2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

Two Teachers Are Better Than One: Leveraging Depth In Training Only For Unsupervised Obstacle Segmentation

Sungmin Eum[1,2], Hyungtae Lee[1], Heesung Kwon[1], Phil Osteen[1], and Andre Harrison[1]

***Abstract*— We present a novel unsupervised obstacle segmentation architecture that follows a novel Relation Distillation (RD) paradigm. Our architecture design was inspired by a self-supervised teacher-student approach that relies on the Semantic Distillation originally devised for representation learning. While the teacher in the Semantic Distillation considers a single patch at a time, the teacher within RD takes a 'pair of patches' instead to transfer the local Semantic Co-occurrence Localization (SCooL) relationship that focuses more on the segmentation-boosting signals. To further improve the proposed architecture, we introduce the utilization of another teacher that leverages the depth information which inherently separates the entities at different physical distances, often tied with the boundaries of the obstacles. As the depth is distilled towards the student network only at the time of training, it adds zero computational/hardware cost at run-time. As no relevant public dataset is available, we have curated the Avoiding Obstacles In unstructured Driving (AvOID) dataset as a new testbed for unsupervised obstacle segmentation. We have validated that both the Relation Distillation and depth contribute to boosting the no-annotation segmentation performance on AvOID and KITTI-Obstacles.**

I. INTRODUCTION

One of the critical perception capabilities for an autonomous robot is to accurately identify obstacles in the vicinity and find its way around them to carry on a navigation. In realistic scenarios with little to no visual cues to guide paths, it is important to have the ability to identify even the obstacles the model had not seen at the time of training. An effective way to acquire such capability is to build a perception model (e.g., segmentation) that does not rely on predefined annotations of certain obstacles in training, thus not confining the target categories to a limited set of potential obstacles.

A recent architecture [1] showed that training a self-supervised teacher-student distillation model, originally devised for representation learning, can provide a capability to segment objects as an unexpected byproduct. However, while this approach did provide a good starting point to explore the new realm of the 'no-annotation obstacle segmentation', the original design was not directly focused on the segmentation task itself. Instead, it was focused on creating generally effective representations (often tied with classifications), suggesting that there is a performance gap that can be filled by adjusting the focus of the model.

While it was shown that having a general-purpose teacher in a distillation process is helpful for various downstream tasks, we claim that *learning from a teacher with a specific expertise leads to a better performance when targeting the corresponding task (in our case, segmentation).* We realize our claim by distilling the segmentation-boosting information through a module that examines the 'localized semantic relationships' between pairs of different patches extracted from the input images. While the baseline was focusing on transferring the semantics in a holistic sense (i.e., Semantic Distillation), our module focuses on determining where the matching semantics are located, even between the non-corresponding locations in the two input patches (i.e., relation distillation), which we define as the *Semantic Co-occurrence Localization (SCooL)*, as shown in Figure 5.

We went one step further and leveraged an additional teacher that learns off of a non-RGB modality, backing up our second claim that *two teachers are better than one*. Paying attention to the fact that a depth map of a scene roughly visualizes the boundaries between perceptual entities, we sought for a model that connects two teachers (i.e., RGB and depth) with a single student, providing complementary

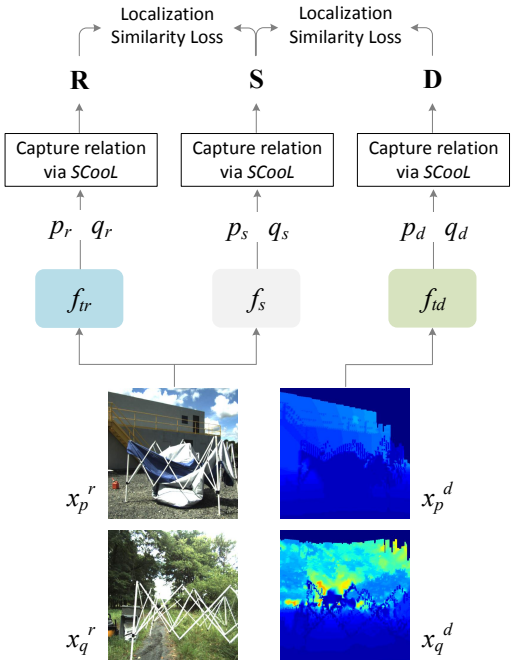

Fig. 1. **Two Teacher Distillation with Depth (T2D2).** f_{tr}, f_{td} and f_s denote the RGB teacher, depth teacher and the student network, respectively. Knowledge from the two teachers are distilled towards one student network within T2D2 framework which probes into the relationship between paired inputs ($\{x_p{}^r, x_q{}^r\}$ or $\{x_p{}^d, x_q{}^d\}$ for RGB and depth, respectively). The segmentation-boosting relationships (**R**, **S**, and **D**) are captured by the *SCooL* modules.

[1] DEVCOM Army Research Laboratory
[2]Booz Allen Hamilton

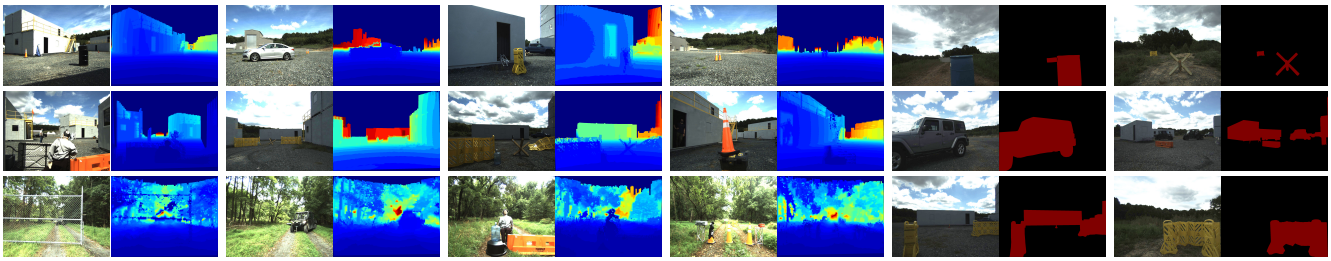

Fig. 2. **Sample images of AvOID train (columns 1-4) and test (columns 5-6) set.** Each pair shows an RGB image and its corresponding depth (for train) or obstacle ground truth (for test) image.

learning via two sources. Note that the two teachers are involved only at the time of training, while the student (RGB) is solely leveraged for inference, adding zero computational complexity and no additional modality at run-time compared to the single-teacher architecture.

To that end, we have devised a novel architecture, namely, Two Teacher Distillation with Depth (T2D2) (Figure 1), with carefully designed modules (i.e., SCooL and the two-teacher distillation) that directly influences the segmentation task. As a means to further improve the quality of the features distilled by T2D2, we have devised a novel module, *In-batch Pressure Generation (IPG)*, that selects certain patch-pairs from a given batch to generate stronger positive and negative pressures for the loss optimization, eventually improving our unsupervised distillation process.

To train and evaluate our model, we needed a dataset that contains an ample amount of obstacles in unstructured outdoor environments. No publicly available datasets were found to be appropriate, motivating us to curate a new obstacle segmentation dataset, which we call the **Av**oiding **O**bstacles **I**n unstructured **D**riving (AvOID) dataset (Figure 2). The dataset is composed of a set of video sequences captured by a camera on an unmanned ground vehicle navigating through an outdoor scene where a diverse set of obstacles were present. Aligned with all the RGB images, depth images (constructed from LiDAR points) and ground truth masks are included to be used as a second modality and for evaluation purposes, respectively.

From the experiments, our model was found to outperform the baselines including the Semantic Distillation architecture in terms of the segmentation accuracy on both AvOID and a publicly available dataset repurposed for our scenario (i.e., KITTI-Obstacles). The performance enhancement clearly demonstrates the benefits of the key contributions of the paper: 1) relation distillation architecture with SCooL that focuses on capturing segmentation-boosting features in an unsupervised paradigm, 2) two-teacher distillation that uses RGB and depth for training, and RGB only for inference, 3) IPG that provides enhanced learning signals for distillation, and 4) a novel dataset (AvOID) for obstacle segmentation in unstructured environments.

II. RELATED WORK

Knowledge distillation. There has been a tremendous amount of literature [2], [3], [4], [5], [6], [7], [8] ever since Hinton et al. [9] introduced the concept of knowledge distillation (KD). We constrain our scope to the methods that share similarity with our architecture either in conceptual or architectural aspects. One line of work focuses on using KD for exchanging information between domains or modalities, e.g., between RGB image and depth [10], between RGB and LiDAR [11], or between human pose and motions [12]. Our model seeks to distill from depth towards an RGB network.

Another line of work leveraged multiple teacher networks instead of a single teacher, in pursuit of more stable distillation. In [13], [14], [15], one student network distills the knowledge from several teachers. [10], [16] used ensemble methods to combine decisions from multiple teachers before the distillation. While our method also leverages two teachers (RGB and depth), it is noteworthy to mention that our approach is the first to distill multiple modalities into a single modality in a completely unsupervised way, without any annotations or pre-training.

Lastly, several approaches [17], [18], [19], [20], [21] have been introduced that distill the ability to capture the relationship among multiple samples in a batch. While we also seek to distill the relationships, our method is unique in that it focuses on transferring the ability to find semantically co-occurring spatial locations between two samples (*semantic co-occurrence localization*) which turned out to be highly beneficial for the task of unsupervised segmentation.

Depth in training only. Depth information is highly beneficial as it complements the information acquired from RGB alone. However, the fact that devices are required to get depth at run-time is a major drawback. To avoid such inconvenience, attempts have been made to leverage depth only in training while addressing tasks such as 3D scene reconstruction [22], [23], visual odometry [24], [25], [26], [27], monocular 3D object detection [28], [11], panoptic segmentation [29], [30], frame interpolation [31]. These methods commonly train a depth generator to infer depth (e.g., depth map [32], point cloud [33], [34]) from the RGB input, to help at inference time. However, such an approach may be problematic as it relies heavily on the depth generator output regardless of its quality.

Our architecture distills the segmentation-boosting depth information at the training stage, *but not for inference.* The advantage is that there is no need to worry about possible deterioration of depth information at runtime.

Dataset with obstacles in unstructured environments. Most of the known datasets (KITTI, CityScapes, NuScenes,

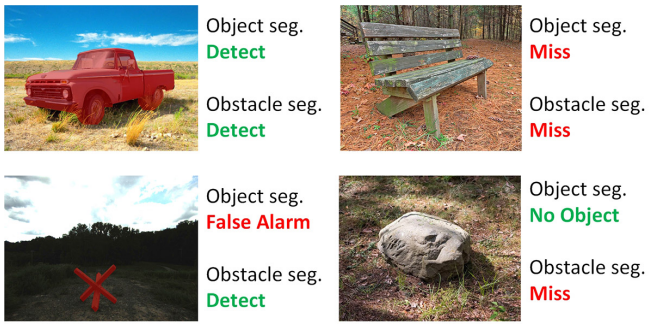

Fig. 3. **Example evaluation for object segmentation and obstacle segmentation.** Red masks represent segmentation results. Evaluation results for the two different tasks are shown on the right side of each image. Text written in green and red indicate whether the result is correct or incorrect, respectively. Note that the pre-defined categories for the object segmentation is assumed to follow the 80 classes from MS COCO dataset which includes 'truck' and 'bench'.

or BDD100K [35], [36], [37], [38]), targeted to assist the perception for autonomous driving, were collected in urban environments, focusing on a narrow set of objects commonly seen during urban driving, such as pedestrians, vehicles, buildings, traffic signs. Although some recent collections have been made within off-road environments [39], [40], they primarily focus on identifying various types of terrain or vegetation for semantic segmentation. IDD [41] dataset sought to include unstructured traffic, but objects were very limited to categories commonly seen in urban driving. While including obstacle-like objects into the dataset, Lost and Found [42] does not carry depth information, and the number of images provided are very small, thus inappropriate for our scenario. In turn, we have constructed a novel dataset, AVoiding Obstacles In unstructured Driving (AvOID).

III. Unsupervised Obstacle Segmentation

Obstacle vs. object. We define an 'obstacle' within the context of autonomous navigation carried out by a vehicle. An obstacle could be *anything* on the paths that hinder the vehicle from moving forward, ranging from man-made barriers to natural objects such as large stones. One may ask why this problem cannot be addressed by a conventional way of identifying objects (detection/segmentation). While such approaches strictly confine the targets to pre-defined labels, we cannot guarantee that an entity always falls into the pre-defined set. For instance, a truck or a bench in Figure 3 can certainly act as an obstacle and its label can easily be identified via running a supervised detection/segmentation algorithm that is trained on a dataset with 'truck' or 'bench' as one of the categories (e.g., MS COCO). However, objects not included in publicly available datasets (e.g., a Czech hedgehog or a large stone in Figure 3) cannot be dealt with by the approaches that solely rely on the pre-defined labels. Because of such differences in definition (predefined objects vs. obstacles), the evaluations for their detection/segmentation must be handled in a different perspective as well.

Unsupervised. There exist myriads of undefined obstacles in the real world, thus, supplying semantic samples of *all possible obstacles* to train a supervised model is almost impossible, and thus not a sustainable solution. To that end, we claim that it is necessary to tackle the problem of the obstacle segmentation with an *unsupervised approach* that can handle open-set/unknown obstacles rather than being limited by a pre-defined set of objects. Note that, however, when considering an unsupervised approach, absolutely no prior information (e.g., obstacle appearance, category) exists, and this makes it inherently more difficult than the object segmentation that can be dealt with supervised approaches.

IV. Methodology

A. Relation Distillation with SCooL

Semantic vs. relation distillation. Distillation is commonly used to transfer knowledge between two networks, often from a teacher to a student. Conventional distillation (called *Semantic Distillation* in Figure 4 (a)) is performed by training a student network to produce similar output that the teacher generates when each network takes in one image patch that shares the same semantic label with the image patch fed into the other network, e.g., two patches cropped from the same image. When training, Cross-entropy loss or Kullback-Leibler divergence are commonly used. Semantic Distillation is effective when transferring categorical knowledge, such as the knowledge required to make predictions on what objects are present in an image, towards the student.

As another type of distillation, we consider *relation distillation* (Figure 4 (b)) where the *relationship* between two input patches are transferred from a teacher to a student network. The difference when compared with Semantic Distillation is that a *pair* of two input patches are fed into *both* student and the teacher instead of feeding one per network.

Capturing relationship: SCooL. We devised a way to directly probe into the *relationship* between the local semantics across a pair of input patches to provide better learning signals for segmentation. Such relationship is captured by our *semantic co-occurrence localization* (SCooL) module that produces the *SCooL map*, $\mathbf{M}_{\text{SCooL}}$, computed by tensor product of embedded features of the two patches, as follows:

$$\mathbf{M}_{\text{SCooL}} = \{p_x^T q_y\}_{x,y}, \tag{1}$$

where p_x and q_y are normalized encoded feature vectors in location x and y for two patches, respectively. Figure 5 shows how a SCooL map is generated. Intuitively, the high value in any bin of $\mathbf{M}_{\text{SCooL}}$ indicates that the two inputs are highly likely to share semantics in their corresponding locations from the perspective of the encoders. In other words, $\mathbf{M}_{\text{SCooL}}$ can reveal where co-occurrences of the two inputs are spatially located in each input.

Optimizing relation distillation. During training, distillation from a teacher to a student network is driven by enforcing the localization similarity loss (i.e., L_{ls}) measuring the similarity between the two relationships (captured by the two SCooL maps):

$$\mathcal{L}_{ls}(\mathbf{T}, \mathbf{S}) = \sum_{x,y} (\mathbf{T} - \delta) \odot \mathbf{S}, \tag{2}$$

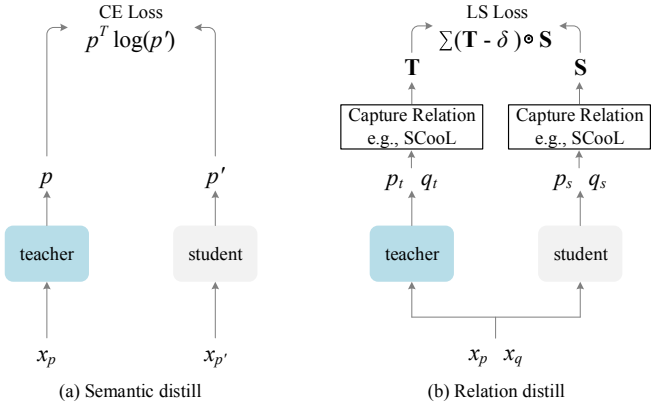

Fig. 4. **Conceptual comparison of two distillations.** The two mechanisms differ in the type of knowledge being transferred. (a) In *Semantic Distillation*, the student distills knowledge of how the teacher understands the semantics of an input image. Here, inputs (x_p and $x_{p'}$) of teacher and student should have the same semantics, e.g., patches cropped at different locations from the same image. (b) In *Relation distillation*, in contrast, the knowledge that focuses on relationship between the paired input images is transferred. To do so, the same image pair (x_p and x_q) needs to be fed into the teacher and the student at the same training instance.

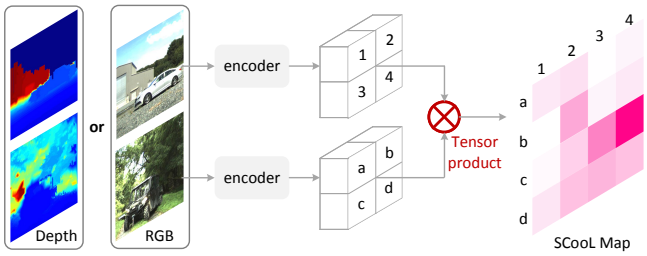

Fig. 5. **Semantic Co-occurrence Localization (SCooL).** Most of the *vehicle area* is located in region **4** of the top image and region **c** of the bottom image. (In this example, each encoded feature is spatially divided into 2×2.) Encoder is responsible for producing a SCooL map in which the semantic co-occurrence coordinate **(4, c)** is highly activated.

where **S** and **T** are the SCooL maps from the student and the teacher, respectively. δ is a hyperparameter pressing negative value in the SCooL relationship for preventing collapse. This loss was also leveraged in [43] but in a different setting where an additional layer on top of a frozen backbone is trained to mimic the capability of the backbone.

Our distillation using SCooL and LS Loss[1] is well suited for the unsupervised segmentation task because the network is encouraged to acquire localization capability, beneficial for segmentation. The two teacher networks in our architecture transfer such capability towards the student while probing into two different modalities (i.e., RGB and depth).

B. Two Teachers of RGB and Depth

Architecture. Our two-teacher relation distillation architecture, which we call the *T2D2* (Two Teacher Distillation with Depth) is equipped with two teacher networks (f_{tr} and f_{td}) to distill the information from RGB and depth modalities towards the student network (f_s).

[1]Hereafter, we will use the term "relation distillation" to represent our own relation distillation architecture that contains SCooL and LS Loss.

The two distillations take the form of a relation distillation that leverages the SCooL process and the LS loss. Similar to the practice done in [44], we augmented the student network with a feed forward network as a head layer. This is to prevent collapse while the student distills two different sources (RGB and depth).

The architecture takes two separate pairs of patches from RGB and depth. While $\{x_p^r, x_q^r\}$ is processed by the RGB teacher and the student network, the corresponding depth pair, $\{x_p^d, x_q^d\}$ is taken by the depth teacher. From these networks, **R**, **D**, and **S** are produced in the form of the SCooL maps and the overall loss ($\mathcal{L}_b$) is computed as:

$$\mathcal{L}_b = \sum_{\mathbf{R},\mathbf{D},\mathbf{S} \in \text{SCooL}(b)} \mathcal{L}_{ls}(\mathbf{R}, \mathbf{S}) + \mathcal{L}_{ls}(\mathbf{D}, \mathbf{S}), \tag{3}$$

where SCooL(b) is the SCooL process that computes for all possible pairs of RGB images and the corresponding depth images in a given batch b.

While f_s and f_{td} get updated when the back-propagation takes place, f_{tr} only gets updated by the exponential moving average (ema) directed from f_s. It has been validated in [45] that slow-progressing teacher updated by the student via ema instead of back-propagation was effective in the teacher-student distillation when they share same architecture. The ema strategy was not used for the depth teacher as it takes non-RGB input, thus not sharing the same architecture.

In-batch pressure generation (IPG). While motivated by [43] where holistic image pairs need to be prepared with an additional pre-processing to generate attractive and repulsive learning signals, we took a completely different/more efficient route to generate self, positive, and negative pressures within each mini-batch at the time of training. For a batch of features (each feature corresponds to a single patch) coming out of a backbone net, we compute cosine similarity between all possible pairs of features and rank them, highest to lowest. Top N pairs, bottom N pairs, and whole self-pairs are used to compute three LS Loss values per modality. These represent positive, negative and self-generated pressures for a given batch, respectively.

Overall loss is modified with the IPG controlling hyper-parameters (i.e., w_{s}^{domain}, w_{p}^{domain} and w_{n}^{domain}, where domain is either r (rgb) or d (depth)), as below:

$$\begin{aligned}\mathcal{L}_b^{ipg} = &\sum_{\substack{\mathbf{R}_\text{s},\mathbf{D}_\text{s},\mathbf{S}_\text{s} \\ \in \text{SCooL(self}(b))}} \left(w_\text{s}^r \mathcal{L}_{ls}(\mathbf{R}_\text{s}, \mathbf{S}_\text{s}) + w_\text{s}^d \mathcal{L}_{ls}(\mathbf{D}_\text{s}, \mathbf{S}_\text{s})\right) \\ &+ \sum_{\substack{\mathbf{R}_\text{p},\mathbf{D}_\text{p},\mathbf{S}_\text{p} \\ \in \text{SCooL(pos}_N(b))}} \left(w_\text{p}^r \mathcal{L}_{ls}(\mathbf{R}_\text{p}, \mathbf{S}_\text{p}) + w_\text{p}^d \mathcal{L}_{ls}(\mathbf{D}_\text{p}, \mathbf{S}_\text{p})\right) \\ &+ \sum_{\substack{\mathbf{R}_\text{n},\mathbf{D}_\text{n},\mathbf{S}_\text{n} \\ \in \text{SCooL(neg}_N(b))}} \left(w_\text{n}^r \mathcal{L}_{ls}(\mathbf{R}_\text{n}, \mathbf{S}_\text{n}) + w_\text{n}^d \mathcal{L}_{ls}(\mathbf{D}_\text{n}, \mathbf{S}_\text{n})\right),\end{aligned} \tag{4}$$

where self(b), pos$_N(b)$ and neg$_N(b)$ are the selection process (within batch b) for self, positive, and negative pairs, respectively. N denotes the number of selected patch pairs

for positive and negative pressures. w_s:w_p:w_n is referred to as the IPG ratio. Note that we have empirically verified that using different IPG ratios for RGB and depth is effective. The optimal IPG parameters (i.e., ratio and N) were determined via the ablation experiments. (Tab. II)

V. Dataset

As we need a dataset that contains both RGB and depth in which obstacles in unstructured environments are present, we collected a new dataset called AVoiding Obstacles In unstructured Driving (AvOID). Although originally collected in urban setting, KITTI [35] also includes objects and depth. In turn, we select a subset to construct KITTI-Obstacle Segmentation (KITTI-OS) tailored for our scenario.

AvOID. The data was collected in a variety of locations, including open flat gravel areas near buildings, dirt paths lined with dense vegetation, forest regions with fallen trees, and a sloped gravel hill. It was acquired at various times (10AM – 5PM) on multiple days to encompass a wide range of daytime lighting/weather conditions. Raw collections which will be publicly released contain full set of proprioceptive and exteroceptive sensing data. A set of peculiar obstacles (along with some general ones) were chosen to create a unique testbed, potentially difficult for the models trained on conventional datasets. Obstacles in the dataset include *person, mannequins, vehicle, building, tent, barrier, barrel, cone, box, metal truss, expandable barrier, Czech Hedgehog, table, chair, jersey barrier, and shovel* along with other obstacles that are difficult to name as shown in Figure 6.

The dataset contains 36,950 pairs of RGB and depth frames (23,070 train and 13,880 test) while obstacles in 614 images (every 20 frames from test set) were manually annotated as groundtruths[2] for evaluation purposes only. To acquire the dense depth images (instead of LiDAR point clouds) necessary for training our architecture, we projected the LiDAR points to the same image plane defined by the RGB camera and interpolated the missing regions in each of the images. Samples of RGB and depth image pairs in the train set, and sample images with their groundtruths for evaluation (test set) are included in Figure 2.

We have used the Clearpath Warthog to collect the dataset which was equipped with Ouster OS1-128 LiDAR with vertical and horizontal resolution of 128 and 1024, respectively, and the FLIR Blackfly S color camera (for RGB) with resolution of 1440×1080. The two sensors were synchronized using Precision Time Protocol (PTP) synchronization, and are calibrated using the technique described in [46].

KITTI-OS. Train set includes RGB images extracted from 54 videos and their corresponding dense depth images, generating 23,977 frame pairs (i.e., RGB & depth) in total. While the RGB images are provided from the original raw dataset, the dense depth images were newly constructed by interpolating the originally provided sparse LiDAR points. To construct the test set, we selected a subset of labeled categories that can be considered as obstacles in the 200 images provided by KITTI Semantic Segmentation Challenge. The selected categories are *sidewalk, building, wall, fence, guard rail, bridge, pole, polegroup, traffic sign, person, rider, car, truck, bus, caravan, trailer, train, motorcycle,* and *bicycle*. Pixel locations for the selected categories in each image were relabeled as 'obstacle' and these were used as the ground truth for the evaluation purposes.

[2]Annotating groundtruth only for the test set is solely in consideration of our task, unsupervised obstacle segmentation. Note that AvOID is bigger than any other comparable obstacle-containing datasets. To account for the potential use of this dataset for various supervised tasks, we plan to include groundtruth for the train set in the same way we did for the test set.

Fig. 6. **Subset of obstacles included in the AvOID dataset.** The obstacles are cropped from the original RGB images for the presentation purpose only.

VI. Experiments

Baseline. We chose DINO [1] as the primary baseline mainly as it does not have any architectural constraint that hinders the usage of a different modality other than RGB. Other unsupervised segmentation methods follow their own design philosophies, which limit the scalability towards the depth. In addition, since DINO contains a flavor of Semantic Distillation, we could demonstrate the advantage of our relation distillation technique. As an augmented baseline, we constructed 'DINO-2Teach', that adds an additional depth teacher on top of vanilla DINO.

While our initial scope was to focus on segmentation models that can be trained via *distillation+NO supervision*, we also compared with two recent knowledge distillation (KD) methods that fall into SD category, StructKD (SKD) [47] and TransKD (TKD) [48]. Note that, SKD and TKD were originally proposed to distill knowledge *w/ supervision*. For a fair comparison with our model with depth-assisted 2-teacher paradigm *w/ NO supervision*, we adapted all modules/losses in SKD and TKD into equivalent 2-teacher+unsupervised framework, using author provided codes.

Measuring accuracy. After summing up the ViT head output features (from multi-headed attention) at each spatial coordinates, it is thresholded to keep 80% of the mass, following

Method	Depth	Distill.	Accuracy
StructKD	✓	sem.	0.17
TransKD	✓	sem.	17.09
DINO	-	sem.	17.21
DINO-2Teach	✓	sem.	18.11
RD-RGB	-	rel.	20.39
+ IPG*	-	rel.	23.29
T2D2	✓	rel.	20.94
+ IPG	✓	rel.	24.83

TABLE I

EFFECTIVENESS OF THE RELATION DISTILLATION AND DEPTH. SEM. AND REL. DENOTE THE SEMANTIC DISTILLATION AND THE RELATION DISTILLATION, RESPECTIVELY. 'RD-RGB' USES THE RELATION DISTILLATION ONLY WITH RGB IMAGES. IPG* USES OPTIMAL RATIO 0:2:1 BETWEEN RGB TEACHER (f_tr) AND THE STUDENT(f_s).

w/ f_{tr}	w/ f_{td}	Acc.
0:2:1	**0**:2:1	23.58
1:2:1	**0**:2:1	21:24
0:2:1	**1**:2:1	**24.83**
1:2:1	**1**:2:1	24.50
0:**1**:1	1:**1**:1	23.01
0:**2**:1	1:**2**:1	**24.83**
0:**3**:1	1:**3**:1	22.89

(a) Varying the IPG ratio

N	Acc.
10	23.83
30	**24.83**
50	20.63

(b) Varying N

TABLE II

IPG PARAMETER STUDY. (A) 'W/ f_{tr}' AND 'W/ f_{td}' CORRESPOND TO THE TWO CONNECTIONS WITHIN T2D2 THAT THE f_s IS INVOLVED WITH. (B) N INDICATES THE NUMBER OF POSITIVE AND NEGATIVE PAIRS USED WITHIN THE IPG. IPG RATIOS WERE FIXED WITH THE OPTIMAL RATIOS IN (A).

w/o pre-training	w/ pre-training	Gain
24.83	19.98	-4.85

TABLE III

THE EFFECT OF PRE-TRAINING. ALL METHODS USE IPG. (FOR 'W/ PRE-TRAINING', IPG IS USED IN THE FINE-TUNING STAGE.)

[1]. This final map is compared with the corresponding ground truth to compute the accuracy using the Intersection-over-Union(IoU).

A. Results on AvOID

Relation Distillation (RD) vs. Semantic Distillation (SD). As shown in Table I, our RD consistently presents better accuracy than SD with or without the depth by a significant margin. This demonstrates distilling the *relation* that captures the semantic co-occurrence localization is indeed effective for the segmentation. Moreover, the addition of IPG improved the accuracy with another significant jump of 3.89 (from T2D2 to T2D2+IPG). The performance advantage using the IPG is made possible because the distillation is changed from SD to RD. The fact that SKD extremely underperformed shows that its modules failed to guide proper learning of ViT attentions as they were specifically designed for CNNs, unlike TKD designed for ViTs.

The depth effect. Table I shows that leveraging depth is consistently effective for better accuracy regardless of the distillation type. Note that supplementing depth on top of RD-RGB *without* the IPG does not bring significant gain (0.55). But the gain increases significantly (i.e., 3.89 from T2D2 and 4.44 from RD-RGB) when supported by the IPG, validating that relation distillation can be better utilized when effective learning signals are provided.

Ablation: inspecting the IPG. IPG is driven by three factors: self, positive, and negative pairs. Ratio between the three can be controlled separately within the two connections in T2D2, i.e., distillation from the RGB teacher (i.d., w/ f_{tr}) and distillation from the Depth teacher (i.d., w/ f_{td}).

The top four results in Table II (a) show that removing the 'self' pairs only from 'w/ f_{tr}' brought the best result, which shares the same spirit with the DINO [1] in which the 'self' patch-pairs were intentionally left out to generate effective self-supervision signals. As the ratio between the positive and negative pairs was fixed at $2:1$, we went further to verify if having more positive pairs would be beneficial, and this is shown in the last 3 rows of Table II (a). Based on these experiments, we use a ratio of 0:2:1 for the student-to-RGB teacher connection and a ratio of 1:2:1 for the student-to-Depth teacher connection throughout all experiments.

Instead of blindly leveraging all the features, only a selected number (N) of features serve as learning signals for a given batch. Using a small N value means that the in-batch pressures are led by strongly coupled features, whereas using a big N value allows loosely tied samples to get involved. Table II (b) shows that $N = 30$ led to the best performance. The performance starts to drop rather abruptly when using a bigger N value. This overall trend suggests that strongly tied feature pairs act as effective pressures within the IPG.

Is it necessary for the teacher to acquire segmentation ability before distillation? We ran an experiment to verify if it is beneficial for the backbone of the teacher to be pre-trained prior to the overall training. We adopted a cascaded approach, first to separately pre-train the RGB-teacher and the depth-teacher, followed by a fine-tuning stage on the T2D2. We used the DINO framework to pre-train each of the RGB and depth teachers, using RGB or depth maps as training data, respectively. The pre-training generates two networks (teacher and student) for each modality. Pre-trained RGB teacher and student is used to finetune the new RGB teacher and student in T2D2 and pre-trained depth student is used for finetuning the depth teacher. In Table III, we see that the additional effort in training the pre-trained models brings an adverse effect towards the accuracy, and that training T2D2 from scratch is more than sufficient.

Masking out non-LiDAR regions. We consider a simple technique to improve the accuracy with minuscule addition to our model, while maintaining the initial motivation of not leveraging any ground truth annotations in either train or test mode. Based on the fact that LiDAR sensors are only capable

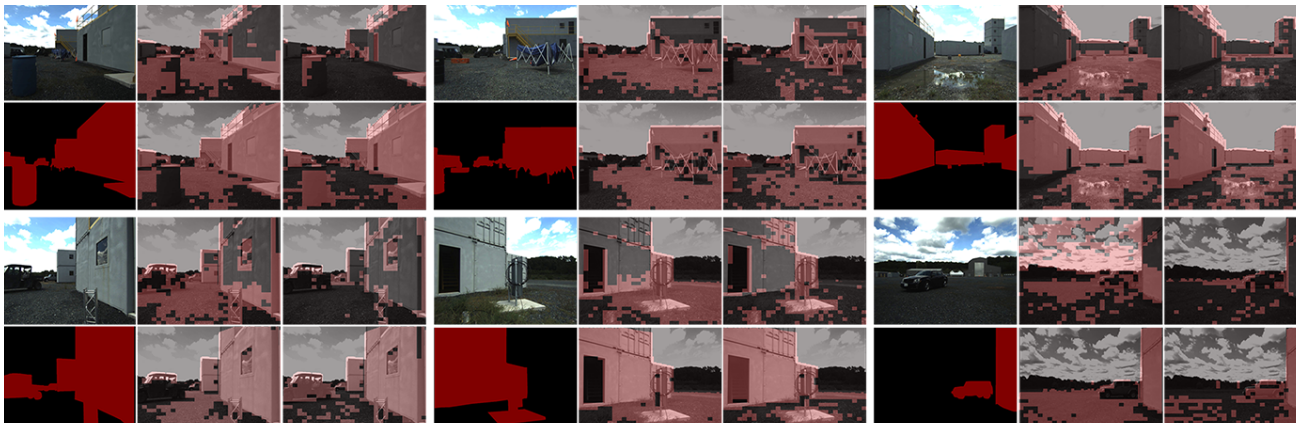

Fig. 7. **Examples of segmentation results.** Each set consists of 2×3 images, with the input image and the corresponding ground truth segmentation map shown at the top and bottom of the 1st column, respectively. Segmentation results of DINO, DINO-2Teach, RD-RGB, and T2D2 are shown in the top of the 2nd column, the top of the 3rd column, the bottom of the 2nd column, and the bottom of the 3rd column, respectively. Results shown in this figure have non-LiDAR regions masked.

Method	NoMask	Mask	Gain
DINO-2Teach	18.11	25.76	+7.65
T2D2	20.94	40.77	+19.83
+ IPG	24.83	42.01	+17.18

TABLE IV

MASKING OUT.

Method	Distill.	Depth	IPG	Acc.	Margin
DINO	sem.	-	-	20.52	
DINO-2Teach	sem.	✓	-	22.61	+2.09
T2D2	rel.	✓	-	21.52	-1.09
T2D2+IPG	rel.	✓	✓	**23.10**	+1.58

TABLE V

RESULTS ON THE KITTI-OS. METHODS ARE LISTED IN THE ORDER OF EVOLVING BY ADDING OR CHANGING ONE COMPONENT AT A TIME FROM DINO TO T2D2. MARGIN REPRESENTS THE ACCURACY DIFFERENCE W.R.T PREVIOUS METHOD.

of acquiring the information only within certain range (120 meters for the case of AvOID), we have observed that the output segmentations can be improved just by masking out the non-LiDAR regions in the image based on the intuition that obstacles do not appear in regions which are void of LiDAR output (e.g., sky). Table IV also shows how much performance can be gained when this simple post-processing is leveraged at run-time. KITTI-OS is not considered here as the test set only contains RGB images.

Qualitative analysis. Figure 7 shows examples of segmentation results from our T2D2 and three comparison methods. First, methods using the relation distillation (i.e., RD-RGB and T2D2) segmented the planar areas with crispier boundaries. Second, methods that distilled from the depth information (i.e., DINO-2Teach and T2D2) demonstrate better ability in removing the ground regions from the positives. Lastly, it is consistently shown in several examples that our T2D2 outperforms all the baselines in a qualitative manner, coinciding with the quantitative analyses.

B. Results on KITTI-OS

Table V compares our T2D2 with other baselines. First, DINO-2Teach performs better than DINO, indicating that the use of depth information in distillation was also effective for unsupervised obstacle segmentation on the relatively urban setting. Second, when proper training signals are not leveraged, the relation distillation framework itself did not demonstrate its advantage and even showed degraded accuracy when compared to the Semantic Distillation. Third, consequently, T2D2 with IPG appropriately showed the best accuracy among all comparison methods by a margin which is not negligible. This demonstrates that the effects of the three technical contributions claimed in this paper (i.e., depth usage, relation distillation, and IPG) are similarly applicable to the urban environments depicted within the KITTI-OS.

C. Computational Efficiency

Table VI shows the computational efficiency for different models while carrying out train or inference using Nvidia RTX 3090. While the train efficiency may vary due to training modules/losses, it is the same for all methods as using the same architecture (ViT-tiny based student networks).

method	train			inference			
	arch	# param. (M)	sec/iter	# param. (M)	sec/im	FLOPs (G)	mem. (GB)
Semantic distill.							
TransKD	+DT, 4 LMs	17.12	1.07				
StructKD	+DT, Discr	19.95	0.91				
DINO	·	11.18	0.88	5.52	0.0057	1.26	2.64
DINO-2Teach	+DT	16.77	0.89				
Relation distill.							
RD-RGB	·	11.18	0.89				
T2D2	+DT	16.77	0.90				

TABLE VI

Efficiency. PRESENCE OF IPG DOES NOT AFFECT THE EFFICIENCY, THUS OMITTED. 'DT', '4 LMS', AND 'DISCR' ARE THE DEPTH TEACHER, 4 LEARNABLE MODULES, AND DISCRIMINATOR, RESPECTIVELY.

VII. Conclusion

We proposed a novel architecture, T2D2, containing three modules to improve the unsupervised obstacle segmentation performance: i) relation distillation with SCooL to capture effective features for segmentation, ii) use of depth information in training only via two teacher framework, and iii) IPG module to generate better learning signals for distillation. Our experiments demonstrate that these contributions are effective for the given task on both AvOID and KITTI-OS. While our current model indirectly acquires the segmentation ability via the relation distillation-based representation learning, we expect to acquire additional gain in accuracy by introducing more direct methods to assist the unsupervised segmentation in our follow-up research.

References

[1] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," in *Proc. ICCV*, 2021.

[2] C. Yang, L. Xie, C. Su, and A. L. Yuille, "Snapshot distillation: Teacher-student optimization in one generation," in *Proc. CVPR*, 2019.

[3] J. H. Cho and B. Hariharan, "On the efficacy of knowledge distillation," in *Proc. ICCV*, 2019.

[4] F. Tung and G. Mori, "Similarity-preserving knowledge distillation," in *Proc. ICCV*, 2019.

[5] Z. Shen, Z. He, and X. Xue, "MEAL: Multi-model ensemble via adversarial learning," in *Proc. AAAI*, 2019.

[6] N. Passalis, M. Tzelepi, and A. Tefas, "Heterogeneous knowledge distillation using information flow modeling," in *Proc. CVPR*, 2020.

[7] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proc. CVPR*, 2020.

[8] S. Shin, J. Lee, J. Lee, Y. Yu, and K. Lee, "Teaching where to look: Attention similarity knowledge distillation for low resolution face recognition," in *Proc. ECCV*, 2022.

[9] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," in *Proc. NeurIPS Workshop*, 2014.

[10] Y. Tian, D. Krishnan, and P. Isola, "Contrastive representation distillation," in *Proc. ICLR*, 2020.

[11] Y. Hong, H. Dai, and Y. Ding, "Cross-modality knowledge distillation network for monocular 3D object detection," in *Proc. ECCV*, 2022.

[12] Y. Mao, W. Zhou, Z. Lu, J. Deng, and H. Li, "CMD: Self-supervised 3D action representation learning with cross-modal mutual distillation," in *Proc. ECCV*, 2022.

[13] S. Zhao, J. Yu, Z. Sun, B. Zhang, and X. Wei, "Enhanced accuracy and robustness via multi-teacher adversarial distillation," in *Proc. ECCV*, 2022.

[14] L. Wang, X. Li, Y. Liao, Z. Jiang, J. Wu, F. Wang, C. Qian, and S. Liu, "HEAD: Hetero-assists distillation for heterogeneous object detectors," in *Proc. ECCV*, 2022.

[15] Z. Xue, S. Ren, Z. Gao, and H. Zhao, "Multimodal knowledge expansion," in *Proc. ICCV*, 2021.

[16] S. Du, S. You, X. Li, J. Wu, F. Wang, C. Qian, and C. Zhang, "Agree to disagree: Adaptive ensemble knowledge distillation in gradient space," in *Proc. NeurIPS*, 2020.

[17] S. Tang, Z. Zhang, Z. Cheng, J. Lu, Y. Xu, Y. Niu, and F. He, "Distilling object detectors with global knowledge," in *Proc. ECCV*, 2022.

[18] X. Li, J. Wu, H. Fang, Y. Liao, F. Wang, and C. Qia, "Local correlation consistency for knowledge distillation," in *Proc. ECCV*, 2020.

[19] Y. Liu, J. Cao, B. Li, C. Yuan, W. Hu, Y. Li, and Y. Duan, "Knowledge distillation via instance relationship graph," in *Proc. CVPR*, 2019.

[20] W. Park, D. Kim, Y. Lu, and M. Cho, "Relational knowledge distillation," in *Proc. CVPR*, 2019.

[21] J. Liu, H. Qin, Y. Wu, J. Guo, D. Liang, and K. Xu, "CoupleFace: Relation matters for face recognition distillation," in *Proc. ECCV*, 2022.

[22] Z. Murez, T. van As, J. Bartolozzi, A. Sinha, V. Badrinarayanan, and A. Rabinovich, "Atlas: End-to-end 3d scene reconstruction from posed images," in *Proc. ECCV*, 2020.

[23] W. Yin, J. Zhang, O. Wang, S. Niklaus, S. Chen, Y. Liu, and C. Shen, "Towards accurate reconstruction of 3D scene shape from a single monocular image," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2022.

[24] B. Li, S. Wang, H. Ye, X. Gong, and Z. Xiang, "Cross-modal knowledge distillation for depth privileged monocular visual odometry," *IEEE Robotics and Automation Letters*, vol. 7, no. 3, pp. 6171–6178, 2022.

[25] H. Zhan, C. S. Weerasekera, J.-W. Bian, and I. Reid, "Visual odometry revisited: What should be learnt?" in *Proc. ICRA*, 2020.

[26] Z. Min, Y. Yang, and E. Dunn, "VOLDOR: Visual odometry from log-logistic dense optical flow residuals," in *Proc. CVPR*, 2020.

[27] S. Li, X. Wu, Y. Cao, and H. Zha, "Generalizing to the open world: Deep visual odometry with online adaptation," in *Proc. CVPR*, 2021.

[28] K.-C. Huang, T.-H. Wu, H.-T. Su, and W. H. Hsu, "MonoDTR: Monocular 3D object detection with depth-aware transformer," in *Proc. CVPR*, 2022.

[29] S. Qiao, Y. Zhu, H. Adam, A. Yuille, and L.-C. Chen, "ViP-DeepLab: Learning visual perception with depth-aware video panoptic segmentation," in *Proc. CVPR*, 2021.

[30] N. Gao, F. He, J. Jia, Y. Shan, H. Zhang, X. Zhao, and K. Huang, "PanopticDepth: A unified framework for depth-aware panoptic segmentation," in *Proc. CVPR*, 2022.

[31] W. Bao, W.-S. Lai, C. Ma, X. Zhang, Z. Gao, and M.-H. Yang, "Depth-aware video frame interpolation," in *Proc. CVPR*, 2019.

[32] H. Fu, M. Gong, C. Wang, K. Batmanghelich, and D. Tao, "Deep ordinal regression network for monocular depth estimation," in *Proc. CVPR*, 2018.

[33] Y. Wang, W.-L. Chao, D. Garg, B. Hariharan, M. Campbell, and K. Q. Weinberger, "Pseudo-LiDAR from visual depth estimation: Bridging the gap in 3D object detection for autonomous driving," in *Proc. CVPR*, 2019.

[34] Y. You, Y. Wang, W.-L. Chao, D. Garg, G. Pleiss, B. Hariharan, M. Campbell, and K. Q. Weinberger, "Peudos-LiDAR++: Accurate depth for 3D object detection in autonomous driving," in *Proc. ICLR*, 2020.

[35] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *International Journal of Robotics Research (IJRR)*, 2013.

[36] 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 *Proc. CVPR*, 2016.

[37] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuscenes: A multimodal dataset for autonomous driving," in *Proc. CVPR*, 2020.

[38] F. Yu, H. Chen, X. Wang, W. Xian, Y. Chen, F. Liu, V. Madhavan, and T. Darrell, "BDD100K: A diverse driving dataset for heterogeneous multitask learning," in *Proc. CVPR*, 2020.

[39] M. Wigness, S. Eum, J. G. Rogers, D. Han, and H. Kwon, "A rugd dataset for autonomous navigation and visual perception in unstructured outdoor environments," in *Proc. IROS*, 2019.

[40] P. Jiang, P. Osteen, M. Wigness, and S. Saripalli, "Rellis-3d dataset: Data, benchmarks and analysis," in *Proc. ICRA*, 2021.

[41] G. Varma, A. Subramanian, A. Namboodiri, M. Chandraker, and C. V. Jawahar, "Idd: A dataset for exploring problems of autonomous navigation in unconstrained environments," in *Proc. WACV*, 2018.

[42] P. Pinggera, S. Ramos, S. Gehrig, U. Franke, C. Rother, and R. Mester, "Lost and found: detecting small road hazards for self-driving vehicles," in *IROS*, 10 2016, pp. 1099–1106.

[43] M. Hamilton, Z. Zhang, B. Hariharan, N. Snavely, and W. T. Freeman, "Unsupervised semantic segmentation by distilling feature correspondences," in *Proc. ICLR*, 2022.

[44] X. Chen and K. He, "Exploring simple siamese representation learning," in *Proc. CVPR*, 2021.

[45] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proc. CVPR*, 2020.

[46] J. L. Owens, P. R. Osteen, and K. Daniilidis, "MSG-cal: Multi-sensor graph-based calibration," in *Proc. ICRA*, 2015.

[47] Y. Liu, K. Chen, C. Liu, Z. Qin, Z. Luo, and J. Wang, "Structured knowledge distillation for semantic segmentation," in *Proc. CVPR*, 2019.

[48] R. Liu, K. Yang, A. Roitberg, J. Zhang, K. Peng, H. Liu, and R. Stiefelhagen, "Transkd: Transformer knowledge distillation for efficient semantic segmentation," 2022.