Few-Shot Panoptic Segmentation With Foundation Models

Markus Käppeler[1*], Kürsat Petek[1*], Niclas Vödisch[1*], Wolfram Burgard[2], and Abhinav Valada[1]

***Abstract*— Current state-of-the-art methods for panoptic segmentation require an immense amount of annotated training data that is both arduous and expensive to obtain posing a significant challenge for their widespread adoption. Concurrently, recent breakthroughs in visual representation learning have sparked a paradigm shift leading to the advent of large foundation models that can be trained with completely unlabeled images. In this work, we propose to leverage such task-agnostic image features to enable few-shot panoptic segmentation by presenting Segmenting Panoptic Information with Nearly 0 labels (SPINO). In detail, our method combines a DINOv2 backbone with lightweight network heads for semantic segmentation and boundary estimation. We show that our approach, albeit being trained with only ten annotated images, predicts high-quality pseudo-labels that can be used with any existing panoptic segmentation method. Notably, we demonstrate that SPINO achieves competitive results compared to fully supervised baselines while using less than 0.3% of the ground truth labels, paving the way for learning complex visual recognition tasks leveraging foundation models. To illustrate its general applicability, we further deploy SPINO on real-world robotic vision systems for both outdoor and indoor environments. To foster future research, we make the code and trained models publicly available at `http://spino.cs.uni-freiburg.de`.**

I. Introduction

Panoptic segmentation [1] poses an important contribution to holistic scene understanding by enabling robots to assign semantic meaning to their environment while delineating individual objects. However, most previous methods addressing panoptic segmentation rely on supervised training [2], [3], hence requiring a large amount of ground truth labels. This hinders their widespread adoption as generating panoptic annotations is both expensive and time-consuming, e.g., manually labeling a single high-resolution image of urban scenarios takes approximately $1.5\,\mathrm{h}$ [4]. Therefore, it is paramount to reduce the number of required labels [5], e.g., by advancing weakly- and unsupervised methods or by leveraging task-agnostic pretraining strategies [6].

Facing similar issues, the domain of natural language processing (NLP) has recently seen a rise of large foundation models [7]. This paradigm shift in NLP also inspired the vision community to propose similar methods such as CLIP [8] or Segment Anything [9]. While both still require some supervision signal, e.g., from image captions or coarse object masks, DINO [10] learns visual representation in a fully unsupervised manner allowing to significantly extend the amount of usable resources. Prior works have shown that one can bootstrap such general representations for several tasks including depth estimation [11], semantic segmentation [11], [12], and object detection [13]. Based on these findings, we argue that it is time for a fundamental paradigm switch for vision tasks that exploit task-agnostic foundation models to enable few-shot training. In contrast to unsupervised techniques [12], [14], we show that such an approach can yield results competitive with fully supervised learning methods.

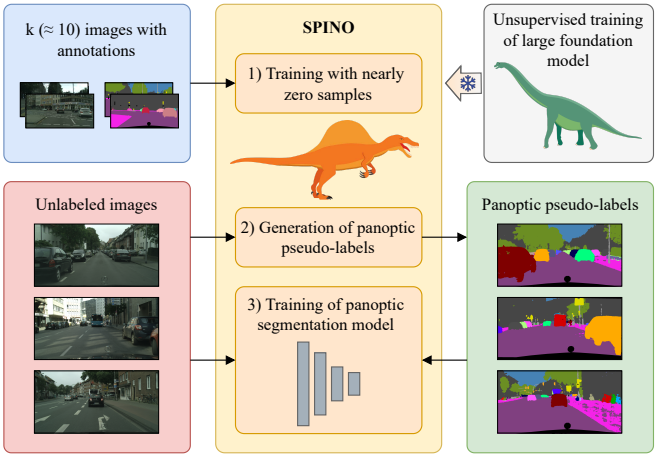

Fig. 1. SPINO enables few-shot panoptic segmentation by exploiting descriptive image features from unsupervised task-agnostic pretraining. We generate panoptic pseudo-labels by learning from only $k \approx 10$ annotated images in an offline manner. We can then leverage these pseudo-labels to train any panoptic segmentation model enabling online deployment.

In this work, we present a method for *Segmenting Panoptic Information with Nearly 0 labels* (SPINO), illustrated in Fig. 1. First, we leverage a frozen DINOv2 [11] backbone to extract visual features. Subsequently, we train two task-specific heads for semantic segmentation and boundary estimation with as few as ten annotated images to perform few-shot panoptic segmentation. To enable real-time inference and to further boost the quality of our predictions, we generate panoptic pseudo-labels in an offline manner for a larger bag of raw images that can then be used to train any existing panoptic segmentation model. We perform extensive evaluations on several public [4], [15] and in-house datasets that demonstrate that our SPINO approach yields results that are highly competitive with fully supervised learning models. In particular, our extensive evaluations suggest that few-shot panoptic segmentation provides the means to soon become on par with supervised state-of-the-art methods.

To summarize, the main contributions are as follows:

1) We propose the first method for few-shot panoptic segmentation based on unsupervised foundation models.

* Equal contribution.
[1] Department of Computer Science, University of Freiburg, Germany.
[2] Department of Eng., University of Technology Nuremberg, Germany.
This work was funded by the German Research Foundation (DFG) Emmy Noether Program grant No 468878300 and the European Union's Horizon 2020 research and innovation program grant No 871449-OpenDR.
Accepted for the 2024 IEEE Int. Conf. on Robotics and Automation.

2) We present a novel pseudo-label generation scheme that can be trained with as few as ten annotated images.
3) We show that SPINO yields results that are competitive to supervised training with ground truth labels.
4) In extensive evaluations, we illustrate the effect of various architectural design choices and apply our method to real-world robotic vision platforms.
5) We make the code and trained models publicly available at `http://spino.cs.uni-freiburg.de`.

II. Related Work

In this section, we present an overview of panoptic segmentation, visual representation learning, and both unsupervised and weakly-supervised image segmentation techniques.

Panoptic Segmentation: Panoptic segmentation [1] combines semantic and instance segmentation into a single task with two categories of scene elements. The static background comprises the so-called "stuff" classes such as *buildings*, whereas dynamic objects such as *cars* belong to the "thing" category. While "stuff" classes only receive a semantic label, "thing" classes are further separated on an instance level. Since the introduction of this task, several deep learning-based methods [2], [16]–[19] have been proposed requiring a large amount of data for training. Recently, the focus has shifted towards more challenging variants, e.g., open-vocabulary methods such as from Ding *et al.* [20] leveraging insights from foundation models [8]. Removing the need for labels, CoDEPS [21] addresses unsupervised domain adaptation from a source to a previously unseen target domain. In this work, we propose a method for few-shot panoptic segmentation requiring as few as ten annotated images.

Visual Representation Learning: Breakthroughs in natural language processing (NLP) [7] have shown that task-agnostic pretraining can yield feature representations that, fine-tuned to specific applications, become competitive with prior state-of-the-art methods [22]. A common approach to obtaining similar representations in the visual domain is contrastive learning [23]. However, although not using human annotations, the choice of the dataset still introduces a significant bias on the learned representation that can be mitigated by extensive data augmentation [24]. Masked autoencoders (MAE) [25] represent another type of self-supervised learners that learn to reconstruct areas in an image that have been masked. After pretraining, MAEs can be fine-tuned for various downstream tasks. More recently, the usage of foundation models in NLP has also started to influence computer vision. For instance, CLIP [8] leverages insights from constrastive learning by exploiting textual supervision to guide the learning of visual features. However, this text-guided supervision strategy limits the choice of training data. SAM [9] removes the need for captions and relies on a self-iterative training scheme starting from coarse object masks. While showing impressive zero-shot performance for semantic segmentation on unseen domains, it lacks the ability to assign class labels to the segments. Finally, DINO [10] represents a new family of foundation models that can be trained only from raw images. In particular, DINO demonstrates that such unsupervised pretraining can achieve even more explicit features for semantic segmentation than their supervised counterparts. Further advances have been shown by DINOv2 [11] that combines several prior insights with training on a curated dataset. In this work, we exploit descriptive image features from a DINOv2 backbone to generate panoptic pseudo-labels.

Unsupervised and Weakly-Supervised Segmentation: Since obtaining pixel-wise annotations for supervised training of image segmentation tasks is expensive, in the last few years research has shifted towards reducing the number of human annotations. Recent methods build on the observation that features from unsupervised pretraining are semantically consistent across images from differing domains [12]. For instance, LOST [26] uses DINO [10] features for bounding box extraction to bootstrap supervised training of an object detector. Objects can be assigned to the same class via k-means clustering in the feature space. Similarly, TokenCut [27] relies on Normalized Cut (NCut) [28] to group self-similar image regions based on DINO features. While these previous methods work well for foreground/background segmentation, FreeSOLO [29] addresses multi-object detection by enhancing coarse masks via one-stage self-training in a weakly supervised manner. However, requiring in-domain data results in a lack of generalization. In contrast, CutLER [13] achieves impressive zero-shot performance leveraging DINO features to generate coarse masks followed by weakly supervised training of a separate instance segmentation network. Although applicable to multi-object scenarios, relying on iterative NCut requires specifying the number of expected objects.

With respect to semantic segmentation, MaskContrast [30] and PiCIE [14] are notable methods from before the advent of large pretraining models. While MaskContrast contrasts learned features within and across saliency masks, PiCIE searches for descriptive image features guided by photometric invariance and geometric equivariance. Recently, both MaskDistill [31] and STEGO [12] leverage features from a frozen DINO [10] backbone. To further refine the pretrained features, STEGO adds a task-specific segmentation head followed by clustering. Other examples of exploiting foundation models include CLIP-ES [32], which relies on contrastive language-image pretraining [8], and SEPL [33] that combines the class-agnostic masks from SAM [9] with class activation maps for class assignment. To the best of our knowledge, our proposed SPINO constitutes the first attempt to directly exploit fully unsupervised representation pretraining for panoptic segmentation.

III. Technical Approach

In this section, we present our proposed approach SPINO for few-shot panoptic segmentation. As illustrated in Fig. 2, we leverage the recent foundation model DINOv2 [11] to extract descriptive image features for both semantic segmentation and boundary estimation. In particular, we propose a novel pseudo-label generation scheme that separates semantic regions of "thing" classes into individual instances

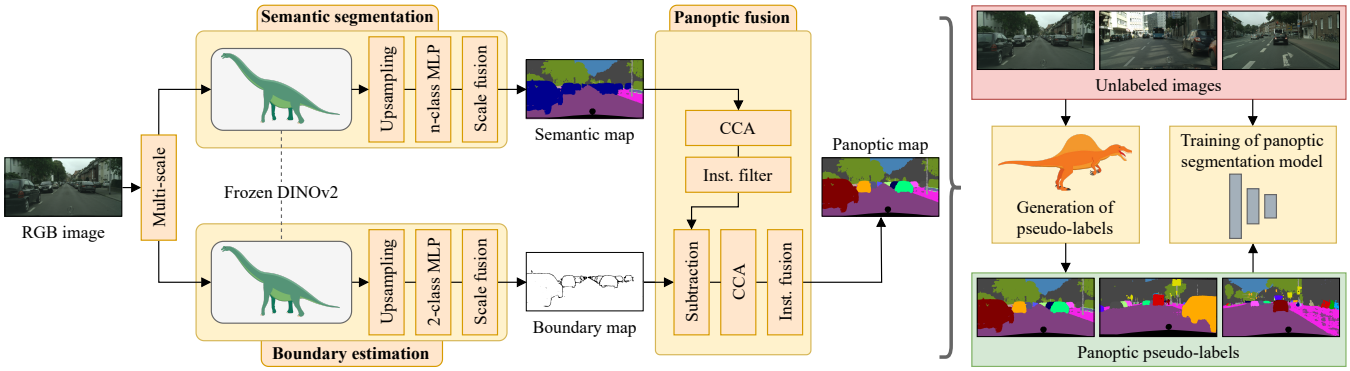

Fig. 2. Overview of our proposed SPINO approach for few-shot panoptic segmentation. SPINO consists of two learning-based modules for semantic segmentation and boundary estimation that leverage features from the recent foundation model DINOv2 [11]. A panoptic fusion scheme combines their outputs using connected component analysis (CCA) and multiple small instance filtering steps. SPINO creates pseudo-labels for a large number of unlabeled images using only $k \approx 10$ images with ground truth annotations. These pseudo-labels can then be utilized to train any panoptic segmentation model.

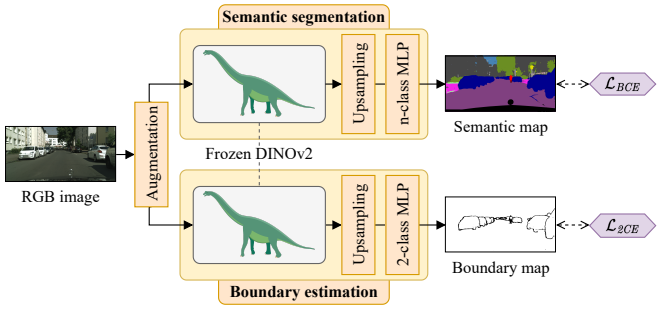

Fig. 3. Our proposed pseudo-label generator comprises two learnable modules for semantic segmentation and boundary estimation that exploit descriptive image features from the recent DINOv2 [11] foundation model, enabling training with only $k \approx 10$ ground truth panoptic annotations.

by predicting object boundaries. With this approach, SPINO can bootstrap very few ground truth annotations for generating high-quality panoptic pseudo-labels. To enable real-time inference and to further boost the quality of our panoptic predictions, we train a panoptic segmentation model using the generated pseudo-labels.

A. Few-Shot Pseudo-Label Generation

We propose a novel panoptic segmentation scheme to generate panoptic pseudo-labels in an offline manner while requiring very few ground truth annotations for training. Our label generator consists of three main building blocks shown in Fig. 2, namely learnable modules for semantic segmentation and boundary estimation as well as a static component to fuse their predictions. The semantic segmentation module is comprised of a frozen DINOv2 [11] backend, a bilinear 14x-upsampling layer, and a final n-class MLP with 4 layers. Here, n denotes the number of semantic classes as specified in Sec. IV-A. In detail, we use the DINOv2 weights of the ViT-B/14 variant provided by the authors. For the boundary estimation module, we employ a similar design but use 4x-upsampling and set $n = 2$ for binary classification.

Training the Label Generator: A key idea of SPINO is to train our proposed pseudo-label generator with only k ground truth annotations, where k denotes numbers as small as 10. Notably, the unsupervised training procedure of DINOv2 does not further increase this number even when considering the pretraining. We illustrate the training of our pseudo-label generator in Fig. 3. First, to stabilize the training with such few samples, we employ various data augmentation techniques on the input RGB image including random cropping, horizontal flipping, and color jitter. Subsequently, we feed the augmented image to the two task-specific heads and compute the respective loss functions.

We supervise the semantic segmentation head with the bootstrapped cross-entropy loss function $\mathcal{L}_{BCE}$ [34] to account for rare classes.

$$\mathcal{L}_{BCE} = -\frac{1}{K}\sum_{i=1}^{N} \mathbb{1}\left[p_{i,y_i} < t_K\right] \cdot \log(p_{i,y_i}), \qquad (1)$$

where p_{i,y_i} denotes the posterior probability of pixel $i \in [1, N]$ for its ground truth class $y_i \in \{1, ..., c\}$ with N and c being the number of pixels and classes, respectively. The indicator function $\mathbb{1}(\cdot)$ is 1 if p_{i,y_i} is below a threshold t_K and 0 otherwise. Following previous works [2], [21], we set $t_K = 0.2$ such that only those pixels with top-K highest losses contribute to $\mathcal{L}_{BCE}$. In order to train the boundary estimation module, we generate ground truth boundary maps as follows: If the instance ID of a pixel is different from any of its eight neighbors, we assign 1 to this pixel. Otherwise, we set the value of the center pixel to 0. During training, we compute the binary cross-entropy loss $\mathcal{L}_{2CE}$ as the supervision signal.

$$\mathcal{L}_{2CE} = -\frac{1}{N}\sum_{i=1}^{N} y_i \cdot \log(p_i) + (1 - y_i) \cdot \log(1 - p_i), \qquad (2)$$

where $y_i \in \{0, 1\}$ is the binary boundary label of pixel i and p_i denotes the probability of the pixel i being a boundary.

Employing the Label Generator: In the next step, we leverage the aforementioned trained modules for semantic segmentation and boundary estimation to generate panoptic pseudo-labels for a large number of unlabeled images. In the following, we describe the procedure as depicted in Fig. 2. Inspired by ensemble learning, we use multi-scale test-time augmentation for both semantic segmentation and boundary

estimation. For instance, for scale $s = 2$, we divide the image into four equally sized regions, upsample each region to the size of the original input image ($s = 1$), and obtain their softmax features. In the scale fusion block, we downsample these feature maps to the original size of the region, join the features of all regions in a single $s = 1$ map, and compute the mean across the considered scales. In detail, we use scales $\{1, 2, 3\}$ for the semantic head and scales $\{3, 4, 5\}$ for the boundary estimation head. Next, we feed the predicted semantic map and the estimated object boundary map to our panoptic fusion module. First, for each "thing" class, we perform connected component analysis (CCA) yielding disconnected blobs. If a blob consists of fewer pixels than a threshold, we assign the semantic *void* class to its pixels. Otherwise, we subtract the predicted border for this blob from the semantic map followed by CCA to detect separate instances within a blob. If the number of pixels of an instance is below another threshold, we add it to its nearest neighbor which fulfills the minimum size requirement. If all instances of a blob are below this threshold, we combine them into a single instance. Finally, due to the top-down approach, the inferred instance maps already contain semantic information leading to the desired pseudo-labels for panoptic segmentation.

B. Training a Panoptic Segmentation Model

After creating pseudo-labels for a large set of unlabeled images, we train a panoptic segmentation model as illustrated in Fig. 2. In contrast to the offline label generator, such a model allows for online panoptic segmentation while further enhancing the overall performance. Although this approach is generally applicable to any panoptic segmentation model, in this work, we follow the spirit of our pseudo-label generator. In detail, our bottom-up panoptic segmentation network consists of a frozen DINOv2 [11] backbone with an adapter module [35] and three task-specific heads [2] for semantic segmentation, instance center prediction, and pixel offset regression, respectively. In Fig. 4, we visualize this architecture. The semantic head predicts a semantic class for each pixel and is trained with the bootstrapped cross-entropy loss with hard pixel mining [2].

$$\mathcal{L}_{BCEH} = -\frac{1}{K}\sum_{i=1}^{N} w_i \cdot \mathbb{1}\left[p_{i,y_i} < t_K\right] \cdot \log(p_{i,y_i}), \quad (3)$$

which builds upon Eq. (1) but adds weights $w_i > 1$ for pixels that belong to small instances. For other instances and "stuff" classes, the pixel weight remains at $w_i = 1$. Similar to Eq. (1), we set $t_K = 0.2$. Addressing instance segmentation, the center head generates a probability map with high values for instance centers and the offset head estimates the 2D offset of a pixel to the nearest instance center. To train these heads, we utilize the MSE loss $\mathcal{L}_{MSE}$ for the center head and the L1 loss $\mathcal{L}_{L1}$ for the offset head. Consequently, we compute the total loss as a weighted sum:

$$\mathcal{L}_{PAN} = \lambda_{sem}\mathcal{L}_{BCEH} + \lambda_{cen}\mathcal{L}_{MSE} + \lambda_{off}\mathcal{L}_{L1} \quad (4)$$

To increase the learning speed, we propose to further exploit the k annotated images, which were used to train the pseudo-label generator, also when training the panoptic segmentation model. In particular, we construct batches that contain both pseudo-labels and one ground truth sample. Formally, a batch $\mathbf{b}$ of size n is given by

$$\mathbf{b} = \{\hat{\mathbf{I}}_1, \ldots, \hat{\mathbf{I}}_{n-1}, \mathbf{I}_{GT}\}, \quad (5)$$

where $\hat{\mathbf{I}}_i$ denote pseudo-labeled images and $\mathbf{I}_{GT}$ is from the set of k images with ground truth labels. We further apply data augmentation via color jitter and horizontal flipping.

During test-time, a panoptic fusion module [2] predicts the final panoptic segmentation map from the output of the individual heads, shown in Fig. 4. In detail, it assigns a semantic label to the class-agnostic instance predictions using majority voting over the semantic predictions of all pixels within an instance.

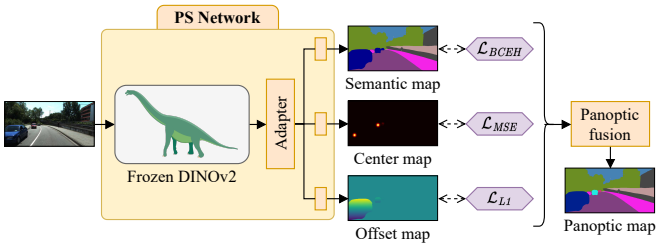

Fig. 4. To enable online predictions and to further boost the performance compared to the pseudo-label generator, we train a bottom-up panoptic segmentation model using our generated pseudo-labels. The network consists of a frozen DINOv2 [11] backbone with an adapter [35] and three task-specific heads, whose output is merged by a panoptic fusion module [2].

IV. Experimental Evaluation

In this section, we demonstrate that our proposed SPINO outperforms unsupervised methods for semantic segmentation and yields competitive results compared to fully supervised setups for panoptic segmentation that require a huge number of ground truth annotations. We provide both quantitative and qualitative results on multiple public and in-house datasets. Finally, we extensively evaluate several design choices for our pseudo-label generator.

A. Datasets

We present results on various datasets including the public Cityscapes [4] and KITTI-360 [15] as well as our in-house data for automated driving and from an indoor office environment.

Cityscapes: The Cityscapes dataset [4] contains RGB images and fine panoptic annotations for automated driving in 50 cities across Germany and bordering regions. We select k images from the *train* split to train our label generator and generate pseudo-labels for the remaining images. In a separate experiment, we also generate pseudo-labels on the entire *train_extra* split. To evaluate the performance, we report metrics on the *val* split. When creating the pseudo-labels, we mask out the hood of the ego car as it remains static and hence can be inferred from the k annotated images [5]. We report metrics using 19 classes as per the official Cityscapes evaluation protocol.

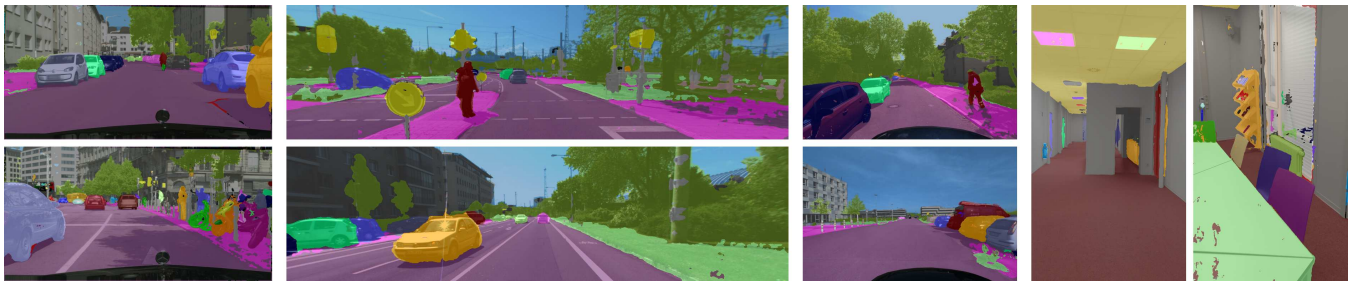

Fig. 5. Qualitative performance of our pseudo-label generator in four diverse domains from both public and in-house data sources. From left to right, we show two samples each for Cityscapes [4], KITTI-360 [15], in-house automated driving, and an in-house office environment.

TABLE I
PANOPTIC/SEMANTIC SEGMENTATION ON CITYSCAPES

Method	**Train. data**	Acc	mIoU	PQ	SQ	RQ
Fully supervised						
DINOv2 + Adapt. + PH	GT	91.9	77.0	51.4	78.9	63.1
Unsupervised						
Modified DC [38]	n/a	35.3	6.8	–	–	–
PiCIE [14]	n/a	72.7	13.8	–	–	–
STEGO [12]	n/a	89.1	38.0	–	–	–
Few-shot supervision						
ResNet-50 + PH	10 GT	74.9	32.1	16.8	45.6	20.8
DINOv2 + PH	10 GT	81.6	49.4	20.6	49.9	25.8
DINOv2 + Adapt. + PH	10 GT	82.8	52.5	22.0	60.9	27.0
Pseudo-labels (*ours*)	10 GT	86.0	61.5	35.9	73.7	45.9
SPINO (*ours*)	PL	86.3	60.6	36.4	73.5	46.7
+ Mixed-batch	PL	86.6	61.2	36.5	74.8	46.3
SPINO (*ours*)	PL++	86.6	61.8	37.2	74.5	47.5

PH refers to the panoptic heads as shown in Fig. 4. GT and PL indicate training with ground truth annotations and pseudo-labels, where the "PL++" marks pseudo-labels on the *train_extra* split. The architecture of SPINO corresponds to "DINOv2 + Adapt. + PH".

KITTI-360: The KITTI-360 dataset [15] was recorded in Karlsruhe, Germany, and provides RGB images and panoptic annotations for sequential data. Following prior works [36], [37], we use sequence 10 for evaluation and the remaining sequences for the pseudo-label generation. We report results using 14 classes as detailed by Vödisch *et al.* [21].

In-House: To illustrate the main benefit of SPINO, i.e., enabling panoptic segmentation on different vision systems with very few reference annotations, we employ our method on two in-house data sources. First, following the spirit of the public datasets, we use an automated driving perception car navigating in Freiburg, Germany. Second, to demonstrate general applicability, we record indoor data in our office environment. For both domains, we prepare annotations for ten images to train the pseudo-label generator.

B. Panoptic Segmentation

To evaluate the performance of SPINO, we measure the pixel accuracy (Acc) and the mean IoU (mIoU) for semantic segmentation as well as the panoptic quality (PQ), the segmentation quality (SQ), and the recognition quality (RQ) for panoptic segmentation. Based on the ablation studies in Sec. IV-C, we train our pseudo-label generator on $k = 10$ human-selected, labeled images with a batch size $b = 1$ and a learning rate $lr = 0.001$.

Few-Shot Training: First, we illustrate the efficacy of our pseudo-label generation scheme. As shown by the metrics in Tab. I, training Panoptic-DeepLab [2] (with a ResNet-50 backbone) on only ten images yields poor results that can be improved by replacing the backbone with a frozen DINOv2 [11]. Following the common methodology for dense prediction tasks, we also add an adapter module [35] to further increase the performance. However, the results remain significantly inferior to the quality of our pseudo-labels with respect to both semantic and panoptic segmentation. Notably, our pseudo-label generator comprises a much simpler design, e.g., estimating object boundaries instead of predicting instance centers and pixel offsets. For the overall SPINO approach, we adopt the network design of DINOv2 plus an adapter module. Naive training on the generated pseudo-labels already yields highly competitive results compared to training with ground truth labels considering that we use less than $0.29\,\%$ of the labels. We further show how the proposed mixed-batch strategy that closely incorporates the ten ground truth labels increases all three semantic metrics.

Next, we also generate pseudo-labels for the unlabeled *train_extra* split of Cityscapes, increasing the amount of training data for the panoptic segmentation model. The results in Tab. I indicate that our approach opens up an avenue for exploiting unlabeled large-scale data recordings for the training of existing panoptic segmentation methods.

Comparison with Unsupervised Segmentation: Second, we compare SPINO to the state-of-the-art for unsupervised semantic segmentation. As we follow the official Cityscapes evaluation protocol, we retrain PiCIE [14] and their modified DeepCluster [14], [38] using the released code on 19 classes. For STEGO [12], we use the provided network weights but reevaluate on 19 classes. Note that, for both PiCIE and STEGO, reducing the number of classes leads to higher metrics than reported by the authors. As SPINO significantly outperforms these baselines, we argue that requiring ten instead of zero annotated images is well justified.

Generalizability: Finally, we extend the evaluation to multiple datasets. In Tab. II, we report quantitative results on both Cityscapes [4] and KITTI-360 [15]. In detail, we compare supervised training with ground truth annotations to our few-shot approach. Similar to Tab. I, we report results for three backbones, namely ResNet-50 [39], DINOv2 [11], and

TABLE II

PANOPTIC SEGMENTATION ON CITYSCAPES AND KITTI-360

Method	Train. data	Cityscapes Acc	Cityscapes mIoU	Cityscapes PQ	Cityscapes SQ	Cityscapes RQ	KITTI-360 Acc	KITTI-360 mIoU	KITTI-360 PQ	KITTI-360 SQ	KITTI-360 RQ
Pseudo-labels	10 GT	86.0	61.5	35.9	73.7	45.9	75.8	54.7	32.5	70.7	42.1
ResNet-50 + PH	GT	89.4	64.9	44.2	75.3	56.1	83.0	64.1	41.0	76.5	50.5
DINOv2 + PH	GT	89.4	71.4	41.0	74.4	51.7	83.5	62.8	39.3	70.5	48.7
DINOv2 + Adapt. + PH	GT	91.9	77.0	51.4	78.9	63.1	86.0	65.6	42.5	72.9	51.2
ResNet-50 + PH	PL	85.4	57.3	33.0	67.8	42.3	76.2	52.1	32.2	67.6	41.0
DINOv2 + PH	PL	84.5	57.1	31.4	70.9	40.3	76.4	54.6	32.7	71.7	42.0
DINOv2 + Adapt. + PH	PL	86.3	60.6	36.4	73.5	46.7	76.6	55.5	33.3	71.9	42.8

PH refers to the panoptic heads shown in Fig. 4. GT and PL indicate ground truth annotations and pseudo-labels. The gray row corresponds to SPINO without mixed-batch training.

TABLE III

ABLATION STUDY: NETWORK ARCHITECTURE

Method	A: k-NN	B: Lin. Layer	C: CNN	D: MLP	E: Upsamling	Acc	mIoU	PQ	SQ	RQ
A	✓					78.7	51.7	26.1	68.5	35.0
B		✓				84.3	60.0	32.6	71.3	42.5
B + E		✓			✓	84.3	60.0	33.6	71.7	43.8
C + E			✓		✓	82.9	55.1	29.7	70.9	38.4
D + E				✓	✓	**86.0**	**61.5**	**35.9**	**73.7**	**45.9**

Due to the high computational complexity, the k-NN is evaluated without training data augmentation.

TABLE IV

ABLATION STUDY: DATA AUGMENTATION

Method	Acc	mIoU	PQ	SQ	RQ
Base	83.2	55.8	29.5	70.8	38.0
Training time					
+ Random flip	83.3	56.1	29.5	70.8	38.0
+ Random crop	83.0	57.2	30.0	70.7	39.1
+ Color jitter	83.1	57.3	30.1	70.9	39.1
Test time					
+ Multi-scale ensemble	**86.0**	**61.5**	**35.9**	**73.7**	**45.9**

TABLE V

ABLATION STUDY: BATCH SIZE

Batch size	Acc	mIoU	PQ	SQ	RQ
1	**86.0**	**61.5**	**35.9**	**73.7**	**45.9**
2	84.9	59.8	34.0	72.6	43.7
4	85.3	59.4	33.6	72.3	43.3
8	84.5	56.8	31.3	71.4	39.9

TABLE VI

ABLATION STUDY: NUMBER OF LABELS

Label count	Acc	mIoU	PQ	SQ	RQ
1	69.8	37.1	19.8	55.4	27.2
3	81.8	49.3	30.3	64.3	38.8
5	82.8	55.0	32.1	65.5	41.3
10	86.0	61.5	35.9	73.7	45.9
25	88.5	66.9	39.6	74.9	50.1
50	89.4	69.1	40.9	74.8	51.6
100	90.3	71.3	42.9	76.3	53.8

DINOv2 with an adapter [35]. Considering that our pseudo-labels are generated based on only ten images, the few-shot methods yield impressive results across the board. Note that ten images correspond to $0.29\,\%$ and $0.02\,\%$ of the utilized ground truth labels for Cityscapes and KITTI-360, respectively. Finally, we provide qualitative visualizations of our pseudo-labels in Fig. 5 for both public datasets as well as our in-house data including outdoor urban and indoor office environments. Further examples are shown in the supplementary video on the project website.

C. Ablation Studies of Pseudo-Label Generation

We extensively evaluate the architectural design of our pseudo-label generator and demonstrate its efficacy in contrast to several alternatives. In Tabs. III, IV, V and VI, we highlight the utilized variant in gray.

Network Architecture: In Tab. III, we compare the architectural design of our pseudo-label generator using MLPs to other network architectures. Similar to other methods [10], [26], we use a k-NN classifier on the DINOv2 feature patches with $k = 5$. Due to the high computational complexity of this approach, we omit training data augmentation for the k-NN. Next, we utilize a linear layer with and without prior upsampling. Compared to the k-NN, these learnable methods yield a significant improvement but remain inferior to the MLPs. Finally, we demonstrate that our design also outperforms a 4-layer CNN with 3×3 convolutions.

Data Augmentation: Next, we gradually activate the data augmentation techniques and list the results in Tab. IV. Utilizing data augmentation during the training enhances mIoU, PQ, and RQ, whereas the accuracy and SQ remain stable. Additionally, our employed test-time augmentation based on multi-scale ensemble prediction vastly improves the metrics across the board.

Batch Size: In Tab. V, provide results for various batch sizes. Note that we scale the learning rate proportionally to the batch size and keep the number of epochs constant. Due to leading to the highest quality of the pseudo-labels, we select a batch size $b = 1$.

Number of Ground Truth Labels: Finally, we investigate the effect of the label count on the quality of the pseudo-labels. In Tab. VI, we report results for increasing k from one-shot to $k = 100$. Note that for up to $k = 10$, we manually select the samples used for training. For $k > 10$, we randomly add further data. We observe a continuous improvement for greater k. Notably, for $k = 100$, our pseudo-label generator is almost on par with Panoptic-DeepLab while using $2.9\,\%$ of the annoations (see ResNet-50 backbone in Tab. II).

V. CONCLUSION

In this work, we introduced SPINO for few-shot panoptic segmentation by exploiting descriptive image representations from the unsupervised foundation model DINOv2. We demonstrated that SPINO can generate high-qualitative pseudo-labels after being trained on as little as ten annotated images. These pseudo-labels can then be used to train any existing panoptic segmentation method yielding results that are highly competitive to fully supervised learning approaches relying on human annotations. Finally, we extensively evaluated several design choices for the proposed pseudo-label generator and employed our SPINO approach to both public and in-house data. To facilitate further research, we made our code publicly available. In the future, we will further enhance the instance separation by refining the boundary estimation and employ SPINO in additional domains.

REFERENCES

[1] A. Kirillov, K. He, R. Girshick, C. Rother, and P. Dollár, "Panoptic segmentation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9396–9405.

[2] B. Cheng, M. D. Collins, Y. Zhu, T. Liu, T. S. Huang, H. Adam, and L.-C. Chen, "Panoptic-DeepLab: A simple, strong, and fast baseline for bottom-up panoptic segmentation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 12 472–12 482.

[3] R. Mohan and A. Valada, "Perceiving the invisible: Proposal-free amodal panoptic segmentation," *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 9302–9309, 2022.

[4] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The Cityscapes dataset for semantic urban scene understanding," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3213–3223.

[5] L.-C. Chen, R. G. Lopes, B. Cheng, M. D. Collins, E. D. Cubuk, B. Zoph, H. Adam, and J. Shlens, "Naive-Student: Leveraging semi-supervised learning in video sequences for urban scene segmentation," in *European Conference on Computer Vision*, 2020, pp. 695–714.

[6] C. Lang, A. Braun, L. Schillingmann, K. Haug, and A. Valada, "Self-supervised representation learning from temporal ordering of automated driving sequences," *IEEE Robotics and Automation Letters*, vol. 9, no. 3, pp. 2582–2589, 2024.

[7] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, *et al.*, "Llama 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:2307.09288*, 2023.

[8] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Conference on Robot Learning*, vol. 139, 2021, pp. 8748–8763.

[9] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick, "Segment anything," *arXiv preprint arXiv:2304.02643*, 2023.

[10] M. Caron, H. Touvron, I. Misra, H. Jegou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," in *International Conference on Computer Vision*, 2021, pp. 9630–9640.

[11] M. Oquab, T. Darcet, T. Moutakanni, H. V. Vo, M. Szafraniec, *et al.*, "DINOv2: Learning robust visual features without supervision," *arXiv preprint arXiv:2304.07193*, 2023.

[12] M. Hamilton, Z. Zhang, B. Hariharan, N. Snavely, and W. T. Freeman, "Unsupervised semantic segmentation by distilling feature correspondences," in *International Conference on Learning Representations*, 2022.

[13] X. Wang, R. Girdhar, S. X. Yu, and I. Misra, "Cut and learn for unsupervised object detection and instance segmentation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 3124–3134.

[14] J. Hyun Cho, U. Mall, K. Bala, and B. Hariharan, "PiCIE: Unsupervised semantic segmentation using invariance and equivariance in clustering," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 16 789–16 799.

[15] Y. Liao, J. Xie, and A. Geiger, "KITTI-360: A novel dataset and benchmarks for urban scene understanding in 2D and 3D," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 3, pp. 3292–3310, 2023.

[16] Y. Xiong, R. Liao, H. Zhao, R. Hu, M. Bai, E. Yumer, and R. Urtasun, "UPSNet: A unified panoptic segmentation network," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8818–8826.

[17] L. Porzi, S. R. Bulo, A. Colovic, and P. Kontschieder, "Seamless scene segmentation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8277–8286.

[18] R. Mohan and A. Valada, "EfficientPS: Efficient panoptic segmentation," *International Journal of Computer Vision*, vol. 129, pp. 1551 – 1579, 2020.

[19] H. Wang, Y. Zhu, B. Green, H. Adam, A. Yuille, and L.-C. Chen, "Axial-DeepLab: Stand-alone axial-attention for panoptic segmentation," in *European Conference on Computer Vision*, 2020, pp. 108–126.

[20] Z. T. Zheng Ding, Jieke Wang, "Open-vocabulary universal image segmentation with maskclip," in *International Conference on Machine Learning*, 2023.

[21] N. Vödisch, K. Petek, W. Burgard, and A. Valada, "CoDEPS: Online continual learning for depth estimation and panoptic segmentation," *Robotics: Science and Systems*, 2023.

[22] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, *et al.*, "Language models are few-shot learners," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 1877–1901.

[23] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9726–9735.

[24] W. Van Gansbeke, S. Vandenhende, S. Georgoulis, and L. V. Gool, "Revisiting contrastive methods for unsupervised learning of visual representations," in *Advances in Neural Information Processing Systems*, vol. 34, 2021, pp. 16 238–16 250.

[25] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 15 979–15 988.

[26] O. Siméoni, G. Puy, H. V. Vo, S. Roburin, S. Gidaris, A. Bursuc, P. Pérez, R. Marlet, and J. Ponce, "Localizing objects with self-supervised transformers and no labels," *British Machine Vision Conference*, 2021.

[27] Y. Wang, X. Shen, Y. Yuan, Y. Du, M. Li, S. X. Hu, J. L. Crowley, and D. Vaufreydaz, "TokenCut: Segmenting objects in images and videos with self-supervised transformer and normalized cut," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–13, 2023.

[28] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 888–905, 2000.

[29] X. Wang, Z. Yu, S. De Mello, J. Kautz, A. Anandkumar, C. Shen, and J. M. Alvarez, "FreeSOLO: Learning to segment objects without annotations," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 14 156–14 166.

[30] W. Van Gansbeke, S. Vandenhende, S. Georgoulis, and L. Van Gool, "Unsupervised semantic segmentation by contrasting object mask proposals," in *International Conference on Computer Vision*, 2021, pp. 10 032–10 042.

[31] W. V. Gansbeke, S. Vandenhende, and L. V. Gool, "Discovering object masks with transformers for unsupervised semantic segmentation," *arXiv preprint arXiv:2206.06363*, 2022.

[32] Y. Lin, M. Chen, W. Wang, B. Wu, K. Li, B. Lin, H. Liu, and X. He, "CLIP is also an efficient segmenter: A text-driven approach for weakly supervised semantic segmentation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 15 305–15 314.

[33] T. Chen, Z. Mai, R. Li, and W. lun Chao, "Segment anything model (sam) enhanced pseudo labels for weakly supervised semantic segmentation," *arXiv preprint arXiv:2305.05803*, 2023.

[34] T. Pohlen, A. Hermans, M. Mathias, and B. Leibe, "Full-resolution residual networks for semantic segmentation in street scenes," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3309–3318.

[35] Z. Chen, Y. Duan, W. Wang, J. He, T. Lu, J. Dai, and Y. Qiao, "Vision transformer adapter for dense predictions," in *International Conference on Learning Representations*, 2023.

[36] R. Mohan and A. Valada, "Amodal panoptic segmentation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 20 991–21 000.

[37] N. Gosala, K. Petek, P. L. Drews-Jr, W. Burgard, and A. Valada, "SkyEye: Self-supervised bird's-eye-view semantic mapping using monocular frontal view images," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 14 901–14 910.

[38] M. Caron, P. Bojanowski, A. Joulin, and M. Douze, "Deep clustering for unsupervised learning of visual features," in *European Conference on Computer Vision*, 2018, pp. 132–149.

[39] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.