Drive&Segment: Unsupervised Semantic Segmentation of Urban Scenes via Cross-modal Distillation

Antonin Vobecky[1,2], David Hurych[2], Oriane Siméoni[2], Spyros Gidaris[2], Andrei Bursuc[2], Patrick Pérez[2], and Josef Sivic[1]

[1] Czech Institute of Informatics, Robotics and Cybernetics, Czech Technical University in Prague
[2] valeo.ai
https://vobecant.github.io/DriveAndSegment

Abstract. This work investigates learning pixel-wise semantic image segmentation in urban scenes without any manual annotation, just from the raw non-curated data collected by cars which, equipped with cameras and LiDAR sensors, drive around a city. Our contributions are threefold. First, we propose a novel method for cross-modal unsupervised learning of semantic image segmentation by leveraging synchronized LiDAR and image data. The key ingredient of our method is the use of an object proposal module that analyzes the LiDAR point cloud to obtain proposals for spatially consistent objects. Second, we show that these 3D object proposals can be aligned with the input images and reliably clustered into semantically meaningful pseudo-classes. Finally, we develop a cross-modal distillation approach that leverages image data partially annotated with the resulting pseudo-classes to train a transformer-based model for image semantic segmentation. We show the generalization capabilities of our method by testing on four different testing datasets (Cityscapes, Dark Zurich, Nighttime Driving and ACDC) without any finetuning, and demonstrate significant improvements compared to the current state of the art on this problem.

Keywords: autonomous driving · unsupervised semantic segmentation

1 Introduction

In this work, we investigate whether it is possible to learn pixel-wise semantic image segmentation of urban scenes without the need for any manual annotation, just from the raw non-curated data collected by cars equipped with cameras and LiDAR sensors while driving in town. This topic is important as current methods require large amounts of pixel-wise annotations over various driving conditions and situations. Such a manual segmentation of images on a large scale is very expensive, time-consuming, and prone to biases.

Currently, the best methods for unsupervised learning of semantic segmentation assume that images contain centered objects [50] rather than full scenes or

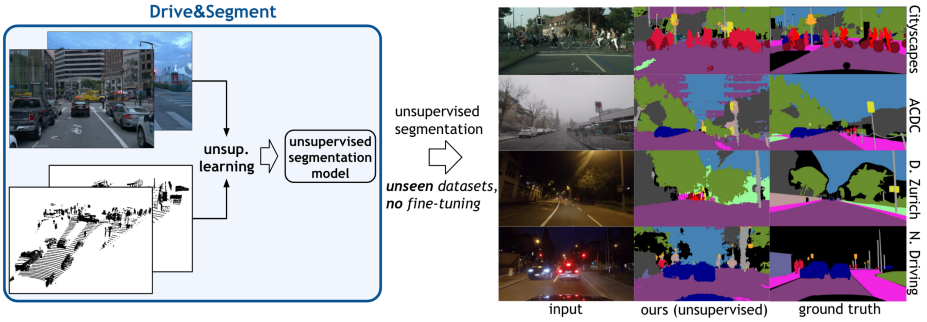

Fig. 1. Proposed fully-unsupervised approach. From uncurated images and LiDAR data, our Drive&Segment approach learns a semantic image segmentation model with no manual annotations. The resulting model performs unsupervised semantic segmentation of new unseen datasets without any human labeling. It can segment complex scenes with many objects, including thin structures such as people, bicycles, poles or traffic lights. Black color denotes the ignored/missing label.

use spatial self-supervision available in the image domain [15]. They do not leverage additional modalities, such as the LiDAR data, available for urban scenes in the autonomous driving set-ups. In this work, we develop an approach for unsupervised semantic segmentation that learns to segment complex scenes containing many objects, including thin structures such as pedestrians or traffic lights, without the need for any manual annotation. Instead, it leverages cross-modal information available in (aligned) LiDAR point clouds and images, see Fig. 1. Exploiting point clouds as a form of supervision is, however, not straightforward: data from LiDAR and camera are rarely perfectly synchronized; moreover, point clouds are unstructured and of much lower resolution compared to images; finally, extracting useful semantic information from LiDAR is still a very hard problem. In this work, we overcome these issues and show that it is nevertheless possible to extract useful pixel-wise semantic supervision from LiDAR data.

The contributions of our work are threefold. First, we propose a novel method for unsupervised cross-modal learning of semantic image segmentation by leveraging synchronized LiDAR and image data. The key ingredient is a module that analyzes the LiDAR point cloud to obtain proposals for spatially consistent objects that can be clearly separated from each other and the ground plane in the 3D scene. Second, we show that these 3D object proposals can be aligned with input images and reliably clustered into semantically meaningful pseudo-classes by using image features from a network trained without supervision. We demonstrate that this approach is robust to noise in point clouds and delivers, without the need for any manual annotation, pseudo-classes with pixel-wise segmentation for various objects present in driving scenes. These classes include objects such as pedestrians or traffic lights that are notoriously hard to segment automatically in the image domain. Third, we develop a novel cross-modal distillation approach that first trains a teacher network with the available partial pseudo

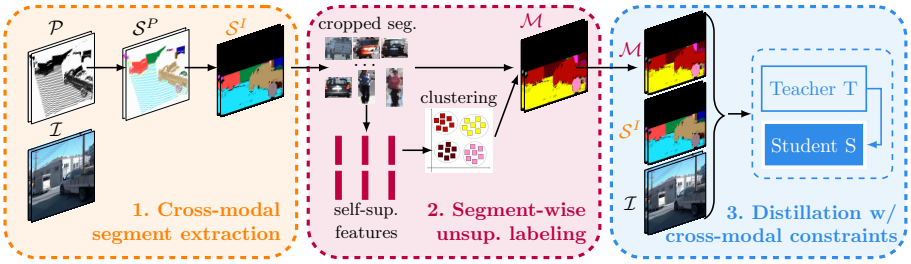

Fig. 2. Overview of Drive&Segment. We first perform cross-modal segment extraction on training dataset by exploiting raw *LiDAR* point clouds $\mathcal{P}$ and raw *images* $\mathcal{I}$. This yields segments $\mathcal{S}^I$ projected onto the image space (§ 3.1). By clustering their self-supervised features, we obtain an unsupervised labeling of these segments (§ 3.2) and, as a consequence, of their pixels. This provides pixel-wise *pseudo ground truth* for the next learning step. Finally, given the pseudo-labels and the segments, we perform distillation with cross-modal constraints (§ 3.3) that conjugates information of the LiDAR and the images to learn a final segmentation model using a teacher-student architecture. The learnt segmentation model S –highlighted in the figure– is used for inference on unseen datasets, yielding compelling results (§ 4).

labels and then exploits its predictions for training the student with pixel-wise pseudo annotations that cover the whole image. Additionally, our approach exploits geometric constraints extracted from the LiDAR point cloud during the teacher-student learning process to refine teacher predictions that are distilled into the student network. Implemented with transformer-based networks, this cross-modal distillation approach results in a trained student model that performs well in various challenging conditions such as day, night, fog, or rain, outside the domain of the original training dataset, as shown in Fig. 1.

We train our proposed unsupervised semantic segmentation method on two datasets, Waymo Open [47] and nuScenes [8] (nuScenes results are in the appendix available in the extended version of the paper [52]), and test it on four different datasets in the autonomous driving domain, Cityscapes [16], Dark-Zurich [44], Nighttime driving [17] and ACDC [45] dataset. We demonstrate significant improvements compared to the current state of the art on this problem, improving the current best published unsupervised semantic segmentation results on Cityscapes from 15.8 to 21.8 and from 4.6 to 14.2 on Dark Zurich, measured by mean intersection over union.

2 Related work

Image semantic segmentation. Semantic segmentation is a challenging key visual perception task, especially for autonomous driving [16,39,45,51,58]. Current top-performing models are based on fully convolutional networks [36] with encoder-decoder structures and a large diversity of designs [12,14,35,43,54,63]. Recent progress in vision transformers (ViT) [19] opened the door to a new wave

of decoders [46,56,59,64] with appealing predictive performance. These methods attain impressive performance by exploiting large amounts of pixel-wise labeled data. Yet, urban scenes are expensive to annotate manually (1.5h-3.3h per image [16,45]). This motivates recent works to rely less on pixel-wise supervision.

Reducing supervision for semantic segmentation. A popular strategy when dealing with limited labeled data is to pre-train some of the blocks of the architecture, e.g., the encoder, on related auxiliary tasks with plentiful labels [18,60]. Pre-training encoder for ImageNet [18] classification has been shown to be a successful recipe for both convnets [12] and ViT-based models [46]. Pre-training can be conducted even without any human annotations on artificially-designed self-supervised pretext tasks [9,22,23,24,26,28] with impressive results on a variety of downstream tasks. Some works also make use of synthetically generated data for pre-training [20,34,53]. Fully unsupervised semantic segmentation [6,13,15,29,31,32,40,50,61] has been recently addressed via generative models to generate object masks [6,13,40] or self-supervised clustering [15,31]. Prior methods are limited to segmenting foreground objects of a single class [6,13] or to *stuff* pixels that far outnumber *things* pixels [31,40]. Others assume that images contain centered objects [50], rely on weak spatial cues from the image domain [13,15,31] or require instance masks during pre-training and annotated data at test time [29]. On the contrary, our approach exploits cross-modal supervision from aligned LiDAR point clouds and images. We show that leveraging this information can considerably improve segmentation performance in complex autonomous driving scenes with multiple classes and strong class imbalance, outperforming PiCIE [15], the current state of the art in unsupervised segmentation. Concurrent work STEGO [25] develops a contrastive formulation for unsupervised semantic segmentation but does not use LiDAR during training.

Cross-modal self-supervised learning. Leveraging language, vision, and/or audio, self-supervised representation learning has seen tremendous progress in recent years [2,3,4,38,41,42,62]. Besides learning useful representations, these approaches show that signals from one modality can help train object detectors in the other, e.g., detecting instruments that sound in a scene [11,41,62], and even other object types [1]. In autonomous driving, a vehicle is equipped with diverse sensors (e.g., camera, LiDAR, radar), and cross-modal self-supervision is often used to generate labels from a sensor for augmenting the perception of another [5,30,48,55]. LiDAR clues [48] have been recently shown to boost unsupervised object detection (for *things* classes). Both our work and [48] use the same prior method [7] to extract object proposals from LiDAR scans. However, we consider a different problem of dense pixel-wise unsupervised semantic segmentation (for both *things* and *stuff*) and design a new approach for both extraction and learning with pixel-level pseudo labels.

3 Proposed unsupervised semantic segmentation

Our goal is to train an *image segmentation model* with *no human annotation*, by exploiting easily-available aligned *LiDAR* and *image* data. To that end, we

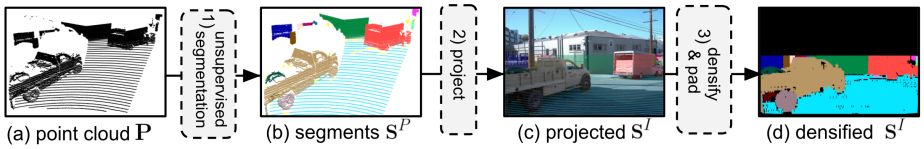

Fig. 3. Cross-modal segment extraction. Input raw point cloud (a) is first segmented with [7] into object segment candidates (b), which are then projected into the image (c); Projected segments are densified to get pixel-level pseudo-labels, with missed pixels being labeled as "ignore", as shown in black (d).

propose a novel method, Drive&Segment, that consists of three major steps and is illustrated in Figure 2. First, as discussed in Section 3.1, we extract *segment* proposals for the objects of interest from 3D LiDAR point clouds and project them to the aligned RGB images. In the second step (Section 3.2), we build *pseudo-labels* by clustering *self-supervised* image features corresponding to these segments. Finally, in Section 3.3, we propose a new teacher-student training scheme that incorporates *spatial constraints* from the LiDAR data and learns an unsupervised segmentation model from the noisy and partial pseudo-annotations generated in the previous two steps.

3.1 Cross-modal segment extraction

Throughout the next sections, we consider a dataset composed of a set $\mathcal{P}$ of 3D point clouds and a set $\mathcal{I}$ of images aligned with the point clouds. In this section, we detail the process of extracting segments of interest in an image $\mathbf{I} \in \mathcal{I}$ using the corresponding aligned LiDAR point cloud $\mathbf{P} \in \mathcal{P}$. The process, illustrated in Fig. 3, consists of three steps. We start by segmenting the LiDAR point cloud $\mathbf{P}$ using its geometrical properties. Then, we project the resulting 3D segments into the image $\mathbf{I}$, and densify the output to obtain pixel-level segments.

Geometric point cloud segmentation. We first extract J non-overlapping object segmentation proposals (*segments*), from the LiDAR point cloud $\mathbf{P}$. Let $\mathbf{S}^P = \{s_j^P\}_{j=1}^J$ be this set, where each segment s_j^P is a subset of the 3D point cloud $\mathbf{P}$ and $\forall j \neq j',\ s_j^P \cap s_{j'}^P = \varnothing$. Additionally, we refer to the set of segments across the entire data set as $\mathcal{S}^P$, with $\mathbf{S}^P \subset \mathcal{S}^P$. The J segments detected in one point cloud should ideally correspond to J individual objects in the scene. To get them, we use the unsupervised 3D point cloud segmentation proposed in [7], which exploits the geometrical properties of point clouds and range images.[1] It is a two-stage process that segments the ground plane and objects using greedy labeling by breadth-first search in the range image domain. Urban scenes are particularly suited to this purely geometry-based method as most objects are spatially well separated and the ground plane is relatively easy to segment out.

[1] Range images are depth maps corresponding to the raw LiDAR measurements. Valid measurements are back-projected to the 3D space to form a point cloud.

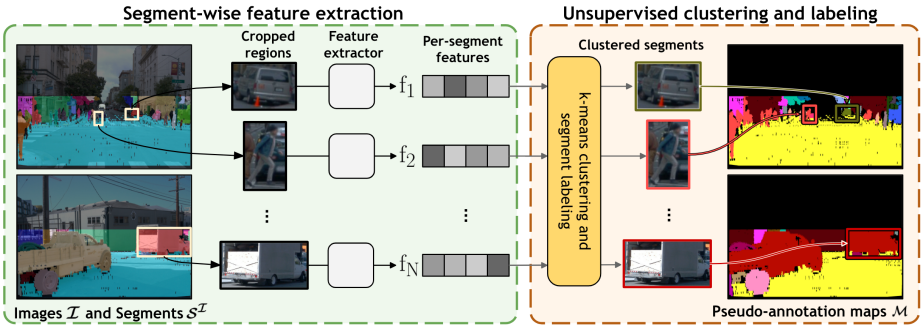

Fig. 4. Segment-wise unsupervised pseudo-labeling. First, given object segments $\mathcal{S}^{\mathcal{I}}$ obtained in the segment extraction stage (left), we take crops around all N objects and feed them to a feature extractor to get a set of N feature vectors. Then, we use the k-means algorithm to cluster the feature vectors into k clusters. Finally, we assign pixel-wise *pseudo-labels* to all pixels belonging to each segment based on the corresponding cluster id. Pixels not covered by a segment are assigned the label "ignore" (black).

Point-cloud-to-image transfer. The next step of the segment extraction is to transfer the set $\mathbf{S}^P$ of point cloud segments to the image $\mathbf{I}$, producing the set $\mathbf{S}^I$. Although LiDAR data and camera images are captured at the same time, one-to-one matching is not straightforward. Indeed, among other difficulties, LiDAR data only covers a fraction of the image plane because of its different field of view, its lower density, and its lack of usable measurements on far away objects or on the sky for instance. To overcome the mismatch between the two modalities, we proceed as follows. First, we project the points from the point cloud to the image using the known sensors' calibration. This gives us the locations of 3D points from the point cloud in the image. We also identify locations with invalid measurements in the LiDAR range image, e.g., reflective surfaces or the sky, and assign an "ignore" label to the respective locations.

Densify & pad. Next, we perform nearest-neighbor interpolation to propagate the $J+1$ segment labels to all pixels, where J is the number of segments (ideally corresponding to objects) and $+1$ denotes the additional "ignore" label. Last, we pad the image with "ignore" label to the input image size.

3.2 Segment-wise unsupervised labeling

Next, we produce *pseudo-labels* for all extracted segments in the image space without using any supervision. To that end, we leverage the recent ViT [19] model pre-trained in a fully unsupervised fashion [10] which has shown impressive results on various downstream tasks. We use this representation for unsupervised learning of pseudo-labels as described next and illustrated in Figure 4.

Considering the image $\mathbf{I}$, we crop a tight rectangular region in the image around each segment $s_j^I \in \mathbf{S}^I$ obtained using the proposal mechanism described in the previous section. We resize it and feed it to the ViT model to extract

the feature $\mathbf{f}_j$ corresponding to the output features of the `CLS` token. To limit the influence of pixels outside the object segment, which may correspond to other objects or the background, we mask out these pixels before computing the features. We repeat this operation for all segments in each image $\mathbf{I}$ in the training dataset and cluster the `CLS` token features using k-means algorithm, thus discovering k clusters of visually similar segments. Therefore, each feature $\mathbf{f}_j$ and its corresponding segment s_j^I, is assigned a cluster id l_j in $[\![1, k]\!]$.

To obtain a dense *segmentation* map $\mathbf{M}$ corresponding to the image $\mathbf{I}$, we assign discovered cluster ids to each pixel belonging to a segment in the image. Additionally, we assign a predefined *ignore* label to pixels not covered by segments, which correspond to missing annotations. This allows us to construct a set $\mathcal{M}$ of dense *maps of pseudo-annotations*, which we later use as a pseudo-ground-truth. Examples of resulting segmentation maps are shown in Figure 4.

3.3 Distillation with cross-modal spatial constraints

After previous steps have a set of pseudo-annotated segmentation maps $\mathbf{M} \in \mathcal{M}$, one for every image $\mathbf{I}$ in the training dataset. However, as explained above, the pseudo-annotations are **only partial**, since the segments that were used to construct them do not cover all pixels of an image. Furthermore, due to imperfections in the segment extraction process or the segment clustering step, these annotations are noisy. Using them to train an image segmentation model directly might be sub-optimal. Instead, we propose a new teacher-student training approach with cross-modal distillation, which is able to learn more accurate unsupervised segmentation models under such partial and noisy pseudo-annotations.

Training the teacher. The first step of our teacher-student approach is to train the teacher T to make pixel-wise predictions only on the pixels for which pseudo-annotations are available, i.e., only for the pixels that belong to a segment. We denote $\mathbf{Y}_\mathrm{T} = \mathrm{T}(\mathbf{I}) \in \mathbb{R}^{H \times W}$ the segmentation predictions made by the teacher model on image $\mathbf{I}$ with a resolution of $H \times W$ pixels. We train the teacher T using loss $\mathcal{L}_\mathrm{T}(\mathbf{I})$ and image $\mathbf{I}$:

$$\mathcal{L}_\mathrm{T}(\mathbf{I}) = \frac{1}{\sum_{h,w} B_{(h,w)}} \sum_{h,w} \mathrm{CE}\left(\mathbf{Y}_{\mathrm{T},(h,w)}, \mathbf{M}_{(h,w)}\right) B_{(h,w)}, \tag{1}$$

where CE is the cross-entropy loss measuring the discrepancy between the predicted labels $\mathbf{Y}_\mathrm{T}$ and target pseudo-labels $\mathbf{M}$ for each pixel (h, w), and B is a $H \times W$ binary mask for filtering out pixels without pseudo-annotations. The loss is normalized by the number of pseudo-labeled pixels in the image. The trained teacher T is then able to predict pixel-wise segmentation for all pixels in an image, even if they do not belong to a segment. Moreover, since the teacher T is trained on a large set of pseudo-annotated segments, it learns to smooth out some of the noise in the raw pseudo-annotations.

Integrating spatial constraints. Considering this smoothing property, we can exploit the trained teacher T for generating new, complete (instead of partial)

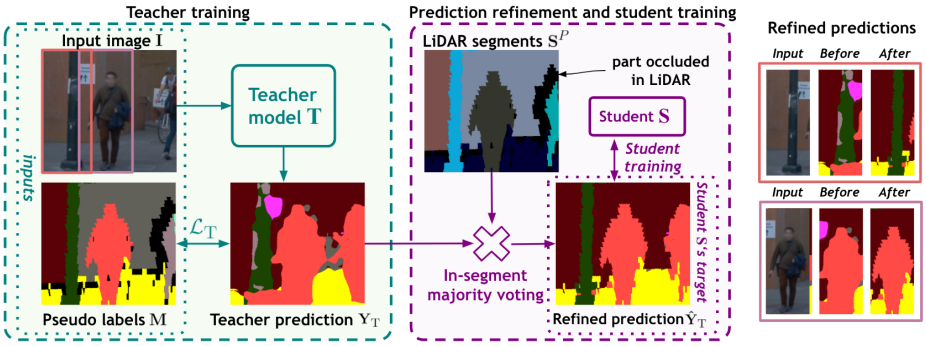

Fig. 5. Teacher prediction refinement using spatial constraints. First, the teacher T is trained using loss $\mathcal{L}_{\mathrm{T}}$ on images in $\mathcal{I}$ together with segmentation maps in $\mathcal{M}$ obtained from segment-wise unsupervised pseudo-labeling. The teacher predictions $\mathbf{Y}_{\mathrm{T}}$ are refined, using LiDAR segments $\mathbf{S}^P$, into maps $\hat{\mathbf{Y}}_{\mathrm{T}}$ that are then used to train the student. Note that teacher's predictions span the whole image, producing outputs even in areas where LiDAR segments $\mathbf{S}^P$ are not available.

and smooth pseudo-segmentation maps for training images. In addition, we propose to refine these teacher-generated pseudo-segmentation maps by using the projected LiDAR segments; indeed, these segments encode useful 3D spatial constraints as they often correspond to complete 3D objects, thus respecting the depth discontinuities and occlusion boundaries. In particular, for each image segment s_j^I in image $\mathbf{I}$, we apply majority voting to pixel-wise teacher predictions $\mathbf{Y}_{\mathrm{T}}$ inside the segment. Then we annotate each pixel belonging to the segment with its most frequently predicted label, giving us a new refined segmentation map $\hat{\mathbf{Y}}_{\mathrm{T}} \in \mathbb{R}^{H \times W}$. This procedure is illustrated in Figure 5.

Training the student. Having computed these complete, teacher-generated, and spatially refined pseudo-segmentation maps $\hat{\mathbf{Y}}_{\mathrm{T}}$, we train a student network S using the following loss

$$\mathcal{L}_{\mathrm{distill}}(\mathbf{I}) = \frac{1}{HW} \sum_{h,w} \mathrm{CE}\left(\hat{\mathbf{Y}}_{\mathrm{T},(h,w)}, \mathbf{Y}_{\mathrm{S},(h,w)}\right), \tag{2}$$

where the cross-entropy is computed between $\hat{\mathbf{Y}}_{\mathrm{T}}$ and the segmentation map $\mathbf{Y}_{\mathrm{S}} \in \mathbb{R}^{H \times W}$ predicted by the student at the same resolution as the teacher. The outputs of the trained student are our final unsupervised image segmentation predictions. Further details about our training can be found in Section 4.1.

4 Experiments

In this section, we give the implementation details, compare our results with the state-of-the-art unsupervised semantic segmentation methods on four different datasets, and ablate the key components of our approach.

Methods and architectures. We investigate the benefits of our approach using two different semantic segmentation models to demonstrate the generality of our method. We implement Drive&Segment with both a classical convolutional model and a transformer-based architecture. For the convolutional architecture, we follow [15] and use a ResNet18 [27] backbone followed by an FPN [35] decoder. For the transformer-based architecture, we use the state-of-the-art Segmenter [46] model. We use the ViT-S/16 [19] model as the Segmenter's encoder and use a single layer of the mask transformer [46] as a decoder. We compare our method to three recent unsupervised segmentation models: IIC [31], modified version of DeepCluster [9] (DC), and PiCIE [15]. Please refer to [15] for implementation details of IIC and DC.

Training. In the following, we first discuss how we obtain segment labels by k-means clustering, then we talk about details of pre-training the model backbones, which is followed by the discussion of the datasets for actual training of the models. Finally, we give details of the training procedure.

K-means. We use $k = 30$ in the k-means algorithm (the ablation of the value of k is in the appendix [52]. To extract segment-wise features used for k-means clustering, we use `CLS` token features of the DINO-trained [10] ViT-S [19] model. Obtained segment-wise labels serve as pseudo-annotations for training the ResNet18-FPN and Segmenter models, as discussed in Section 3.2.

Pre-training data and networks. To be comparable to [15], in our experiments with the ResNet18+FPN model, we initialize its backbone with a ResNet18 trained with supervision on the ImageNet-1k [18] classification task, exactly as all the compared prior methods (PiCIE, DC, and IIC). However, as we aim for a completely unsupervised setup, we initialize the ViT-S backbone of the Segmenter model with weights learned with the self-supervised approach DINO [10] on ImageNet-1k [18]. The decoders of our models are randomly initialized.

Training datasets. We train our models on about 7k images from the Waymo Open [47] dataset, which has both image and LiDAR data available. For the baseline methods (IIC [31], modified DC [9] and PiCIE [15]), we take models from PiCIE [15] codebase, i.e., models that are trained on all available images of Cityscapes [16], meaning the 24.5k images from the *train*, *test*, and *train_extra* splits. Note that those models then do not face the problem of domain gap when evaluated on the Cityscapes [16] dataset. To be directly comparable with our approach, we also train a variant of modified DC [9] and PiCIE [15] on the same subset of the Waymo Open [47] dataset as used in our approach. Furthermore, to test the generalizability of our method to other training datasets, we provide results when training on the nuScenes [8] dataset in the appendix [52].

Optimization. To train IIC [31], modified DC [9], and PiCIE [15], we use the setup provided in [15]. For our Drive&Segment, we train the teacher and student models with batches of size 32 and with a learning rate of $2e{-}4$ with a polynomial schedule on a single V100 GPU. During training, we perform data augmentation consisting of random image resizing in the $(0.5, 2.0)$ range, random cropping to 512×512 pixels, random horizontal flipping, and photometric distortions.

Evaluation protocol. *Mapping.* To evaluate our models in the unsupervised setup, we run trained models on every image, thus getting segmentation predictions with values from 1 to k. Then, we compute the confusion matrix between the C ground-truth classes of the target dataset and the $k \geq C$ pseudo-classes. We map the C ground-truth classes to C out of the k pseudo-classes using Hungarian matching [33]. The pixel predictions for the $k-C$ unmapped pseudo-classes are considered as false negatives.

Test datasets. We evaluate our fully-unsupervised models on the *full-resolution images* of Cityscapes [16], Dark Zurich [44], Nighttime driving [17] and ACDC [45] datasets, *without any finetuning* (no samples from these datasets are ever seen during training). Cityscapes [16] is a well-established dataset with 500 validation images that we use for evaluation. Dark Zurich [44] and Nighttime driving [17] are two nighttime datasets, each with 50 validation images annotated for semantic segmentation that we use for evaluation. ACDC [45] is a recent dataset providing four different adverse weather conditions with 400 training and 100 validation samples per weather condition. We test our approach on the validation images annotated for semantic segmentation. The Cityscapes dataset defines 30 different semantic classes for the pixel-wise semantic segmentation task. Unless stated otherwise, we follow prior work and evaluate our approach on the pre-defined subset of 19 classes [16] for all datasets.

Metrics. Using the mapping, we evaluate the results using two standard metrics for the semantic segmentation task, the mean Intersection over Union, mIoU, and the pixel accuracy, PA, as done in prior work [15]. The mIoU is the mean intersection over union averaged over all classes, while PA defines the percentage of pixels in the image that are segmented correctly, averaged over all images.

4.1 Comparison to state of the art

Here we evaluate our trained models in the unsupervised setup using the evaluation protocol described above. We compare our method using both the Segmenter [46] and ResNet18+FPN models to three recent unsupervised segmentation models: IIC [31], modified version of DeepCluster [9] (mod. DC), and PiCIE [15]. In the appendix [52], we assess the utility of the features learned by our model in other settings, such as k-NN pixel-wise classification, and linear probing and fine-tuning for semantic segmentation.

We provide results on the Cityscapes, Dark Zurich, and Nighttime Driving datasets in Table 1, and show qualitative results in Figure 6. As shown in the first two columns of Table 1, our approach (D&S) outperforms [15] on the Cityscapes dataset by a large margin in both the 19-class and 27-class set-ups. Improvements are visible for both architectures, but the best results are usually obtained with the distilled Segmenter architecture using the ViT-S/16 backbone. Our models again outperform [15] in all setups. In addition, we observe a better performance of our models compared to [15] when evaluating on the nighttime scenes. For example, on the Dark Zurich [44] dataset, the mIoU of PiCIE [15] decreases by 71% compared to the results on Cityscapes (15.8 $\rightarrow$ 4.6), while the mIoU of our Segmenter-based model decreases only by 35% (21.8 $\rightarrow$ 14.2). This suggests that

Table 1. Comparison to the state of the art for unsupervised semantic segmentation on Cityscapes [16] (CS), DarkZurich [44] (DZ) and Nighttime driving [17] (ND) datasets measured by the mean IoU (mIoU). The colored differences are reported w.r.t. the SoTA approach of [15] denoted by ⚓. The *sup. init.* abbreviation stands for supervised initialization of the *encoder*, and the column *train. data* indicates whether Cityscapes (CS) or Waymo Open (WO) dataset was used for training.

architecture, method	sup. init.	train. data	CS19 [16] mIoU	CS27 [16] mIoU	DZ [44] mIoU	ND [17] mIoU
RN18+FPN						
IIC† [31]	yes	CS	-	6.4 (-4.8)	-	-
Modified DC‡ [9]	yes	CS	11.3 (-4.5)	7.9 (-3.3)	7.5 (+2.9)	8.2 (-1.3)
⚓ PiCIE‡ [15]	yes	CS	15.8	11.2	4.6	9.5
Modified DC⋆	yes	WO	11.4 (-4.4)	7.0 (-4.1)	5.9 (+1.3)	8.2 (-1.3)
PiCIE⋆	yes	WO	13.7 (-2.1)	9.7 (-1.5)	4.9 (+0.3)	9.3 (-0.2)
D&S (Ours, S)	yes	WO	19.5 (+3.7)	**16.2** (**+5.1**)	10.9 (+6.3)	14.4 (+4.9)
Segmenter, ViT-S/16						
D&S (Ours, S)	no	WO	**21.8** (**+6.0**)	15.3 (+4.1)	**14.2** (**+9.6**)	**18.9** (**+9.3**)

† Results reported in [15]. ‡ Models provided by the PiCIE [15] authors.
⋆ Trained by PiCIE code base.

Table 2. Comparison to the state-of-the-art for unsupervised semantic segmentation on the ACDC [45] dataset. Please refer to Table 1 for the used symbols.

arch., method	sup. init.	train. data	night mIoU	fog mIoU	rain mIoU	snow mIoU	average mIoU
RN18+FPN							
mod. DC‡ [9]	yes	CS	8.1 (+3.7)	8.3 (-4.0)	6.9 (-5.6)	7.4 (-4.7)	7.7 (-2.6)
⚓ PiCIE‡ [15]	yes	CS	4.4	12.2	12.5	12.1	10.3
mod. DC⋆	yes	WO	5.9 (+1.5)	11.7 (-0.5)	9.6 (-2.9)	9.8 (-2.3)	9.2 (-1.0)
PiCIE⋆	yes	WO	4.7 (+0.3)	14.4 (+2.1)	13.7 (+1.2)	14.3 (+2.2)	11.7 (+1.5)
D&S (Ours,S)	yes	WO	11.2 (+6.8)	14.5 (+2.3)	14.9 (+2.5)	14.6 (+2.6)	13.8 (+3.5)
Segmenter, ViT-S/16							
D&S (Ours,S)	no	WO	**13.8** (**+9.4**)	**18.1** (**+5.9**)	**16.4** (**+3.9**)	**18.7** (**+6.6**)	**16.7** (**+6.5**)

our models generalize significantly better to out-of-distribution scenes. These findings hold for PiCIE models trained on both Cityscapes and on Waymo Open.

Finally, Table 2 shows results on the ACDC dataset in four different weather conditions. Results follow a similar trend as in Table 1 and show the superiority of our approach measured by mIoU compared to the current SoTA unsupervised semantic segmentation method of [15] on images out of the training distribution, such as images at night or in snow. Please see the appendix available in the extended version of the paper [52] for the complete set of results, including results using nuScenes, pixel accuracy, per-class results, and confusion matrices.

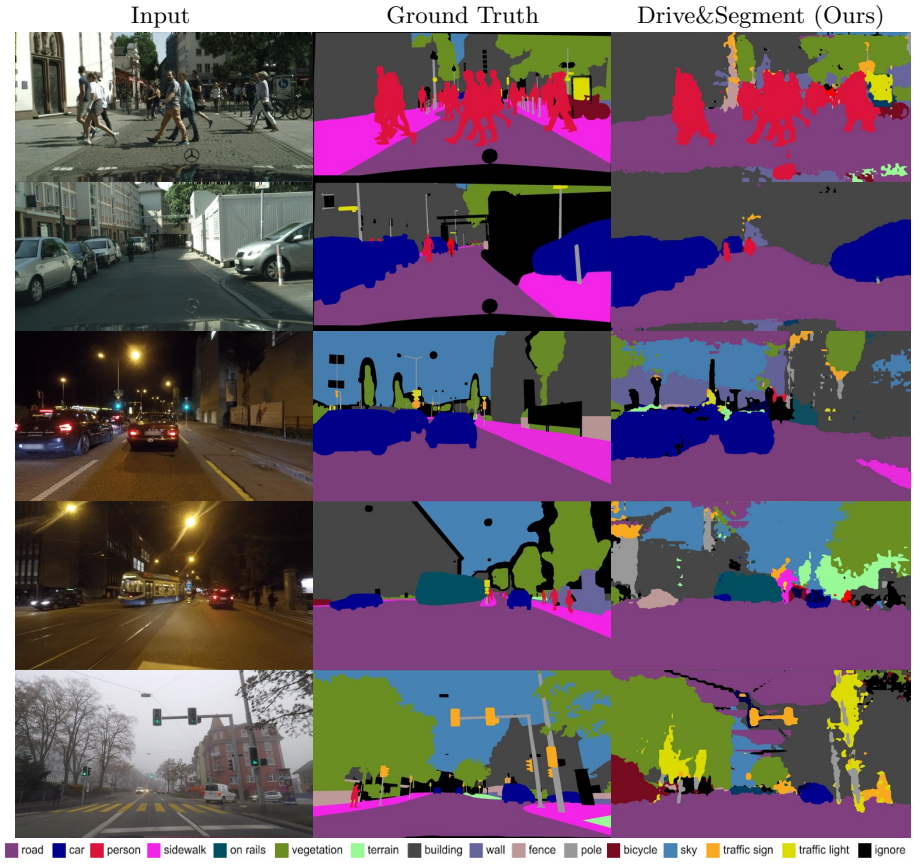

Fig. 6. Qualitative results for *unsupervised* semantic segmentation using our Drive&Segment approach. To obtain the matching between our pseudo-classes and the set of ground-truth classes, we use the Hungarian algorithm. The first two rows show samples from Cityscapes [16], and the three bottom rows show samples from the night and fog splits of the ACDC [45] dataset. See appendix in [52] for more results.

4.2 Ablations

In this section, we ablate the main components of our approach, which we present in Table 3, and discuss them in more detail below.

Segment extraction approach. To evaluate the benefits of our cross-modal segment extraction module, we investigate using segment proposals generated with a purely image-based segmentation approach by Felzenszwalb and Huttenlocher (FH) [21]. It groups pixels into segments based on similar color and texture properties. We use the same set of hyperparameters as [28]. The results are shown in Table 3a and demonstrate clear benefits of our LiDAR-based cross-modal segment extraction method despite the difficulties of using LiDAR data discussed in Section 3.1. We attribute the better results of our approach to the

Table 3. Ablations on the Cityscapes dataset. (a) Benefits of our segment extraction method over segment proposals from [21]. **(b)** Benefits of our distillation approach showing an improvement of the student (S) over the the teacher (T) and benefits of our LiDAR cross-modal spatial constraints (LiD). **(c)** Ablation of different feature extractors for the k-means clustering.

(a) Segment extraction

arch. / seg. prop.	mIoU	PA
RN18+FPN		
FH [21]	15.5	52.8
Ours	**17.4** (+1.9)	**55.9** (+3.1)
Segmenter		
FH [21]	15.8	51.8
Ours	**20.4** (+4.6)	**65.4** (+13.6)

(b) Distillation

model	LiD.	mIoU	PA
RN18+FPN			
PiCIE (T)		13.7	48.6
PiCIE (S)		14.8 (+1.1)	64.1 (+15.5)
PiCIE (S)	✓	15.1 (+1.4)	**68.4** (+19.8)
Ours (T)		17.4	55.9
Ours (S)		18.8 (+1.4)	63.4 (+7.5)
Ours (S)	✓	**19.5** (+2.1)	**66.4** (+10.5)
Segmenter			
Ours (T)		20.4	65.4
Ours (S)		20.8 (+0.4)	68.5 (+3.1)
Ours (S)	✓	**21.8** (+1.4)	**69.5** (+4.1)

(c) Feature extractors

arch. / method	mIoU	PA
ViT-S/16		
DeiT [49]	21.7	73.0
DINO [10]	20.2	64.4
ResNet18		
supervised [27]	19.6	70.0
ResNet50		
supervised [27]	21.3	67.6
OBOW [22]	20.7	65.9
PixPro [57]	20.7	65.9
MaskCon. [50]	19.1	68.0

fact that LiDAR data segmentation operates with range information, which is much stronger at separating objects from the background and from each other compared to the purely image-based approach of FH [21]. Indeed, FH relies only on color/texture and is therefore much more likely to join multiple objects into one segment or separate a single object into multiple segments. The benefits of our cross-modal segment extraction are observed for both studied architectures.

Distillation with cross-modal spatial constraints. To evaluate the benefits of our teacher-student distillation method with cross-modal spatial constraints (Section 3.3), we compare the predictions of the teacher T (before distillation) and the student S (after distillation). Table 3b presents results on the Cityscapes dataset using both convolutional- and transformer-based architectures. The results show consistent improvements using our distillation technique, particularly regarding the pixel accuracy metric. We believe that this could be attributed to improvements in predictions for classes such as vegetation and buildings. They often occupy large areas of the image and benefit most from the distillation as they are usually not well covered by the LiDAR scans. Furthermore, the results show clear benefits of using this distillation step both with and without cross-modal spatial constraints (LiD) by Student S outperforming Teacher T in both scenarios. Please also note that our distillation technique works well even in combination with another training approach (PiCIE [15]).

Sensitivity to the initialization. To study the influence of initialization, we take the features extracted by DINO [10] and run the k-means clustering (Section 3.2) four times. For each k-means clustering outcome, we run the segmentation model training four times with different initializations. The variance over all k-means and training runs is only 0.5 for mIoU and 1.5 for pixel accuracy

(i.e., $20.4 \pm 0.5/65.4 \pm 1.5$). These results clearly show that our method is not very sensitive to k-means initialization or to the network initialization.

Feature extractors. An ablation of different convolutional and ViT feature extractors for the task of segment-wise unsupervised labelling is shown in Table 3c. The results on the Cityscapes [16] dataset using our Segmenter model demonstrate that our approach works well with several different feature extractors.

LiDAR resolution and number of clusters. Ablation of the influence of LiDAR resolution is in the appendix [52] and demonstrates that our method is robust to LiDAR's sparsity. Furthermore, we study the choice of the number of clusters for the k-means clustering in the appendix [52].

4.3 Limitations and failure modes

Our approach has the following three main limitations. First, LiDAR point clouds do not provide information about very distant or even infinitely distant objects, e.g., the sky, which our approach cannot learn to segment. Second, LiDAR point clouds paired with geometric segmentation can not correctly distinguish road from sidewalk or grass, when all surfaces are similarly flat. Both the above limitations might be possibly tackled by pairing our LiDAR-based segment proposals with an unsupervised image-based method such as [21], or by introducing simple heuristics. Also, the LiDAR points must not be too sparse (e.g., only 4 beams), since otherwise the LiDAR-based segments would be of poor quality. However, this is not an overly restricting requirement as it is common to use LiDAR sensors with sufficient beam resolution, e.g., as in the recent Waymo Open [47] or ONCE [37] datasets. Finally, we encounter semantically similar objects appearing in multiple pseudo-classes, a natural side effect of clustering. This issue may be mitigated by using different feature clustering methods that would allow the measurement of similarities on manifolds in the feature space.

5 Conclusion

We have developed Drive&Segment, a fully unsupervised approach for semantic image segmentation in urban scenes. The approach relies on novel modules for (i) cross-modal segment extraction and (ii) distillation with cross-modal constraints that leverage LiDAR point clouds aligned with images. We evaluate our approach on four different autonomous driving datasets in challenging weather and illumination conditions and demonstrate major gains over prior work. This work opens up the possibility of large-scale autonomous learning of embodied perception models without explicit human supervision.

Acknowledgments. This work was supported by the European Regional Development Fund under the project IMPACT (no. CZ.02.1.010.00.015_0030000468), by the Ministry of Education, Youth and Sports of the Czech Republic through the e-INFRA CZ (ID:90140), and by CTU Student Grant SGS21184OHK33T37.

References

1. Afouras, T., Asano, Y.M., Fagan, F., Vedaldi, A., Metze, F.: Self-supervised object detection from audio-visual correspondence. In: arXiv (2021) 4
2. Alayrac, J.B., Recasens, A., Schneider, R., Arandjelovic, R., Ramapuram, J., De Fauw, J., Smaira, L., Dieleman, S., Zisserman, A.: Self-supervised multimodal versatile networks. In: NeurIPS (2020) 4
3. Alwassel, H., Mahajan, D., Korbar, B., Torresani, L., Ghanem, B., Tran, D.: Self-supervised learning by cross-modal audio-video clustering. In: NeurIPS (2020) 4
4. Arandjelovic, R., Zisserman, A.: Look, listen and learn. In: ICCV (2017) 4
5. Bartoccioni, F., Zablocki, É., Pérez, P., Cord, M., Alahari, K.: Lidartouch: Monocular metric depth estimation with a few-beam lidar. In: arXiv (2021) 4
6. Bielski, A., Favaro, P.: Emergence of object segmentation in perturbed generative models. In: NeurIPS (2019) 4
7. Bogoslavskyi, I., Stachniss, C.: Efficient online segmentation for sparse 3d laser scans. PFG (2017) 4, 5
8. Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: nuscenes: A multimodal dataset for autonomous driving. In: CVPR (2020) 3, 9
9. Caron, M., Bojanowski, P., Joulin, A., Douze, M.: Deep clustering for unsupervised learning of visual features. In: ECCV (2018) 4, 9, 10, 11
10. Caron, M., Touvron, H., Misra, I., Jegou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging Properties in Self-Supervised Vision Transformers. In: ICCV (2021) 6, 9, 13
11. Chen, H., Xie, W., Afouras, T., Nagrani, A., Vedaldi, A., Zisserman, A.: Localizing visual sounds the hard way. In: CVPR (2021) 4
12. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: ECCV (2018) 3, 4
13. Chen, M., Artières, T., Denoyer, L.: Unsupervised object segmentation by redrawing. In: NeurIPS (2019) 4
14. Cheng, B., Collins, M.D., Zhu, Y., Liu, T., Huang, T.S., Adam, H., Chen, L.C.: Panoptic-deeplab: A simple, strong, and fast baseline for bottom-up panoptic segmentation. In: CVPR (2020) 3
15. Cho, J.H., Mall, U., Bala, K., Hariharan, B.: PiCIE: Unsupervised semantic segmentation using invariance and equivariance in clustering. In: CVPR (2021) 2, 4, 9, 10, 11, 13
16. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: CVPR (2016) 3, 4, 9, 10, 11, 12, 14
17. Dai, D., Van Gool, L.: Dark model adaptation: Semantic image segmentation from daytime to nighttime. In: IEEE ITSC (2018) 3, 10, 11
18. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A Large-Scale Hierarchical Image Database. In: CVPR (2009) 4, 9
19. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR (2021) 3, 6, 9
20. Dosovitskiy, A., Fischer, P., Ilg, E., Häusser, P., Hazirbas, C., Golkov, V., van der Smagt, P., Cremers, D., Brox, T.: Flownet: Learning optical flow with convolutional networks. In: ICCV (2015) 4

21. Felzenszwalb, P.F., Huttenlocher, D.P.: Efficient graph-based image segmentation. IJCV (2004) 12, 13, 14
22. Gidaris, S., Bursuc, A., Puy, G., Komodakis, N., Cord, M., Pérez, P.: Obow: Online bag-of-visual-words generation for self-supervised learning. In: CVPR (2021) 4, 13
23. Gidaris, S., Singh, P., Komodakis, N.: Unsupervised representation learning by predicting image rotations. In: ICLR (2018) 4
24. Grill, J., Strub, F., Altché, F., Tallec, C., Richemond, P.H., Buchatskaya, E., Doersch, C., Pires, B.Á., Guo, Z., Azar, M.G., Piot, B., Kavukcuoglu, K., Munos, R., Valko, M.: Bootstrap your own latent - A new approach to self-supervised learning. In: NeurIPS (2020) 4
25. Hamilton, M., et al.: Unsupervised semantic segmentation by distilling feature correspondences. In: ICLR (2022) 4
26. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.B.: Momentum contrast for unsupervised visual representation learning. In: CVPR (2020) 4
27. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016) 9, 13
28. Hénaff, O.J., Koppula, S., Alayrac, J.B., Oord, A.v.d., Vinyals, O., Carreira, J.: Efficient visual pretraining with contrastive detection. In: ICCV (2021) 4, 12
29. Hwang, J.J., Yu, S.X., Shi, J., Collins, M.D., Yang, T.J., Zhang, X., Chen, L.C.: Segsort: Segmentation by discriminative sorting of segments. In: ICCV. pp. 7334–7344 (2019) 4
30. Jaritz, M., Vu, T.H., Charette, R.d., Wirbel, E., Pérez, P.: xmuda: Cross-modal unsupervised domain adaptation for 3d semantic segmentation. In: CVPR (2020) 4
31. Ji, X., Henriques, J.F., Vedaldi, A.: Invariant information clustering for unsupervised image classification and segmentation. In: ICCV (2019) 4, 9, 10, 11
32. Kanezaki, A.: Unsupervised image segmentation by backpropagation. In: ICASSP (2018) 4
33. Kuhn, H.W., Yaw, B.: The hungarian method for the assignment problem. NRLQ (1955) 10
34. Li, D., Yang, J., Kreis, K., Torralba, A., Fidler, S.: Semantic segmentation with generative models: Semi-supervised learning and strong out-of-domain generalization. In: CVPR. pp. 8300–8311 (2021) 4
35. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: CVPR (2017) 3, 9
36. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: CVPR (2015) 3
37. Mao, J., Niu, M., Jiang, C., Liang, H., Liang, X., Li, Y., Ye, C., Zhang, W., Li, Z., Yu, J., et al.: One million scenes for autonomous driving: Once dataset. NeurIPS (2021) 14
38. Miech, A., Alayrac, J.B., Smaira, L., Laptev, I., Sivic, J., Zisserman, A.: End-to-end learning of visual representations from uncurated instructional videos. In: CVPR (2020) 4
39. Neuhold, G., Ollmann, T., Rota Bulo, S., Kontschieder, P.: The mapillary vistas dataset for semantic understanding of street scenes. In: ICCV (2017) 3
40. Ouali, Y., Hudelot, C., Tami, M.: Autoregressive unsupervised image segmentation. In: ECCV (2020) 4
41. Owens, A., Efros, A.A.: Audio-visual scene analysis with self-supervised multisensory features. In: ECCV (2018) 4

42. Recasens, A., Luc, P., Alayrac, J.B., Wang, L., Strub, F., Tallec, C., Malinowski, M., Pătrăucean, V., Altché, F., Valko, M., Grill, J.B., van den Oord, A., Zisserman, A.: Broaden your views for self-supervised video learning. In: ICCV (2021) 4
43. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: MICCAI (2015) 3
44. Sakaridis, C., Dai, D., Van Gool, L.: Map-guided curriculum domain adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation. IEEE TPAMI (2020) 3, 10, 11
45. Sakaridis, C., Dai, D., Van Gool, L.: ACDC: The adverse conditions dataset with correspondences for semantic driving scene understanding. In: ICCV (2021) 3, 4, 10, 11, 12
46. Strudel, R., Garcia, R., Laptev, I., Schmid, C.: Segmenter: Transformer for semantic segmentation. In: ICCV (2021) 4, 9, 10
47. Sun, P., Kretzschmar, H., Dotiwalla, X., Chouard, A., Patnaik, V., Tsui, P., Guo, J., Zhou, Y., Chai, Y., Caine, B., et al.: Scalability in perception for autonomous driving: Waymo open dataset. In: CVPR (2020) 3, 9, 14
48. Tian, H., Chen, Y., Dai, J., Zhang, Z., Zhu, X.: Unsupervised object detection with lidar clues. In: CVPR (2021) 4
49. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention. In: ICML (2021) 13
50. Van Gansbeke, W., Vandenhende, S., Georgoulis, S., Van Gool, L.: Unsupervised semantic segmentation by contrasting object mask proposals. In: ICCV (2021) 1, 4, 13
51. Varma, G., Subramanian, A., Namboodiri, A., Chandraker, M., Jawahar, C.: Idd: A dataset for exploring problems of autonomous navigation in unconstrained environments. In: WACV (2019) 3
52. Vobecky, A., Hurych, D., Siméoni, O., Gidaris, S., Bursuc, A., Pérez, P., Sivic, J.: Drive&segment: Unsupervised semantic segmentation of urban scenes via cross-modal distillation. https://arxiv.org/abs/2203.11160 (2022) 3, 9, 10, 11, 12, 14
53. Vobeckỳ, A., Hurych, D., Uřičář, M., Pérez, P., Sivic, J.: Artificial dummies for urban dataset augmentation. In: AAAI. vol. 35, pp. 2692–2700 (2021) 4
54. Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., Liu, D., Mu, Y., Tan, M., Wang, X., et al.: Deep high-resolution representation learning for visual recognition. IEEE TPAMI (2020) 3
55. Weston, R., Cen, S., Newman, P., Posner, I.: Probably unknown: Deep inverse sensor modelling radar. In: ICRA (2019) 4
56. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: Segformer: Simple and efficient design for semantic segmentation with transformers. arXiv preprint arXiv:2105.15203 (2021) 4
57. Xie, Z., Lin, Y., Zhang, Z., Cao, Y., Lin, S., Hu, H.: Propagate yourself: Exploring pixel-level consistency for unsupervised visual representation learning. In: CVPR (2021) 13
58. Yu, F., Chen, H., Wang, X., Xian, W., Chen, Y., Liu, F., Madhavan, V., Darrell, T.: Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In: CVPR (2020) 3
59. Yuan, Y., Chen, X., Wang, J.: Object-contextual representations for semantic segmentation. In: ECCV (2020) 4
60. Zamir, A.R., Sax, A., Shen, W., Guibas, L.J., Malik, J., Savarese, S.: Taskonomy: Disentangling task transfer learning. In: CVPR (2018) 4

61. Zhang, X., Maire, M.: Self-supervised visual representation learning from hierarchical grouping. Advances in Neural Information Processing Systems **33**, 16579–16590 (2020) 4
62. Zhao, H., Gan, C., Rouditchenko, A., Vondrick, C., McDermott, J., Torralba, A.: The sound of pixels. In: ECCV (2018) 4
63. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: CVPR (2017) 3
64. Zheng, S., Lu, J., Zhao, H., Zhu, X., Luo, Z., Wang, Y., Fu, Y., Feng, J., Xiang, T., Torr, P.H., Zhang, L.: Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In: CVPR (2021) 4