Weakly Supervised Approach for Joint Object and Lane Marking Detection

Pranjay Shyam[1], Kuk-Jin Yoon[2], Kyung-Soo Kim[1]
[1]Mechatronics Systems and Control Lab, [2]Visual Intelligence Lab
Korea Advanced Institute of Science and Technology (KAIST)

Abstract

Understanding the driving scene is critical for the safe operation of autonomous vehicles with state-of-the-art (SoTA) systems relying upon a combination of different algorithms to perform tasks for mathematically representing an environment. Amongst these tasks, lane and object detection are highly popular and have been extensively researched independently. However, their joint operation is rarely studied primarily due to the lack of a dataset that captures these attributes together, resulting in increased redundant computations that can be eliminated simply by performing these tasks together. To overcome this, we propose a weakly-supervised approach wherein, given an image from the lane detection dataset, we use a pretrained network to label different objects within a scene, generating pseudo bounding boxes used to train a network that jointly detects objects and lane lines. With an emphasis on inference speed and performance, we utilize prior works to construct two architectures based on Convolutional Neural Networks (CNNs) and Transformers. The CNN-based approach uses row-based pixel classification to detect and cluster lane lines alongside a single-stage anchor free object detector while sharing the same encoder backbone. Alternatively, using dual decoders, the transformer-based approach directly estimates bounding boxes and polynomial coefficients of lane lines. Through extensive qualitative and quantities experiments, we demonstrate the efficacy of the proposed architectures on leading datasets for object and lane detections and report state-of-the-art (SoTA) performance per GFLOPs. Codes with trained model will be available at https://github.com/PS06/JOLD

1. Introduction

Object and Lane detection form a core component within modern advanced driving assistance systems (ADAS) and find application in features such as lane-keeping, collision avoidance, visual positioning, adaptive cruise control, autonomous navigation (in autonomous vehicles AVs), etc. with vision sensors being the primary data source. To ensure consistent performance in diverse weather and illumination conditions, algorithms performing these tasks are constructed using convolution neural networks that provide SoTA performance while being robust, unlike traditional computer vision algorithms that work well, only in certain scenarios. However, present approaches formulate these tasks independently, resulting in multiple repetitious computations that increase the computational complexity while bypassing information sharing that could boost the performance of these tasks hence reducing latency.

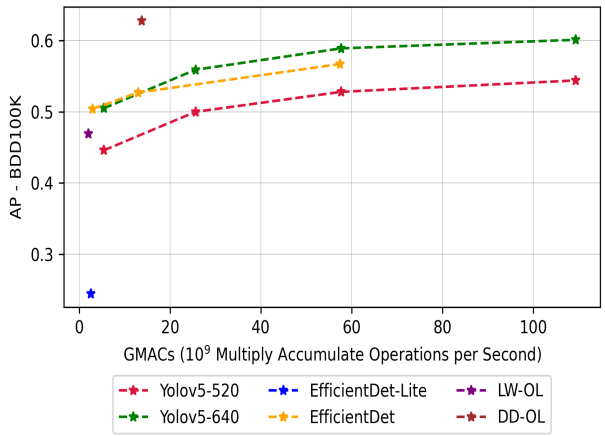


Figure 1: Performance Landscape of SoTA 2D Object Detection Algorithms on BDD100K dataset.

To avoid redundant computations and better leverage the common features, MultiNet [53] proposed a mechanism to classify road scenes, perform object detection, and segment road areas via a single encoder and triple decoder architecture. However, the computational requirements were considered expensive for embedded systems to overcome, following which [48] proposed a ResNet-10 [18] based encoder and dual decoders for performing bounding box regression and segmentation of road and sidewalks on KITTI dataset [12]. Despite these advances, such approaches are limited by attributes available in training dataset and hence cannot capture finer details such as lane structure, traffic signs and lights, different types of objects (traffic cones, debris, animals, etc.), and vehicles (Trucks, Pickups, etc.)

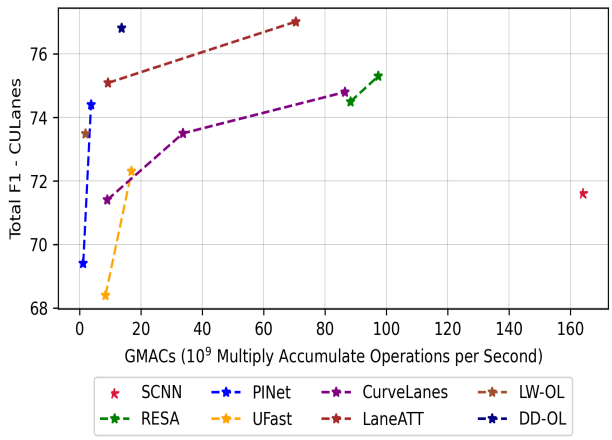


Figure 2: Performance Landscape of SoTA Lane Detection Algorithms on CULanes dataset.

present on the road. A simplistic approach to overcome this could be extending the labels within the dataset, but this might be unfavorable due to large labeling time requiring knowledge expertise of label formats.

Hence an alternative mechanism is needed to train multi-task networks to exploit their full potential vis-a-vis their task-specific counterparts for lane and object detection. Towards this goal, we propose a weakly supervised mechanism wherein we utilize bounding box labels generated using a pretrained object detector to train a multi-task network for jointly performing object and lane detection, assuming prior lane information in the training dataset. Thus, we extend commonly used lane detection datasets to contain object bounding box information additionally. This approach is motivated by the fact that current lane detection algorithms require access to at least one lane segment to extrapolate it into lane lines and are thus sensitive at intersections and occluded conditions. Contrarily, object detection algorithms are more robust in diverse weather and occlusion conditions, hence bounding boxes generated using a pretrained network can be used as pseudo ground truth for training the joint network.

With specific goals of maximizing performance in terms of latency and accuracy, we examine two distinct architectural configurations focusing on specific deployment objectives. To reduce latency, we emphasize upon a CNN based architecture wherein multi-level features could be utilized to jointly perform lane and object detection. To ensure reduced computations and small model for deployment on edge devices, we utilize the formulation proposed in UFAST [41] for performing lane detection, wherein given predefined row anchors, the task is to perform row-wise lane classification. We leverage anchor-free object detection formulation to bypass computations needed to perform non-maximum-suppression (NMS) for object detection. While this approach ensures low latency, it comes at the cost of reduced performance, still being comparable to SoTA for both object (Fig. 1) and lane detection (Fig. 2), that might not be desired in situations where computational limitations are non-existent. For such conditions, we propose a transformer based architecture having dual decoders with each decoder performing direct set predictions for object and lane detection. Our motivation for using transformers stems from its strength to effectively capture long-range dependencies that are useful for detecting both occluded objects and lane lines. Thus we summarize our contributions as,

- Propose a weakly supervised framework that can be used to jointly train object and lane detectors.
- For resource constraint devices, we propose CNN based object and lane detection algorithm.
- For situations with relaxed computational requirements, we propose a dual decoder transformer based architecture that leverage long range dependencies to improve performance in occluded conditions.
- Compare performance of proposed architectures with task specific SoTA algorithms to demonstrate viability of the training mechanism as well as efficacy of the proposed architectures.

2. Related Works

Lane Detection : Classical approaches for lane detection developed different masking operations [56, 60, 21] to segment and extract lane markers which are extrapolated to generate lane lines using curve fitting techniques such as Hough transform [30, 23] and RANSAC [3]. However sensitivity of these operations towards illumination and occlusion resulted in multiple failure scenarios. To overcome these limitations CNN based algorithms are proposed that formulate this task either as regression by predicting polynomial coefficients of lane lines [59, 13, 55, 40, 26, 49, 41, 64, 63, 31] or semantic segmentation by pixel wise labeling of lane segments [38, 20, 14, 68]. LaneNet [36] utilized this segmentation as intermediate representation to further improve estimation of lane lines by integrating perspective transformation information through a separate CNN. While these works relied on lane marker information for determining lane lines, several works focused on improving performance by extracting additional features for providing geometric and structural cues. Specifically, VPGNet [28] proposed a joint mechanism for extracting road and lane information such as lane estimation, road marking detection and classification and vanishing point detection using a single encoder and multiple (four) decoders. Subsequently different works relied on additional approaches such as key point estimation i.e. PINet [26], neural architecture search i.e. CurveLane-NAS [63].

Object Detection : Common object detection pipeline can be categorized either as single or two stage wherein a single stage network predicts bounding boxes directly which are then refined either using post processing techniques such as Non Maximum Suppression (NMS for anchor based object detectors) or using CNNs (for two stage object detectors [16, 17, 5]). While a complete review of SoTA object detectors is beyond the scope of this work, we provide a brief overview on single stage object detection algorithms are designed with an emphasis on speed. Notably, Yolo [42] proposed to regress bounding boxes directly whereas SSD [34] proposed to regress deviations for predefined anchor boxes, requiring a non maximum suppression operation to remove weak detections. Taking a different approach, CenterNet [11] proposed to detect object centers using which the width and height of the bounding boxes could be estimated. Apart from different formulations to estimate object bounding boxes, emphasis was also given to the loss function being minimized such as focal loss [29], Side Aware Boundary Loss [58] or Non maximum suppression with Soft-NMS [2] or DANet [45] proposing a dynamic anchor selection mechanism to automatically reduce the number of bounding boxes using IoU and class probabilities.

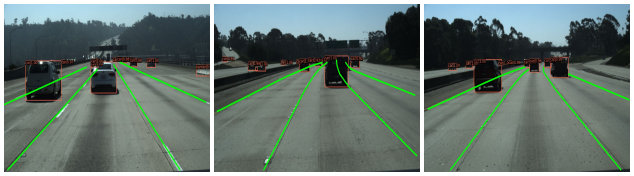
Multi-Task Frameworks : Despite its advantages, multi-task networks are not well studied with initial approach MultiNet [53] using VGG-16 [47] as the encoder and three decoders, with detection decoder using regression mechanism whereas segmentation decoder follows FCN architecture [35] and classification decoder is constructed using fully connected layer with softmax. To reduce computational overhead [48] used ResNet-10 while using two decoders for performing the task of detection and segmentation. However limitations of training dataset hinder in reaching peak performance, which we aim to overcome using pseudo labels.

3. Methodology

3.1. Problem Formulation

We underline lack of dataset with diverse attributes as one of the bottlenecks hindering research into multi-task learning of lane and object detection. To avoid this we propose a weakly supervised approach wherein we use a pretrained object detection algorithm to generate bounding boxes of objects of interest on current lane detection datasets. The combined labels can then be used to train networks that jointly identifies and localized lane and objects within an image while sharing information between them. As computational resources dictate the peak performance of an algorithm, we design two architectures focusing on either latency or accuracy.

(a) CULane Dataset



(b) TUSimple Dataset

(c) CurveLanes Dataset

Figure 3: Extended datasets containing pseudo bounding box labels (red) with prior lane information (green).

3.2. Generating Pseudo Bounding Boxes

While there are multiple datasets focused on detecting different objects that can be encountered in real driving conditions, we choose NuScenes [4], and BDD100K [65] datasets that collectively contain attributes such as car, truck, trailer, bus, construction vehicle, bicycle, motorcycle, pedestrian, traffic cone, traffic sign, traffic light, and barrier. We used pretrained NuScenes weights from mmdet3d [10] for Hybrid Task Cascade Network (HTC) [8], Cascade Mask-RCNN [6] algorithms with ResNeXt-101 [62] as backbone, however, for BDD100K dataset we retrained the networks following training methodology mentioned in mmdet3D[10]. These pretrained models are then inferenced on lane detection datasets to generate bounding box labels, and boxes with a $Confidence \geq 0.8$ are saved to ensure high-quality labels. As we run models trained using different datasets, there would be conditions wherein different bounding boxes enclose the same object. Hence for deduplication, we first identify duplicates by computing IoU, and when $IoU \geq 0.75$, these boxes would be considered as duplicates. We then discard the bounding box with a lower confidence score. While model ensembling is preferred to obtain high-quality labels, we observed this approach not to work as effectively while requiring higher computational resources. An alternative mechanism to use semantic labels to generate high-quality bounding boxes is ineffective since current segmentation models are performance limited by domain and illumination changes, generating higher false positives since current SoTA semantic segmentation models cannot identify unidentified road objects. Hence we found

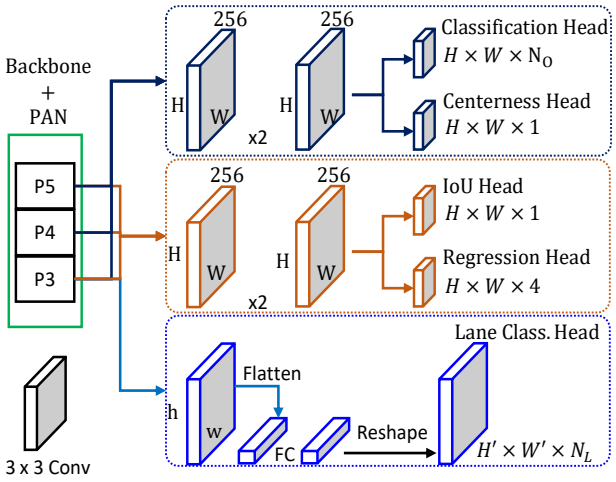


Figure 4: Architecture of the proposed lightweight object and lane detector.

the proposed approach as ideal to generate pseudo labels. The final pseudo bounding boxes along with existing lane information are visualized for different datasets in Fig. 3. They are subsequently used for training the networks to detect object and lane lines jointly.

3.3. Light Weight Joint Architecture

For edge devices where computational resources are limited, it's desirable to have a lightweight network that can perform these tasks with sufficient accuracy. To ensure such performance characteristics, we utilize CNN as a feature extractor and integrate a feature enhancement network that ensures features across all levels have high semantic and spatial information. To perform this, we use Path Aggregation Networks [33] and use enhanced features from different levels for predicting (1) bounding boxes that enclose an object, (2) its category, and (3) segmenting the presence of lane markers for predefined row anchors.

To perform object detection, we follow decoupled head approach wherein two branches are tasked to perform classification and regression independently, resulting in improved performance, and convergence speed [61, 29, 54]. Hence, we first reduce the number of channels from each layer to 256 using 1×1 convolutions followed by 2 3×3 convolutions. The classification branch has N_O channels for O number of objects (12 objects + 1 background). Furthermore, we additionally perform center-ness detection proposed in FCOS [54] to classify whether an object center is within the cell of the feature map. This allows us to reduce the number of computations by assigning one cell the task of predicting a single bounding box where the values to be predicted are coordinates of the top-left point of the bounding box and width and height. Following this approach allows us to reduce the number of parameters and aids in performance as we can select the regression results from detections that have a center lying in the cell. However, in the case of occluded objects, filtering predictions based on center location might result in dropping bounding boxes with an object. To avoid such a scenario, we include an IoU head that estimates the IoU of a regressed bounding box. Combining these two allows us to filter out bounding boxes without performing non-maximum suppression.

For performing lane detection, we follow the approach proposed in UFAST and construct a lane classification head that uses features from the backbone network to classify lane presence on predefined row anchors. Unlike the original UFAST network that relied on global features obtained at the backbone base, we use features from P3 layers as they provide richer spatial and semantic content providing improved performance. For our implementation, we assume constant row anchors to 60 (H') and a number of gridding cells to 200 (W') with N_L set to 6 for CULane and TuSimple datasets and 11 for CurveLanes dataset.

To train the proposed framework (Fig. 4, referred hereafter as LW-OL), we use binary cross-entropy (BCE) loss for training the classification, IoU and centerness branch, and IoU loss [67] for training the regression branch, whereas for the lane detection branch, we use cross-entropy loss (CE). We then train the complete network for 100 epochs for TuSimple, 300 epochs for CULanes, and CurveLanes dataset with an initial learning rate of 0.001, image size of 512 $\times$ 512, batch size of 4, and ADAM optimizer [24] on a system equipped with RTX3090 GPU. Furthermore, we utilize random flipping, rotating, color jitter with probabilities of 0.5 and copy-paste [15, 44] with maximum number of samples as 5, as data augmentations techniques.

3.4. Performance focused Transformer Architecture

CNN-based object and lane detectors observe a performance drop in occluded conditions wherein the same object or lane might be considered different, increasing the number of false positives. To ensure consistent performance in such scenarios, we require effective modeling of features present across complete feature space issued to directly predict a single bounding box or lane line. For this, DETR [7] reformulated object detection using a transformer-based encoder-decoder architecture that model feature relationship along a sequence and leverage it to predict bounding boxes. Subsequently, object detection is treated as set prediction tasks wherein a single predicted bounding box is assigned to ground truth, eliminating NMS's need. Following similar motivation, LSTR [32] demonstrated that lane detection could also be performed via a similar approach of predicting lane line parameters, thereby avoiding any clustering or post-processing. To further improve lane detection performance LSTR reparamterized the curve lanes to also

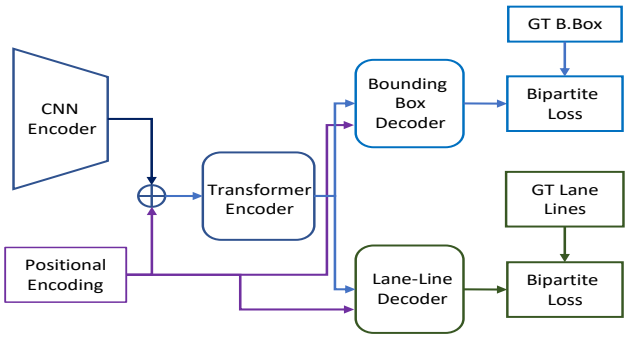


Figure 5: Architecture of the proposed dual decoder object and lane detector.

account for camera position w.r.t road surface.

As these two approaches largely have a common architecture to a great extent, they are ideal candidates for designing a transformer-based model that jointly estimates object bounding boxes and lane line parameters using two decoupled decoders relying on feature relationship modelled by the encoder. However, we observe that relying on commonly used curve parameter approach assumes a common horizon line where all lane lines end. This is an inaccurate model as it cannot capture lane information at intersections or horizontal lane lines wherein the number of horizon lines increases. To avoid this limitation, we instead estimate all the coefficients of lane lines following the polynomial model of degree 3 following PolyLaneNet [49]. Apart from the polynomial coefficient, a starting point is also estimated resulting in a total of 6 parameters to be estimated per lane line (4 for lane model and 1 each for start point, end point). As we use bipartite matching loss, the model would ensure a single lane line predicted for corresponding ground truth. Hence we train the dual decoder transformer (Fig. 5, referred hereafter as DD-OL) for 500 epochs for TuSimple, 1000 epochs for CULanes, and CurveLanes dataset with an initial learning rate of 0.001, image size of 512×512, batch size of 4, and ADAM optimizer [24] on a system equipped with RTX Titan GPU following the same data augmentation techniques as mentioned above. The larger epoch requirement for training transformer-based models compared to CNN-based ones arises from the optimization of a larger parameter space.

4. Experimental Analysis

4.1. Datasets and Evaluation Metrics

To evaluate the performance on lane detection task, we use CULanes [39], TUSimple [1] and CurveLanes [63] datasets that have been widely used in lane detection literature. These have {88880; 9675; 34680}, {3268; 358; 2782} and {100000; 20000; 30000} train, val and test images with at most 5, 5, and 9 concurrent lane lines being present in an image. Furthermore image resolution for these datasets is 1280×720, 1640×540 and 2650×1440 for TUSimple, CULanes and CurveLanes respectively. For evaluating lane detection performance, we follow dataset-specific metrics such as,

- TuSimple follows three metrics namely false positive (FP), false negative (FN) and accuracy that is calculated following the relation $Acc. = \frac{\sum_c lipC_{clip}}{\sum_c lipS_{clip}}$ wherein C_{clip} refers to number of correctly estimated lane points and S_{clip} refers to total number of ground truth points in each clip.

- CULanes measures the IoU (Intersection-over-Union) for a 30 pixel-width predicted and ground truth lane line with predictions having an IoU larger than 0.5 are considered as true positive (TP). F1 measure is subsequently calculated as $F1 = \frac{2 \times Precision \times Recall}{Precision + Recall}$ where $Precision = \frac{TP}{TP+FP}$ and $Recall = \frac{TP}{TP+FN}$ where FP and FN refer to false positive and false negative respectively. In addition, CULanes dataset also provide performance in diverse conditions such as normal, crowded, night, no-line, shadow, arrow, dazzle, curve and crossroads.

- CurveLanes dataset follows similar approach of CULanes and uses F1, Precision and Recall to evaluate different algorithms.

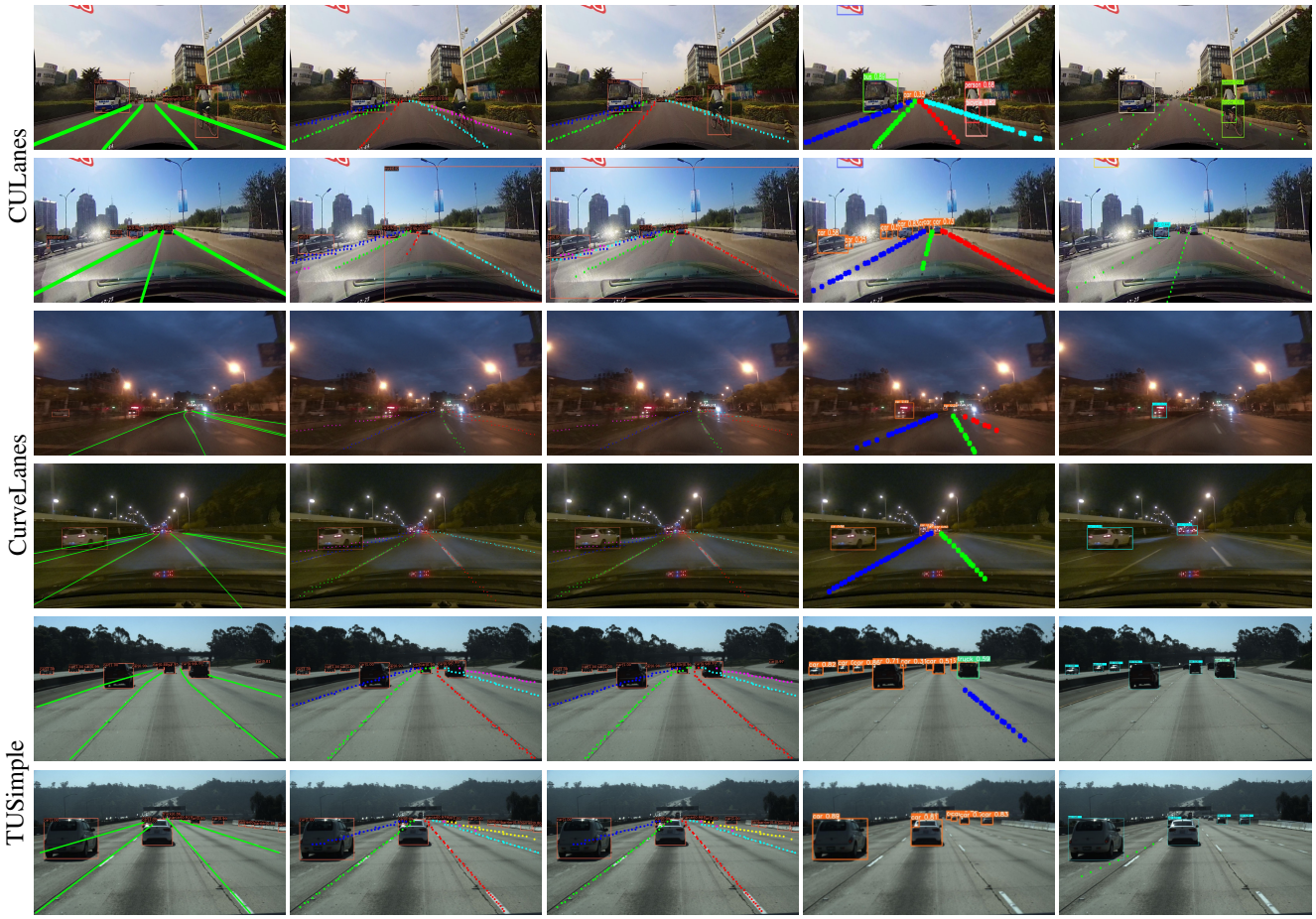
As we use CNN for generating pseudo labels, evaluating object detection performance on lane detection datasets might not provide accurate results due to missed detections, despite its accuracy on training dataset due to domain shift. Hence we utilize BDD100K dataset to evaluate the performance of SoTA object detection algorithms along with the proposed architecture and use standard metrics such as average precision (AP) and average precision at 0.5 IoU.

4.2. Lane Detection

We summarize the performance landscape of SoTA lane detection algorithms on CurveLanes, CULanes and TuSimple datasets in Tab.1, Tab.2 and Tab.3 respectively with qualitative results in Fig. 6. Apart from dataset specific metrics we also summarize the backbones, input image resolution and corresponding GMACs[1] to provide a better overview of associated computational cost. For our evaluations we choose lane detection algorithms such as SCNN [38], SAD [19], RESA [66], HESA [27], E2E [64], FastDraw [40], PINet [25], UFAST [41], CurveLanes [63], LaneATT [50], PRNet [57].

To check performance on curved roads wherein multiple lane lines are present, we utilize CurveLanes dataset that

[1] Where source code is available



	GT	LW-OL	DD-OL	PINet (1-HG) + Yolov5m	UFAST (R-34) + E. Det Lite0
GMACs		1.98	13.71	26.8	19.48

Figure 6: Visual performance comparison of proposed algorithms with joint operation of SoTA lane and object detection algorithms on different datasets along with pseudo bounding box labels.

Method	F1	Precision	Recall	GFLOPs
SCNN [38]	65.02	76.13	56.74	328.4
Enet-SAD [19]	50.31	63.60	41.60	3.9
PointLaneNet [9]	78.47	86.33	72.91	14.8
CurveLane-S [63]	81.12	93.58	71.59	7.4
CurveLane-M [63]	81.80	93.49	72.71	11.6
CurveLane-L [63]	82.29	91.11	75.03	20.7
LW-OL (E-Lite-b0)	80.27	90.38	70.49	3.96
DD-OL (E-b4)	81.42	91.38	73.66	27.42

Table 1: Evaluation SoTA lane detection algorithms on CurveLane dataset

has a large number of lanes ($= 9$) with more than 90% having curves. Thus based on performance summarized in Tab. 1 we can conclude PointLaneNet, Curvelane and proposed algorithms to perform well with $F1 > 75.00$. Comparing the performance of proposed networks w.r.t neural architecture search based algorithm (CurveLane) that emphasize on lower GMACs, we observe our lightweight network LW-OL to achieve comparable performance to CurveLane-S (-0.85 F1 score), taking only 53.5 % of the FLOPs (Floating Point Operations Per Second) while simultaneously performing object detection, whereas the performance-focused network DD-OL achieves comparable performance to CurveLane-M with a difference of -0.38 F1 score but surpassed by the deeper variant of Curvelane model. However, on CULane and TuSimple, the performance of LW-OL is comparable to CurveLane-S with DD-OL surpassing CurveLane-L, persuading us to believe that performance of the proposed algorithms is highly sensitive towards number of lanes that are available in the training cycle. Despite the performance difference between NAS-based Curvelanes we maintain the proposed approach to be computationally efficient since our approach also provides bounding boxes that localize an object. In contrast, due to the formulation of Curvelanes, an-

Method	Backbone	Normal	Crowded	Night	No-Line	Shadow	Arrow	Dazzle	Curve	Cross	Total	GMACs
SCNN [38]	VGG-16	90.6	69.7	66.1	43.4	66.9	84.1	58.5	64.4	1990	71.6	164.2
SAD [19]	ENet	90.1	68.8	66.0	41.6	65.9	84.0	60.2	65.7	1998	70.8	-
	R-18	89.8	68.1	64.2	42.5	67.5	83.9	59.8	65.5	1995	70.5	-
	R-101	90.7	70.0	66.3	43.5	67.0	84.4	59.9	65.7	2052	71.8	-
RESA [66]	R-34	91.9	72.4	69.8	46.3	72.0	88.1	66.5	68.6	1896	74.5	88.31
	R-50	92.1	73.1	69.9	47.7	72.8	88.3	69.2	70.3	1503	75.3	97.38
HESA [27]	R-34	90.2	67.7	65.2	42.0	64.3	84.7	59.1	60.8	1665	70.7	-
	ERFNet	92.0	73.1	69.2	45.0	75.0	88.2	63.8	67.9	2028	74.2	-
E2E-LMD [64]	ERFNet	91.0	73.1	67.9	46.6	74.1	85.8	64.5	71.9	2022	74.0	-
	R-18	90.0	69.9	63.2	43.2	62.5	83.2	60.2	70.3	2296	70.8	-
	R-101	90.1	71.2	65.2	44.9	68.1	84.3	60.9	70.2	2333	71.9	-
FastDraw [40]	R-50	85.9	63.6	57.8	40.6	59.9	79.4	57.0	65.2	7013	-	-
PINet [25]	1H	85.8	67.1	61.7	44.8	63.1	79.6	59.4	63.3	1534	69.4	1.19
	4H	90.3	72.3	67.7	49.8	68.4	83.4	66.3	65.6	1427	74.4	3.73
UFAST [41]	R-18	87.7	66.0	62.1	40.2	62.8	81.0	58.4	57.9	1743	68.4	8.46
	R-34	90.7	70.2	66.7	44.4	69.3	85.7	59.5	69.5	2037	72.3	16.97
CurveLanes [63]	NAS-S	88.3	68.6	66.2	47.9	68.0	82.5	63.2	66.0	2817	71.4	9.0
	NAS-M	90.2	70.5	68.2	48.8	69.3	85.7	65.9	67.5	2359	73.5	33.7
	NAS-L	90.7	72.9	68.9	49.4	70.1	85.8	67.7	68.4	1746	74.8	86.5
LaneATT [50]	R-18	91.11	72.96	68.95	48.35	70.91	85.49	63.37	65.72	1170	75.09	9.3
	R-122	91.74	76.16	70.81	50.46	76.31	86.29	64.05	69.47	1264	77.02	70.5
PRNet [57]	BiSeNet	90.8	72.3	69.2	47.6	70.6	85.2	64.2	67.2	1113	74.8	-
	ERFNet	92.0	74.7	70.5	51.7	76.0	87.8	68.4	70.0	2114	76.4	-
LW-OL	R-18	88.13	64.67	63.50	41.09	63.14	82.42	60.08	58.16	1682	66.01	10.38
	E-Lite-b0	90.21	71.67	68.69	46.67	71.96	85.31	60.38	71.62	1522	73.49	1.98
DD-OL	R-50	90.05	70.42	66.46	45.08	68.19	86.12	57.44	68.32	1994	72.09	23.67
	E-b4	91.08	76.69	69.72	50.98	75.84	87.22	67.58	70.08	1438	76.81	13.71

Table 2: Evaluation SoTA lane detection algorithms on CULane dataset (F1 score) under diverse conditions.

other object detection algorithm is required, increasing repetitious computations.

Method	Backbone	Resolution	Acc	FP	FN	GMACs
SCNN [38]	VGG-16	288 × 800	96.53	0.0617	0.0180	164.2
End2End [55]	ERF	256 × 512	95.80 -	-	-	8.52
FastDraw [40]	R-50	352 × 640	95.20	0.076	0.045	-
PINet [25]	1×HG	256 × 512	95.81	0.0585	0.0330	1.19
UFAST [41]	R-18	288 × 800	95.87	-	-	8.38
E2E-LMD [64]	ERF	256 × 512	96.02	0.0321	0.0428	-
PolyLaneNet [49]	Eff. b0	360 × 640	93.36	0.0942	0.0933	1.748
LSTR [31]	R-18	360 × 640	96.18	0.0291	0.0338	0.574
LaneATT [50]	R-18	360 × 640	95.57	0.0356	0.0301	9.3
LaneNet [36]	Custom	256 × 512	96.4	0.0780	0.0244	-
SAD [19]	ENet	364 × 640	96.64	0.0602	0.0205	-
RESA [66]	R-34	368 × 640	96.82	0.0363	0.0248	100.28
HESA [27]	ERF	368 × 640	96.01	0.0329	0.0458	-
LW-OL	E-Lite-b0	512 × 512	96.67	0.0301	0.0441	1.98
DD-OL	E-b4	512 × 512	97.23	0.0267	0.0357	13.71

Table 3: Evaluation of SoTA algorithms on TUSimple Dataset

We observe the performance of our lightweight approach (LW-OL) to surpass that of UFAST [41] with the same backbone, i.e., ResNet-18 [18] wherein our formulation is based on the latter. This demonstrates that using the same encoded features for multiple tasks can improve performance on all the tasks. However, since this approach utilizes large computations, we switch to a more efficient backbone, i.e., EfficientNet-Lite [51] that can achieve higher performance on both object and lane detection while minimizing the computational requirement. A closer analysis of model performance reveals, current SoTA lane detection algorithms to be sensitive in the absence of lane line (No-Line), in which condition performance of all lane detection algorithms is peaked at 50.7 (by PRNet) in comparison to 92.1 (by RESA) under normal conditions. This strengthens our approach of relying on pseudo bounding box labels instead of pseudo-lane lines as the absence of lane markers due to occlusion could generate many false negatives, reducing model performance.

Furthermore, we observe shadows, dazzle, night, and curves conditions limiting the performance of all lane detection algorithms compelling us to believe a lot of work is still accomplished for visual perception-based lane detection. That being said, performance in shadows, dazzle, and night conditions can be significantly improved by simply tweaking the camera signal processing pipeline, wherein the objective would be to generate well-illuminated images irrespective of outdoor conditions. To verify if such a minor tweak could improve the performance of object and lane detection algorithms, we preprocessed the images using pretrained AFNet [46] and summarize the visual results in Fig. 8. Since the networks were not jointly trained, there is a scope for performance improvement by joint optimization, however, it is beyond the scope of this work. Nevertheless, the results show that a well-tuned camera ISP can improve the performance of lane and object detection algorithms.

Figure 7: Joint Object, Lane and Drivable area detection using LW-OL model on CULane dataset.

Figure 8: Performance in shadow and no-line conditions after image enhancement.

Moreover, we additionally examined pushing the limits of lightweight architecture to also perform drivable road segmentation based on labels generated by a model trained on the BDD100K dataset with visual results in Fig. 7.

As TuSimple was among the first lane detection dataset, we can examine the performance landscape of SoTA in terms of image resolutions and backbone to obtain a fair performance overview. As latency is critical in lane detection, majority of works focused on lightweight backbones such as ResNet-18, Hourglass [37] and ERFNet [43]. Furthermore, inferring from image resolution, larger emphasis was given to information along the width rather than height. This however would also warp the objects and would thus require modification of anchor-based object detectors. However, anchor-based object detectors such as SSD [34] also require additional post-processing (NMS) to filter out weak detections, encouraging us to use anchor-free object detectors.

4.3. 2D Object Detection

Method	Resolution	mAP@0.5	mAP	GMACs
Yolov5-s	512 × 512	0.44	0.23	5.44
Yolov5-m	512 × 512	0.50	0.27	25.61
Yolov5-l	512 × 512	0.52	0.29	57.73
Yolov5-x	512 × 512	0.54	0.30	109.4
Yolov5-s	640 × 640	0.50	0.27	5.44
Yolov5-m	640 × 640	0.55	0.31	25.61
Yolov5-l	640 × 640	0.58	0.33	57.73
Yolov5-x	640 × 640	0.60	0.34	109.4
EfficientDet-D0	640 × 640	0.50	0.28	2.93
EfficientDet-D3	640 × 640	0.52	0.30	12.95
EfficientDet-D7	640 × 640	0.56	0.31	57.46
LW-OL-R18	512 × 512	0.48	0.30	10.38
LW-OL-EL0	512 × 512	0.52	0.31	1.98
DD-OL-R50	640 × 640	0.51	0.31	23.67
DD-OL-EB3	640 × 640	0.62	0.35	13.71

Table 4: Evaluation on SoTA object detectors on BDD100K Dataset

We subsequently examine object detection performance of pretrained LW-OL and DD-OL networks and compare it with Yolov5 [22], and EfficientDet [52] models trained on BDD100K datasets following training approach used to train proposed algorithms, thus providing us with peak performance. We observe LW-OL and DD-OL with efficientnet backbones to perform on par with different variants of Yolov5 and EfficientDet when the input resolution is 512 × 512 (Fig. 4). Despite domain gaps between lane detection datasets and BDD100K, the comparable performance is due to copy-paste data augmentation process, wherein different objects based on their masks are randomly cropped and pasted onto training images. This approach, while originally proposed for image segmentation and restoration algorithms, aids the performance of object detection algorithms as well improves the generalization ability of both object and lane detection algorithms as inferred from results summarized in Tab. 5. Specifically LW-OL and DD-OL observe a performance boost of 4.15 AP and 2.76 AP on object detection and 1.49 Acc and 1.31 Acc on lane detection when trained on TUSimple dataset and evaluated on BDD100K dataset. While CNN based LW-OL witnessed higher improvement compared to DD-OL, we believe this to be due to better modelling of feature relationship in transformer based algorithms.

Method	Acc	FP	FN	AP	AP_{50}
Yolov5-m	-	-	-	52.03	61.39
Yolov5-l	-	-	-	53.22	67.99
LW-OL	95.18	0.0343	0.0501	48.34	60.47
DD-OL	95.92	0.0355	0.0414	59.63	71.35

Table 5: Object Detection Performance of algorithms before copy-paste augmentation.

As SoTA is currently held by task-specific networks for lane detection tasks, we require an additional task-specific network for generating bounding box proposals. Hence to ensure comparable GMACs, we club a lightweight object detector with SoTA lane detector and present the visual results in Fig. 5. From model-specific GMACs we can observe the performance of joint networks to surpass prior works while requiring 9.83× and 1.42× fewer computations for LW-OL and DD-OL algorithms, respectively.

5. Conclusion

In this paper, we presented a weakly-supervised approach for generating pseudo ground truth that can aid training multitask networks focused on jointly performing object and lane detection. Focused on latency and performance, we proposed two architectures using CNNs and transformers that used the same feature encoder to avoid redundant computations and quantitatively and qualitatively demonstrated them to perform similar to SoTA while requiring fewer computations, thus being ideal candidates to be deployed on edge devices.

References

[1] Tusimple benchmark. https://github.com/TuSimple/tusimple-benchmark/, 2017. [Online]. 5

[2] Navaneeth Bodla, Bharat Singh, Rama Chellappa, and Larry S Davis. Soft-nms–improving object detection with one line of code. In *Proceedings of the IEEE international conference on computer vision*, pages 5561–5569, 2017. 3

[3] Amol Borkar, Monson Hayes, and Mark T Smith. Robust lane detection and tracking with ransac and kalman filter. In *2009 16th IEEE International Conference on Image Processing (ICIP)*, pages 3261–3264. IEEE, 2009. 2

[4] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. *arXiv preprint arXiv:1903.11027*, 2019. 3

[5] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *CVPR*, 2018. 3

[6] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6154–6162, 2018. 3

[7] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229. Springer, 2020. 4

[8] Kai Chen, Jiangmiao Pang, Jiaqi Wang, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. Hybrid task cascade for instance segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 3

[9] Zhenpeng Chen, Qianfei Liu, and Chenfan Lian. Pointlanenet: Efficient end-to-end cnns for accurate real-time lane detection. In *IEEE Intelligent Vehicles Symposium (IV)*, pages 2563–2568, 2019. 6

[10] MMDetection3D Contributors. MMDetection3D: OpenMMLab next-generation platform for general 3D object detection. https://github.com/open-mmlab/mmdetection3d, 2020. 3

[11] Kaiwen Duan, Song Bai, Lingxi Xie, Honggang Qi, Qingming Huang, and Qi Tian. Centernet: Keypoint triplets for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6569–6578, 2019. 3

[12] Jannik Fritsch, Tobias Kuehnl, and Andreas Geiger. A new performance measure and evaluation benchmark for road detection algorithms. In *International Conference on Intelligent Transportation Systems (ITSC)*, 2013. 1

[13] Noa Garnett, Rafi Cohen, Tomer Pe'er, Roee Lahav, and Dan Levi. 3d-lanenet: end-to-end 3d multiple lane detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2921–2930, 2019. 2

[14] Mohsen Ghafoorian, Cedric Nugteren, Nóra Baka, Olaf Booij, and Michael Hofmann. El-gan: Embedding loss driven generative adversarial networks for lane detection. In *proceedings of the european conference on computer vision (ECCV) Workshops*, pages 0–0, 2018. 2

[15] Golnaz Ghiasi, Yin Cui, Aravind Srinivas, Rui Qian, Tsung-Yi Lin, Ekin D Cubuk, Quoc V Le, and Barret Zoph. Simple copy-paste is a strong data augmentation method for instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2918–2928, 2021. 4

[16] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014. 3

[17] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 3

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1, 7

[19] Yuenan Hou, Zheng Ma, Chunxiao Liu, and Chen Change Loy. Learning lightweight lane detection cnns by self attention distillation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1013–1021, 2019. 5, 6, 7

[20] Yen-Chang Hsu, Zheng Xu, Zsolt Kira, and Jiawei Huang. Learning to cluster for proposal-free instance segmentation. In *2018 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2018. 2

[21] Junhwa Hur, Seung-Nam Kang, and Seung-Woo Seo. Multi-lane detection in urban driving environments using conditional random fields. In *2013 IEEE Intelligent Vehicles Symposium (IV)*, pages 1297–1302. IEEE, 2013. 2

[22] Glenn Jocher, Alex Stoken, Jirka Borovec, NanoCode012, Ayush Chaurasia, TaoXie, Liu Changyu, Abhiram V, Laughing, tkianai, yxNONG, Adam Hogan, lorenzomammana, AlexWang1900, Jan Hajek, Laurentiu Diaconu, Marc, Yonghye Kwon, oleg, wanghaoyang0106, Yann Defretin, Aditya Lohia, ml5ah, Ben Milanko, Benjamin Fineran, Daniel Khromov, Ding Yiwei, Doug, Durgesh, and Francisco Ingham. ultralytics/yolov5: v5.0 - YOLOv5-P6 1280 models, AWS, Supervise.ly and YouTube integrations, Apr. 2021. 8

[23] ZuWhan Kim. Robust lane detection and tracking in challenging scenarios. *IEEE Transactions on intelligent transportation systems*, 9(1):16–26, 2008. 2

[24] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 4, 5

[25] Yeongmin Ko, Jiwon Jun, Donghwuy Ko, and Moongu Jeon. Key points estimation and point instance segmentation approach for lane detection. *arXiv preprint arXiv:2002.06604*, 2020. 5, 7

[26] Yeongmin Ko, Younkwan Lee, Shoaib Azam, Farzeen Munir, Moongu Jeon, and Witold Pedrycz. Key points estimation and point instance segmentation approach for lane detection. *IEEE Transactions on Intelligent Transportation Systems*, 2021. 2

[27] Minhyeok Lee, Junhyeop Lee, Dogyoon Lee, Woojin Kim, Sangwon Hwang, and Sangyoun Lee. Robust lane detection via expanded self attention. *arXiv preprint arXiv:2102.07037*, 2021. 5, 7

[28] Seokju Lee, Junsik Kim, Jae Shin Yoon, Seunghak Shin, Oleksandr Bailo, Namil Kim, Tae-Hee Lee, Hyun Seok Hong, Seung-Hoon Han, and In So Kweon. Vpgnet: Vanishing point guided network for lane and road marking detection and recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 1947–1955, 2017. 2

[29] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 3, 4

[30] Guoliang Liu, Florentin Wörgötter, and Irene Markelić. Combining statistical hough transform and particle filter for robust lane detection and tracking. In *2010 IEEE Intelligent Vehicles Symposium*, pages 993–997. IEEE, 2010. 2

[31] Ruijin Liu, Zejian Yuan, Tie Liu, and Zhiliang Xiong. End-to-end lane shape prediction with transformers. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3694–3702, 2021. 2, 7

[32] Ruijin Liu, Zejian Yuan, Tie Liu, and Zhiliang Xiong. End-to-end lane shape prediction with transformers. In *WACV*, 2021. 4

[33] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. Path aggregation network for instance segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8759–8768, 2018. 4

[34] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016. 3, 8

[35] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 3

[36] Davy Neven, Bert De Brabandere, Stamatios Georgoulis, Marc Proesmans, and Luc Van Gool. Towards end-to-end lane detection: an instance segmentation approach. In *2018 IEEE Intelligent Vehicles Symposium (IV)*, pages 286–291. IEEE, 2018. 2, 7

[37] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *European conference on computer vision*, pages 483–499. Springer, 2016. 8

[38] Xingang Pan, Jianping Shi, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Spatial as deep: Spatial cnn for traffic scene understanding. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018. 2, 5, 6, 7

[39] Xingang Pan, Jianping Shi, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Spatial as deep: Spatial cnn for traffic scene understanding. In *AAAI Conference on Artificial Intelligence (AAAI)*, February 2018. 5

[40] Jonah Philion. Fastdraw: Addressing the long tail of lane detection by adapting a sequential prediction network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11582–11591, 2019. 2, 5, 7

[41] Zequn Qin, Huanyu Wang, and Xi Li. Ultra fast structure-aware deep lane detection. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIV 16*, pages 276–291. Springer, 2020. 2, 5, 7

[42] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016. 3

[43] Eduardo Romera, José M Alvarez, Luis M Bergasa, and Roberto Arroyo. Erfnet: Efficient residual factorized convnet for real-time semantic segmentation. *IEEE Transactions on Intelligent Transportation Systems*, 19(1):263–272, 2017. 8

[44] Pranjay Shyam, Sandeep Singh Sengar, Kuk-Jin Yoon, and Kyung-Soo Kim. Evaluating copy-blend augmentation for low level vision tasks. *arXiv preprint arXiv:2103.05889*, 2021. 4

[45] Pranjay Shyam, Kuk-Jin Yoon, and Kyung-Soo Kim. Dynamic anchor selection for improving object localization. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9477–9483, 2020. 3

[46] Pranjay Shyam, Kuk-Jin Yoon, and Kyung-Soo Kim. Lightweight hdr camera isp for robust perception in dynamic illumination conditions via fourier adversarial networks. 2021. 7

[47] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 3

[48] Ganesh Sistu, Isabelle Leang, and Senthil Yogamani. Real-time joint object detection and semantic segmentation network for automated driving. *arXiv preprint arXiv:1901.03912*, 2019. 1, 3

[49] Lucas Tabelini, Rodrigo Berriel, Thiago M Paixao, Claudine Badue, Alberto F De Souza, and Thiago Oliveira-Santos. Polylanenet: Lane estimation via deep polynomial regression. In *Proceedings of the International Conference on Pattern Recognition*, 2020. 2, 5, 7

[50] Lucas Tabelini, Rodrigo Berriel, Thiago M Paixão, Claudine Badue, Alberto F De Souza, and Thiago Olivera-Santos. Keep your eyes on the lane: Attention-guided lane detection. *arXiv preprint arXiv:2010.12035*, 2020. 5, 7

[51] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pages 6105–6114. PMLR, 2019. 7

[52] Mingxing Tan, Ruoming Pang, and Quoc V Le. Efficientdet: Scalable and efficient object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10781–10790, 2020. 8

[53] Marvin Teichmann, Michael Weber, Marius Zoellner, Roberto Cipolla, and Raquel Urtasun. Multinet: Real-time joint semantic reasoning for autonomous driving. In *2018 IEEE Intelligent Vehicles Symposium (IV)*, pages 1013–1020. IEEE, 2018. 1, 3

[54] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9627–9636, 2019. 4

[55] Wouter Van Gansbeke, Bert De Brabandere, Davy Neven, Marc Proesmans, and Luc Van Gool. End-to-end lane detection through differentiable least-squares fitting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019. 2, 7

[56] Thomas Veit, Jean-Philippe Tarel, Philippe Nicolle, and Pierre Charbonnier. Evaluation of road marking feature extraction. In *2008 11th International IEEE Conference on Intelligent Transportation Systems*, pages 174–181. IEEE, 2008. 2

[57] Bingke Wang, Zilei Wang, and Yixin Zhang. Polynomial regression network for variable-number lane detection. In *European Conference on Computer Vision*, pages 719–734. Springer, 2020. 5, 7

[58] Jiaqi Wang, Wenwei Zhang, Yuhang Cao, Kai Chen, Jiangmiao Pang, Tao Gong, Jianping Shi, Chen Change Loy, and Dahua Lin. Side-aware boundary localization for more precise object detection. In *ECCV*, 2020. 3

[59] Ze Wang, Weiqiang Ren, and Qiang Qiu. Lanenet: Real-time lane detection networks for autonomous driving. *arXiv preprint arXiv:1807.01726*, 2018. 2

[60] Tao Wu and Ananth Ranganathan. A practical system for road marking detection and recognition. In *2012 IEEE intelligent vehicles symposium*, pages 25–30. IEEE, 2012. 2

[61] Yue Wu, Yinpeng Chen, Lu Yuan, Zicheng Liu, Lijuan Wang, Hongzhi Li, and Yun Fu. Rethinking classification and localization for object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10186–10195, 2020. 4

[62] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. *arXiv preprint arXiv:1611.05431*, 2016. 3

[63] Hang Xu, Shaoju Wang, Xinyue Cai, Wei Zhang, Xiaodan Liang, and Zhenguo Li. Curvelane-nas: Unifying lane-sensitive architecture search and adaptive point blending. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16*, pages 689–704. Springer, 2020. 2, 5, 6, 7

[64] Seungwoo Yoo, Hee Seok Lee, Heesoo Myeong, Sungrack Yun, Hyoungwoo Park, Janghoon Cho, and Duck Hoon Kim. End-to-end lane marker detection via row-wise classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1006–1007, 2020. 2, 5, 7

[65] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 3

[66] Tu Zheng, Hao Fang, Yi Zhang, Wenjian Tang, Zheng Yang, Haifeng Liu, and Deng Cai. Resa: Recurrent feature-shift aggregator for lane detection. *arXiv preprint arXiv:2008.13719*, 2020. 5, 7

[67] Dingfu Zhou, Jin Fang, Xibin Song, Chenye Guan, Junbo Yin, Yuchao Dai, and Ruigang Yang. Iou loss for 2d/3d object detection. In *2019 International Conference on 3D Vision (3DV)*, pages 85–94. IEEE, 2019. 4

[68] Qin Zou, Hanwen Jiang, Qiyu Dai, Yuanhao Yue, Long Chen, and Qian Wang. Robust lane detection from continuous driving scenes using deep neural networks. *IEEE transactions on vehicular technology*, 69(1):41–54, 2019. 2