Y_{rand}}$ from the feature buffer. The segmentation head h distills the features into a low-dimensional space and is trained using $\mathcal{L}_{\text{dist}}$.

the reduced spatial dimension of \mathbf{F}_t , we employ the down-sampler ψ proposed by Fu *et al.* [25] to our rendered features $\hat{\mathbf{F}}_t$. While current 2D SSL features capture semantics, they lack multi-view consistency, *i.e.*, due to positional encodings used in ViTs [111], leading to different features for identical visual content at two distinct positions in an image. As we aim for multi-view consistency, we compensate for this by learning a constant decomposition $\bar{\mathbf{F}} \in \mathbb{R}^{D \times \frac{H}{p} \times \frac{W}{p}}$ of features induced by positional encodings. Our feature loss is defined per feature as

$$\mathcal{L}_f = 1 - \text{cos-sim}(\mathbf{F}_t, \psi(\hat{\mathbf{F}}_t)) + \bar{\mathbf{F}}, \quad (6)$$

where cos-sim is the cosine similarity between two features.

As image edges correlate with semantic edges and to further impose consistency, we regularize the rendered features $\hat{\mathbf{F}}_t$ using an edge-aware smoothness loss per feature

$$\mathcal{L}_{fs} = |\nabla_x \hat{\mathbf{F}}_t| e^{-|\nabla_x \mathbf{I}_t|} + |\nabla_y \hat{\mathbf{F}}_t| e^{-|\nabla_y \mathbf{I}_t|}. \quad (7)$$

Our final loss for training SceneDINO is a weighted sum of the photometric loss, the feature loss, and both smoothness losses $\mathcal{L}_{\text{SceneDINO}} = \lambda_p \mathcal{L}_p + \lambda_s \mathcal{L}_s + \lambda_f \mathcal{L}_f + \lambda_{fs} \mathcal{L}_{fs}$, averaged over all pixels and features.

3.3. 3D feature distillation for unsupervised SSC

Given the expressive feature field representation, we aim to obtain unsupervised semantic predictions for SSC. While naïve k -means [68, 70] can yield meaningful pseudo semantics, distilling features into a lower-dimensional space has been shown to be more effective in 2D semantic segmentation [32, 54]. To this end, we present a novel 3D feature distillation approach (*cf.* Fig. 3). We train a point-wise segmentation head h , mapping $f_X \in \mathbb{R}^D$ to a lower-dimensional distilled representation $z_X \in \mathbb{R}^K$, with $K \ll D$. The resulting distilled space is clustered to obtain pseudo-semantics $p_X \in [0, 1]^C$, with C pseudo classes.

Existing work in 2D unsupervised semantic segmentation has shown that SSL feature correspondence captures

semantic class co-occurrence [32]. This correspondence between two batches of N sample points $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_N]$ and $\mathbf{Y} = [\mathbf{y}_1, \dots, \mathbf{y}_N]$ can be expressed by pairwise feature similarity $S_{i,j} = \text{cos-sim}(f_{\mathbf{x}_i}, f_{\mathbf{y}_j}) \in [-1, 1]$. Similarly, we can express the correspondence in the distilled feature space by $S_{i,j}^h = \text{cos-sim}(h(f_{\mathbf{x}_i}), h(f_{\mathbf{y}_j})) \in [-1, 1]$. We describe the sampling of the \mathbf{x}_i and \mathbf{y}_j below.

Feature distillation. We aim to distill features such that similar features align while dissimilar features are separated. To this end, we use the contrastive correlation loss $\mathcal{L}_{\text{corr}}$, introduced by STEGO [32] and defined as

$$\mathcal{L}_{\text{corr}}(f_{\mathbf{X}}, f_{\mathbf{Y}}, b) = - \sum_{i,j} (S_{i,j} - b) \max(S_{i,j}^h, 0), \quad (8)$$

where $f_{\mathbf{X}}$, $f_{\mathbf{Y}}$ are the features of the two sample batches. This loss pushes $S_{i,j}^h$ towards 1 in case $S_{i,j}$ exceeds the threshold b . Otherwise, $\mathcal{L}_{\text{corr}}$ pushes the $S_{i,j}^h$ below 0.

The correlation loss $\mathcal{L}_{\text{corr}}$ requires informative pairs of sampled features, balancing attractive and repulsive signals. Following STEGO [32], we consider three different relations: (1) feature pairs from the same image ($f_{\mathbf{X}}, f_{\mathbf{X}}$), (2) feature pairs from an image and its k -nearest neighbors in feature space ($f_{\mathbf{X}}, f_{\mathbf{Y}_{k\text{NN}}}$), and (3) feature pairs from an image and a randomly sampled other image ($f_{\mathbf{X}}, f_{\mathbf{Y}_{\text{rand}}}$). Note that each pair is obtained from SceneDINO’s 3D feature field, see below. Equipped with the three feature sample pairs, we compute the full distillation loss as

$$\begin{aligned} \mathcal{L}_{\text{dist}} = & \lambda_{\text{self}} \mathcal{L}_{\text{corr}}(f_{\mathbf{X}}, f_{\mathbf{X}}, b_{\text{self}}) \\ & + \lambda_{k\text{NN}} \mathcal{L}_{\text{corr}}(f_{\mathbf{X}}, f_{\mathbf{Y}_{k\text{NN}}}, b_{k\text{NN}}) \\ & + \lambda_{\text{rand}} \mathcal{L}_{\text{corr}}(f_{\mathbf{X}}, f_{\mathbf{Y}_{\text{rand}}}, b_{\text{rand}}), \end{aligned} \quad (9)$$

where λ_{self} , $\lambda_{k\text{NN}}$, and λ_{rand} denote the scalar loss weights. b_{self} , $b_{k\text{NN}}$, and b_{rand} are the contrastive thresholds.

Feature sampling in 3D. While obtaining feature pairs using 2D rendered features is straightforward [32], we aim to take advantage of our learned 3D geometry of the scene. To this end, we introduce a novel 3D feature sampling approach for the distillation loss $\mathcal{L}_{\text{dist}}$ from Eq. (9). Our goal is to sample features both similar and dissimilar in terms of the encoded semantic concept, which should capture *rich semantics* as well as *different semantic concepts*.

First, we obtain all G visible 3D surface points $\mathbf{V} \in \mathbb{R}^{3 \times G}$ and their depth $d_{\mathbf{V}} \in \mathbb{R}_+^G$ from the camera. To sample points that cover different semantic concepts, we use depth as a cue and sample different depth ranges. In particular, we sort the surface points \mathbf{V} based on $d_{\mathbf{V}}$. The sorted surface points $\hat{\mathbf{V}}$ are partitioned into M equally-sized chunks; we uniformly sample a single 3D point from each chunk, resulting in M center points $\mathbf{X} \in \mathbb{R}^{3 \times M}$.

Equipped with the center points \mathbf{X} , we aim to extract rich semantic features from the feature field. While we could just obtain the features for \mathbf{X} , we query positions in the

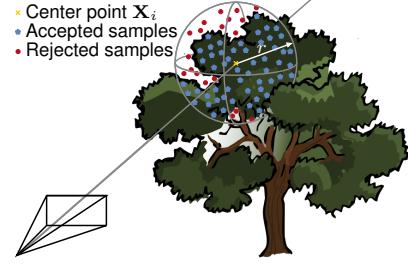


Figure 4. **3D feature sampling.** We first sample a center point \mathbf{X}_i from all visible surface points. Further points are sampled within the radius r around the center point \mathbf{X}_i . Sampled points with sufficient density are accepted; otherwise rejected. The accepted points are used to obtain the feature batch $f_{\mathbf{X}}$.

neighborhood of \mathbf{X} to increase semantic richness and better capture the 3D structure of the scene for distillation. In particular, for each center point, we randomly sample a point within a radius of $r = 0.5$ m. To account for samples falling into unoccupied regions in our feature field, we only keep samples with a sufficient density $\sigma > 0.5$. We repeat this sampling process until we obtain N valid samples per center point. Using these samples, we query our feature field, resulting in a feature batch $f_{\mathbf{X}} \in \mathbb{R}^{D \times N}$ for each of the G center points in each scene (cf. Fig. 4).

To obtain $f_{\mathbf{Y}_{k\text{NN}}}$ and $f_{\mathbf{Y}_{\text{rand}}}$, we utilize a feature buffer that efficiently stores the sampled features of multiple scenes. Given a new input image, we obtain G feature batches $f_{\mathbf{X}}$ as just described. For each $f_{\mathbf{X}}$, we randomly sample another feature batch from the buffer to obtain $f_{\mathbf{Y}_{\text{rand}}}$. To obtain $f_{\mathbf{Y}_{k\text{NN}}}$, we search in the feature buffer for the k -nearest neighbors of $f_{\mathbf{X}}$, using the average feature of each batch. From these k -nearest neighbors, we randomly pick a feature batch to obtain $f_{\mathbf{Y}_{k\text{NN}}}$ and compute the distillation loss $\mathcal{L}_{\text{dist}}$. After repeating this process for each of the current G feature batches, we add the current feature batches to the feature buffer and remove the oldest batches.

Unsupervised probing. To obtain semantic predictions, we probe the distilled feature space using k -means [68, 70]. In particular, we iteratively update cluster centers $\theta \in \mathbb{R}^{K \times C}$ using cosine distance-based mini-batch k -means [90] during distillation. To infer the final semantic prediction, we compute $p_{\mathbf{x}} = \text{softmax}(\text{cos-sim}(h(f_{\mathbf{X}}), \theta))$.

4. Experiments

We evaluate SceneDINO on SSC and compare it to a simple unsupervised SSC baseline (Sec. 4.1). We also report results for 2D unsupervised segmentation, including domain generalization results (Sec. 4.2). Finally, we explore multi-view feature consistency (Sec. 4.3) and present an analysis of SceneDINO and our 3D distillation (Sec. 4.4).

Datasets. We train using KITTI-360 [65], composed of clips from a moving vehicle equipped with cameras. For

consistency, we follow S4C [37] by sampling eight views and using the dataset camera poses. We also provide results with estimated poses. We also show experiments for training on RealEstate10k [118], composed of monocular videos. Here, we follow the setup of BTS [107], obtaining three views. If not noted differently, we report results obtained with training on KITTI-360. For SSC and 2D semantic segmentation validation, we use the SSCBench-KITTI-360 test split [63]. Cityscapes [19] and BDD100K [114] val are used for domain generalization results. To enable evaluation in 3D and 2D, we use the 19-class taxonomy of Cityscapes and perform 2D evaluation on Cityscapes, BDD100K, and KITTI-360 on 19 classes. For SSCBench, we combine classes to adhere to the 15 SSCBench classes.

3D evaluation. Given our unsupervised setup, we predict pseudo-semantics classes that must be aligned with the ground truth for evaluation. We follow standard practice in 2D unsupervised semantic segmentation [17, 31, 32, 50, 91, 94] by applying Hungarian matching [56] to align our pseudo semantics. For validating the aligned semantics, we follow the standardized SSCBench [63] protocol and report both semantic performance using the mean Intersection-over-Union (mIoU) and geometric performance using IoU, precision, and recall. We report all metrics on SSCBench ranges 12.8 m, 25.6 m, and 51.2 m.

2D evaluation. Following the established evaluation protocol in 2D unsupervised semantic segmentation [17, 31, 32, 50, 91, 94], we use the all-pixel accuracy (Acc) and mean Intersection-over-Union (mIoU) metrics. Likewise, in line with prior work, 2D segmentation predictions of all models are refined using a dense Conditional Random Field [55] before computing Acc and mIoU.

Multi-view feature consistency evaluation. We aim to evaluate the multi-view consistency of our feature field. As we are not aware of any general feed-forward 3D feature field approach, we compare against 2D SSL models. To measure multi-view consistency in 2D, we use two video frames and estimate optical flow and occlusions with RAFT [98]. We backward warp 2D features of the second frame to the first. On the aligned features, we compute the feature similarity using absolute error (L_1), the Euclidean distance (L_2), and the cosine similarity, ignoring occlusions.

Baselines. We are not aware of any existing unsupervised SSC approach. To allow for comparisons, we construct a competitive baseline for unsupervised SSC. In particular, we train the S4C approach with unsupervised semantics of the established STEGO [32] approach. For 2D semantic segmentation, we use U2Seg [77] and STEGO [32] as established unsupervised baselines. Note U2Seg is trained on ImageNet [21] and COCO [66] using STEGO pseudo-labels. We use STEGO [32] with DINO [11] (ViT-B/8), DINOv2 [79] (ViT-B/14), and FiT3D [115] (ViT-B/14)

Table 1. **SSCBench-KITTI-360 results.** Semantic results using mIoU and per class IoU, and geometric results using IoU, Precision, and Recall (all in %, ↑) on SSCBench-KITTI-360 test using three depth ranges. We compare our baseline S4C + STEGO to our SceneDINO. We report S4C as a 2D supervised baseline.

Method	S4C + STEGO			SceneDINO (Ours)			S4C		
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
car	18.57	14.09	9.22	21.24	15.94	11.21	22.58	18.64	11.49
bicycle	0.01	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00
motorcycle	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
truck	0.11	0.04	0.02	0.00	0.00	0.00	7.51	4.37	2.12
other-v.	0.01	0.05	0.02	0.00	0.00	0.00	0.00	0.01	0.06
person	0.01	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00
road	61.97	52.47	38.15	51.10	49.12	39.82	69.38	61.46	48.23
sidewalk	18.74	20.95	18.21	20.26	22.31	18.97	45.03	37.12	28.45
building	14.75	24.44	17.81	12.33	18.27	14.32	26.34	28.48	21.36
fence	1.41	0.20	0.11	1.91	0.90	0.58	9.70	6.37	3.64
vegetation	15.83	16.58	11.30	31.22	25.57	19.85	35.78	28.04	21.43
terrain	26.49	9.95	4.17	23.26	18.02	15.22	35.03	22.88	15.08
pole	0.08	0.04	0.04	0.05	0.05	0.05	1.23	0.94	0.65
traffic-sign	0.00	0.00	0.00	0.00	0.00	0.00	1.57	0.83	0.36
other-obj.	0.05	0.04	0.02	0.00	0.00	0.00	0.00	0.00	0.00
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

features. FiT3D offers multi-view refined DINOV2 features [115]. Note that FiT3D reports results, concatenating the refined features with DINOV2 features. We report results using both plain features only and the concatenation. We also use rendered 2D segmentations of our S4C + STEGO baseline for 2D validation. For multi-view feature consistency, we utilize DINO [11], DINOV2, and FiT3D [115] features as a baseline.

Implementation details. Our encoder-decoder uses a DINO-B/8 [11] backbone and a dense prediction decoder [83]. The MLP decoder ϕ entails two layers with 128 hidden features. As rendering features is expensive, ϕ predicts 64 features. We employ another MLP to up-project again to the full dimensionality $D = 768$. If not stated differently, our target features are obtained from DINO-B/8 [11]. We train using a batch size of 4 and extract 32 patches of size 8×8 from each image to compute $\mathcal{L}_{\text{SceneDINO}}$. Volume rendering samples each ray at $L = 32$ uniformly spaced points in inverse depth within [3 m, 80 m]. We train for 100 k steps using Adam [51] with a base learning rate of 10^{-4} . Training takes ca. 2 days on a single V100 GPU. We distill using a batch size of 4, 5 center points, a feature batch of size 576, and cluster with $K = 19$. For k NN sampling, we use $k = 4$. The feature buffer holds 256 feature batches. Refer to the supplement for more details.

4.1. 3D semantic scene completion

We assess the unsupervised SSC and geometric accuracy of SceneDINO with our 3D feature distillation approach on SSCBench-KITTI-360. In particular, Tab. 1 com-

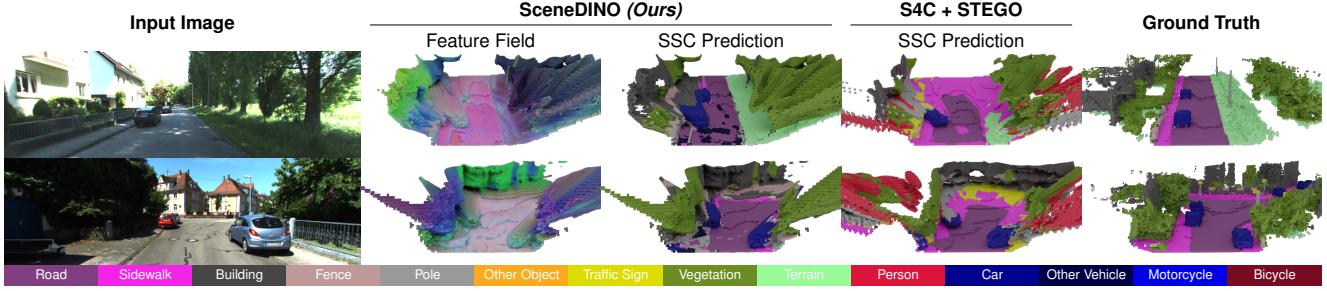


Figure 5. **Qualitative SSC comparison on KITTI-360.** We show the input image, SceneDINO’s feature field using the first three principal components and SSC prediction, the prediction of our baseline S4C + STEGO, and the ground truth. We only visualize surface voxels. Qualitative results show the expressiveness of our feature field and SceneDINO’s capabilities to accurately reconstruct and label a scene.

Table 2. **2D unsupervised semantic segmentation results on KITTI-360.** Comparing SceneDINO to existing 2D methods and our S4C + STEGO 3D baseline, using Accuracy and mean IoU (in %, ↑) on the SSCBench-KITTI-360 test split. † denotes the use of plain FiT3D features. ‡ denotes training on ImageNet and COCO.

Method	Features	Acc	mIoU
U2Seg [‡] [77]	—	72.89	23.43
STEGO [32]	DINO [11]	73.32	23.57
STEGO [32]	DINOv2 [79]	64.54	24.82
STEGO [32]	FiT3D [115]	54.19	22.29
STEGO [32]	FiT3D [†] [115]	57.25	18.95
S4C [37] + STEGO [32]	DINO [11]	65.16	21.67
SceneDINO (Ours)	DINO [11]	77.74	25.81

pares SceneDINO against our unsupervised SSC baseline S4C [37] + STEGO [32]. SceneDINO achieves a (semantic) mIoU of 8.0 % for the range of 51.2 m, significantly improving over our unsupervised baseline (6.6 %). This demonstrates that SceneDINO effectively lifts DINO features into 3D. In terms of geometric accuracy, SceneDINO moderately improves over S4C + STEGO. Despite being *fully unsupervised*, SceneDINO comes within 2.2 % points mIoU of the 2D-supervised S4C.

Fig. 5 provides qualitative samples on SSCBench-KITTI-360. SceneDINO’s unsupervised SSC predictions are less noisy and capture finely resolved semantics compared to S4C + STEGO. Compared to the ground truth, we observe, SceneDINO captures both the geometry and general semantics of the scene. We visualize SceneDINO’s feature field (before distillation) using the first three principal components. In PCA space, we observe that our feature field captures semantically meaningful regions.

4.2. 2D semantic segmentation

Table 2 compares the semantic predictions of SceneDINO to recent 2D approaches and our 3D baseline. We obtain 2D semantic segmentations from SceneDINO and our S4C + STEGO baseline using semantic rendering [37]. SceneDINO with our 3D distillation approach outperforms STEGO with DINO features, an established 2D unsupervised

Table 3. **2D unsupervised semantic segmentation domain generalization results.** Comparing SceneDINO to existing 2D unsupervised semantic segmentation methods and S4C + STEGO 3D baseline, using Accuracy and mean IoU (in %, ↑). We train on KITTI-360 images and report domain generalization results on Cityscapes and BDD-100K val. † denotes plain FiT3D features.

Method	Features	Cityscapes		BDD-100K	
		Acc	mIoU	Acc	mIoU
U2Seg [77]	—	75.57	18.62	69.00	17.99
STEGO [32]	DINO [11]	71.21	19.42	75.02	21.41
STEGO [32]	DINOv2 [79]	68.41	19.73	65.72	21.77
STEGO [32]	FiT3D [115]	66.94	21.01	65.96	20.99
STEGO [32]	FiT3D [†] [115]	64.76	17.17	60.83	19.09
S4C [37] + STEGO [32]	DINO [11]	54.80	14.04	44.98	11.62
SceneDINO (Ours)	DINO [11]	73.17	22.81	72.28	22.09

vised semantic segmentation approach. In particular, the mIoU of SceneDINO is 2.24 % points higher than for STEGO (w/ DINO). Utilizing 3D refined features from FiT3D deteriorates the baseline relative to DINO, indicating that the FiT3D refinement reduces feature expressiveness. Notably, our unsupervised 3D baseline S4C + STEGO transfers significantly worse to 2D than SceneDINO.

We also validate SceneDINO, trained on KITTI-360, on Cityscapes and BDD10K, demonstrating domain generalization. The results are reported in Tab. 3. SceneDINO outperforms all baselines in mIoU on both datasets while only falling short in Acc. S4C + STEGO falls short in generalization. We suspect this poor generalization is caused by the fact that S4C does not rely on general SSL features in the final model, while our feature field generalizes.

4.3. Multi-view feature consistency

We analyze the multi-view consistency of our feature field against existing 2D SSL features in Tab. 4. We report the results of SceneDINO trained on KITTI-360 and RealEstate10K. SceneDINO trained using DINO features exhibits significant improvements in multi-view feature consistency over standard DINOv2 features. We also train SceneDINO using target features from DINOv2 [79]. Compared to standard DINOv2 and FiT3D fea-

Table 4. **Multi-view consistency results.** Comparing multi-view consistency of SceneDINO to existing 2D SSL features, using L_1 distance (\downarrow), L_2 distance (\downarrow), and cosine similarity (\uparrow) on KITTI-360 and RealEstate10K. We compare DINO (*top*) and DINOv2-based (*bottom*) features. \dagger denotes plain Fit3D features.

Method	KITTI-360			RealEstate10K		
	L_1	L_2	Cos-Sim	L_1	L_2	Cos-Sim
DINO [11]	16.06	0.74	0.70	14.41	0.66	0.75
SceneDINO (w/ DINO)	6.45	0.33	0.93	5.87	0.28	0.95
DINOv2 [79]	15.83	0.73	0.70	14.20	0.66	0.75
Fit3D [115]	22.86	0.81	0.82	19.88	0.72	0.85
Fit3D \dagger [115]	7.02	0.33	0.93	5.67	0.27	0.95
SceneDINO (w/ DINOv2)	5.24	0.24	0.96	4.87	0.22	0.97

Table 5. **SceneDINO analysis.** We analyze the role of decomposing positional encodings, the choice of downsampling features during training, the effectiveness of the feature smoothness loss, the effect of estimated camera poses, and the choice of target features. We report the mean IoU (in %, \uparrow) using a range of 51.2 m on SSCBench-KITTI-360 test. Δ mIoU reports the absolute difference in % points to our standard model with DINO target features.

Δ mIoU	mIoU	Configuration
-1.18	6.82	No downsample (bilinear up. + aug.)
-1.17	6.83	No feature smoothness loss ($\lambda_{fs} = 0$)
-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)

tures, SceneDINO’s feature field yields significantly better multi-view consistency. Notably, compared against plain 3D refined features of Fit3D, SceneDINO shows a better multi-view consistency on both datasets and all metrics while also offering more expressiveness (*cf.* Tab. 2).

4.4. Analyzing SceneDINO

To understand what core components contribute to obtaining an expressive feature field of SceneDINO, we omit or replace individual components and report the results in Tab. 5. Replacing the downsampling approach with bilinear upsampling and multi-crop augmentations, similar to [1], to obtain high-resolution target features leads decrease SSC mIoU by 1.18 %. Omitting the feature smoothness loss leads to a similar mIoU drop. Abolishing the constant decomposition of positional encodings leads to a mIoU drop of 0.74 %. Training using unsupervised camera poses estimated by ORB-SLAM3 [7] results in an insignificant mIoU drop of only 0.12 %, over using KITTI-360 poses. Going from DINO to DINOv2 target features leads to an increased mIoU of 1.08 %, demonstrating, SceneDINO can benefit from more expressive 2D target features.

In Tab. 6, we analyze our 3D distillation. Performing no distillation at all, just clustering our features, decreases mIoU by 1.61 %. Omitting the k NN-correlation loss leads to a mIoU drop of 1.35 %. Distilling only with center points,

Table 6. **Feature distillation analysis.** We analyze the effectiveness of distilling SceneDINO’s features, the k NN-correlation loss, our neighborhood sampling, and our 3D sampling approach over standard 5-crop sampling. We report the mean IoU (in %, \uparrow) using a range of 51.2 m on SSCBench-KITTI-360 test.

Δ mIoU	mIoU	Configuration
-1.61	6.39	No distillation
-1.35	6.65	No k NN-correlation loss ($\lambda_{kNN} = 0$)
-0.97	7.03	No neighborhood sampling (<i>cf.</i> Fig. 4)
-0.47	7.53	5-crop sampling [32] (instead 3D sampling)
—	8.00	Full framework (SceneDINO)

Table 7. **Probing analysis.** We analyze linear and unsupervised probing of our distilled SceneDINO features on SSCBench-KITTI-360 test using mean IoU (in %, \uparrow). For reference, we also report S4C (2D supervised). Linear probing uses 2D annotations.

Probing approach	Target features	mIoU		
		12.8 m	25.6 m	51.2 m
Unsupervised	DINO [11]	10.76	10.01	8.00
	DINOv2 [79]	13.76	11.78	9.08
Linear	DINO [11]	13.63	12.07	9.34
	DINOv2 [79]	15.85	13.70	10.57
S4C (full training)	n/a	16.94	13.94	10.19

i.e., not performing neighborhood sampling (*cf.* Fig. 4), reduces mIoU by 0.97 %. Using 5-crop feature sampling [32], instead of our proposed 3D sampling, leads to a reduced mIoU of 0.47 %. This demonstrates the effectiveness of performing distillation in 3D using our novel approach.

While focusing on unsupervised SSC, we can also linearly probe our distilled feature field (*cf.* Tab. 7). In particular, we train SceneDINO using different target features (DINO [11] and DINOv2 [11]), perform distillation, and probe the resulting distilled features. Using linear probing, *i.e.*, training a *single* linear layer using 2D semantic labels, leads to a consistent mIoU increase over unsupervised probing. SceneDINO trained using DINOv2 target features even closes the gap to S4C, trained using 2D ground-truth semantic labels. We even surpass 2D supervised S4C slightly on the full range (51.2 m), suggesting the effectiveness of SceneDINO also for weakly-supervised tasks.

5. Conclusion

We presented SceneDINO, to our knowledge, the first approach for unsupervised semantic scene completion. Trained using multi-view images and 2D DINO features without human supervision, SceneDINO is able to predict an expressive 3D feature field using a single input image during inference. Our novel 3D distillation approach yields state-of-the-art results in unsupervised SSC. While we focus on unsupervised SSC, our multi-view feature consistency, linear probing, and domain generalization results highlight the potential of SceneDINO as a strong foundation for various 3D scene-understanding tasks.

Acknowledgments. This project was partially supported by the European Research Council (ERC) Advanced Grant SIMULACRON, DFG project CR 250/26-1 “4D-YouTube”, and GNI Project “AICC”. This project has also received funding from the ERC under the European Union’s Horizon 2020 research and innovation programme (grant agreement No. 866008). Additionally, this work has further been co-funded by the LOEWE initiative (Hesse, Germany) within the emergenCITY center [LOEWE/1/12/519/03/05.001(0016)/72] and by the Excellence Cluster EXC3066 “The Adaptive Mind”. Christoph Reich is supported by the Konrad Zuse School of Excellence in Learning and Intelligent Systems (ELIZA) through the DAAD programme Konrad Zuse Schools of Excellence in Artificial Intelligence, sponsored by the Federal Ministry of Education and Research. Christian Rupprecht is supported by an Amazon Research Award. Finally, we acknowledge the support of the European Laboratory for Learning and Intelligent Systems (ELLIS) and thank Mateo de Mayo as well as Igor Cvišić for help with estimating camera poses.

References

- [1] Nikita Araslanov and Stefan Roth. Self-supervised augmentation consistency for adapting semantic segmentation. In *CVPR*, pages 15384–15394, 2021. [8](#)
- [2] Yuki Markus Asano, Christian Rupprecht, and Andrea Vedaldi. Self-labelling via simultaneous clustering and representation learning. In *ICLR*, 2020. [2](#)
- [3] Philip Bachman, R. Devon Hjelm, and William Buchwalter. Learning representations by maximizing mutual information across views. In *NeurIPS*2019*, pages 15509–15519. [2](#)
- [4] Adrien Bardes, Jean Ponce, and Yann LeCun. VI-CRegL: Self-supervised learning of local visual features. In *NeurIPS*2022*, pages 8799–8810. [2](#)
- [5] Adrien Bardes, Jean Ponce, and Yann LeCun. VI-CReg: Variance-invariance-covariance regularization for self-supervised learning. In *ICLR*, 2022. [2](#)
- [6] Yingjie Cai, Xuesong Chen, Chao Zhang, Kwan-Yee Lin, Xiaogang Wang, and Hongsheng Li. Semantic scene completion via integrating instances and scene in-the-loop. In *CVPR*, pages 324–333, 2021. [2](#)
- [7] Carlos Campos, Richard Elvira, Juan J. Gómez Rodríguez, José M. M. Montiel, and Juan D Tardós. ORB-SLAM3: An accurate open-source library for visual, visual-inertial and multi-map SLAM. *IEEE Trans. Robot.*, 37(6):1874–1890, 2021. [4, 8, vi](#)
- [8] Anh-Quan Cao and Raoul de Charette. Monoscene: Monocular 3D semantic scene completion. In *CVPR*, pages 3981–3991, 2022. [2](#)
- [9] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *NeurIPS*2020*, pages 9912–9924. [2](#)
- [10] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *ECCV*, pages 132–149, 2018. [2](#)
- [11] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, pages 9650–9660, 2021. [1, 2, 3, 4, 6, 7, 8, vi](#)
- [12] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv:2003.04297 [cs.CV]*, 2020. [2](#)
- [13] Xiaokang Chen, Kwan-Yee Lin, Chen Qian, Gang Zeng, and Hongsheng Li. 3d sketch-aware semantic scene completion via semi-supervised structure prior. In *CVPR*, pages 4192–4201, 2020. [2](#)
- [14] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *CVPR*, pages 9640–9649, 2021. [2](#)
- [15] Yuedong Chen, Chuanxia Zheng, Haofei Xu, Bohan Zhuang, Andrea Vedaldi, Tat-Jen Cham, and Jianfei Cai. MVSplat360: Feed-forward 360 scene synthesis from sparse views. In *NeurIPS*2024*, pages 107064–107086. [2](#)
- [16] Ran Cheng, Christopher Agia, Yuan Ren, Xinhai Li, and Bingbing Liu. S3CNet: A sparse semantic scene completion network for LiDAR point clouds. In *CoRL*, pages 2148–2161, 2020. [2](#)
- [17] Jang Hyun Cho, Utkarsh Mall, Kavita Bala, and Bharath Hariharan. PiCIE: Unsupervised semantic segmentation using invariance and equivariance in clustering. In *CVPR*, pages 16794–16804, 2021. [2, 6, i](#)
- [18] Özgür Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3D U-Net: Learning dense volumetric segmentation from sparse annotation. In *MICCAI*, pages 424–432, 2016. [1](#)
- [19] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The Cityscapes dataset for semantic urban scene understanding. In *CVPR*, pages 3213–3223, 2016. [6, i](#)
- [20] Timothée Darcet, Federico Baldassarre, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Cluster and predict latents patches for improved masked image modeling. *arXiv:2502.08769 [cs.CV]*, 2025. [2](#)
- [21] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009. [6](#)
- [22] Carl Doersch, Abhinav Gupta, and Alexei A. Efros. Unsupervised visual representation learning by context prediction. In *ICCV*, pages 1422–1430, 2015. [2](#)
- [23] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16×16 words: Transformers for image recognition at scale. In *ICLR*, 2021. [2](#)
- [24] Linus Ericsson, Henry Gouk, Chen Change Loy, and Timothy M. Hospedales. Self-supervised representation learning: Introduction, advances, and challenges. *IEEE Trans. Signal Process.*, 39(3):42–62, 2022. [2](#)
- [25] Stephanie Fu, Mark Hamilton, Laura E. Brandt, Axel Feldmann, Zhoutong Zhang, and William T. Freeman. FeatUp:

- A model-agnostic framework for features at any resolution. In *ICLR*, 2024. 4
- [26] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The KITTI dataset. *Int. J. Robot. Res.*, 32(11):1231–1237, 2013. 1
- [27] Clément Godard, Oisin Mac Aodha, and Gabriel J. Brostow. Unsupervised monocular depth estimation with left-right consistency. In *CVPR*, pages 270–279, 2017. 4
- [28] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, et al. Bootstrap your own latent: A new approach to self-supervised learning. In *NeurIPS*2020*, pages 21271–21284. 2
- [29] Agrim Gupta, Jiajun Wu, Jia Deng, and Li Fei-Fei. Siamese Masked Autoencoders. In *NeurIPS*2023*, pages 40676–40693. 2
- [30] Huy Ha and Shuran Song. Semantic abstraction: Open-world 3D scene understanding from 2D vision-language models. In *CoRL*, pages 643–653, 2023. 2
- [31] Oliver Hahn, Nikita Araslanov, Simone Schaub-Meyer, and Stefan Roth. Boosting unsupervised semantic segmentation with principal mask proposals. *Trans. Mach. Learn. Res.*, 2024. 3, 6, i, iv
- [32] Mark Hamilton, Zhoutong Zhang, Bharath Hariharan, Noah Snavely, and William T. Freeman. Unsupervised semantic segmentation by distilling feature correspondences. In *ICLR*, 2022. 1, 2, 3, 4, 5, 6, 7, 8, i, iv
- [33] Keonhee Han, Dominik Muhle, Felix Wimbauer, and Daniel Cremers. Boosting self-supervision for single-view scene completion via knowledge distillation. In *CVPR*, pages 9837–9847, 2024. 1, 2
- [34] Xian-Feng Han, Hamid Laga, and Mohammed Benamoun. Image-based 3D object reconstruction: State-of-the-art and trends in the deep learning era. *IEEE Trans. Pattern Anal. Mach. Intell.*, 43(5):1578–1604, 2019. 2
- [35] Robert Harb and Patrick Knöbelreiter. InfoSeg: Unsupervised semantic image segmentation with mutual information maximization. In *GCPR*, pages 18–32, 2021. 2
- [36] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge University Press, 2003. 2
- [37] Adrian Hayler, Felix Wimbauer, Dominik Muhle, Christian Rupprecht, and Daniel Cremers. S4C: Self-supervised semantic scene completion with neural fields. In *3DV*, pages 409–420, 2024. 1, 2, 6, 7, i, iv, vi
- [38] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, pages 9729–9738, 2020. 2
- [39] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, pages 16000–16009, 2022. 2
- [40] Olivier Henaff. Data-efficient image recognition with contrastive predictive coding. In *ICML*, pages 4182–4192, 2020. 2
- [41] R. Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. In *ICLR*, 2019. 2
- [42] Mu Hu, Wei Yin, Chi Zhang, Zhipeng Cai, Xiaoxiao Long, Hao Chen, Kaixuan Wang, Gang Yu, Chunhua Shen, and Shaojie Shen. Metric3D v2: A versatile monocular geometric foundation model for zero-shot metric depth and surface normal estimation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 46(12):10579–10596, 2024. 2
- [43] Rui Huang, Songyou Peng, Ayca Takmaz, Federico Tombari, Marc Pollefeys, Shiji Song, Gao Huang, and Francis Engelmann. Segment3D: Learning fine-grained class-agnostic 3D segmentation without manual labels. In *ECCV*, pages 278–295, 2024. 2
- [44] Yuanhui Huang, Wenzhao Zheng, Yunpeng Zhang, Jie Zhou, and Jiwen Lu. Tri-perspective view for vision-based 3D semantic occupancy prediction. In *CVPR*, pages 9223–9232, 2023. 2
- [45] Yuanhui Huang, Wenzhao Zheng, Borui Zhang, Jie Zhou, and Jiwen Lu. SelfOcc: Self-supervised vision-based 3D occupancy prediction. In *CVPR*, pages 19946–19956, 2024. 1, 2
- [46] Joel Janai, Fatma Güney, Aseem Behl, and Andreas Geiger. Computer vision for autonomous vehicles: Problems, datasets and state of the art. *Found. Trends Comput. Graph. Vis.*, 12(1–3):1–308, 2020. 1
- [47] Xu Ji, Joao F. Henriques, and Andrea Vedaldi. Invariant information clustering for unsupervised image classification and segmentation. In *ICCV*, pages 9865–9874, 2019. 2
- [48] Haoyi Jiang, Liu Liu, Tianheng Cheng, Xinjie Wang, Tianwei Lin, Zhizhong Su, Wenyu Liu, and Xinggang Wang. GaussTR: Foundation model-aligned gaussian transformer for self-supervised 3D spatial understanding. *arXiv:2412.13193 [cs.CV]*, 2024. 2
- [49] Justin Kerr, Chung Min Kim, Ken Goldberg, Angjoo Kanazawa, and Matthew Tancik. LERF: Language embedded radiance fields. In *ICCV*, pages 19729–19739, 2023. 2
- [50] Chanyoung Kim, Woojung Han, Dayun Ju, and Seong Jae Hwang. EAGLE: Eigen aggregation learning for object-centric unsupervised semantic segmentation. In *CVPR*, pages 3523–3533, 2024. 3, 6, i
- [51] Diederik P. Kingma and Jimmy Lei Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 6
- [52] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment Anything. In *ICCV*, pages 4015–4026, 2023. 1, 2
- [53] Sosuke Kobayashi, Eiichi Matsumoto, and Vincent Sitzmann. Decomposing NeRF for editing via feature field distillation. In *NeurIPS*2022*, pages 23311–23330. 2
- [54] Alexander Koenig, Maximilian Schambach, and Johannes Otterbach. Uncovering the inner workings of STEGO for safe unsupervised semantic segmentation. In *CVPRW*, pages 3789–3798, 2023. 4

- [55] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected CRFs with Gaussian edge potentials. In *NIPS*2011*, pages 109–117. 6
- [56] Harold W. Kuhn. The hungarian method for the assignment problem. *Nav. Res. Logist. Q.*, 2:83–97, 1955. 6, i
- [57] Jie Li, Yu Liu, Dong Gong, Qinfeng Shi, Xia Yuan, Chunxia Zhao, and Ian D. Reid. RGBD based dimensional decomposition residual network for 3D semantic scene completion. In *CVPR*, pages 7693–7702, 2019. 2
- [58] Jie Li, Kai Han, Peng Wang, Yu Liu, and Xia Yuan. Anisotropic convolutional networks for 3D semantic scene completion. In *CVPR*, pages 3348–3356, 2020.
- [59] Jie Li, Yu Liu, Xia Yuan, Chunxia Zhao, Roland Siegwart, Ian Reid, and Cesar Cadena. Depth based semantic scene completion with position importance aware loss. *IEEE Robotics Autom. Lett.*, 5(1):219–226, 2020. 2
- [60] Junnan Li, Pan Zhou, Caiming Xiong, and Steven Hoi. Prototypical contrastive learning of unsupervised representations. In *ICLR*, 2021. 2
- [61] Pengfei Li, Yongliang Shi, Tianyu Liu, Hao Zhao, Guyue Zhou, and Ya-Qin Zhang. Semi-supervised implicit scene completion from sparse lidar. *arXiv:2111.14798 [cs.CV]*, 2021. 2
- [62] Yiming Li, Zhiding Yu, Christopher B. Choy, Chaowei Xiao, José M. Álvarez, Sanja Fidler, Chen Feng, and Anima Anandkumar. VoxFormer: Sparse voxel transformer for camera-based 3D semantic scene completion. In *CVPR*, pages 9087–9098, 2023. 2, iii, iv
- [63] Yiming Li, Sihang Li, Xinhao Liu, Moonjun Gong, Kenan Li, Nuo Chen, Zijun Wang, Zhiheng Li, Tao Jiang, Fisher Yu, Yue Wang, Hang Zhao, Zhiding Yu, and Chen Feng. SSCBench: A large-scale 3D semantic scene completion benchmark for autonomous driving. In *IROS*, pages 13333–13340, 2024. 1, 2, 6, i, iii
- [64] Zhiqi Li, Zhiding Yu, David Austin, Mingsheng Fang, Shiyi Lan, Jan Kautz, and José M. Álvarez. FB-OCC: 3D occupancy prediction based on forward-backward view transformation. *arXiv:2307.01492 [cs.CV]*, 2023. 2
- [65] Yiyi Liao, Jun Xie, and Andreas Geiger. KITTI-360: A novel dataset and benchmarks for urban scene understanding in 2D and 3D. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(3):3292–3310, 2023. 1, 5, i, v
- [66] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, pages 740–755, 2024. 6
- [67] Shice Liu, Yu Hu, Yiming Zeng, Qiankun Tang, Beibei Jin, Yinhe Han, and Xiaowei Li. See and think: Disentangling semantic scene completion. In *NeurIPS*2018*, pages 261–272. 2
- [68] Stuart Lloyd. Least squares quantization in PCM. *IEEE Trans. Inf. Theory*, 28(2):129–137, 1982. 4, 5
- [69] Zhiliang Ma and Shilong Liu. A review of 3D reconstruction techniques in civil engineering and their applications. *Adv. Eng. Inform.*, 37:163–174, 2018. 1
- [70] James MacQueen. Some methods for classification and analysis of multivariate observations. In *Berkeley Symp. on Math. Statist. and Prob.*, pages 281–298, 1967. 4, 5
- [71] Nelson Max. Optical models for direct volume rendering. *IEEE Trans. Vis. Comput. Graph.*, 1(2):99–108, 1995. 3
- [72] Kirill Mazur, Edgar Sucar, and Andrew J Davison. Feature-realistic neural fusion for real-time, open set scene understanding. In *ICRA*, pages 8201–8207, 2023. 2
- [73] Ruihang Miao, Weizhou Liu, Mingrui Chen, Zheng Gong, Weixin Xu, Chen Hu, and Shuchang Zhou. Occdepth: A depth-aware method for 3D semantic scene completion. *arXiv:2302.13540 [cs.CV]*, 2023. 1, 2
- [74] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. *Commun. ACM*, 65(1):99–106, 2021. 2, 3
- [75] Yue Ming, Xuyang Meng, Chunxiao Fan, and Hui Yu. Deep learning for monocular depth estimation: A review. *Neurocomputing*, 438:14–33, 2021. 2
- [76] Duy Kien Nguyen, Yanghao Li, Vaibhav Aggarwal, Martin R. Oswald, Alexander Kirillov, Cees G. M. Snoek, and Xinlei Chen. R-MAE: Regions meet masked autoencoders. In *ICLR*, 2024. 2
- [77] Dantong Niu, Xudong Wang, Xinyang Han, Long Lian, Roei Herzig, and Trevor Darrell. Unsupervised universal image segmentation. In *CVPR*, pages 22744–22754, 2024. 6, 7, i
- [78] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *ECCV*, pages 69–84, 2016. 2
- [79] Maxime Oquab, Timothée Darcret, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *Trans. Mach. Learn. Res.*, 2024. 2, 3, 6, 7, 8, iv
- [80] Martin R Oswald, Eno Töppe, Claudia Nieuwenhuis, and Daniel Cremers. A review of geometry recovery from a single image focusing on curved object reconstruction. *Innovations for Shape Analysis: Models and Algorithms*, pages 343–378, 2013. 2
- [81] Mingjie Pan, Jiaming Liu, Renrui Zhang, Peixiang Huang, Xiaoqi Li, Hongwei Xie, Bing Wang, Li Liu, and Shanghang Zhang. RenderOcc: Vision-centric 3D occupancy prediction with 2D rendering supervision. In *ICRA*, pages 12404–12411, 2024. 2
- [82] Songyou Peng, Kyle Genova, Chiyu “Max” Jiang, Andrea Tagliasacchi, Marc Pollefeys, and Thomas Funkhouser. OpenScene: 3D scene understanding with open vocabularies. In *CVPR*, pages 815–824, 2023. 2
- [83] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *ICCV*, pages 12179–12188, 2021. 6
- [84] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, et al. SAM 2: Segment anything in images and videos. *arXiv:2408.00714 [cs.CV]*, 2024. 1
- [85] Stephan R. Richter and Stefan Roth. Matryoshka networks: Predicting 3D geometry via nested shape layers. In *CVPR*, pages 1936–1944, 2018. 2

- [86] Christoph B. Rist, David Emmerichs, Markus Enzweiler, and Dariu M. Gavrila. Semantic scene completion using local deep implicit functions on lidar data. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(10):7205–7218, 2022. 2
- [87] Luis Roldão, Raoul de Charette, and Anne Verroust-Blondet. Lmscnet: Lightweight multiscale 3D semantic completion. In *3DV*, pages 111–119, 2020. 1, 2
- [88] Luis Roldao, Raoul De Charette, and Anne Verroust-Blondet. 3D semantic scene completion: A survey. *Int. J. Comput. Vis.*, 130(8):1978–2005, 2022. 1
- [89] Johannes L. Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *CVPR*, pages 4104–4113, 2016. 2
- [90] David Sculley. Web-scale k -means clustering. In *WWW*, page 1177–1178, 2010. 5
- [91] Hyun Seok Seong, WonJun Moon, SuBeen Lee, and Jae-Phil Heo. Leveraging hidden positives for unsupervised semantic segmentation. In *CVPR*, pages 19540–19549, 2023. 3, 6, i
- [92] Nur Muhammad Mahi Shafiullah, Chris Paxton, Lerrel Pinto, Soumith Chintala, and Arthur Szlam. CLIP-Fields: Weakly supervised semantic fields for robotic memory. In *ICRA Workshop on Pretraining for Robotics*, 2023. 2
- [93] William Shen, Ge Yang, Alan Yu, Jansen Wong, Leslie Pack Kaelbling, and Phillip Isola. Distilled feature fields enable few-shot language-guided manipulation. In *CoRL*, pages 405–424, 2023. 2
- [94] Leon Sick, Dominik Engel, Pedro Hermosilla, and Timo Ropinski. Unsupervised semantic segmentation through depth-guided feature correlation and sampling. In *CVPR*, pages 3637–3646, 2024. 3, 6, i
- [95] Shuran Song, Fisher Yu, Andy Zeng, Angel X. Chang, Manolis Savva, and Thomas A. Funkhouser. Semantic scene completion from a single depth image. In *CVPR*, pages 190–198, 2017. 1, 2, iii, iv
- [96] Stanislaw Szymanowicz, Eldar Insafutdinov, Chuanxia Zheng, Dylan Campbell, Joao F. Henriques, Christian Rupprecht, and Andrea Vedaldi. Flash3D: Feed-forward generalisable 3D scene reconstruction from a single image. *arXiv:2406.04343 [cs.CV]*, 2024. 2
- [97] Ayça Takmaz, Elisabetta Fedele, Robert W Sumner, Marc Pollefeys, Federico Tombari, and Francis Engelmann. OpenMask3D: Open-vocabulary 3D instance segmentation. In *NeurIPS*2023*, pages 68367–68390. 2
- [98] Zachary Teed and Jia Deng. RAFT: Recurrent all-pairs field transforms for optical flow. In *ECCV*, pages 402–419, 2020. 6, ii
- [99] Wenwen Tong, Chonghao Sima, Tai Wang, Li Chen, Silei Wu, Hanming Deng, Yi Gu, Lewei Lu, Ping Luo, Dahua Lin, and Hongyang Li. Scene as occupancy. In *ICCV*, pages 8372–8381, 2023. 2
- [100] Nikolaos Tsagkas, Oisin Mac Aodha, and Chris Xiao-xuan Lu. VL-Fields: Towards language-grounded neural implicit spatial representations. In *ICRA Workshop on Representations, Abstractions, and Priors for Robot Learning*, 2023. 2
- [101] Vadim Tschernezki, Iro Laina, Diane Larlus, and Andrea Vedaldi. Neural feature fusion fields: 3D distillation of self-supervised 2D image representations. In *3DV*, pages 443–453, 2022. 2
- [102] Shubham Tulsiani, Tinghui Zhou, Alexei A. Efros, and Jitendra Malik. Multi-view supervision for single-view reconstruction via differentiable ray consistency. In *CVPR*, 2017. 2
- [103] Xudong Wang, Rohit Girdhar, Stella X. Yu, and Ishan Misra. Cut and learn for unsupervised object detection and instance segmentation. In *CVPR*, pages 3124–3134, 2023. 1, 2
- [104] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Trans. Image Process.*, 13(4):600–612, 2004. 4
- [105] Silvan Weder, Hermann Blum, Francis Engelmann, and Marc Pollefeys. LabelMaker: Automatic semantic label generation from RGB-D trajectories. In *3DV*, pages 334–343, 2024. 2
- [106] Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan Yuille, and Christoph Feichtenhofer. Masked feature prediction for self-supervised visual pre-training. In *CVPR*, pages 14668–14678, 2022. 2
- [107] Felix Wimbauer, Nan Yang, Christian Rupprecht, and Daniel Cremers. Behind the scenes: Density fields for single view reconstruction. In *CVPR*, pages 9076–9086, 2023. 1, 2, 3, 4, 6, i
- [108] Yiheng Xie, Towaki Takikawa, Shunsuke Saito, Or Litany, Shiqin Yan, Numair Khan, Federico Tombari, James Tompkin, Vincent Sitzmann, and Srinath Sridhar. Neural fields in visual computing and beyond. In *Comput. Graph. Forum*, pages 641–676, 2022. 2
- [109] Xu Yan, Jiantao Gao, Jie Li, Ruimao Zhang, Zhen Li, Rui Huang, and Shuguang Cui. Sparse single sweep lidar point cloud segmentation via learning contextual shape priors from scene completion. In *AAAI*, pages 3101–3109, 2021. 2
- [110] Jiawei Yang, Boris Ivanovic, Or Litany, Xinshuo Weng, Seung Wook Kim, Boyi Li, Tong Che, Danfei Xu, Sanja Fidler, Marco Pavone, and Yue Wang. EmerneRF: Emergent spatial-temporal scene decomposition via self-supervision. In *ICLR*, 2024. 2
- [111] Jiawei Yang, Katie Z Luo, Jiefeng Li, Congyue Deng, Leonidas Guibas, Dilip Krishnan, Kilian Q Weinberger, Yonglong Tian, and Yue Wang. Denoising vision transformers. In *ECCV*, pages 453–469, 2024. 4
- [112] Zhuoyue Yang, Ju Dai, and Junjun Pan. 3D reconstruction from endoscopy images: A survey. *Comput. Biol. Med.*, 175:108546, 2024. 1
- [113] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelNeRF: Neural radiance fields from one or few images. In *CVPR*, pages 4578–4587, 2021. 2
- [114] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashishth Madhavan, and Trevor Darrell. BDD100K: A diverse driving dataset for heterogeneous multitask learning. In *CVPR*, pages 2633–2642, 2020. 6, i

- [115] Yuanwen Yue, Anurag Das, Francis Engelmann, Siyu Tang, and Jan Eric Lenssen. Improving 2D feature representations by 3D-aware fine-tuning. In *ECCV*, pages 57–74, 2024. [2](#), [6](#), [7](#), [8](#)
- [116] Pingping Zhang, Wei Liu, Yinjie Lei, Huchuan Lu, and Xiaoyun Yang. Cascaded context pyramid for full-resolution 3D semantic scene completion. In *ICCV*, pages 7800–7809, 2019. [2](#)
- [117] Yunpeng Zhang, Zheng Zhu, and Dalong Du. OccFormer: Dual-path transformer for vision-based 3D semantic occupancy prediction. In *ICCV*, pages 9433–9443, 2023. [2](#), [iii](#), [iv](#)
- [118] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. *ACM Trans. Graph.*, 37(4):65, 2018. [6](#), [ii](#)
- [119] Onur Özyeşil, Vladislav Voroninski, Ronen Basri, and Amit Singer. A survey of structure from motion. *Acta Numer.*, 26:305–364, 2017. [2](#)

Feed-Forward SceneDINO for Unsupervised Semantic Scene Completion

Supplementary Material

Aleksandar Jevtić^{* 1} Christoph Reich^{* 1,2,4,5} Felix Wimbauer^{1,4}
Oliver Hahn² Christian Rupprecht³ Stefan Roth^{2,5,6} Daniel Cremers^{1,4,5}
¹TU Munich ²TU Darmstadt ³University of Oxford ⁴MCML ⁵ELIZA ⁶hessian.AI * equal contribution
<https://visinf.github.io/scenedino>

In this appendix, we provide further implementation details, including dataset properties and an overview of SceneDINO’s computational complexity (*cf.* Sec. A). We discuss our multi-view feature consistency evaluation approach (*cf.* Sec. B). Next, we provide additional qualitative and quantitative results (*cf.* Sec. C), including failure cases. Finally, we discuss the limitations of SceneDINO and suggest future research directions (*cf.* Sec. D).

A. Reproducibility

Here, we provide further implementation details, information about the utilized dataset, and computational complexity details to ensure reproducibility. Note that our code is available at <https://github.com/tum-vision/scenedino>.

A.1. Implementation details

We implement SceneDINO in PyTorch [122] and build on the code of BTS [107], STEGO [32], and S4C [37]. Our encoder-decoder (pre-trained DINO-B/8 and randomly initialized dense prediction decoder) produces per-pixel embeddings of dimensionality $D_E = 256$. Based on these embeddings, the two-layer MLP ϕ (hidden dimension 128) predicts 64 features. As rendering features is expensive, requiring multiple forward passes through the MLP, ϕ predicts 64 features. We employ another MLP to up-project again to the full dimensionality $D = 768$, this MLP is learned with SceneDINO and can up-project both 3D features and 2D rendered features. We train for 100k steps with a base learning rate of 10^{-4} , dropping to 10^{-5} after 50k steps. We train using a batch size of 4, extracting 32 patches of size 8×8 per image. These patches align with the per-patch DINO target features. For our feature field loss formulation (*cf.* Sec. 3.2), we use the loss weights $\lambda_p = 1$, $\lambda_s = 0.001$, $\lambda_f = 0.2$, $\lambda_{fs} = 0.25$.

The MLP head h (hidden dimension 768) produces 64 distilled features. We perform distillation for 1000 steps with a learning rate of $5 \cdot 10^{-4}$. We train using a batch size of 4, 5 center points, a feature batch of size 576, and cluster with $K = 19$. For kNN sampling, we use $k = 4$. The feature buffer holds 256 feature batches. The loss term in Eq. (9) is parameterized with $\lambda_{self} = 0.08$ $\lambda_{kNN} = 0.43$

$\lambda_{rand} = 0.67$, and $b_{rand} = 0.87$. For the similarity thresholds we use $b_{self} = 0.44$, $b_{kNN} = 0.18$, and $b_{rand} = 0.87$.

We follow standard practice in 2D unsupervised semantic segmentation [17, 31, 32, 50, 77, 91, 94] by applying Hungarian matching [56] to align our pseudo semantics. For SSC validation, we map down to 15 semantic classes while following existing work [31, 32] for 2D validation and map to 19 semantic classes.

A.2. Datasets

We provide additional details about the datasets utilized to train and evaluate SceneDINO.

KITTI-360 [63, 65] provides video sequences from a moving vehicle equipped with a forward-facing stereo pair and two side-facing fisheye cameras. In future frames, the fisheye views capture additional geometric and semantic cues of regions occluded in the forward-facing view. For training, we resample the fisheye images into perspective projection. We focus on an area approximately 50 meters ahead of the ego vehicle. Assuming an average velocity of 30–50 km/h, side views are randomly sampled 1–4 seconds into the future. Given a frame rate of 10 Hz, this translates to 10–40 time steps. Each training sample consists of eight images: four forward-facing views (including the input image) and four side-facing views.

To evaluate our predicted field in SSCBench-KITTI-360, we follow the evaluation procedure of S4C [37]. The voxel predictions are evaluated in three different ranges: $12.8 \text{ m} \times 12.8 \text{ m} \times 6.4 \text{ m}$, $25.6 \text{ m} \times 25.6 \text{ m} \times 6.4 \text{ m}$, and the full range $51.2 \text{ m} \times 51.2 \text{ m} \times 6.4 \text{ m}$. For each voxel, multiple evenly distributed points are sampled from the semantic field. The predictions are aggregated per voxel by taking the maximum occupancy and weighting the class predictions accordingly.

Cityscapes [19] consists of 500 high-resolution and densely annotated validation images of ego-centric driving scenes. For validation, Cityscapes uses a 19-class taxonomy. We leverage the Cityscapes validation samples at a resolution of 640×192 for our domain generalization experiments (2D semantic segmentation).

BDD-100K [114] is a driving scene dataset obtained from urban areas in the US. BDD-100K contains 1000 semantic

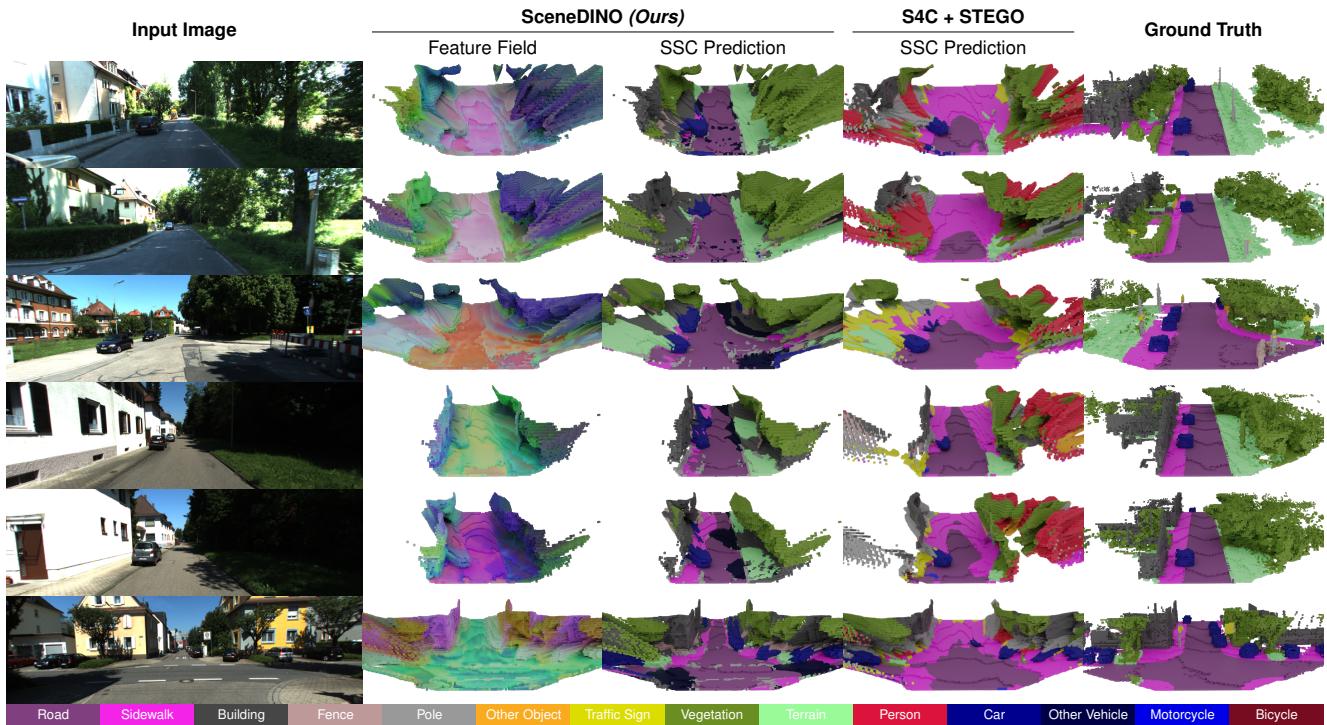


Figure 6. **3D qualitative SSC comparison on KITTI-360.** We provide additional qualitative results, visualizing the input image, SceneDINO’s predicted feature field using the first three principal components, and SSC prediction, the SSC prediction of our baseline S4C+STEGO, and the SSC ground truth. We only visualize surface voxels within the field of view for the sake of clarity.

segmentation validation images. The semantic taxonomy follows the 19-class Cityscapes definition. For domain generalization experiments, we utilize BDD-100K images at a resolution of 640×192 .

RealEstate10K [118] is a large-scale dataset containing videos of real-world indoor and outdoor scenes, primarily sourced from YouTube. For our experiments, we train with a resolution of 512×288 . Each training sample consists of three frames, separated by a randomly sampled time offset. There are no semantic annotations provided with the dataset. We evaluate the multi-view consistency of our model in this setting.

A.3. Computational complexity

SceneDINO requires only a *single* GPU for training and inference. In SSCBench (51.2 m range), SceneDINO requires 0.76 ± 0.1 s to infer a full scene on a V100 GPU. The peak VRAM usage during inference is 11 GB. For reference, S4C requires 0.32 ± 0.13 s. Considering our expressive and high-dimensional feature field and ViT encoder, this is a moderate runtime increase. SceneDINO has 100 M parameters and is trained for approximately 2 days on a *single* V100 32 GB GPU. All results are reported using automatic mixed precision.

B. Multi-View Feature Consistency Evaluation

We aim to measure the multi-view consistency of 2D and 3D features. Note, we are not aware of any standardized approach for multi-view feature consistency. To this end, we employ a straightforward approach. Given two video frames with a temporal stride of 3, forward optical flow is computed using RAFT large [98]. We estimate occlusion by forward-backward consistency [124]; for this, we also compute the backward optical flow. The 2D feature maps obtained using the second frame are backward warped to the features of the first frame. We compute different similarity metrics between the aligned features (L_1 , L_2 , and cos-sim). Note that we ignore occlusions. While features from DINO, DINOv2, and FiT3D possess a lower resolution than our 2D rendered SceneDINO features, we upscale these features to the image resolution before warping. This evaluation approach utilizes optical flow correspondences and captures both ego motion as well as object motion, offering a simple way to evaluate multi-view feature consistency.

C. Additional Results

Here we provide additional qualitative and quantitative results, extending our results reported in the main paper.

Qualitative results. In Fig. 6, we present additional qualitative results of SceneDINO using our 3D feature distilla-

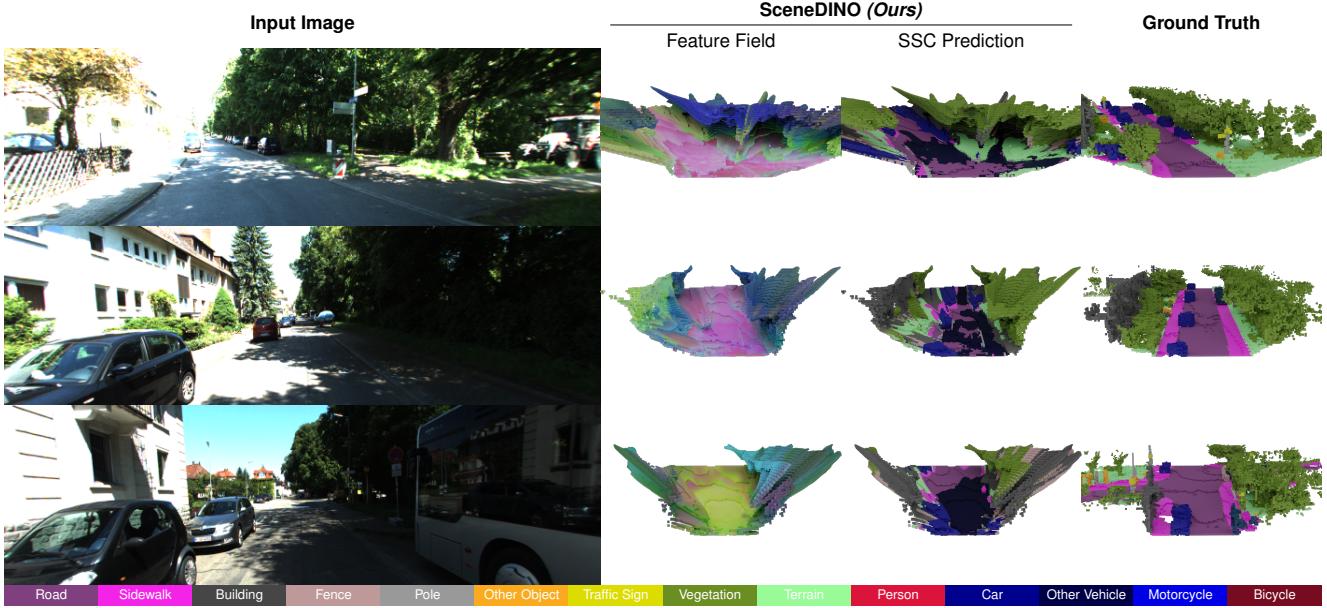


Figure 7. **Failure cases of SceneDINO on KITTI-360.** We provide failure cases of SceneDINO. We visualize the input image, the predicted feature field using the first three principal components, the SSC prediction, and the SSC ground truth. We observe that our semantic predictions struggle in shaded regions. We only visualize surface voxels within the field of view for the sake of clarity.

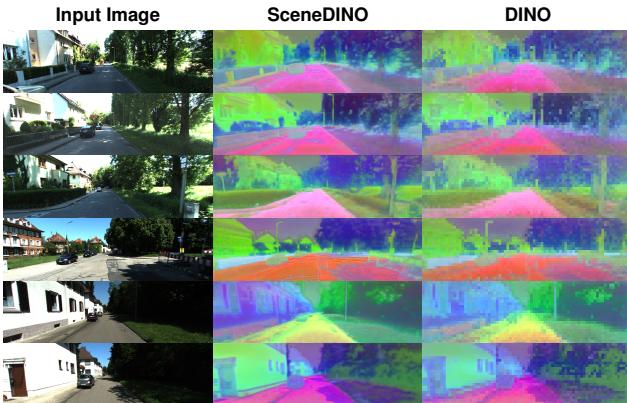


Figure 8. **2D SceneDINO features on KITTI-360.** We visualize our 2D rendered features and DINO features for a given input image (*left*). We use the first three principal components for feature visualization. Notably, SceneDINO’s features (*middle*) are smoother and capture finer structures than DINO (*right*). Additionally, SceneDINO’s features are high-resolution, while DINO generates features with a stride of 8.

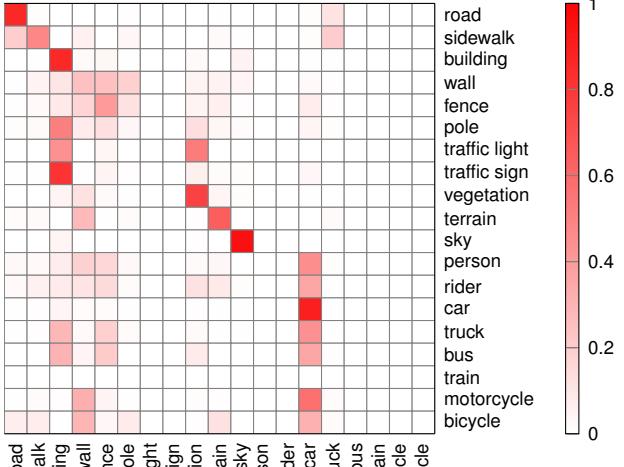
tion approach on unsupervised semantic scene completion. We also provide visualizations of our unsupervised SSC baseline, S4C + STEGO. Qualitatively, our approach obtains more accurate SSC results and is able to segment far-away objects, such as cars, better than the S4C + STEGO baseline. This observation aligns with the quantitative results presented in Tab. 1 of the main paper.

Figure 8 qualitatively analyzes our 2D rendered fea-

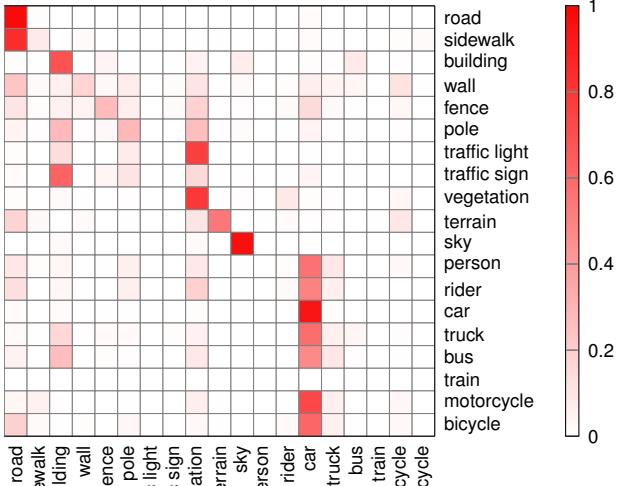
tures against DINO. Our features exhibit a smooth appearance for uniform regions, such as sidewalks. Additionally, SceneDINO’s features better capture fine structures like poles than DINO features. 2D rendered SceneDINO features are also high resolution in contrast to DINO features that exhibit a lower resolution.

Failure cases. In Fig. 7, we provide failure cases of SceneDINO’s SSC predictions. Our predictions exhibit two common failure cases. First, shadowed regions often lead to wrong semantic predictions. Regions affected by significant brightness changes are breaking the brightness consistency, subsequently offering a poor learning signal during training, thus impeding accurate predictions of shadowed regions. Second, objects such as cars can entail tail-like artifacts, not accurately capturing the geometry. As our multi-view image and feature reconstruction training cannot handle dynamic objects, tail-like artifacts could be caused by the poor learning signal for dynamic objects.

Quantitative results. In Tab. 8, we provide additional semantic scene completion results of 3D supervised approaches as an additional point of comparison. In particular, we report official SSCBench [63] results of VoxFormer-S [62] and OccFormer [117]. Both utilize 3D supervision, including both semantic and geometric annotations. We also report the results of SSCNet [95]. This approach trains using 3D supervision but utilizes a depth image during inference. While SceneDINO achieves state-of-the-art segmentation accuracy in the unsupervised setting, supervised approaches are significantly more accurate.



(a) SceneDINO



(b) STEGO

Figure 9. Confusion matrices for 2D unsupervised semantic segmentation on KITTI-360. Rows represent ground-truth class labels (normalized to 1), while columns correspond to predicted class labels. We report results for (a) SceneDINO and (b) STEGO on the SSCBench-KITTI-360 test split.

generally observe that SceneDINO performs well in segmenting frequent classes, such as “road”, “building”, and “sky”. Less frequent classes, such as “fence” and “pole”, are less well segmented. Classes including very small and fine structures (*e.g.*, “pole”) are completely missed by SceneDINO. This trend can also be observed for our 3D unsupervised baseline S4C + STEGO and 2D STEGO. We also observe that class-wise metrics strongly correlate between 2D and 3D.

Table 11. Linear probing results on SSCBench-KITTI-360.

We extend Tab. 7 and report detailed results of SceneDINO using 2D supervised linear probing. Semantic results using mIoU and class IoU, and geometric results using IoU, Precision, and Recall, and (all in %, ↑) on SSCBench-KITTI-360 test using three depth ranges.

Method	SceneDINO w/ DINO (Ours)			SceneDINO w/ DINOv2 (Ours)		
	Supervision			Unsupervised		
Range	12.8 m	25.6 m	51.2 m	12.8 m	25.6 m	51.2 m
<i>Semantic validation</i>						
mIoU	13.63	12.07	9.34	15.85	13.70	10.57
car	16.77	12.37	8.42	20.35	15.04	10.16
bicycle	1.10	0.70	0.47	0.00	0.00	0.00
motorcycle	0.00	0.00	0.00	0.00	0.00	0.00
truck	3.80	2.21	1.52	11.48	7.46	4.63
other-v.	0.13	0.08	0.06	0.00	0.00	0.00
person	0.01	0.00	0.00	0.00	0.00	0.00
road	66.63	62.21	49.99	69.92	63.06	50.49
sidewalk	29.46	25.17	18.85	42.35	37.13	29.13
building	18.64	22.82	17.66	23.03	27.05	21.40
fence	9.29	6.03	3.96	8.82	6.40	4.61
vegetation	32.76	26.49	20.89	30.42	24.96	19.75
terrain	24.80	22.43	18.00	30.73	23.85	17.93
pole	0.25	0.24	0.14	0.46	0.40	0.28
traffic-sign	0.50	0.17	0.09	0.00	0.00	0.00
other-obj.	0.26	0.07	0.04	0.00	0.00	0.00
<i>Geometric validation</i>						
IoU	49.34	42.26	37.61	49.77	43.19	38.55
Precision	52.83	45.95	41.55	52.76	46.46	42.11
Recall	88.21	84.05	79.88	89.76	85.99	82.02

Figure 9 reports confusion matrices of SceneDINO and STEGO for 2D semantic segmentation on KITTI-360. Both approaches share a similar confusion pattern. We attribute this to the fact that both approaches rely on the feature representation of DINO. In particular, we observe confusion between semantically close classes, such as “pole”, “traffic light”, and “traffic sign”. Interestingly, for the semantic classes “person”, “rider”, “car”, “truck”, “bus”, “motorcycle”, and “bicycle”, we see a strong confusion. We suspect this correlation is potentially caused by the fact that these classes often appear on the “road” and “sidewalk” and are rare in KITTI-360.

We also provide class-wise SSC results of SceneDINO using 2D supervised linear probing in Tab. 11. Linear probing provides an upper bound for clustering our features, improving the segmentation accuracy for almost all classes. However, rare classes like “motorcycle” are still not captured using linear probing. This suggests that the DINO feature space fails to express these classes accurately, limiting the segmentation accuracy of SceneDINO. Still, our approach is agnostic to the utilized target features and can potentially profit from better 2D features.

Camera pose analysis. Training SceneDINO, requires accurate camera poses. While KITTI-360 offers ground truth camera poses, these poses are obtained using additional cues, including LiDAR data [65]. To adhere to our fully unsupervised setting, we provide results training with unsupervised camera poses, estimated using stereo visual SLAM. In particular, Tab. 5 reports results of SceneDINO

Table 12. **Camera pose analysis on SSCBench-KITTI-360.** We extend the camera pose analysis in Tab. 5 and report detailed results of SceneDINO with unsupervised camera poses estimated by SOFT2 [121] and ORB-SLAM3 [7]. For reference, we also provide results obtained using the KITTI-360 dataset poses. Semantic results using mIoU and class IoU, and geometric results using IoU, Precision, and Recall, and (all in %, ↑) on SSCBench-KITTI-360 test using three depth ranges.

Method	SceneDINO (Ours)									
	Poses			SceneDINO (Ours)			KITTI-360			
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	10.58	9.58	7.72	10.88	9.86	7.88	10.76	10.01	8.00	
car	18.47	13.98	10.44	19.37	14.09	9.72	21.24	15.94	11.21	
bicycle	0.04	0.03	0.03	0.06	0.03	0.02	0.00	0.00	0.00	
motorcycle	0.00	0.00	0.00	0.01	0.01	0.00	0.00	0.00	0.00	
truck	0.00	0.00	0.00	0.05	0.02	0.01	0.00	0.00	0.00	
other-v.	0.01	0.02	0.04	0.08	0.06	0.05	0.00	0.00	0.00	
person	0.02	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	
road	44.48	44.50	36.06	44.74	40.58	31.86	51.10	49.12	39.82	
sidewalk	16.55	16.79	14.38	21.45	23.56	19.88	20.26	22.31	18.97	
building	19.40	23.40	18.56	19.19	24.87	20.02	12.33	18.27	14.32	
fence	1.79	1.00	0.68	1.62	1.21	0.91	1.91	0.90	0.58	
vegetation	32.10	25.65	20.67	32.60	24.91	19.49	31.22	25.57	19.85	
terrain	25.59	18.11	14.79	23.98	18.41	16.16	23.26	18.02	15.22	
pole	0.18	0.11	0.09	0.00	0.00	0.00	0.05	0.05	0.05	
traffic-sign	0.00	0.01	0.00	0.03	0.03	0.02	0.00	0.00	0.00	
other-obj.	0.08	0.05	0.03	0.08	0.05	0.03	0.00	0.00	0.00	
Geometric validation										
IoU	49.91	41.85	37.25	45.42	40.21	36.65	49.54	42.27	37.60	
Precision	54.74	45.66	40.79	54.42	45.54	40.98	53.27	46.10	41.59	
Recall	84.98	83.40	81.12	73.33	77.46	77.62	87.61	83.59	79.67	

trained using unsupervised camera poses estimated by ORB-SLAM3 [7]. Table 12, extends these results and reports detailed SSC results using two different unsupervised stereo visual SLAM approaches—SOFT2 [121] and ORB-SLAM3 [7]. Using unsupervised and visually estimated poses leads to a minor drop in both semantic and geometric SSC validation. While ORB-SLAM3 poses lead to slightly better semantic accuracy than SOFT2 poses, SOFT2 estimated poses result in higher geometric accuracy. Still, both SOFT2 and ORB-SLAM3 provide poses accurate enough for train SceneDINO, reaching a similar accuracy to employing KITTI-360 poses.

Out-of-domain results. We illustrate on out-of-domain prediction in Fig. 10. While our SceneDINO model is trained on the KITTI-360 dataset, we still obtain plausible features when inferring 2D features for vastly different scenes. The 2D rendered features still show a strong correlation with semantically uniform regions, showcasing the generalization of our feature field.

D. Limitations and Future Work

Target features. Our method builds on DINO [11] to obtain target features. While we learn to lift these features into 3D and improve multi-view feature consistency, we cannot improve the discriminative power of the target features *per se*. However, SceneDINO can be trained using arbitrary 2D target features and can profit from future advances in SSL

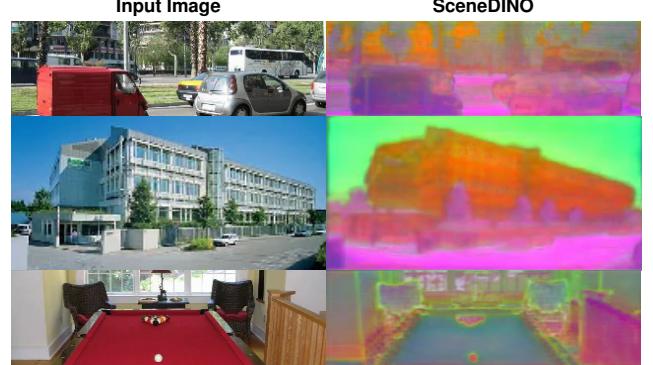


Figure 10. **2D SceneDINO features on out-of-domain images.** We visualize our 2D rendered features (right) given an out-of-domain image (left) from ADE20K [126]. We use the first three principal components for feature visualization. While not trained on such scenes, SceneDINO still produces plausible feature maps.

representation. Note that training SceneDINO requires only 2 days on a single GPU and our training transfers seamlessly to different target features (*e.g.*, DINOv2), thus, utilizing SceneDINO differently is straightforward.

Dynamic objects. Our loss does not model dynamic objects and relies on a static scene assumption. This can potentially cause inaccurate predictions for dynamic classes such as *person* in our experiments. Recent works in depth estimation have explicitly modeled the probability of areas being dynamic [125] and even their motion within the scene [123], which might be extended to SceneDINO.

View sampling and camera poses. For sampling views during training, we rely on the sampling scheme of S4C [37]. This is not directly applicable to other non-driving datasets, where the sampling needs to be tuned. In addition, our approach requires accurate camera poses for each view. We demonstrated that these can be obtained in an unsupervised way for KITTI-360 (*cf.* Tab. 5 & Tab. 12). However, obtaining unsupervised camera poses in more challenging scenarios and conditions is still challenging [120].

Future work. SceneDINO is only trained using a single dataset to be comparable to existing SSC approaches. However, scaling our approach to multiple datasets of more variable scenes could lead to more general feature representations. Ultimately, scaling SceneDINO to internet-scale videos might enable strong zero-shot and cross-domain 3D scene understanding.

References

- [120] Lucas R. Agostinho, Nuno M. Ricardo, Maria I. Pereira, Pinto Antoine, and Andry M. Pinto. A practical survey on visual odometry for autonomous driving in challenging

- scenarios and conditions. *IEEE Access*, 10:72182–72205, 2022. [vi](#)
- [121] Igor Cvišić, Ivan Marković, and Ivan Petrović. SOFT2: Stereo visual odometry for road vehicles based on a point-to-epipolar-line metric. *IEEE Trans. Robot.*, 39(1):273–288, 2023. [vi](#)
- [122] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Z. Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An imperative style, high-performance deep learning library. In *NeurIPS*2019*, pages 8024–8035. [i](#)
- [123] Yihong Sun and Bharath Hariharan. Dynamo-Depth: Fixing unsupervised depth estimation for dynamical scenes. In *NeurIPS*2023*, pages 54987–55005. [vi](#)
- [124] Narayanan Sundaram, Thomas Brox, and Kurt Keutzer. Dense point trajectories by GPU-accelerated large displacement optical flow. In *ECCV*, pages 438—451, 2010. [ii](#)
- [125] Sungmin Woo, Wonjoon Lee, Woo Woo Jin, Dogyoon Lee, and Sangyoun Lee. ProDepth: Boosting self-supervised multi-frame monocular depth with probabilistic fusion. In *ECCV*, pages 201–217, 2024. [vi](#)
- [126] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ADE20K dataset. In *CVPR*, pages 5122–5130, 2017. [vi](#)