Geometry-Aware Instance Segmentation with Disparity Maps

Cho-Ying Wu[1], Xiaoyan Hu[2], Michael Happold[2], Qiangeng Xu[1], and Ulrich Neumann[1]
[1]University of Southern California[2]Argo AI

No Institute Given

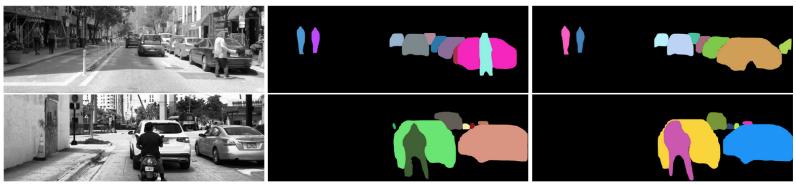

Fig. 1. GAIS-Net results on HQDS dataset. Left column shows stereo left images with histogram equalization to enhance contrast for better visualization. Middle and last column show Mask-RCNN and GAIS-Net results, respectively. Each instance has different colors. With the aid of geometric information, GAIS-Net can segment out the person from the overlapping area in the first row example. In the second row scenario, Mask-RCNN generates distorted mask for the smoking motorcyclist because of cigarette plume and in contrast GAIS-Net displays a more robust shape control capability.

Abstract. Instance segmentation for autonomous driving aims at identifying each object of interest to facilitate environment understanding on roads. Most previous works of instance segmentation for images only use color information. We explore a novel direction of sensor fusion to exploit disparity modality from stereo cameras. Our work fuses images and geometric scene priors and further directly regresses masks from disparity maps. The geometric information helps separate overlapping objects of the same or different classes. Moreover, geometric information penalizes region proposals with unlikely 3D shapes thus suppressing false positive detections. Mask regression is based on 2D, 2.5D, and 3D ROI using the pseudo-lidar and image-based representations with a self-supervised correspondence loss. These mask predictions are fused by a mask scoring process. However, public datasets only adopt stereo systems with shorter baseline and focal legnth, which limit measuring ranges and produce disparities with larger error. We collect and utilize High-Quality Driving Stereo (HQDS) dataset, using much longer baseline and focal length with higher resolution. Our performance attains state of the art. Codes will be released.

1 Introduction

Instance segmentation, which segments out every object of interest, is an elemental and important task for computer vision. It is crucial for autonomous driving because it is vital to know positions for every object instance on roads. Instance masks are

widely used in object detection [1, 5, 37, 50], object tracking [6, 19, 44], and HD-maps constructions [4].

In the context of instance segmentation on images, previous approaches only operate on RGB imagery, such as Mask-RCNN [16]. However, image data could be affected by illumination, color change, shadows, or optical defects. These factors can degrade the performance of image-based instance segmentation. By utilizing another modality that provides geometric cues of scenes, and since object shapes are independent of object texture and color change, these strong priors add more robust information of the scenes. A prior work [52] that goes beyond the dominant paradigm to incorporate depth information only uses it for naive ordering rather than directly regressing masks.

Also, there are some works on RGBD semantic segmentation [15, 39, 46], which work on indoor scenes with low resolutions and limited ranges. Instance segmentation is arguably harder than semantic segmentation since we need to separate every instance out rather than only regressing pixel classes. In addition, *sensor fusion for outdoor scenes is much harder and has been less explored than indoors*, since much longer range sensing is required to align information from images and depth. Such as vehicles at distances showing in images but undetected by a depth sensor would bring ambiguity into RGBD methods.

In outdoor scenes, stereo cameras or lidar sensors are commonly used for depth acquisition. Lidars are precise, although they have several disadvantages compared with stereo cameras. The performance of lidars is restricted by their power and measuring range of sensors, and the nature of lidar scans leads to limited spatial resolution and produces sparse depth maps. From a practical perspective, lidars are commonly much more expensive than stereo cameras as well.

Stereo cameras are low-cost, and their adjustable parameters, such as *longer baselines* (b) and *focal lengths* (f), favor stereo matching at far fields. Relationship of depth and disparity is given by

$$depth = \frac{f \times b}{disparity}. \tag{1}$$

1-disparity (the minimal pixel difference showing the ideal longest range a stereo system could detect) represents farther distance if using longer f and b. Likewise, a far object could have larger disparity if using longer f and b. Next, longer baselines and focal lengths favor more precise geometric estimations [14, 36], since longer baselines produce smaller triangulation error, and longer focal lengths project objects on images with more pixels and thus enhance the robustness of stereo matching and show more complete shapes. Using longer baselines and focal lengths to reconstruct 3D from multiview attracts lots of research interest before deep learning [12, 13, 21, 28, 48].

Recent deep learning based stereo matching algorithms are capable of generating high-quality and dense disparity estimations that rival the accuracy of lidar measurements and achieve much higher angular resolutions [3,51,53]. By using longer baselines and focal lengths, stereo cameras can exceed lidars' working distance.

In this work, we explore *a novel direction of sensor fusion* to develop an end-to-end trainable Geometry-Aware Instance Segmentation Network (GAIS-Net) that takes the advantages of both the semantic information from image domain and geometric information from disparity maps. GAIS-Net firstly extracts features and generates proposals

from images. Disparity maps are then introduced at the ROI heads, where the encoded geometry helps the network control shapes to regress more complete shapes. We use the proposals to crop out the ROI in disparity maps. To fully utilize representation advantages, we adopt both the *pseudo-lidar representation*, which pops up the disparity from images and extracts features using a point cloud based-network, and the *image-based representation* using a CNN. We regress masks from images, image-based disparity, and pseudo lidar-based disparity features. In addition to mask loss for shape control, a self-supervised correspondence loss is used to self-guide the training from different representations, and a mask continuity loss reduces the shape distortion problem in pseudo-lidar sampling. At inference time, we fuse the masks using mask scoring.

GAIS-Net's motivation for sensor fusion exploits the stereo camera modality. It is practical in autonomous driving since stereo cameras are commonly used as depth acquisition sensors on self-driving cars and they are much cheaper than lidars. Instance segmentation benchmarks currently lack stereo pairs with longer baselines, longer focal lengths and higher resolutions. We collect a High-Quality Driving Stereo (HQDS) dataset, with a total of 8.8K stereo pairs with $f \times b$ **4 times larger** than the current best dataset, Cityscapes [7]. We compare GAIS-Net with recent benchmark algorithms on HQDS. The results show that GAIS-Net achieves state-of-the-art performance. We also validate on Cityscapes but with shorter f and b.

Our contributions are summarized as follows:

1. To our knowledge, we are the first to perform instance segmentation on imagery by fusing images and disparity information to regress object masks.

2. We collect HQDS dataset with longer baseline and longer focal length, which favors far-field stereo matching.

3. We present GAIS-Net, an aggregation of representation design for instance segmentation using images, image-based, and point cloud-based networks. We train GAIS-Net with different losses, and fuse these predictions using the mask scoring. GAIS-Net achieves the state of the art.

2 Related Work

2.1 Instance Segmentation

Instance segmentation has attracted much research interest in computer vision. Methods of instance segmentation on imagery fall into 2 categories: segmentation-based, object detection-based.

Segmentation-based methods usually perform per-pixel semantic labeling followed by clustering to develop instances. DWT [1] learns a watershed transform to segment out each object by cutting at an energy level, and resulting connected components represent instances. Zhang *et al.* [54,55] use networks to predict instance labels first and then adopt MRF on the patch level to ensure local and global consistency. InstanceCut [26] combines semantic segmentation and edge detection to separate out each instance.

Detection-based methods usually perform object detection to generate proposals first and then regress a mask within a region of interest. Mask-RCNN [16], based on a 2-stage detection network with Region Proposal Network (RPN) from Faster-RCNN [42], further introduces a mask head to regress masks for detected bounding

boxes. PANet [34] adds a bottom-up path augmentation and adaptive pooling in Mask-RCNN for better information flow. However, its inference frames-per-second is much lower than Mask-RCNN and thus not suitable for real applications. MS-RCNN [20] introduces a MaskIoU head to directly regress MaskIoU scores, which could calibrate misalignments between mask quality and confidence scores and prioritize more accurate mask predictions. However, their regressed mask scores are not used at inference time to help mask shape control. Recent research focuses on detection-based methods for their high performance and robustness.

There are some hybrid methods combining semantic segmentation and instance segmentation using both pixel-level and instance-level labeling to enhance scene understanding. Recently, HTC [5] adopts a multi-stage cascade detection network [2] with semantic information flow. Semantic regression is used to refine cascaded masks. UPSNet [49] and SSAP [10] focus on panoptic segmentation [25] combining both semantic and instance segmentation with multi-task learning. These methods require both *pixel-level* and *instance-level* segmentation groundtruth. More complete scene understanding could be learned such as relationships of classes *road and vehicle* or *street and pedestrian*, which could not be learned from solely using instance-level labeling since road and street are uncountable and thus unlabeled. However, acquiring both per-pixel and per-instance groundtruth labeling is expensive for real applications. By contrast, detection-based methods such as Mask-RCNN and our work, do not require extra per-pixel semantic labeling but only need instance-level labeling.

Neven *et al.* [35], which is not categorized into segmentation or detection-based methods, predict per-pixel class-specific seed maps and sigma maps followed by clustering to regress masks. Recently, there is some work [8, 9] using metric learning for instance segmentation.

Instance segmentation with depth. Ye *et al.* [52] adopt a simple depth ordering technique, which uses depth information to reconcile overlapping areas between two objects. An overlapping area is assigned to objects depending on which object has much closer depth values to this area. However, they neither use depth information to regress mask shapes, nor build an end-to-end trainable model to propagate depth information. Besides, their depth maps are predicted from monocular images, making the depth ordering unreliable.

2.2 Sensor Fusion for Detection

Recent work for 3D object detection usually combines different data modalities such as images and lidar point cloud [27, 29, 30, 37]. However, the recently introduced pseudo-lidar work [47] demonstrates that with appropriate processing, stereo depth data could produce 3D detection results on par with results obtained using both cameras and lidars, or with lidars only. In particular, pseudo-lidar displays that by simply popping the stereo disparity map into depth point cloud, lidar-based detectors could be used to great effect.

These works show that in addition to depth domain information, image domain information is semantic and thus could help 3D bounding box detection. Different from them, our work demonstrates that using disparity, which contains geometric information as priors, can greatly benefit 2D instance segmentation. The final mask inference is

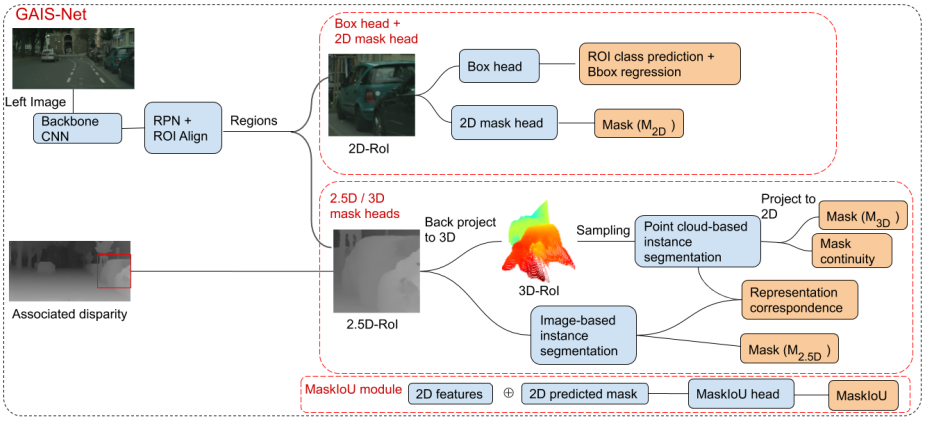

Fig. 2. Network design of our GAIS-Net. Bbox is for bounding box. We color modules in blue and outputs or loss parts in orange. In the MaskIoU module, the 2D features and 2D predicted mask are from the 2D mask head. They are fed into MaskIoU head to regress MaskIoU scores. We draw the MaskIoU head separately for clear visualization. $\oplus$ stands for concatenation.

obtained by using mask scoring to fuse predictions from geometric and semantic information, making the output more accurate than using single modality. Besides, our work only requires shape and geometric information, which is embedded in disparity maps, and does not need actual depth values. Therefore, we do not convert disparity to depth as done in the pseudo-lidar work [47].

3 Method

Our goal is to construct an end-to-end trainable network to perform instance segmentation for autonomous driving. Our system segments out each instance and outputs confidence scores for bounding boxes and masks for each instance. To exploit geometric information, we adopt PSMNet [3], the state-of-the-art stereo matching network, and introduce disparity information at ROI heads. The whole network design is in Fig. 2.

Two-stage networks on object detection, such as Faster-RCNN [42] and Mask-RCNN [16], are generally more precise than single-stage networks. We build a two-stage detector with a backbone network (such as ResNet50-FPN [17, 31]) and a region proposal network (RPN) with non-maximum suppression. Object proposals are collected by feeding a stereo left image into the backbone network and RPN. Same as Faster-RCNN and Mask-RCNN, we perform bounding box regression, class prediction for proposals, and mask prediction based on image domain features. Corresponding losses are denoted as $\mathcal{L}_{box}$ (error of regressed box parameters), $\mathcal{L}_{cls}$ (cross-entropy with groundtruth class), and $\mathcal{L}_{2Dmask}$ (cross-entropy with groundtruth mask) and are identified in [16].

We then introduce geometry-aware mask prediction, which bundles image-based and point cloud-based representation to aggregate features for mask prediction.

3.1 Geometry-Aware Mask Prediction

2.5D ROI and 3D ROI. Conventional methods of stereo matching, such as block matching or SGM [18], usually produce sparse and incomplete disparity maps. Recently, neural networks-based approaches have demonstrated capabilities to predict dense disparity maps and outperform conventional methods. We use PSMNet [3] to predict dense disparity maps, projected onto the left stereo frame. Next, RPN outputs numerous region proposals. We collect proposals and crop out these areas from the disparity map. We call these cropped out disparity areas as *2.5D ROI*.

Based on the observations from pseudo-lidar work, which describes the advantage of back-projecting 2D grid structured data into 3D point cloud and processing with point cloud networks, we back-project the disparity map into $\mathbb{R}^3$ space, where for each point, the first and second components describe its 2D grid coordinates, and the third component stores its disparity value. We name this representation as *3D ROI*.

Pseudo-lidar work further converts disparity to depth. However, based on the metric error in depth ΔZ, pixel matching error Δm, $|\Delta Z| = Z^2 \frac{\Delta m}{bf}$, the error would quadratically increase with depth Z. For instance segmentation on images, using the disparity representation is better than converting to depth, since disparity maps already contain shape information of objects and do not suffer from the quadratically increasing error issue. See the supplementary for validation.

Instance Segmentation Networks. Each 3D ROI contains different number of points. To facilitate training, we uniformly sample the 3D ROI to 1024 points, and collect all the 3D ROI into a tensor. We develop a PointNet [38] structured instance segmentation network to extract point features and perform per-point mask probability prediction. We re-project the 3D feature onto the 2D grid to calculate the mask prediction and its loss $\mathcal{L}_{3Dmask}$. The re-projection is efficient because we do not break the point order in the point cloud-based instance segmentation. $\mathcal{L}_{3Dmask}$, same as $\mathcal{L}_{2Dmask}$, is a cross-entropy loss between a predicted probability mask and its matched groundtruth.

To fully utilize advantages of different representations, we further do 2.5D ROI instance segmentation with an image-based CNN. Similar to instance segmentation on 2D ROI, this network extracts local features of 2.5D ROI, and later performs per-pixel mask probability prediction. The mask prediction loss is denoted as $\mathcal{L}_{2.5Dmask}$. In the later ablation study, we find that 2.5D ROI and 3D ROI have advantages at different IoU levels. Network architectures are detailed in the supplementary.

3.2 Mask Continuity

We sample 3D ROI to 1024 points uniformly. However, predicted mask outlines are sensitive to pseudo-lidar sampling strategies. An undesirable sampling is illustrated in Fig. 3. If a sampled point just lies outside a foreground object, its occupied cell would represent background. After sampling, the point cloud shows a skewed contour. Using this point cloud as an input to the point cloud-based instance segmentation network, we would obtain a mask with irregular outlines. The predicted probability mask is denoted as M_{3D}. To compensate the undesirable effect, we introduce a mask continuity loss.

We address the outline irregularity issue at M_{3D}. Since objects are structured and continuous, we calculate a *mask Laplacian* as $\nabla^2 M = \frac{\partial^2 M}{\partial x^2} + \frac{\partial^2 M}{\partial y^2}$, where x and y

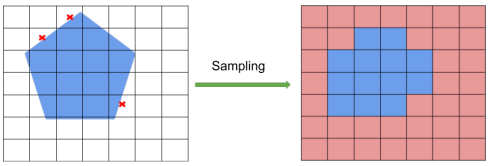

Fig. 3. Undesirable sampling example. The blue areas represent the foreground. Suppose we uniformly sample every grid center point in the left figure, resulting in the point cloud showing in the occupancy grid in the right. Red crosses are undesirable sampling points, which just lie outside the foreground object, making the shape after sampling different from the original one.

denote the dimensions of M. Mask Laplacian computes continuity of M. Further, the mask continuity loss is calculated as $\mathcal{L}_{cont} = \|\nabla^2 M\|^2$ for penalizing discontinuities of M.

3.3 Representation Correspondence

We use the point cloud-based network and the image-based network to extract features and regress M_{3D} and $M_{2.5D}$. The regressed M_{3D} and $M_{2.5D}$ should be similar because they are from the same source of data, disparity, as shown in Fig. 2. To evaluate the similarity, cross-entropy is calculated between M_{3D} and $M_{2.5D}$, and serves as a correspondence loss $\mathcal{L}_{corr}$.

L_{corr} is a *self-supervised* loss. Minimizing this loss term will let the networks of different representations supervise and guide each other to extract more descriptive features for mask regressing, resulting in similar probability distribution between $M_{2.5D}$ and M_{3D}. Thus, the correspondence between different representations from the same data source is ensured through optimizing L_{corr}.

Mask-RCNN uses a 14×14 feature grid after ROI pooling to regress masks. We also use this size at the mask heads of different representations. For 1024-point 3D ROI, after re-projection onto image grids with a size of 32×32, we bilinearly downsample to 14×14 in order to have a uniform mask size.

3.4 Mask Scores and Mask Fusion

MS-RCNN [20] introduces mask scoring, which directly regresses MaskIoU scores based on a predicted mask and its associated matched groundtruth. They first concatenate extracted image features and the predicted mask, and the concatenation is later introduced to the MaskIoU head to regress a MaskIoU score, which represents the quality of the predicted mask. However, their regressed MaskIoU score is not directly used at the inference time to help manipulate mask shapes.

We adopt the mask scoring mechanism and further exploit regressed mask scores to fuse mask predictions from different representations. A predicted mask should score high if its shape is a good fit with the associated groundtruth, and should score low if there is a misalignment between the prediction and the groundtruth. The mask scoring process should not be different for each representation. Therefore, at the training stage, we only use 2D image features and M_{2D} to train a single MaskIoU head instead of constructing 3 MaskIoU heads for each representation. In this way, the MaskIoU module

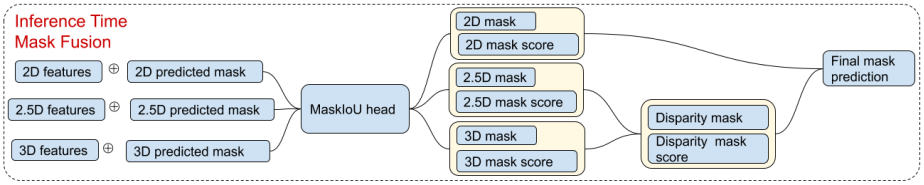

Fig. 4. Inference time mask fusion of predictions from different representations. We fuse the 2.5D mask and 3D mask first because they are from the same source. We then fuse the mask predictions from the image domain and disparity. $\oplus$ represents concatenation. Masks are linearly combined using their associated mask scores. Therefore, a mask with higher score, *i.e.* having a better mask shape, contributes to the final mask more.

would not add much more memory use and the training is also effective. The MaskIoU loss is denoted as $\mathcal{L}_{miou}$.

The total training loss function is formulated as follows.

$$\begin{aligned}\mathcal{L}_{total} = &\overbrace{\mathcal{L}_{cls} + \mathcal{L}_{box} + \mathcal{L}_{2Dmask}}^{\text{from Mask-RCNN}} + \\ &w_D(\mathcal{L}_{2.5Dmask} + \mathcal{L}_{3Dmask}) + w_{corr}\mathcal{L}_{corr} + w_{cont}\mathcal{L}_{cont} + w_m\mathcal{L}_{miou},\end{aligned} \tag{2}$$

where w_D is the weight controlling the disparity mask loss, w_{corr} is for the 2.5D/3D correspondence loss, w_{cont} is for the 3D continuity loss, and w_m is for MaskIoU loss. The mask fusion process is illustrated in Fig. 4. During the inference time, we concatenate features and predicted masks of different representations respectively as inputs to the MaskIoU head. Masks of $M_{2D}, M_{2.5D}$, and M_{3D} and scores of $s_{2D}, s_{2.5D}$, and s_{3D} are outputs from the MaskIoU head. We fuse these mask predictions using their corresponding mask scores. We first linearly combine ($M_{2.5D}$, $s_{2.5D}$) and (M_{3D}, s_{3D}) to obtain (M_D, s_D) for the disparity. The formulation is as follows.

$$M_D = M_{2.5D} \times \frac{s_{2.5D}}{s_{2.5D} + s_{3D}} + M_{3D} \times \frac{s_{3D}}{s_{2.5D} + s_{3D}}, \tag{3}$$

$$s_D = s_{2.5D} \times \frac{s_{2.5D}}{s_{2.5D} + s_{3D}} + s_{3D} \times \frac{s_{3D}}{s_{2.5D} + s_{3D}}. \tag{4}$$

Later, we linearly fuse (M_{2D}, s_{2D}) and (M_D, s_D) likewise to obtain the final probability mask M_f and its corresponding final mask score. The inferred mask is created by binarizing M_f.

4 Experiments

4.1 HQDS Dataset

Outdoor RGBD scene understanding is still less explored since information of images and depth are usually misaligned with low-quality depth acquisition as discussed in Section 1. Current instance segmentation benchmarks lack stereo pairs with long baselines, long focal lengths, and high resolutions for driving scenes.

Dataset	Stereo	Resolution (megapixels)	Stereo Pairs #	Baseline (m)	f_x (pixels)	Measuring distance (km)
COCO	✗	<0.5	-	-	-	-
Mapillary	✗	7.99	-	-	-	-
Cityscapes	✓	2.09	2.7K	0.2	2.2K	up to 0.44
KITTI	✓	0.71	0.2K	0.5	0.7K	up to 0.35
HQDS	✓	**3.15**	**6K**	**0.5**	**3.3K**	up to **1.65**

Table 1. Datasets comparison between collected HQDS and other public datasets for instance segmentation training set. Only COCO [32] is for common objects, and the others are for driving scenes. Stereo pairs # means number of training stereo pairs. Stereo cameras baseline is in meters. f_x is for horizontal focal length in pixels. Measuring distance in kilometers is calculated by $f_x \times$ baseline divided by 1-pixel disparity, showing the ideal farthest possible operating ranges of stereo systems.

To conduct exploration of outdoor RGBD sensor fusion, and provide a high quality platform to fairly evaluate RGBD methods and reveal advantages of sensor fusion, we collect High-Quality Driving Stereo (HQDS) dataset in urban environments. Table 1 shows a comparison with other public datasets for instance segmentation. From the table and Eq. 1. HQDS has the longest $f \times b$. Measuring range by the configuration is up to **1650 meters** with 1-pixel disparity, which is only *440* and *350* for Cityscapes and KITTI. Note that produced disparity maps are computed by stereo matching methods so actual working distances are associated with methods' robustness and image noise. However, as discussed in Section 1, *longer baselines and focal lengths* still favor *far-field stereo matching* for the same stereo matching method. A system with longer baseline and focal length, which produces larger disparity for the same far object compared with a system using shorter parameters (see Eq.1), could show *better geometry and more complete shapes* for objects at distances [14, 36].

Image resolution of HQDS is 1024×3072. The collected images are monochromatic based on the hardware choice and each pixel has a 12-bit response, which could generate more descriptive and robust features for networks compared with normal 8-bit images. Note that ordinary CMOS sensors only have a 8-bit response for each pixel, and later a color filter array is applied to interpolate RGB colors for every pixel.

HQDS contains 6K and 1.2K images for training and testing. We follow a half-automation process to perform data annotations with a group of supervised annotators. Our internal large-scale labeling system produces preliminaries, and the annotators adjust yielded bounding boxes and mask shapes or filter out false predictions to produce groundtruth for HQDS. We also do quality check and is described in the supplementary.

There are 60K instances in the training set and 11K in the testing set. Other datasets on driving adopt more instance classes such as Cityscapes uses 8. However, from MaskR-CNN's study [16] and others [1, 26, 33], their classes are with much inter-class ambiguity and thus cause biased results, such as truck/train/bus/cars or person/rider. Therefore we choose to use 3 classes as human, bicycle/motorcycle, and vehicle. Human includes pedestrians or riders and vehicle includes transportation with covers. Associated number of instances in the training and testing set are (5.5K, 1.5K, 52.8K) and (2.4K, 1K, 8.4K) respectively. Most of non-synthetic datasets encounter class-imbalanced issue,

and compared with Cityscapes, we focus on more crowded vehicle scenes. To remedy the imbalance, we adopt COCO dataset [32], which is a large-scale common objects dataset of instance segmentation with 81 classes, and use their pretrained weights with class pruning in our implementations and comparison methods.

We further collect an additional and more difficult 1.6K testing set under different days and places for more evaluations and it is detailed in the supplementary.

Implementation. We implement GAIS-Net with PyTorch. Our training settings follow the Mask-RCNN experiment on Cityscapes. We adopt ResNet50-FPN as the backbone with pretrained weights on the COCO dataset. We find deeper backbone networks perform similarly, which is also reported in Mask-RCNN on Cityscapes. We use batch size 8 with 8 GPUs and pick the best model within 1x training schedule (50K iterations). During the training and testing, we do not rescale the input image size, since downsizing could cause aliasing or blurring effects to downgrade image quality. **Evaluation and Metrics.** We fairly compare with recent state-of-the-art methods validated on large-scale COCO instance segmentation [32], including Mask-RCNN [16], MS-RCNN [20], Cascade Mask RCNN [5], and HTC [5] (without semantics), by using their publicly released codes and their COCO pretrained weights. We follow their training procedures to conduct comparison experiments.

For evaluation metrics, same as most previous works [5,16,20], we report numerical results in the **standard COCO-style** . Average precision (AP) averages across *different IoU levels*, from 0.5 to 0.95 with 0.05 as an interval. AP_{50} and AP_{75} are 2 typical IoU levels. AP_S and AP_L are AP at small and large scale objects. The units are %. Both bounding box and mask results are reported.

From the comparison results in Table 2, we conclude that our GAIS-Net attains the state of the art compared with other works in nearly all metrics. We exceed Mask-RCNN using the same backbone by 9.7% and 6.8% in bounding box and mask AP, respectively. We further compare with Mask-RCNN using ResNet101 and ResNeXt101 as backbones in Table 3. One can observe that performance of Mask-RCNN with ResNeXt101 is on par with GAIS-Net. However, its number of parameters is nearly double of GAIS-Net. This result validates GAIS-Net design.

Ablation Studies. The importance of individual module is analyzed on HQDS. We follow the fashion of ablation study of previous works [5, 29, 45] and compare GAIS-Net final model with variants including stripping the following modules sequentially. 1) mask scoring and fusion (In this setting, at the inference time the masks are fused with equal weights,) 2) mask continuity loss, 3) representation correspondence loss, 4) 2.5D module (the 2D and 3D masks are fused with equal weights,) and 5) w/ 2.5D module but w/o 3D module (2D and 2.5D masks are fused with equal weights. The results are shown in Table 4. From this study one could find that using 2D and 3D representations has better AP_{50}, and using 2D and 2.5D has better AP_{75}. The performance is further improved by adopting all 3 representations. The other modules all contribute to the final model performance.

It is worth noting that various representations on mask regression, representation correspondence, and mask scoring/ fusion all boost bounding box AP. This result highlights the multi-task learning advantage, which shares information and performs several related tasks simultaneously [24, 40, 43]. In GAIS-Net, the bounding box head, mask

Table 2. Quantitative comparison on HQDS testing set. The first table is for bounding box evaluation. The second table is for mask evaluation. # params means number of parameters

Bbox Evaluation	Backbone	AP	AP_{50}	AP_{75}	AP_S	AP_L	# params
Mask-RCNN	ResNet50+FPN	36.3	57.4	38.8	19.1	51.9	44.1M
MS-RCNN	ResNet50+FPN	42.2	65.1	46.6	20.8	59.6	60.8M
Cascade Mask-RCNN	ResNet50+FPN	37.4	55.8	38.9	18.0	54.7	77.4M
HTC	ResNet50+FPN	39.4	58.3	43.1	18.5	57.9	77.6M
GAIS-Net	ResNet50+FPN	**46.0**	**67.7**	**53.3**	**23.6**	**66.2**	62.6M
Mask Evaluation	Backbone	AP	AP_{50}	AP_{75}	AP_S	AP_L	# params
Mask-RCNN	ResNet50+FPN	33.9	53.2	35.5	14.4	49.7	44.1M
MS-RCNN	ResNet50+FPN	39.2	61.3	40.4	**18.8**	56.4	60.8M
Cascade Mask-RCNN	ResNet50+FPN	33.4	54.4	34.8	11.7	49.5	77.4M
HTC w/o semantics	ResNet50+FPN	34.5	56.9	36.7	11.6	52.0	77.6M
GAIS-Net	ResNet50+FPN	**40.7**	**65.9**	**43.5**	18.3	**59.2**	62.6M

Table 3. Quantitative comparison on HQDS testing set with Mask-RCNN using different backbones.

Bbox Evaluation	Backbone	AP	AP_{50}	AP_{75}	AP_S	AP_L	# params
Mask-RCNN	ResNet50+FPN	36.3	57.4	38.8	19.1	51.9	44.1M
Mask-RCNN	ResNet101+FPN	40.6	62.0	45.9	21.0	59.8	63.1M
Mask-RCNN	ResNeXt101+FPN	43.5	64.7	49.3	22.7	62.4	107M
GAIS-Net	ResNet50+FPN	**46.0**	67.7	53.3	23.6	66.2	62.6M
Mask Evaluation	Backbone	AP	AP_{50}	AP_{75}	AP_S	AP_L	# params
Mask-RCNN	ResNet50+FPN	33.9	53.2	35.5	14.4	49.7	44.1M
Mask-RCNN	ResNet101+FPN	36.4	58.9	35.2	17.6	53.1	63.1M
Mask-RCNN	ResNeXt101+FPN	40.6	62.0	47.9	18.7	59.3	107M
GAIS-Net	ResNet50+FPN	**40.7**	65.9	43.5	18.3	59.2	62.6M

head with different representations, and MaskIoU head share the backbone. Through designing more advanced network modules and introduction of disparity information at the mask heads, and by the backprop, the backbone network could extract more informative features, which are also beneficial for bounding box regression. Therefore, disparity information not only contributes to mask regression, but also helps constrain backbone feature extraction to improve performance on bounding box regression.

Mask-RCNN, MS-RCNN, and HTC also perform multi-task learning. Their bounding box AP performances also increase compared with their baseline approaches, but not prominently. However, our GAIS-Net, compared with Mask-RCNN using ResNet50-FPN, increases the AP by **9.7%** and **6.8%** for box and mask evaluations. This is because our GAIS-Net is a *multi-modal* and *multi-task* learning framework, but other comparing methods are multi-task learning with single modality. GAIS-Net demonstrates the advantage of our multi-task and multi-modal learning design.

We also study different sensor fusion strategies. There are early fusion and late fusion as baseline methods in the sensor fusion context [11, 22, 41]. The former fuses information before information encoding, and the latter is opposite. Here we compare

Table 4. Ablation study on HQDS dataset. Symbol '-' denotes excluding the following module. Repr. Corr. means representation correspondence.

Bbox Evaluation	AP	AP_{50}	AP_{75}	AP_S	AP_L
GAIS-Net	**46.0**	67.7	53.3	23.6	66.2
-Scoring/ Fusion	45.5	67.4	50.7	23.7	65.3
-Mask Continuity	45.5	67.4	50.8	23.2	65.2
-Repr. Corr.	44.6	67.3	49.6	22.4	64.5
-2.5D(w/ 2D&3D)	44.7	67.6	50.8	22.7	64.3
-3D(w/ 2.5D&2D)	44.2	66.4	50.5	23.4	62.6
Mask Evaluation	AP	AP_{50}	AP_{75}	AP_S	AP_L
GAIS-Net	**40.7**	65.9	43.5	18.3	59.2
-Scoring/ Fusion	40.3	65.6	43.5	17.8	59.0
-Mask Continuity	40.0	65.3	43.2	17.6	58.4
-Repr. Corr.	39.7	65.1	41.9	17.8	57.9
-2.5D(w/ 2D&3D)	36.0	66.2	32.1	17.2	52.0
-3D(w/ 2.5D&2D)	39.8	61.6	43.6	17.7	57.1

Table 5. Network structure comparison with early fusion and late fusion. We use only 2D and 3D representations in these methods for this comparison.

Bbox Evaluation	AP	AP_{50}	AP_{75}	AP_S	AP_L
Early Fusion	33.5	51.8	36.9	16.0	49.8
Late Fusion	34.4	54.2	36.4	15.6	49.2
GAIS-Net	**44.7**	**67.6**	**50.8**	**22.7**	**64.3**
Mask Evaluation	AP	AP_{50}	AP_{75}	AP_S	AP_L
Early Fusion	29.9	50.2	28.6	11.7	45.3
Late Fusion	29.2	48.8	28.9	11.1	42.1
GAIS-Net	**36.0**	**66.2**	**36.1**	**17.2**	**52.0**

with early fusion and late fusion network designs. The examined structures are illustrated in Fig. 5. We also conduct an ablation study on the architecture of fusion in the supplementary.

Qualitative Results. We demonstrate qualitative comparison with other works in Fig. 6. These examples show the advantage of using both images and disparity maps. We give more visual results in the supplementary.

4.2 Cityscapes Dataset

We also conduct experiments on Cityscapes dataset. Cityscapes is currently the public dataset on instance segmentation with stereo pairs and enough training data, as seen in Table 1. Cityscapes contains 2725 and 500 images with fine annotations in their training and validation set. Besides, coarse annotations for training are provided. Their script evaluates only mask AP.

However, Cityscape's baseline and focal length are shorter than HQDS, and the maximal measuring distance is only **1/4** of HQDS. Much shorter focal length and baseline **limit the working distance** of stereo matching and produce disparity maps only focusing at near fields with *poor shapes and geometry* [14, 36], as discussed in Section 4.1 on the dataset comparison.

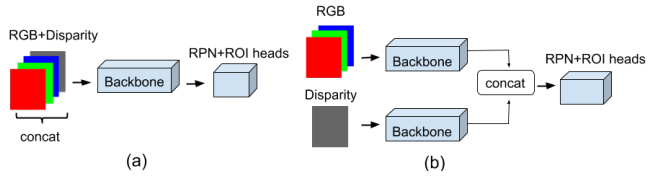

Fig. 5. Early fusion framework and late fusion framework as baseline methods.

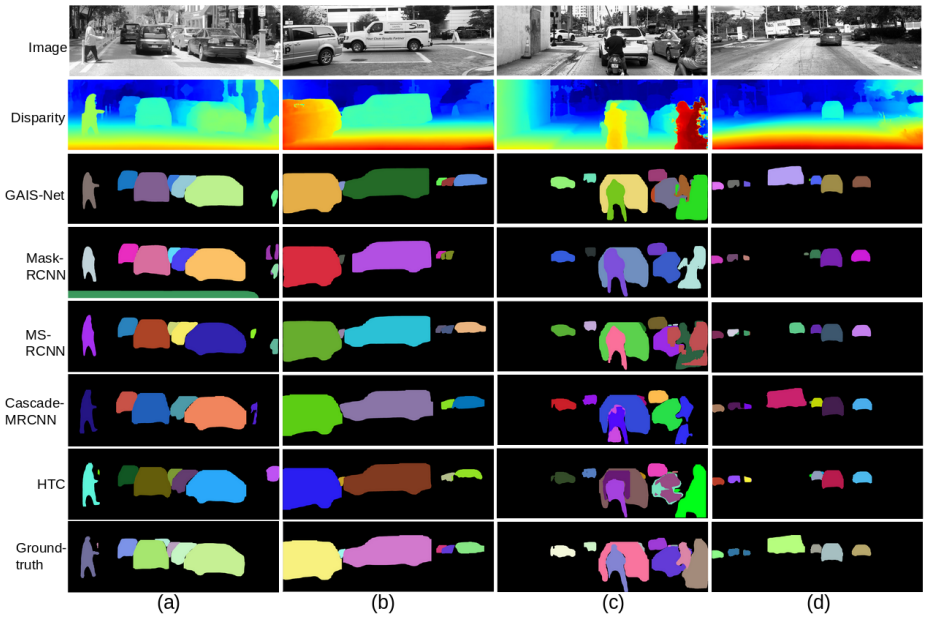

Fig. 6. Qualitative comparison on HQDS. In column (a), there is an area of false prediction of a car on the lower side in the Mask-RCNN's result, which stems from false predictions (engine hood of self car) of preliminary labeling during the half-automation annotation process. During the manual stage, we inspect and deliberately leave 60 false car predictions on the lower side in the training data to examine robustness to these adversarial training examples representing different data processing pipelines. (Some datasets for driving do not crop out self-car engine hood from images.) Compared with MaskRCNN, GAIS-Net uses another modality to suppress the false detection based on the scene geometry. We show more results using geometric information to suppress false detections in the supplementary. In column (b), GAIS-Net has the best segmentation result with the most similar mask shapes to the groundtruth. In column (c), there is an infrequent human pose. If using only the image information, Mask-RCNN could detect the rider at the rear, but the rider ahead has a poor mask shape. Using another modality, GAIS-Net is able to further regress both shapes of the two riders. In column (d), some methods do not predict the truck since its intensity values at the head and body are different. Besides, MS-RCNN only predicts the truck head.

We show the numerical results on validation set in Table 6 and compare with other methods on Cityscapes using numbers reported in their paper. From Table 6, one can see that GAIS-Net has better performances than Mask-RCNN using ResNet50-FPN. (Mask-RCNN only reports Cityscapes results with ResNet50-FPN and claims using a deeper backbone gets similar results.) The improvement gap between HQDS and Cityscapes is mainly caused by the latter's much shorter baseline and focal length. We search over all Cityscapes's disparity from PSMNet and found 9-disparity is the minimum where a rough object shape could be shown, corresponding to about *50 meters*, checked with distances on city maps and their GPS. Cityscapes also officially provides sparse disparity maps using SGM [18], and their analysis shows a similar result. We illustrate the analysis in Fig. 7. By contrast, HQDS could show rough object shapes *over 150 meters* referring to distances on city maps and our GPS. Moreover, quality of disparities is adversely affected by Cityscapes lower resolutions of images and optical defocus for objects at distances. The experiment further validates the importance and our sensor fusion exploration of incorporating geometric information for instance segmentation using longer baseline and focal length, since providing GAIS-Net with higher quality disparity maps significantly improves performance in Table 2.

5 Conclusion and Future Work

From the motivation for sensor fusion, with the aid of disparity information from stereo pairs, which gives a geometry prior of scenes, GAIS-Net improves instance segmentation performance and attains the state of the art. We list 3D bounding box and shape inference from 2D boxes and masks as the future work. In this work, we choose disparity as the representation to avoid quadratic error issue in depth. 3D box inference

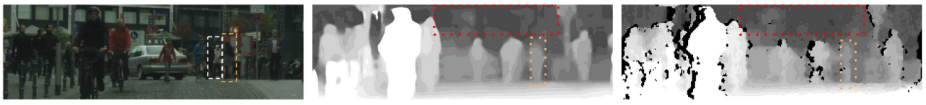

Fig. 7. Disparity maps of Cityscapes. For better visualization, we multiply disparity values by **15**. Left: image; middle: disparity from PSMNet; right: officially provided disparity using SGM. Women in the orange box could be roughly seen in results from PSMNet and SGM (9-disparity), but people in the white box couldn't be seen in both. (No human-shape beside the orange boxes in PSMNet and SGM.) Also, the background shows irregular shapes in red boxes.

Table 6. Instance segmentation results on Cityscapes datset.

Evaluation	Training data	Backbone	Mask AP
DWT [1]	fine + coarse	-	19.8
SGN [33]	fine + coarse	-	29.2
BshapeNet [23]	fine only	-	32.1
Mask-RCNN [16]	fine only	ResNet50-FPN	31.5
Our GAIS-Net	fine only	ResNet50-FPN	**32.5**
Mask-RCNN [16]	fine + COCO	ResNet50-FPN	36.4
Our GAIS-Net	fine + COCO	ResNet50-FPN	**37.1**

requires depth values. Longer baseline and focal also have the advantage to suppress depth error from the error relationship $|\Delta Z| = Z^2 \frac{\Delta m}{bf}$ in Section 3.1.

References

1. Bai, M., Urtasan, R.: Deep watershed transform for instance segmentation. In: CVPR. pp. 5221–5229 (2017) 2, 3, 9, 14
2. Cai, Z., Vasconcelos, N.: Cascade r-cnn: Delving into high quality object detection. In: CVPR. pp. 6154–6162 (2018) 4
3. Chang, J.R., Chen, Y.S.: Pyramid stereo matching network. In: CVPR. pp. 5410–5418 (2018) 2, 5, 6
4. Chang, M.F., Lambert, J., Sangkloy, P., Singh, J., Bak, S., Hartnett, A., Wang, D., Carr, P., Lucey, S., Ramanan, D., et al.: Argoverse: 3d tracking and forecasting with rich maps. In: CVPR. pp. 8748–8757 (2019) 2
5. Chen, K., Pang, J., Wang, J., Xiong, Y., Li, X., Sun, S., Feng, W., Liu, Z., Shi, J., Ouyang, W., et al.: Hybrid task cascade for instance segmentation. In: CVPR. pp. 4974–4983 (2019) 2, 4, 10
6. Cheng, J., Tsai, Y.H., Hung, W.C., Wang, S., Yang, M.H.: Fast and accurate online video object segmentation via tracking parts. In: CVPR. pp. 7415–7424 (2018) 2
7. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: CVPR. pp. 3213–3223 (2016) 3
8. De Brabandere, B., Neven, D., Van Gool, L.: Semantic instance segmentation with a discriminative loss function. In: CVPRW (2017) 4
9. Fathi, A., Wojna, Z., Rathod, V., Wang, P., Song, H.O., Guadarrama, S., Murphy, K.P.: Semantic instance segmentation via deep metric learning. arXiv preprint arXiv:1703.10277 (2017) 4
10. Gao, N., Shan, Y., Wang, Y., Zhao, X., Yu, Y., Yang, M., Huang, K.: Ssap: Single-shot instance segmentation with affinity pyramid. In: ICCV (2019) 4
11. Giancola, S., Zarzar, J., Ghanem, B.: Leveraging shape completion for 3d siamese tracking. In: CVPR. pp. 1359–1368 (2019) 11
12. Goesele, M., Snavely, N., Curless, B., Hoppe, H., Seitz, S.M.: Multi-view stereo for community photo collections. In: 2007 IEEE 11th International Conference on Computer Vision (ICCV). pp. 1–8. IEEE (2007) 2
13. Gool, V., et al.: Dense matching of multiple wide-baseline views. In: Proceedings Ninth IEEE International Conference on Computer Vision (ICCV). pp. 1194–1201. IEEE (2003) 2
14. Hartley, R., Zisserman, A.: Multiple view geometry in computer vision. Cambridge university press (2003) 2, 9, 12
15. Hazirbas, C., Ma, L., Domokos, C., Cremers, D.: Fusenet: Incorporating depth into semantic segmentation via fusion-based cnn architecture. In: ACCV. pp. 213–228 (2016) 2
16. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: ICCV. pp. 2961–2969 (2017) 2, 3, 5, 9, 10, 14
17. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016) 5
18. Hirschmuller, H.: Stereo processing by semiglobal matching and mutual information. IEEE Transactions on pattern analysis and machine intelligence (TPAMI) **30**(2), 328–341 (2007) 6, 14
19. Hu, Y.T., Huang, J.B., Schwing, A.: Maskrnn: Instance level video object segmentation. In: NIPS. pp. 325–334 (2017) 2
20. Huang, Z., Huang, L., Gong, Y., Huang, C., Wang, X.: Mask scoring r-cnn. In: CVPR. pp. 6409–6418 (2019) 4, 7, 10
21. IJsselsteijn, W.A., de Ridder, H., Vliegen, J.: Subjective evaluation of stereoscopic images: effects of camera parameters and display duration. IEEE Transactions on Circuits and Systems for Video Technology (TCSVT) **10**(2), 225–233 (2000) 2

22. Jaritz, M., De Charette, R., Wirbel, E., Perrotton, X., Nashashibi, F.: Sparse and dense data with cnns: Depth completion and semantic segmentation. In: 3DV. pp. 52–60 (2018) 11
23. Kang, B.R., Kim, H.Y.: Bshapenet: Object detection and instance segmentation with bounding shape masks. arXiv preprint arXiv:1810.10327 (2018) 14
24. Kendall, A., Gal, Y., Cipolla, R.: Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In: CVPR. pp. 7482–7491 (2018) 10
25. Kirillov, A., He, K., Girshick, R., Rother, C., Dollar, P.: Panoptic segmentation. In: CVPR (June 2019) 4
26. Kirillov, A., Levinkov, E., Andres, B., Savchynskyy, B., Rother, C.: Instancecut: from edges to instances with multicut. In: CVPR. pp. 5008–5017 (2017) 3, 9
27. Ku, J., Mozifian, M., Lee, J., Harakeh, A., Waslander, S.L.: Joint 3d proposal generation and object detection from view aggregation. In: IROS. pp. 1–8. IEEE (2018) 4
28. Lhuillier, M., Quan, L.: A quasi-dense approach to surface reconstruction from uncalibrated images. IEEE transactions on pattern analysis and machine intelligence (TPAMI) **27**(3), 418–433 (2005) 2
29. Liang, M., Yang, B., Chen, Y., Hu, R., Urtasun, R.: Multi-task multi-sensor fusion for 3d object detection. In: CVPR. pp. 7345–7353 (2019) 4, 10
30. Liang, M., Yang, B., Wang, S., Urtasun, R.: Deep continuous fusion for multi-sensor 3d object detection. In: ECCV. pp. 641–656 (2018) 4
31. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: CVPR. pp. 2117–2125 (2017) 5
32. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: ECCV. pp. 740–755. Springer (2014) 9, 10
33. Liu, S., Jia, J., Fidler, S., Urtasun, R.: Sgn: Sequential grouping networks for instance segmentation. In: ICCV. pp. 3496–3504 (2017) 9, 14
34. Liu, S., Qi, L., Qin, H., Shi, J., Jia, J.: Path aggregation network for instance segmentation. In: CVPR. pp. 8759–8768 (2018) 4
35. Neven, D., Brabandere, B.D., Proesmans, M., Gool, L.V.: Instance segmentation by jointly optimizing spatial embeddings and clustering bandwidth. In: CVPR. pp. 8837–8845 (2019) 4
36. Okutomi, M., Kanade, T.: A multiple-baseline stereo. IEEE Transactions on pattern analysis and machine intelligence (TPAMI) **15**(4), 353–363 (1993) 2, 9, 12
37. Qi, C.R., Liu, W., Wu, C., Su, H., Guibas, L.J.: Frustum pointnets for 3d object detection from rgb-d data. In: CVPR. pp. 918–927 (2018) 2, 4
38. Qi, C.R., Su, H., Mo, K., Guibas, L.J.: Pointnet: Deep learning on point sets for 3d classification and segmentation. In: CVPR. pp. 652–660 (2017) 6
39. Qi, X., Liao, R., Jia, J., Fidler, S., Urtasun, R.: 3d graph neural networks for rgbd semantic segmentation. In: ICCV. pp. 5199–5208 (2017) 2
40. Qi, X., Liu, Z., Chen, Q., Jia, J.: 3d motion decomposition for rgbd future dynamic scene synthesis. In: CVPR. pp. 7673–7682 (2019) 10
41. Qiu, J., Cui, Z., Zhang, Y., Zhang, X., Liu, S., Zeng, B., Pollefeys, M.: Deeplidar: Deep surface normal guided depth prediction for outdoor scene from sparse lidar data and single color image. In: CVPR. pp. 3313–3322 (2019) 11
42. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: NIPS. pp. 91–99 (2015) 3, 5
43. Vandenhende, S., De Brabandere, B., Van Gool, L.: Branched multi-task networks: Deciding what layers to share. arXiv preprint arXiv:1904.02920 (2019) 10
44. Wang, Q., Zhang, L., Bertinetto, L., Hu, W., Torr, P.H.: Fast online object tracking and segmentation: A unifying approach. In: CVPR. pp. 1328–1338 (2019) 2

45. Wang, W., Ceylan, D., Mech, R., Neumann, U.: 3dn: 3d deformation network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1038–1046 (2019) 10
46. Wang, W., Neumann, U.: Depth-aware cnn for rgb-d segmentation. In: ECCV. pp. 135–150 (2018) 2
47. Wang, Y., Chao, W.L., Garg, D., Hariharan, B., Campbell, M., Weinberger, K.Q.: Pseudo-lidar from visual depth estimation: Bridging the gap in 3d object detection for autonomous driving. In: CVPR. pp. 8445–8453 (2019) 4, 5
48. Williamson, T., Thorpe, C.: A specialized multibaseline stereo technique for obstacle detection. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 238–244. IEEE (1998) 2
49. Xiong, Y., Liao, R., Zhao, H., Hu, R., Bai, M., Yumer, E., Urtasun, R.: Upsnet: A unified panoptic segmentation network. In: CVPR. pp. 8818–8826 (2019) 4
50. Yang, B., Liang, M., Urtasun, R.: Hdnet: Exploiting hd maps for 3d object detection. In: CoRL. pp. 146–155 (2018) 2
51. Yang, G., Manela, J., Happold, M., Ramanan, D.: Hierarchical deep stereo matching on high-resolution images. In: CVPR. pp. 5515–5524 (2019) 2
52. Ye, L., Liu, Z., Wang, Y.: Depth-aware object instance segmentation. In: ICIP. pp. 325–329 (2017) 2, 4
53. Zhang, F., Prisacariu, V., Yang, R., Torr, P.H.: Ga-net: Guided aggregation net for end-to-end stereo matching. In: CVPR. pp. 185–194 (2019) 2
54. Zhang, Z., Fidler, S., Urtasun, R.: Instance-level segmentation for autonomous driving with deep densely connected mrfs. In: CVPR. pp. 669–677 (2016) 3
55. Zhang, Z., Schwing, A.G., Fidler, S., Urtasun, R.: Monocular object instance segmentation and depth ordering with cnns. In: CVPR. pp. 2614–2622 (2015) 3