MonoCInIS: Camera Independent Monocular 3D Object Detection using Instance Segmentation

Jonas Heylen[1] Mark De Wolf[1] Bruno Dawagne[1] Marc Proesmans[1,2] Luc Van Gool[2,3]
Wim Abbeloos[4] Hazem Abdelkawy[4] Daniel Olmeda Reino[4]

[1]TRACE vzw [2]KU Leuven/ESAT-PSI [3] ETHZ/CVL [4]Toyota Motor Europe
{jonas.heylen,mark.dewolf,bruno.dawagne}@trace.vision
{marc.proesmans,luc.vangool}@esat.kuleuven.be
{Wim.Abbeloos,Hazem.Abdelkawy,Daniel.Olmeda.Reino}@toyota-europe.com

Abstract

Monocular 3D object detection has recently shown promising results, however there remain challenging problems. One of those is the lack of invariance to different camera intrinsic parameters, which can be observed across different 3D object datasets. Little effort has been made to exploit the combination of heterogeneous 3D object datasets. In contrast to general intuition, we show that more data does not automatically guarantee a better performance, but rather, methods need to have a degree of 'camera independence' in order to benefit from large and heterogeneous training data. In this paper we propose a category-level pose estimation method based on instance segmentation, using camera independent geometric reasoning to cope with the varying camera viewpoints and intrinsics of different datasets. Every pixel of an instance predicts the object dimensions, the 3D object reference points projected in 2D image space and, optionally, the local viewing angle. Camera intrinsics are only used outside of the learned network to lift the predicted 2D reference points to 3D. We surpass camera independent methods on the challenging KITTI3D benchmark and show the key benefits compared to camera dependent methods.

1. Introduction

Predicting accurate 3D object position and orientation is crucial in the context of autonomous systems that interact with a set of objects in a common environment. A particularly relevant application is pose estimation of vehicles in autonomous driving. Where most of the initial efforts have been based on high precision LiDAR and stereo vision, simpler setups based on monocular vision have gained interest. Nevertheless, 3D pose estimation from monocular views remains a challenging task, as it is largely an ill-posed problem. Recently, a number of large 3D detection datasets for autonomous driving have been made available. Starting with KITTI3D [1], others have followed, namely CityScapes3D [2], nuScenes [3], Waymo [4], or Lyft [5]. This provides an opportunity to exploit training data from heterogeneous sources, but it also suggest the need for methods which are able to handle this variety of cameras, with different intrinsics and viewing characteristics. We refer to these methods as **camera independent**. To the best of our knowledge, we are the first to investi-

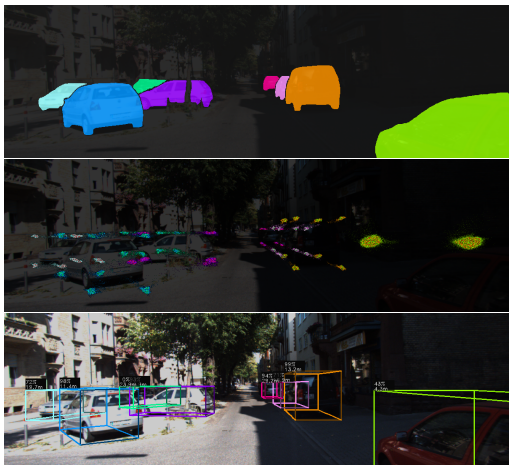


Figure 1: Our method leverages instance segmentation predictions (top), in which every pixel predicts the object's dimensions and projected 3D reference points (mid). Camera independent geometric reasoning lifts these to 3D (bottom).

gate the effects of combining large datasets in the context of monocular 3D object detection for automated driving applications.

Many state-of-the-art (SOTA) methods depend on regressing depth directly, either through estimating a full depth map or individual object distances. This results in the network learning an internal representation of the camera intrinsics, by linking position and scale in the image to real world depth. We show that, for such camera dependent methods, more data doesn't necessarily result in better performance. These methods lack the ability to handle an arbitrary number of views, but rather learn a few different camera models, while not being able to scale and generalise to any camera model. An extensive overview on camera models can be found in Sturm *et al*. [6]. On the other hand, some methods rely solely on camera independent geometric reasoning, leveraging known camera intrinsics outside the learned network. However, existing camera independent methods lack performance.

In this work, we propose a pose estimation method which is able to take advantage of the high scene variability of multiple datasets. It is based on proposal free **instance segmentation** which avoids the need for Non-Maximum-Suppression (NMS). Every pixel of an object's instance predicts multiple camera independent representation attributes, such as the object dimensions, the 3D reference points (RPs) projected into 2D image space and, optionally, a local viewing angle. Each pixel's vote contributes to the final prediction. Camera intrinsics are used outside the learned network along with simple geometric reasoning to uplift the predicted 2D RPs to 3D. We show that our method outperforms existing camera independent methods and achieves similar results as last year's camera dependent methods on the challenging KITTI3D benchmark. Recent works however surpass the performance of our method by strongly focusing on the KITTI3D dataset, but they lack the ability to generalise over heterogeneous cameras. We further investigate the trade-off between accuracy and speed, which is an important aspect for practical applications in the context of autonomous driving. Our main contributions can be summarised as follows:

- We propose a category-level 3D pose estimation method based on instance segmentation. Each pixel of an instance votes for all attributes, resulting in distributions with confidence estimates.
- To the best of our knowledge, we are the first to investigate the effects of combining different datasets in the context of monocular 3D object detection.
- We show qualitatively that our approach generalises well over different camera types such as fisheyes, even without fisheye training data.
- We release instance segmentation annotations for the KITTI3D dataset.

2. Related work

This section briefly reviews related works on 3D object detection using LiDAR data, stereo images, depth, 3D shape information and monocular images.

LiDAR data and stereo images. In the field of autonomous driving, best results on 3D object detection challenges [1, 3, 4] are achieved by methods using LiDAR data [7, 8], which can benefit from having reliable depth information. Stereo images can also provide depth information [9, 10], to even mimic point cloud data based on RGB images only, leveraging the possibility to use existing LiDAR-based methods on so-called Pseudo-LiDAR point clouds [11, 12].

Monocular images. Having no depth sensor data available, 3D object detection based on monocular images only is very challenging. Different strategies have been used to tackle this ill-posed problem. One of the first approaches in monocular 3D object detection was introduced by Deep-Box3D [13], which solves 3D translation using *geometric constraints* by predicting 3D orientation and dimensions for each 2D proposal. Several other works follow this approach and extend it in several ways, for example by visual cues [14], solving a closed form solution [15], or integrating the 3D reconstruction into the network by reprojecting the predicted 3D box in both image space as Bird's Eye View (BEV) [16]. Also the use of segmentation masks leads to improved results for this approach [17, 18]. Other methods adopt a *BEV* to predict bounding boxes [19, 20, 21]. CaDDN[22] uses categorical depth distribution for each pixel to project contextual feature information to the appropriate depth interval in 3D space, and then uses a BEV projection to produce the final output bounding boxes.

Another approach is to *use a 2D detector*, and predict a 3D bounding box for each proposal. These works usually predict direct depth per detected 2D bounding box [23, 24]. MonoGRNet [25] consists of four sub-networks: 2D detection, instance depth estimation, 3D location estimation and local corner regression. ROI-10D [26] proposes a novel loss by lifting 2D detections, orientation and scale estimation into 3D space. MonoDIS [27] proposes a two-stage method, disentangling dependencies of different parameters by introducing a novel loss enabling to handle them separately. SS3D [28] and M3D-RPN [29] are single-stage and feed the predictions to a 3D bounding box optimizer. MoVi-3D [30] generates virtual views where the object appearance is normalized with respect to distance from camera, reducing the visual appearance variability and relieving the model from learning depth-specific representations. M3DSSD[31] introduces feature alignment to avoid mismatching, and asymmetric non-local attention. MonoDLE [32] investigates the misalignment between the center of the 2D bounding box and the projected center of the 3D object, and argues to remove distant objects since they mislead the net-

work. Other methods use *3D Shape Information* [33]. Deep MANTA [34] consists of three levels of box refinement and predicts template similarity to known 3D shapes. 3D-RCNN [35] learns a low-dimensional shape-space from a collection of 3D shape models. Mono3D++ [36] jointly optimises the 3D-2D consistency and task priors like monocular depth, ground plane projection and morphable shape model pose estimation.

In [37], *2D keypoints* of interest are predicted together with 2D bounding boxes, coarse orientations and dimensions. 3D boxes are reconstructed after polling predefined ground planes. RTM3D [38] predicts 2D keypoints of the eight corners and 3D center as heatmaps. The 3D bounding box is computed using geometric constraints of the perspective projection. In [39], a structured polygon of 2D keypoints is predicted and projected in 3D using the object height prior. KeypointNet [40] describes a method to extract 3D keypoints from a single image using geometric reasoning. SMOKE [41] predicts a 3D bounding box for each detected object by combining single keypoint estimates with regressed 3D variables. In addition, MonoPair [42] predicts virtual pairwise constraint keypoints, the middle point of any two objects if they are the nearest neighbors. [43] predicts keypoints using heatmaps. LiteFPN [44] proposes a generic Lite-FPN module that conducts multi-scale feature fusion for their keypoint-based detectors.

In order to overcome the lack of real 3D information, several methods have included the prediction of *depth from monocular images* in a sub-network [45, 46, 47, 48, 49]. One work uses Structure from Motion (SfM) cues together with 2D object detection to predict 3D bounding boxes [50]. DDMP [51] predicts depth-dependent filter weights and affinity matrices for information propagation. D4LCN [52] introduces depth-guided filters which are learned from image-based depth maps. In [53], additional clues are identified from the ground plane and inserted in the depth reasoning to improve the positioning of the bounding boxes. MonoEF [54] provides a mechanism to cope with ego-car pose changes w.r.t. ground plane.

Some of the more recent methods include a form of *uncertainty reasoning* to improve robustness. MonoRUn [55] proposes a robust KL loss that minimizes the uncertainty-weighted reprojection error of the 3D coordinates onto the image plane. MonoFlex[56] uses edge fusion to decouple the feature learning and to predict truncated objects. Object depth is estimated using an uncertainty-guided ensemble of directly regressed depth and solved depths from different groups of keypoints.

Segmentation based 6D pose estimation. A number of works propose the use of semantic or instance segmentation as a means to estimate the 6D object pose. PoseCNN [57] uses segmentation as detector, and predicts the object center by having each pixel of an instance vote for the direction. The distance is computed as the average of all pixels' votes. ConvPoseCNN [58] extends the method by predicting also the rotation for every pixel. LieNet [59] proposes the decoupling of pose parameters so that the rotation can be regressed via Lie algebra representation. Several works use a similar approach to predict 2D keypoints, and use a PnP solution to compute the 3D pose [60, 61, 62]. Optionally depth can be predicted as well [63]. Most of these works assume prior knowledge on object dimensions, shape or 3D cad models. They are designed to operate on specific objects, not on whole categories. A broad overview on the current state-of-the-art on 6D pose detection and tracking is presented by Fan et al. [64]. They cover the topic in the more general context, for applications ranging from autonomous driving, robotics and augmented reality.

Camera independence. In this work, we define camera independence as 'not using depth or indirectly learning camera intrinsics from data', or alternatively, focusing on features that can be learned from appearance such as object size or 2D reference points. Within the monocular methods, several older methods meet this definition such as DeepBox3D [13], Deep MANTA [34] and MonoGRNet2 [33]. Most of them use geometric reasoning, prior shape information or 2D keypoints. Other methods could possibly also become camera independent by leaving out the camera dependent part of their system [15]. However, these methods mostly have poor performance, or don't report results.

3. Method

We propose a novel camera independent approach using instance segmentation as a detection mechanism. Section 3.1 covers the general concepts and geometric reasoning to achieve camera independence. Section 3.2 illustrates how those concepts are applied in our instance-based approach.

3.1. Towards camera independence

The main advantage of camera independence is the ability to combine training data from different datasets. In the scenario of autonomous vehicles, this means one network can handle a combination of multiple viewpoints and cameras, ranging from simple pinhole models to fisheye cameras and cylindrical projections [6] with different fields of view (FOV). Our method is able to generalize pose estimates for never-seen cameras (e.g. Fig. 7). This is useful for data where no ground truth is available during training, e.g. when a long-range camera view is out of LiDAR sensor range.

Object Dimensions. As described in Section 2, convolutional neural networks (CNNs) are able to estimate object dimensions (h, w and l in Figure 4) based on visual appearance. Training a network with images containing different viewpoints and crops, forces the network to become inde-

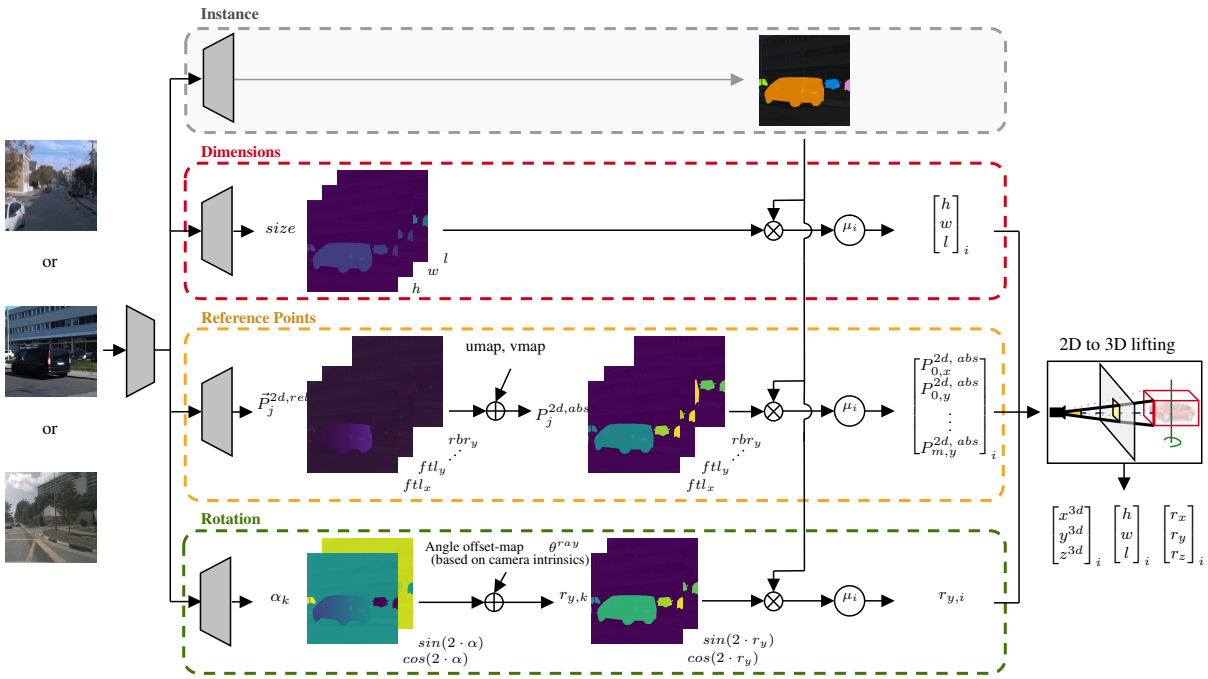


Figure 2: Overview of our method. $\oplus$ denotes an addition: adding each pixel's coordinate to the predicted relative offset map or adding the angle offset for every pixel to the estimated viewing angles. $\otimes$ denotes masking out the predictions for each instance. μ_i represents the operation for averaging the predictions over all pixels belonging to each instance.

pendent of camera-specific features or some internal representation of depth. Based on appearance only, the network can estimate the size of an object, regardless of the perspective of the image.

2D Reference Points. Most recent approaches directly predict the 3D distance for each object. Regressing this distance violates the above explained camera independence goal. We overcome this issue by predicting *Reference Points* (RPs) in the 2D image. These RPs are the 2D projections of predefined RPs in 3D, related to the object. A CNN can estimate the 2D RPs based on visual appearance rather than by learning a representation of the camera intrinsics, again contributing to the camera independence. This work explores two variants of predefined RPs. The first, *8RP*, contains the 8 corners of the 3D bounding box surrounding the object. The second, *2RP*, contains the top center and the bottom center of the 3D bounding box. Note that other combinations are possible.

Object Rotation. Multiple works have discussed the advantages of estimating the allocentric rather than the egocentric pose [26, 14]. We predict the viewing angle, which constraints cars to be parallel to the ground-plane. Figure 3 shows the relationship between the *viewing angle* α_{center} (allocentric) and the *yaw* r_y (egocentric) of an object. As in [13], estimating the sine and cosine avoids the discontinuity between $0°$ and $360°$. In some cases, the network is not able to distinguish the left and right side or the front and

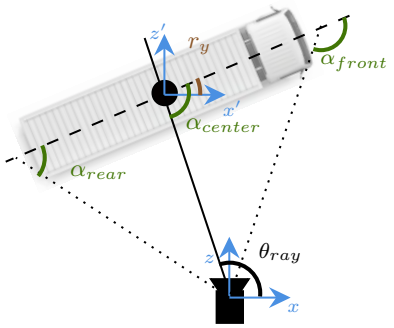


Figure 3: Unique viewing angle (per pixel).

rear side of an object. When the network is confused between these $180°$ alternatives, it might predict the average viewing angle, resulting in a $90°$ offset. We disambiguate these cases by estimating the sine and cosine of $2 \cdot \alpha$, resulting in a correct rotation without heading. To determine the correct heading, the sine and cosine of α could additionally be predicted and projected onto the $2 \cdot \alpha$ vector. In this work, object rotation is required for the 2RP variant and optional for the 8RP variant.

Uplifting 2D to 3D. Several existing methods can be used to map the 2D predicted RPs to a 3D object pose [60, 65, 66, 61, 62]. We propose to calculate the 3D location

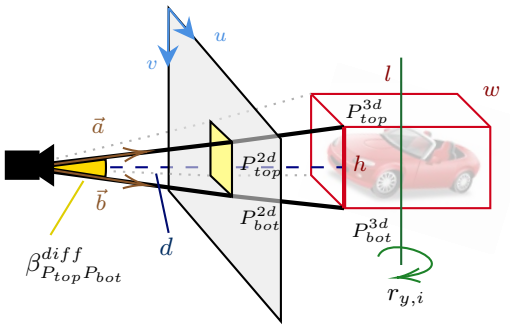


Figure 4: Used notations and 2D to 3D lifting.

of a pair of top-bottom RPs using simple trigonometry, as shown in Figure 4. First, the angle $\beta^{diff}_{P_{top}P_{bot}}$ between the pixel rays $\vec{a}$, $\vec{b}$ is calculated for reference points P^{2d}_{top} and P^{2d}_{bot}: $cos(\beta) = \vec{a} \cdot \vec{b} / (\|\vec{a}\| \cdot \|\vec{b}\|)$ where $\vec{a}$ and $\vec{b}$ can be computed using camera intrinsics. As follows from Figure 4, the distance d between the camera and the center of P^{2d}_{top} and P^{2d}_{bot} is calculated as follows: $d = h/(2 \cdot tan(\frac{\beta}{2}))$. Note that these equations assume the camera view is approximately perpendicular to the top-bottom line, which is a valid assumption in an autonomous vehicle scenario. Using the distance d and the pixel rays, the 2D RPs can be projected to 3D providing an initial estimate of the 3D bounding box. In case of the 8RP variant, we can use the redundancy of the reference points to optimise the final 3D bounding box. This is done by updating the 3D box while minimising the distance between the 2D projections of the RPs in 3D, and the predicted 2D RPs. Similar to [38] we use the Levenberg–Marquardt algorithm (LM) to solve the optimisation problem. Note that camera intrinsics are needed during the uplifting, but since this is a post-processing step, the intrisincs are not embedded anywhere in the network.

3.2. Using instance segmentation

We propose to leverage instance segmentation masks to improve pose estimation. The key idea is that every pixel of an instance votes for each of the parameters described in Section 3.1. This results in vote distributions, which can be leveraged as a measure of confidence. This confidence can be used in a further stage of a broader pipeline, such as a tracking mechanism which requires confidence values. The use of instances has two other advantages. First, the prediction does not suffer from unreliable estimates caused by pixels which do not belong to the object itself, as is often the case for occluded objects in other approaches. Second, since our instance segmentation is proposal free, no NMS is needed. Figure 2 shows an overview of our multi-task CNN. The encoder shares its weights for all tasks, while branched decoders have unique weights for each task. This section describes in more detail how the concepts of Section 3.1 are applied in this network.

Instance Segmentation. The first branch outputs the instance segmentation. Note that any method can be used, even an external network or ground truth instances.

Object Dimensions. The dimensions of an object are directly regressed for every pixel belonging to the instance mask of that object. Subsequently, all estimates belonging to that object's instance mask are averaged, as shown in the second branch of Figure 2.

2D Reference Points. Estimating the absolute coordinates for 2D RPs is not ideal in a CNN since the network has little knowledge of its absolute position within the image. We propose to estimate full offset vectors between a pixel's coordinates (u, v) and each 2D RP, for every pixel belonging to an object's instance mask i. Further on, we will call these offset vectors *relative 2D RPs* $\vec{\mathbf{P}}^{2d,rel}_{i,j}$, in contrast to the previously described *absolute 2D RPs* $P^{2d,abs}_{i,j}$, where j refers to a RP. Once the estimated relative RPs are converted to absolute RPs using Equation (1), they are averaged over all pixels of each instance mask. This is shown in the third branch of Figure 2. Note that the 2D positions of the absolute RPs are not limited by the image boundaries, as opposed to methods which predict heatmaps [43].

$$
\begin{aligned}
P^{2d,abs}_{i,j_u,u,v} &= \vec{\mathbf{P}}^{2d,rel}_{i,j_u,u,v} + umap_{u,v} \\
P^{2d,abs}_{i,j_v,u,v} &= \vec{\mathbf{P}}^{2d,rel}_{i,j_v,u,v} + vmap_{u,v} \\
umap_{u,v} &= u \\
vmap_{u,v} &= v
\end{aligned}
\qquad (1)
$$

where

Object Rotation. Let us consider the example of the truck depicted in Figure 3. The front part of the truck represents a considerably different viewing angle α_{front} compared to the rear part α_{rear}. Based on this insight, we propose to predict a unique local viewing angle α for every pixel of an instance. As described in 3.1, $cos(2 \cdot \alpha)$ and $sin(2 \cdot \alpha)$ are predicted. Every pixel's viewing angle is subsequently compensated by the pixel's light ray offset θ_{ray} to obtain the global yaw r_y. The instance masks are used to average the yaw for every instance, as shown in Figure 2.

Uplifting 2D to 3D. The dimensions, 2D RPs and rotation, can be combined to obtain full 6D object pose as described in Section 3.1 and shown in Figure 4.

4. Experiments

4.1. Setup

Our method. This section discusses the implementation and training details for our method described in Section 3. The method is implemented in PyTorch with off-the-shelf encoder-decoder architectures: ERFNet [67], ResNet50 and ResNet101 [68]. Since the two ResNet backbones do not contain a decoder, we concatenate an ERFNet decoder to scale up the output to full image resolution. We initialise the

ResNet backbones with pre-trained weights on ImageNet provided by PyTorch [69]. The first branch in Figure 2 implements the instance segmentation method proposed in [70], since this method combines good accuracy with low inference time. All other branches are regressed with an $L1$ loss for pixels belonging to an instance, while background pixels are ignored. The rotation branch is only used for the $2RP$ variant. We use following weights on the losses: $[1, 45, 1]$ for the $8RP$ variant, and $[1, 40, 3, 10]$ for the $2RP$ variant. We use NVIDIA GTX 1080 Ti for training, evaluation and timing experiments. All used data is cropped and resized to a resolution of 720×360 or 1200×360.

Other methods. We compare our method with two other representative SOTA methods: SMOKE [41] and M3D-RPN [29]. They both rank high on the KITTI3D benchmark and provide code to train on the KITTI3D dataset only [71, 72], so we implemented additional dataloaders to be able to train on multiple datasets.

	Images (train+val)	Cams	Resolution	FOV	3D boxes	Inst. Segm.	Year	Locations
KITTI3D	3712+3769[1]	1	1242*375[4]	81	Yes	Yes[2]	2012	Germany
VirtKITTI	21260	1	1242*375	81	Yes	Yes	2016	Germany
CityScapes3D	2975+500	1	2048*1024	48	Yes	Yes	2016	Germany
SemKITTI	200	1	1242*375	81	No	Yes	2018	Germany
nuScenes	168780+36114	5 1	1600*900	70 110	Yes	Yes[3]	2019	Boston, Singapore
Waymo	77499+13676 54196+9564	3 2	1920*1280 1920*1040	50	Yes	Yes[3]	2019	USA
Lyft	136080	6	1224*1024 1920*1080	70 82	Yes	Yes[3]	2019	Palo Alto

Table 1: Overview of used datasets. [1]Chen split [18], [2]Manually annotated, [3]Generated with SOTA method [73], [4]Resolution slightly varies.

4.2. Datasets and Benchmarks

In the field of autonomous driving, several datasets and benchmarks exist. Table 1 summarises the datasets we use in our experiments. **KITTI3D** [1] is arguably the most popular benchmark for monocular 3D object detection. We follow the train-validation split described in [18]. We report on *Car* only, both on the validation set and on the test set, using the official 3D object detection and BEV Average Precision (AP) metrics. The metric uses an Intersection-over-Union (IoU) of 0.7 for *Car*, which makes it especially hard for monocular object detection methods. Following [27], we report using the $AP_{3D,BEV|R_{40}}$ metrics, unless stated otherwise. **VirtKITTI** [74] was created by virtually cloning real-world video sequences from KITTI3D, and providing automatic dense labeling. **CityScapes** [2] and **SemKITTI** [75] both provide instance segmentation annotations. The recently launched CityScapes3D dataset adds 3D object annotations. **nuScenes** [3], **Waymo** [4] and **Lyft** [5] are recent big-scale 3D object detection datasets, captured by five or six cameras mounted on top of the car. Both nuScenes and Lyft cover the complete 360 surround view. We will refer to these three datasets as **NWL**.

Generation of missing instance annotations. Our method requires instance segmentation annotations. CityScapes, SemKITTI and VirtKITTI provide these annotations. For KITTI3D, we manually annotated all images with instance segmentation and will release them publicly. Since NWL do not provide these annotations, we generated pseudo-ground truth instances using a SOTA instance segmentation method [73].

Annotation differences. We address two observations in comparing different datasets. First, 3D bounding boxes annotations are very tight in KITTI3D and CityScapes3D, while they are wider in NWL. Second, there is an imbalance in car size between the datasets, as shown in Figure 5 and described earlier in a LiDAR context [76]. KITTI3D and CityScapes3D have semi-overlapping distributions, while NWL is quite different.

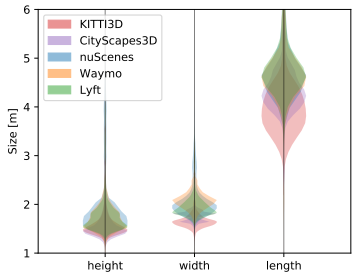


Figure 5: Imbalance in annotated car dimensions: KITTI3D and CityScapes3D overlap, while NWL is very different.

5. Results and Discussion

Method		Training datasets	Evaluation set results $AP_{3D\|R_{40}}$ Easy	Moderate	Hard	$AP_{BEV\|R_{40}}$ Easy	Moderate	Hard
	BL	K3D	12.51	7.53	6.21	18.34	11.22	9.34
Ours (8RP)	E1	CS3D	2.78	1.13	0.82	6.21	3.21	2.55
CI: Yes	E2	K3D, CS3D	**16.16**	8.80	7.43	**23.14**	12.78	10.98
	E3	K3D, NWL	16.09	**9.19**	**7.90**	22.90	**13.86**	**11.53**
	BL	K3D	**6.97**	**4.37**	**3.96**	**12.01**	**8.03**	**6.94**
SMOKE [41]	E1	CS3D	0.00	0.00	0.00	0.00	0.00	0.00
CI: No	E2	K3D, CS3D	4.81	3.80	3.03	9.24	7.10	6.05
	E3	K3D, NWL	0.24	0.19	0.19	0.98	0.76	0.70
	BL	K3D	14.68	**10.76**	8.28	21.52	15.59	12.44
M3D-RPN [29]	E1	CS3D	0.07	0.03	0.03	0.27	0.18	0.17
CI: No	E2	K3D, CS3D	**14.95**	10.71	**8.50**	**22.55**	**16.34**	**12.79**
	E3	K3D, NWL	-	-	-	-	-	-

Table 2: Results on *Car* (0.7 IoU) for the KITTI3D evaluation set. **Bold** refers to best performance within each method. CI: Camera Independent, BL: baseline, E: experiment, K3D: KITTI3D (train), CS3D: CityScapes3D, NWL: nuScenes (train), Waymo and Lyft. Note that the R_{40} variant of AP is used.

In this section we discuss some results and insights of our experiments. Unless stated otherwise, all experiments on our method start from a model pre-trained for instance

segmentation on KITTI3D (train), VirtKITTI, CityScapes (train), SemKITTI (train) and NWL (see Section 4.2).

Multiple datasets and other methods. We verify two hypotheses on combining multiple datasets, comparing our camera independent method with two camera dependent SOTA methods: SMOKE [41] and M3D-RPN [29]. Table 2 summarises the results of experiments **E1**-**E3** on the KITTI3D evaluation set, in comparison with the baseline (BL), trained on KITTI3D only.

Hypothesis 1*: Training on dataset A and evaluating on dataset B works for camera independent methods, while camera dependent methods will fail to provide correct depth estimations when the FOV's of A and B differ substantially.*

E1: trained on CityScapes3D only: our method gives reasonable results, especially since KITTI3D is never seen during training. Both SMOKE and M3D-RPN fail to provide accurate depth predictions.

Hypothesis 2*: Camera dependent methods could possibly learn a few different camera models, but are not able to scale and generalise to any different camera model.*

E2: trained on joint KITTI3D, CityScapes3D: all methods can cope, however SMOKE already deteriorates.

E3: trained on joint KITTI3D, NWL: while SMOKE fails, our method can handle the wide variety of viewpoints, and even benefits from it. Note that the M3D-RPN implementation requires the *occluded* and *truncated* tags, which are not provided by NWL. This results in non-convergent trainings.

For our method, the ResNet101 backbone and $8RP$ variant are used. For both SMOKE and M3D-RPN, we use the provided backbone in their implementations [71, 72]. Note that we should only compare within a method, since the training schemes are not optimised between methods. For SMOKE, we used a training scheme different from the original paper which explains the lower results. Nevertheless, the trainings with more data follow a longer training scheme, and thus are comparable in a fair way.

Comparison to Related Work. Table 4 reports results on the KITTI3D benchmark on *Car*, both on the test set and the evaluation set. For this experiment, we use the ResNet101 backbone. We train only on KITTI3D for both instance segmentation pre-training and 3D object detection, with no additional datasets. For the test set results, we use the full dataset. For the evaluation set results, we use only the train split. We compare only to methods which use RGB, without any additional LiDAR or depth data. We outperform camera independent methods significantly. When comparing with camera dependent methods, which over-fit to the specific camera intrinsics of the KITI3D dataset, we perform similar to last year's works. Recent methods however surpass our performance, at the cost of lacking the ability to generalise well over multiple cameras.

		Evaluation set results						Inference time [ms]	
		$AP_{3D\|R_{40}}$			$AP_{BEV\|R_{40}}$				
		Easy	Moderate	Hard	Easy	Moderate	Hard	PyT	TRT
ERFNet	8RP	6.34	4.14	3.22	10.79	6.96	5.63	**29**	**19**
	2RP	5.26	3.19	2.61	8.21	5.04	4.02	34	21
ResNet50	8RP	14.66	7.71	6.63	20.73	11.19	9.66	101	89
	2RP	14.31	8.08	6.46	20.01	11.68	10.20	112	103
ResNet101	8RP	12.51	7.53	6.21	18.34	11.22	9.34	151	129
	2RP	**17.22**	**9.36**	**7.43**	**22.84**	**13.03**	**11.29**	159	140

Table 3: Results on *Car* (0.7 IoU) for the KITTI3D evaluation set. **Bold** refers to best performance. PyT: PyTorch, TRT: TensorRT. Note that the R_{40} variant of AP is used.

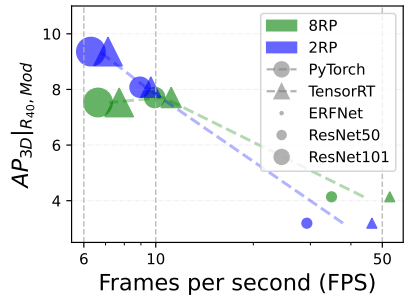


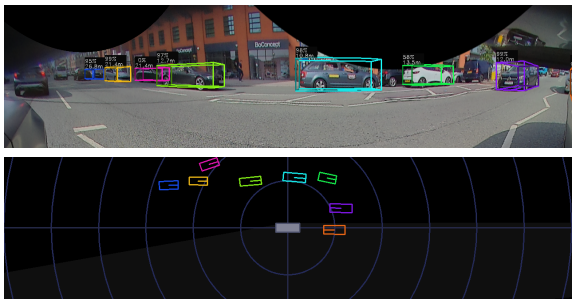
Figure 6: Speed-accuracy trade-off: ResNet achieves higher accuracy, while ERFNet is faster.

Inference time. Table 3 compares results on the KITTI3D evaluation set for all three architectures described in Section 4.1. We use only KITTI3D (train) during training. As expected, smaller network architectures achieve lower performance. However, the inference time also decreases. This leads to higher Frames per second (FPS). We report inference time using both PyTorch and TensorRT [77]. Figure 6 shows the trade-off between speed and accuracy.

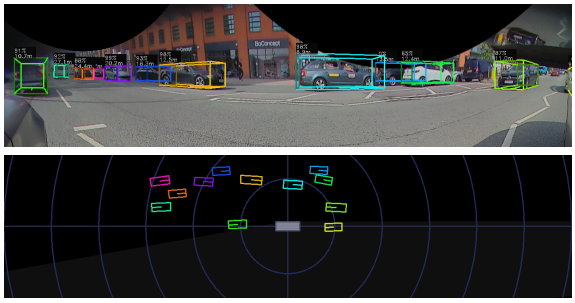
Observations. First, our method is reasonably robust to occlusions. In the presence of comparatively large occluders, the predicted RPs tend to shift towards them. Second, our method can be sensitive to small variations in 2D RPs or dimensions, which possibly lead to larger variations in 3D at far distance. Third, experiments on NWL show that our method is able to generalise easily to viewpoints within the scope of the training data. Although it is not meant to extrapolate to viewpoints that are out of the scope of the available datasets, Figure 7 shows qualitatively that our method even copes with unseen fisheye images from WoodScape [78]. Note that we're not able to train on the data since no 3D bounding box annotations are released yet. Figure 1 shows qualitative results of the $8RP$ variant of our method on a sample from the KITTI3D test set. The proposed method is able to predict correct 3D bounding boxes under heavy occlusion and can recover from imperfect instance masks. More examples are provided in the supplementary material.

Method	Year	Camera independent	Test set results						Evaluation set results					
			$AP_{3D\|R_{40}}$			$AP_{BEV\|R_{40}}$			$AP_{3D\|R_{11}}$			$AP_{BEV\|R_{11}}$		
			Easy	Moderate	Hard	Easy	Moderate	Hard	Easy	Moderate	Hard	Easy	Moderate	Hard
OFTNet [19]	2018	No	1.61	1.32	1.00	7.16	5.69	4.61	4.07	3.27	3.29	11.06	8.79	8.91
MonoGRNet [25]	2018	No	9.61	5.74	4.25	18.19	11.17	8.73	13.88	10.19	7.62	19.72	12.81	10.15
ROI-10D [26]	2019	No	4.32	2.02	1.46	9.78	4.91	3.74	9.61	6.63	6.29	14.50	9.91	8.73
MonoDIS [27]	2019	No	10.37	7.94	6.40	17.23	13.19	11.12	18.05	14.98	13.42	24.26	18.43	16.95
SS3D [28]	2019	No	10.78	7.68	6.51	16.33	11.52	9.93	14.52	13.15	11.85	-	-	-
M3D-RPN [29]	2019	No	14.76	9.71	7.42	21.02	13.67	10.23	20.40	16.48	13.34	**26.86**	21.15	17.14
Shift R-CNN [15]	2019	No	8.13	5.22	4.78	13.32	8.49	6.40	13.84	11.29	11.08	18.61	14.71	13.57
MoVi-3D [30]	2020	No	15.19	10.90	9.26	22.76	17.03	14.85	14.28	11.13	9.68	22.36	17.87	15.73
SMOKE [41]	2020	No	14.03	9.76	7.84	20.83	14.49	12.75	14.76	12.85	11.50	19.99	15.61	15.28
RTM3D [38]	2020	No	14.41	10.34	8.77	19.17	14.20	11.99	20.77	16.86	16.63	25.56	**22.12**	**20.91**
MonoPair [42]	2020	No	13.04	9.99	8.65	19.28	14.83	12.89	-	-	-	-	-	-
MonoDLE [32]	2021	No	17.23	12.26	10.29	24.79	18.89	16.00	-	-	-	-	-	-
MonoFlex [56]	2021	No	**19.94**	**13.89**	**12.07**	**28.23**	**19.75**	**16.89**	**28.17**	**21.92**	**19.07**	-	-	-
DeepBox3D [13]	2017	**Yes**	5.85	4.10	3.84	9.99	7.71	5.30	5.49	3.96	2.92	9.33	6.71	5.11
Barabanau et al. [33]	2019	**Yes**	-	-	-	-	-	-	13.96	7.37	4.54	-	-	-
Linear System [15]	2019	**Yes**	6.80	4.14	3.50	11.75	8.34	6.80	7.24	5.98	5.54	14.74	12.48	11.22
Ours (8RP)	2020	**Yes**	15.82	7.94	6.68	22.28	11.64	9.95	19.92	15.75	13.46	25.32	18.09	17.42
Ours (2RP)	2020	**Yes**	15.21	7.66	6.24	20.42	10.96	9.23	18.68	12.35	11.54	25.06	16.88	13.58

Table 4: Results on *Car* (0.7 IoU) for both the KITTI3D test and evaluation set. **Bold** refers to best performance across all methods, blue refers to best performance across camera independent methods. Note that the R_{40} variant of AP is used for the test set, while the R_{11} variant is used for the evaluation set for comparison purposes. Our method outperforms the other camera independent methods.



(a) 8RP model trained on KITTI3D only



(b) 8RP model trained on KITTI3D and NWL

Figure 7: Qualitative results on unseen camera views: our model predicts 3D boxes on equirectangular fisheye images from [78], though it was never trained for it. (a) vs. (b): generalisation improves when training on more varied data.

6. Conclusion

In this work we present a novel approach for monocular 3D object detection. Leveraging camera independent 2D reference points enables us to handle different camera types in one multi-task CNN. We use a proposal free instance based method which eliminates the need for NMS. This work is the first to take advantage of its camera independence by combining large scale public datasets for better generalisation across viewpoints. We show the benefits by comparing to other methods in a cross-dataset context. We outperform other camera independent methods on the challenging KITTI3D benchmark and show qualitatively how our model can cope with fisheye images even without fisheye training data. We briefly discuss trade-offs between accuracy and speed, which is important for practical applications. Further we will release KITTI3D instance segmentation annotations.

We believe camera independence is key in exploiting multiple datasets, and hope to see more research in this direction. For future work, there are multiple interesting research topics. First, the impact of the imbalance of annotation dimensions between the different datasets should be investigated. Also tracking methods can be explored which leverage the distributions of RPs provided by our method. Finally, clever practices from recent works (e.g. ensembles [56], discarding distant samples [32] and pairwise constraints [42]), can be combined with our method to close the gap between camera dependent and independent methods.

References

[1] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2012.

[2] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.

[3] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 11621–11631, 2020.

[4] P. Sun, H. Kretzschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. Chai, B. Caine, *et al.*, "Scalability in perception for autonomous driving: Waymo open dataset," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2446–2454, 2020.

[5] R. Kesten, M. Usman, J. Houston, T. Pandya, K. Nadhamuni, A. Ferreira, M. Yuan, B. Low, A. Jain, P. Ondruska, S. Omari, S. Shah, A. Kulkarni, A. Kazakova, C. Tao, L. Platinsky, W. Jiang, and V. Shet, "Lyft level 5 av dataset 2019." urlhttps://level5.lyft.com/dataset/, 2019.

[6] P. Sturm, S. Ramalingam, J.-P. Tardif, S. Gasparini, and J. Barreto, "Camera models and fundamental concepts used in geometric computer vision," *Foundations and Trends® in Computer Graphics and Vision*, vol. 6, no. 1–2, pp. 1–183, 2011.

[7] Y. Ye, H. Chen, C. Zhang, X. Hao, and Z. Zhang, "Sarpnet: Shape attention regional proposal network for lidar-based 3d object detection," *Neurocomputing*, vol. 379, pp. 53–63, 2020.

[8] C. He, H. Zeng, J. Huang, X.-S. Hua, and L. Zhang, "Structure aware single-stage 3d object detection from point cloud," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 11873–11882, 2020.

[9] Y. Chen, S. Liu, X. Shen, and J. Jia, "Dsgn: Deep stereo geometry network for 3d object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 12536–12545, 2020.

[10] Z. Xu, W. Zhang, X. Ye, X. Tan, W. Yang, S. Wen, E. Ding, A. Meng, and L. Huang, "Zoomnet: Part-aware adaptive zooming neural network for 3d object detection.," in *AAAI*, pp. 12557–12564, 2020.

[11] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8445–8453, 2019.

[12] Y. You, Y. Wang, W.-L. Chao, D. Garg, G. Pleiss, B. Hariharan, M. Campbell, and K. Q. Weinberger, "Pseudo-lidar++: Accurate depth for 3d object detection in autonomous driving," in *International Conference on Learning Representations*, 2019.

[13] A. Mousavian, D. Anguelov, J. Flynn, and J. Kosecka, "3d bounding box estimation using deep learning and geometry," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7074–7082, 2017.

[14] L. Liu, J. Lu, C. Xu, Q. Tian, and J. Zhou, "Deep fitting degree scoring network for monocular 3d object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1057–1066, 2019.

[15] A. Naiden, V. Paunescu, G. Kim, B. Jeon, and M. Leordeanu, "Shift r-cnn: Deep monocular 3d object detection with closed-form geometric constraints," in *2019 IEEE International Conference on Image Processing*, pp. 61–65, IEEE, 2019.

[16] H. Min Choi, H. Kang, and Y. Hyun, "Multi-view reprojection architecture for orientation estimation," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pp. 0–0, 2019.

[17] J. Ku, A. D. Pon, and S. L. Waslander, "Monocular 3d object detection leveraging accurate proposals and shape reconstruction," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 11867–11876, 2019.

[18] X. Chen, K. Kundu, Z. Zhang, H. Ma, S. Fidler, and R. Urtasun, "Monocular 3d object detection for autonomous driving," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2147–2156, 2016.

[19] T. Roddick, A. Kendall, and R. Cipolla, "Orthographic feature transform for monocular 3d object detection," *British Machine Vision Conference*, 2019.

[20] Y. Kim and D. Kum, "Deep learning based vehicle position and orientation estimation via inverse perspective mapping image," in *2019 IEEE Intelligent Vehicles Symposium*, pp. 317–323, IEEE, 2019.

[21] S. Srivastava, F. Jurie, and G. Sharma, "Learning 2d to 3d lifting for object detection in 3d for autonomous vehicles," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 4504–4511, IEEE, 2019.

[22] C. Reading, A. Harakeh, J. Chae, and S. L. Waslander, "Categorical depth distribution network for monocular 3d object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8555–8564, 2021.

[23] J. Zhu and Y. Fang, "Learning object-specific distance from a monocular image," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3839–3848, 2019.

[24] X. Zhou, D. Wang, and P. Krähenbühl, "Objects as points," *CoRR*, vol. abs/1904.07850, 2019.

[25] Z. Qin, J. Wang, and Y. Lu, "Monogrnet: A geometric reasoning network for monocular 3d object localization," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 8851–8858, 2019.

[26] F. Manhardt, W. Kehl, and A. Gaidon, "Roi-10d: Monocular lifting of 2d detection to 6d pose and metric shape," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2069–2078, 2019.

[27] A. Simonelli, S. R. Bulo, L. Porzi, M. López-Antequera, and P. Kontschieder, "Disentangling monocular 3d object detection," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1991–1999, 2019.

[28] E. Jörgensen, C. Zach, and F. Kahl, "Monocular 3d object detection and box fitting trained end-to-end using intersection-over-union loss," *arXiv preprint arXiv:1906.08070*, 2019.

[29] G. Brazil and X. Liu, "M3d-rpn: Monocular 3d region proposal network for object detection," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 9287–9296, 2019.

[30] A. Simonelli, S. R. Bulo, L. Porzi, E. Ricci, and P. Kontschieder, "Towards generalization across depth for monocular 3d object detection," in *16th European Conference on Computer Vision–ECCV 2020*, pp. 767–782, 2020.

[31] S. Luo, H. Dai, L. Shao, and Y. Ding, "M3DSSD: monocular 3d single stage object detector," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021.

[32] X. Ma, Y. Zhang, D. Xu, D. Zhou, S. Yi, H. Li, and W. Ouyang, "Delving into localization errors for monocular 3d object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2021.

[33] I. Barabanau, A. Artemov, E. Burnaev, and V. Murashkin, "Monocular 3d object detection via geometric reasoning on keypoints," *arXiv preprint arXiv:1905.05618*, 2019.

[34] F. Chabot, M. Chaouch, J. Rabarisoa, C. Teulière, and T. Chateau, "Deep manta: A coarse-to-fine many-task network for joint 2d and 3d vehicle analysis from monocular image," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2040–2049, 2017.

[35] A. Kundu, Y. Li, and J. M. Rehg, "3d-rcnn: Instance-level 3d object reconstruction via render-and-compare," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3559–3568, 2018.

[36] T. He and S. Soatto, "Mono3d++: Monocular 3d vehicle detection with two-scale 3d hypotheses and task priors," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 8409–8416, 2019.

[37] A. Rangesh and M. M. Trivedi, "Ground plane polling for 6dof pose estimation of objects on the road," *IEEE Transactions on Intelligent Vehicles*, 2020.

[38] P. Li, H. Zhao, P. Liu, and F. Cao, "Rtm3d: Real-time monocular 3d detection from object keypoints for autonomous driving," in *16th European Conference on Computer Vision–ECCV 2020*, pp. 644–660, 2020.

[39] Y. Cai, B. Li, Z. Jiao, H. Li, X. Zeng, and X. Wang, "Monocular 3d object detection with decoupled structured polygon estimation and height-guided depth estimation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, pp. 10478–10485, 2020.

[40] S. Suwajanakorn, N. Snavely, J. J. Tompson, and M. Norouzi, "Discovery of latent 3d keypoints via end-to-end geometric reasoning," in *Advances in Neural Information Processing Systems*, pp. 2059–2070, 2018.

[41] Z. Liu, Z. Wu, and R. Tóth, "Smoke: Single-stage monocular 3d object detection via keypoint estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 996–997, 2020.

[42] Y. Chen, L. Tai, K. Sun, and M. Li, "Monopair: Monocular 3d object detection using pairwise spatial relationships," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 12093–12102, 2020.

[43] G. Pavlakos, X. Zhou, A. Chan, K. G. Derpanis, and K. Daniilidis, "6-dof object pose from semantic keypoints," in *2017 IEEE International Conference on Robotics and Automation*, pp. 2011–2018, IEEE, 2017.

[44] L. Yang, X. Zhang, L. Wang, M. Zhu, and J. Li, "Lite-fpn for keypoint-based monocular 3d object detection," *ArXiv*, vol. abs/2105.00268, 2021.

[45] B. Xu and Z. Chen, "Multi-level fusion based 3d object detection from monocular images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2345–2353, 2018.

[46] X. Ma, Z. Wang, H. Li, P. Zhang, W. Ouyang, and X. Fan, "Accurate monocular 3d object detection via color-embedded 3d reconstruction for autonomous driving," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 6851–6860, 2019.

[47] X. Weng and K. Kitani, "Monocular 3d object detection with pseudo-lidar point cloud," in *2019 IEEE International Conference on Computer Vision Workshop*, pp. 857–866, IEEE Computer Society, 2019.

[48] J. M. U. Vianney, S. Aich, and B. Liu, "Refinedmpl: Refined monocular pseudolidar for 3d object detection in autonomous driving," *arXiv preprint arXiv:1911.09712*, 2019.

[49] R. Qian, D. Garg, Y. Wang, Y. You, S. Belongie, B. Hariharan, M. Campbell, K. Q. Weinberger, and W.-L. Chao, "End-to-end pseudo-lidar for image-based 3d object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5881–5890, 2020.

[50] S. Song and M. Chandraker, "Joint sfm and detection cues for monocular 3d localization in road scenes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3734–3742, 2015.

[51] L. Wang, L. Du, X. Ye, Y. Fu, G. Guo, X. Xue, J. Feng, and L. Zhang, "Depth-conditioned dynamic message propagation for monocular 3d object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021.

[52] M. Ding, Y. Huo, H. Yi, Z. Wang, J. Shi, Z. Lu, and P. Luo, "Learning depth-guided convolutions for monocular 3d object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1000–1001, 2020.

[53] Y. Liu, Y. Yixuan, and M. Liu, "Ground-aware monocular 3d object detection for autonomous driving," *IEEE Robotics and Automation Letters*, vol. 6, pp. 919–926, 2021.

[54] Y. Zhou, Y. He, H. Zhu, C. Wang, H. Li, and Q. Jiang, "Monocular 3d object detection: An extrinsic parameter free approach," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021.

[55] H. Chen, Y. Huang, W. Tian, Z. Gao, and L. Xiong, "Monorun: Monocular 3d object detection by reconstruction and uncertainty propagation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 10379–10388, 2021.

[56] Z. Yunpeng, L. Jiwen, and Z. Jie, "Objects are different: Flexible monocular 3d object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021.

[57] Y. Xiang, T. Schmidt, V. Narayanan, and D. Fox, "Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes," in *Proceedings of Robotics: Science and Systems*, (Pittsburgh, Pennsylvania), June 2018.

[58] C. Capellen., M. Schwarz., and S. Behnke., "Convposecnn: Dense convolutional 6d object pose estimation," in *Proceedings of the 15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, pp. 162–172, 2020.

[59] T. Do, T. Pham, M. Cai, and I. Reid, "Lienet: Real-time monocular object instance 6d pose estimation," in *British Machine Vision Conference 2018, BMVC 2018, Newcastle, UK, September 3-6, 2018*, p. 2, 2018.

[60] O. H. Jafari, S. K. Mustikovela, K. Pertsch, E. Brachmann, and C. Rother, "ipose: instance-aware 6d pose estimation of partly occluded objects," in *Asian Conference on Computer Vision*, pp. 477–492, Springer, 2018.

[61] Z. Li, G. Wang, and X. Ji, "Cdpn: Coordinates-based disentangled pose network for real-time rgb-based 6-dof object pose estimation," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 7678–7687, 2019.

[62] S. Zakharov, I. Shugurov, and S. Ilic, "Dpod: 6d pose object detector and refiner," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1941–1950, 2019.

[63] H. Wang, S. Sridhar, J. Huang, J. Valentin, S. Song, and L. J. Guibas, "Normalized object coordinate space for category-level 6d object pose and size estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2642–2651, 2019.

[64] Z. Fan, Y. Zhu, Y. He, Q. Sun, H. Liu, and J. He, "Deep learning on monocular object pose detection and tracking: A comprehensive overview," *ArXiv*, vol. abs/2105.14291, 2021.

[65] S. Peng, Y. Liu, Q. Huang, X. Zhou, and H. Bao, "Pvnet: Pixel-wise voting network for 6dof pose estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4561–4570, 2019.

[66] Y. Hu, J. Hugonot, P. Fua, and M. Salzmann, "Segmentation-driven 6d object pose estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3385–3394, 2019.

[67] E. Romera, J. M. Alvarez, L. M. Bergasa, and R. Arroyo, "Erfnet: Efficient residual factorized convnet for real-time semantic segmentation," *IEEE Transactions on Intelligent Transportation Systems*, vol. 19, no. 1, pp. 263–272, 2017.

[68] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[69] PyTorch, "Documentation: torchvision.models." `https://pytorch.org/docs/stable/torchvision/models.html`. Accessed: 2020-11.

[70] D. Neven, B. D. Brabandere, M. Proesmans, and L. V. Gool, "Instance segmentation by jointly optimizing spatial embeddings and clustering bandwidth," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8837–8845, 2019.

[71] Z. Liu, Z. Wu, and R. Tóth, "SMOKE: Single-Stage Monocular 3D Object Detection via Keypoint Estimation." `https://github.com/lzccccc/SMOKE`, 2020. Accessed: 2020-09.

[72] G. Brazil and X. Liu, "M3D-RPN: Monocular 3D Region Proposal Network for Object Detection." `https://github.com/garrickbrazil/M3D-RPN`, 2019. Accessed: 2020-09.

[73] F. Massa and R. Girshick, "maskrcnn-benchmark: Fast, modular reference implementation of Instance Segmentation and Object Detection algorithms in PyTorch." `https://github.com/facebookresearch/maskrcnn-benchmark`, 2018. Accessed: 2020-11.

[74] A. Gaidon, Q. Wang, Y. Cabon, and E. Vig, "Virtual worlds as proxy for multi-object tracking analysis," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4340–4349, 2016.

[75] H. A. Alhaija, S. K. Mustikovela, L. Mescheder, A. Geiger, and C. Rother, "Augmented reality meets computer vision: Efficient data generation for urban driving scenes," *International Journal of Computer Vision*, vol. 126, no. 9, pp. 961–972, 2018.

[76] Y. Wang, X. Chen, Y. You, L. E. Li, B. Hariharan, M. Campbell, K. Q. Weinberger, and W.-L. Chao, "Train in germany, test in the usa: Making 3d object detectors generalize," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 11713–11723, 2020.

[77] NVIDIA Corporation, "NVIDIA TensorRT: Programmable Inference Accelerator." `https://developer.nvidia.com/tensorrt`. Accessed: 2020-11.

[78] S. Yogamani, C. Hughes, J. Horgan, G. Sistu, P. Varley, D. O'Dea, M. Uricár, S. Milz, M. Simon, K. Amende, *et al.*, "Woodscape: A multi-task, multi-camera fisheye dataset

for autonomous driving," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 9308–9318, 2019.