IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 27 March 2023, accepted 7 April 2023, date of publication 11 April 2023, date of current version 19 April 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3266284

RESEARCH ARTICLE

Joint Multiclass Object Detection and Semantic Segmentation for Autonomous Driving

SHAKHBOZ ABDIGAPPOROV[1], (Student Member, IEEE), SHOKHRUKH MIRALIEV[1], (Student Member, IEEE), VIJAY KAKANI[2], (Member, IEEE), AND HAKIL KIM[1], (Member, IEEE)

[1]Department of Electrical and Computer Engineering, Inha University, Incheon 402751, South Korea
[2]Department of Integrated System Engineering, Inha University, Incheon 402751, South Korea

Corresponding author: Hakil Kim (hikim@inha.ac.kr)

This research was supported by the BK21 Four Program funded by the Ministry of Education (MOE) and by the National Research Foundation of Korea (NRF), South Korea.

ABSTRACT Object detection and semantic segmentation are two fundamental problems in autonomous driving systems. As recent studies have illustrated the strong correlation between the two tasks, the joint development of object detection and semantic segmentation tasks by utilizing end-to-end encoder-decoder architecture has gained popularity in recent years. However, context information loss is a common problem for simple encoder-decoder systems. Considering this problem, this study proposes a joint multi-class object detection and semantic segmentation method with the addition of a modern feature fusion mechanism to prevent context information loss. The experiments are conducted on our proposed large-scale Inha Computer Vision 2022 (ICV22) dataset that was specifically collected and annotated for object detection and semantic segmentation tasks. The proposed model achieves 87.2 mIoU performance on the introduced ICV22 dataset with 92.1 accuracy for the segmentation task, which is far superior to that of the DeeplabV3++ semantic segmentation method. Additionally, joint object detection and semantic segmentation model illustrated 40.2 mAP for object detection task and 56.4 mIoU for semantic segmentation task, outperforming previously introduced methods on publicly available Cityscapes dataset with real-time inference speed of 42 FPS on NVIDIA RTX 3090 GPU.

INDEX TERMS Object detection, semantic segmentation, multitask learning, feature fusion.

I. INTRODUCTION

Scene understanding is a controversial topic in the field of computer vision. Its research results have been widely used for developing various autonomous driving applications. However, the development of a reliable autonomous driving system that is applicable to the real world is challenging. This is because of the planning strategy of system development when it needs to perceive, predict and execute critical decisions in often uncontrolled as well as complex environments. For generalization, most of the perception systems in driverless cars are required to have (1) performance accuracy: precise timely decisions with reliable collected information in real time; (2) an adjustable algorithm that should work in diverse weather conditions(e.g., rainy, cloudy days) and (3) computational efficiency which provides a fast inference speed for real-time: specifically for cars driving on the road at high speed.

For autonomous driving, multi-class object detection and semantic segmentation are the two core techniques used to understand the surrounding environment(e.g., the location of pedestrians, vehicles, and the region of drivable areas). Both tasks are mutually beneficial because they are highly correlated. For instance, object detection can serve as a prior knowledge of semantic segmentation task where semantic segmentation task predicts whether a pixel belongs to the foreground or the background, indicating that the objects should be on the road rather than the sky. Considering the

The associate editor coordinating the review of this manuscript and approving it for publication was Maurice J. Khabbaz.

FIGURE 1. Overview of the proposed ICV22 dataset examples for joint object detection and semantic segmentation.

similarity between tasks, we jointly address both problems with one end-to-end network architecture which resulted in a computationally efficient, robust and generally applicable model. The main goal of joint object detection and semantic segmentation is to simultaneously predict and locate objects in a frame and segment them into pixel-wise semantic regions, as can be shown in Figure 1. To address joint object detection and semantic segmentation problems, many multi tasking approaches have been proposed. The DLT-Net [1] constructs context tensors between sub-task decoders to share information among implemented tasks. BlitzNet [2] utilizes an encoder-decoder neural network as the backbone of the proposed architecture. The extracted feature maps in each layer of the decoder network were used to output the object detection task, and a concatenated feature map was used for the semantic segmentation task. In Addition, TripleNet [3] uses an encoder-decoder network as a backbone with the difference of adding class agnostic supervision and inner-connected modules on the feature maps of each layer of the decoder network. Generally, current pipelines adopt a backbone network with an encoder-decoder structure to extract feature maps using CNNs and relied on skip connections to propagate information across the network ([2], [3], [41]). These extracted feature maps convey more high level semantic information and fewer low-level fine-grained details, which can negatively affect the prediction outputs. The main task of the decoder network is to recover both high and low-level features to correctly estimate jointly implemented tasks. However, context information loss is a common problem for a simple encoder-decoder architecture, as suggested by the previous studies. To handle this issue correctly, this study proposes joint multi-class object detection and semantic segmentation network architecture with the addition of weighted bi-directional feature pyramid network(BiFPN [4])to fuse the extracted multi-scale features. To conduct the experiments, this study proposes new, diverse and large-scale Inha Computer Vision 2022(ICV22) dataset that is specifically tailored for the two implemented tasks(i.e., object detection and semantic segmentation). Experiments are conducted on proposed ICV22 dataset as well as publicly open Cityscapes [32] dataset. The proposed ICV22 dataset consists of 36,960 images collected in multiple distinctive locations of South Korea. All the collected images are annotated for both the implemented tasks in different weather and time conditions(e.g., rainy, cloudy, daytime, nighttime annotations). The contributions of this study as follows:

- The addition of feature a fusion network for fusing extracted high and low level feature maps for the decoder heads of the architecture ensures to avoid the context information loss problem and jointly learns both tasks for maintaining state-of-the-art performance on both publicly available Cityscapes [32] and proposed ICV22 datasets. Proposed multi tasking network outputs the two crucial tasks for practical autonomous driving applications separately: multi-class object detection and semantic segmentation.
- A new large scale Inha Computer Vision (ICV22) dataset is proposed which is collected in South Korean roads and annotated for the jointly implemented vision-based autonomous driving tasks.

The remainder of the paper is organized as follows: Section II presents related work; Section III details about the network architecture; Section IV describes the newly introduced dataset information; Section V presents the experiments on the proposed network model which includes the data preprocessing, implementation details as well as both quantitative and qualitative performance results on both datasets; and Section VII concludes the paper with limitations and summary of the study.

II. RELATED WORKS

A. OBJECT DETECTION

The main purpose of the object detection is to classify and locate target objects using bounding boxes.

With continuous emergence of convolutional neural networks(CNNs), researchers have categorized the object detection methods into two and one-stage object detection.

Two-stage object detection methods first propose candidate objects before classifying them into precise categories. Among the two-stage object detection architectures, R-CNN [5] is one of the most representative models that extracts the features of the entire image and generates regional features through the spacial pyramid pool. Another member of the R-CNN family, faster R-CNN [6], introduced an end-to-end object detection architecture utilizing a region proposal network. Based on this architecture, Cascade R-CNN [7] proposed a multi-level detector via a cascade architecture. To improve the performance of existing object detection methods, Lin et al. [8] proposed feature pyramid network(FPN). This network outputs multi-scale feature maps and fuse multi-layer semantic information through a skip-layer connection. The recently proposed task-aware spatial disentanglement(TSD) [9] model efficiently utilizes the FPN [8] architecture as the backbone and generates two disentangled proposals for both classification and regression tasks in object detection. For traffic sign detection task, recent study by Liang et al. [10] proposed an improved sparse R-CNN algorithm that integrates coordinate attention block with ResNeSt and builds a feature pyramid to modify the backbone to tackle the mismatch problem between the existing detection algorithm and its practical application. Similarly, Cao et al. [11] improved the multi-scale representation ability of the backbone by constructing hierarchical residual-like connections within each single radix block in the original ResNest.

One-stage methods focus on mapping feature maps directly to the classification scores and bounding boxes. YOLO(from [12], [13], [14], [15], [16], [17], [18]) series(e.g., YOLOV5 [16], YOLOV6 [17], and YOLOV7 [18] are the most recent object detection networks) are the most widely used architectures. This network regresses bounding boxes directly from the DarkNet model in a more completely unified detector manner. DSSD [19] uses an encoder-decoder network to add context information for multi-scale object detection, thereby indicating performance improvement. To address the class imbalance problem, RetinaNet [20] proposed the focal loss to down-weight the contribution of a large number of easy samples. The recently proposed EfficientDet [4] utilizes the FPN [8] as the backbone and uses a weighted bi-directional feature pyramid network(BiFPN). Proposal-free one-stage detectors tend to narrow the accuracy gap with two-stage methods, while presenting the advantage of computational efficiency.

B. SEMANTIC SEGMENTATION

Semantic segmentation can be identified as a pixel-wise image classification task that has achieved the significant progress over the years with the introduction of fully convolutional networks(FCN [21]). After the FCN [21] network, researchers efficiently use either an encoder-decoder structure or spatial pyramid methods to improve the base performance of the semantic segmentation task. The encoder part of the architecture extracts semantic features and reduces the spacial resolution of feature maps that are based on CNN networks(e.g., ResNet50 [22] and HrNet [23]) pre-trained on a large dataset. The decoder network gradually up-samples the feature maps of the encoder network. The SegNet [24] and UNet [25] models adopt the same encoder-decoder architecture, wherein the longer range information with gradually decreasing spatial dimension is encoded, and detailed spatial information of the objects is decoded. DeepLab [26] model uses astrous convolutions to capture multi-scale contextual information from an image, which allows the model to better understand the relationship between different parts of the image. However, using multi-scale context information does not always improve the performance of semantic segmentation tasks. PSPNet [27] introduced a pyramid pooling module that down-samples and up-samples feature maps in parallel while efficiently utilizing multi-scale context information. Additionally, studies by Fan et al. [28] and Han and Fan [29] focused on handling the problems of lacking spatial information and the gap of combination between high-level and low-level features in segmentation models by implementing attention select module and a tailored backbone which is modified from Resnet. The model showed excellent performance by achieving 71.5% mIoU score on the Cityscapes test set.

C. JOINT OBJECT DETECTION AND SEMANTIC SEGMENTATION

Multitask learning is gaining increasing popularity owing to the computational efficiency of the proposed methods while maintaining a performance similar to that of single task methods. By utilizing the relationship between the outputted tasks, multi-tasking architectures ([2], [31], [37], [38]) can enable the models to be more robust and applicable in real world. The goal of joint object detection and semantic segmentation is to simultaneously detect objects and predict pixel-wise semantic labels by using a single end-to-end network. Various multi-tasking architectures have been proposed for outputting object detection and semantic segmentation tasks separately or training for combining both outputs as panoptic segmentation result ([46], [47]).

While panoptic segmentation is a popular approach that combines both tasks, separating object detection and semantic segmentation outputs allows more flexibility in post-processing and downstream tasks for practical applications. For instance, a semantic segmentation output can be used for image understanding and an object detection output can be utilized for counting objects on the road and tracking them. Additionally, separation of the two tasks can lead a better performance on each individual task when a model needs to solely focus on either object detection or semantic segmentation task only for required practical usage cases. One of the encoder-decoder type methods that outputted the two tasks separately is the DLT-Net [1] that constructs context tensors

between sub-task decoders to share information among the implemented tasks. BlitzNet [2] and DspNet [30] are both joint object detection and semantic segmentation networks that share simple encoder decoder structure. In BlitzNet [2], each layer of the decoder is used to detect objects of different scales. UberNet [31] efficiently implemented multiple vision tasks, including semantic segmentation and object detection, using a single deep neural network. TripleNet [3] used attention skip layer fusion to expand the feature map, an inner-connected module to increase the correlation between the two predicted tasks, and class-agnostic segmentation supervision to add a deep level of supervision.

D. MULTILABELED AUTONOMOUS DRIVING DATASETS

The main goal of autonomous driving datasets is to understand the challenges faced by computer vision systems in the context of self-driving. Most datasets that are used for single-task models focus on particular object annotations (e.g., vehicles and pedestrians), such as COCO dataset [45]. However datasets with multiple task annotations are required for developing a more complete self-driving system.

1) CITYSCAPES DATASET

Citiscapes [32] dataset was one of the first vision based datasets that addresses the issue of semantic urban scene understanding that is introduced in 2016. Cityscapes [32] provides instance-level semantic segmentation annotations on the sequential frames of videos collected. The dataset was collected from streets in 50 cities. It comprised 5,000 images with high-quality pixel-level annotations and 20,000 images with coarse annotations. This multi-labeled dataset mainly focuses on combination of detection and segmentation problems.

2) MAPILLARY VISTAS DATASET

Subsequently, the Mapillary Vistas [33] dataset was introduced as more large scale street level image dataset. Mapillary Vistas [33] provides fine-grained annotations for user-uploaded data, which are more diverse with respect to location. However, these images were one-off frames that were not placed in the context of videos with a temporal structure. The dataset contains 25000 high resolution images annotated into 66 object categories. Additionally, the dataset contained 37 annotated instance-specific classes. Sample images are collected from different devices (e.g., tablets, action cameras, mobile phones, and professional capturing rigs) in distinct locations.

3) BERKELEY DEEP DRIVE(BDD100K) DATASET

Berkeley Deep Drive(BDD100K) [34] dataset is gaining popularity because it is the largest driving video dataset with 100K videos and annotated image samples for a number of distinctive tasks for autonomous driving. The dataset was annotated for image tagging, object detection, lane marking, drivable area segmentation, semantic segmentation and multiple object tracking. The videos were collected in populous areas such as New York, Berkeley, San Francisco, and thee Bay Area in the US. The tenth second frame of each video was annotated. The annotated 100K images are divided into a training set, which comprised 75 % of the annotated images, as well as validation set which is 15 % and test sets, which comprise 10 % of all annotated images.

III. NETWORK ARCHITECTURE

The introduced network architecture is based on the encoder-decoder model and a bidirectional feature pyramid pooling scheme(BiFPN [4]), as shown in Figure 2. A single encoder network that serves as the backbone of the architecture extracts multi-scale features from the given input frame in four layers. Features extracted in deeper level layers are then fed into the BiFPN for fusion (e.g., a modified version of the feature pyramid pooling scheme introduced by [4]). To obtain the multi-class predictions of the object detection task, only fused features by BiFPN were utilized. For multi-class semantic segmentation, however, additional lower-level features that avoid feature fusion network are provided directly to the decoder network.

A. ENCODER

Encoder of the network is the most essential part, as it extracts multi-scale feature maps from a given input image for feature fusion. Many studies on encoder-decoder architecture currently use pre-trained networks in the ImageNet dataset [35]. Two recently released ImageNet pre-trained networks were utilized as the encoder network for the experiment. The first encoder network Xception65 [36] is selected due to the excellent performance in segmentation-related tasks, and the second encoder network, CSPResNet50 [42], performed well on detection type tasks. The extracted feature maps from the encoder network are then fed into the BiFPN [4] network for feature fusion. BiFPN [4] fuses features at different resolutions based on cross-scale connection for each node by each top-down and bottom-up path as well as adding a weight for each feature to learn the importance of each level.

B. DECODERS

The two decoders were utilized for extracting multi-class object detection and semantic segmentation outputs. Object detection decoder includes classifier and regressor networks as many other typical object detection networks. Both the regressor and classifier included a set of 3×3 and 1×1 convolutional layers, batch normalization and a swish activation function. To predict the precise location of the objects, the regressor branch utilizes non max suppression whereas the classifier includes a sigmoid function to predict the class of the object. The segmentation head consisted a BiFPN decoder, a 1×1 convolutional layer to reduce the computation and SoftMax function for predicting the output masks(as UNET [25]). The BiFPN decoder mainly handles up-sampling of the fused multi-scale high level features from BiFPN network to match the size of the lower-level features from the encoder network.

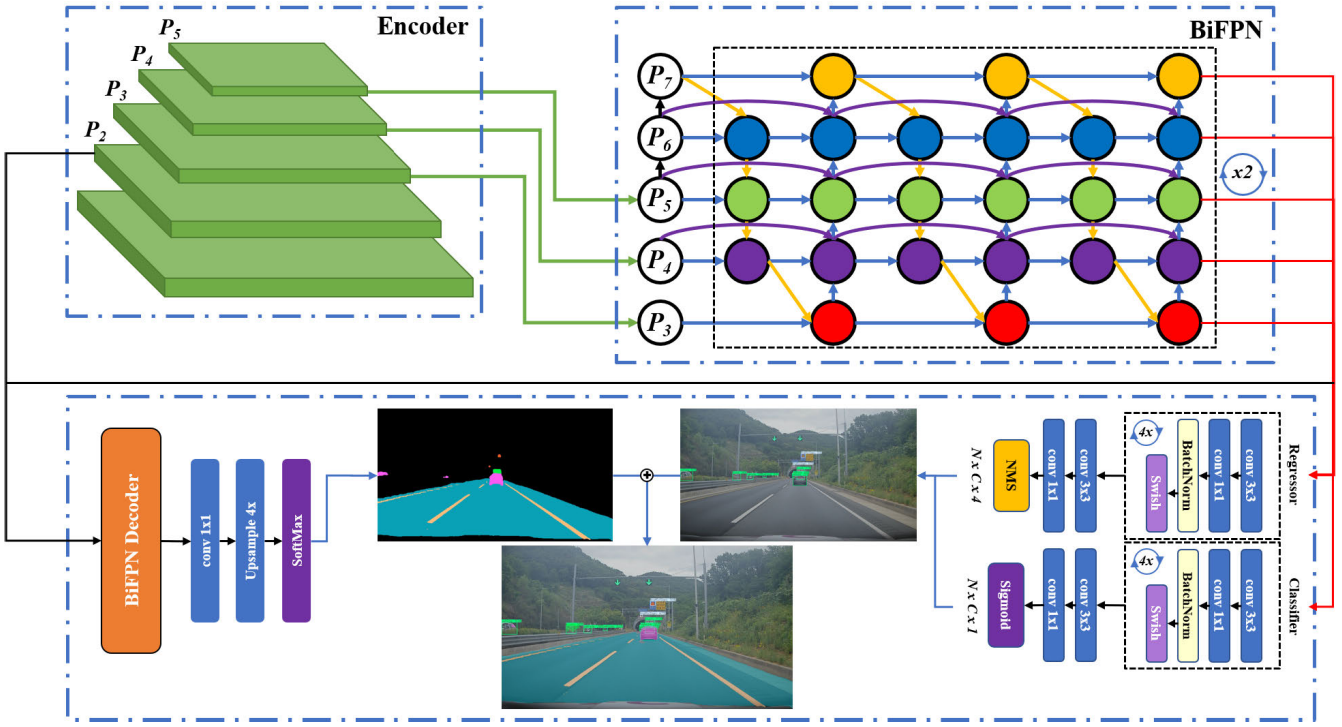

FIGURE 2. Proposed network architecture consists of an encoder for multi-scaled feature extraction, feature fusion network(BiFPN) and two separate decoders for object detection and semantic segmentation tasks.

C. LOSS FUNCTIONS

The semantic segmentation task utilizes the weighted sum (1) of Dice loss and focal loss for efficiency, both of which aim to minimize the classification errors between the ground truth and the predicted output by leading the model to focus on difficult examples. Additionally, the focal loss (2) forces the model to learn incorrectly classified voxels and the Dice loss (3) improves the voxel problem by learning the class distribution.

$$L_{SemanticSegmentation} = \gamma L_{Focal} + L_{Dice} \tag{1}$$

where, γ is a parameter to maintain balance between the focal and Dice loss.

$$L_{Focal} = -\frac{1}{N}\sum_{m=0}^{m-1}\sum_{i=1}^{I} g_i(m)(1-p_i(m))^1 \log(p_i(m)) \tag{2}$$

$$L_{Dice} = M - \sum_{m=0}^{m-1} \frac{TP_p(m)}{TP_p(m) + \alpha FN_p(m) + \beta FP_p(m)} \tag{3}$$

where, m is total number of classes in semantic segmentation task.

The average sum of two types of losses, classification loss and regression loss, with tuning parameters for balance, was implemented for the object detection task. Classification loss is common in detection type-tasks for penalizing classifications. Accurate predictions for object detection strongly depend on the prediction confidence; hence, classification loss improves the prediction confidence. Finally, regression loss was used to correctly estimate the distance of the overlap rate, aspect ratio, and scale similarity between ground truths and predicted outputs.

$$L_{detection} = \alpha_1 L_{classification} + \alpha_2 L_{regression} \tag{4}$$

IV. PROPOSED DATASET

As self-driving systems have gained popularity in recent years, autonomous driving datasets have gained popularity among researchers. The main goal was to solve the issues of autonomous vehicle technology by introducing newly annotated challenging datasets for real-life applications. Some of the datasets focused on particular tasks and objects such as vehicles on the road or pedestrians. Cityscapes [32] is a large-scale multi-labeled dataset that provides instance-level semantic segmentation of sampled frames of videos collected. KITTI [40] dataset is composed of data from multiple sources(e.g., vision based and Lidar sensor-based data). Because diversification of the collected data samples can be a complex task, it is challenging to collect data samples on a variety of road conditions in different times of day under different weather conditions. The newly introduced ICV22 dataset was collected throughout both crowded streets and remote areas of South Korean cities throughout different weather conditions during both day and night.

A. DATA SPECIFICATIONS

Total of 36,960 image frames were acquired during the span of several months covering spring, summer and fall

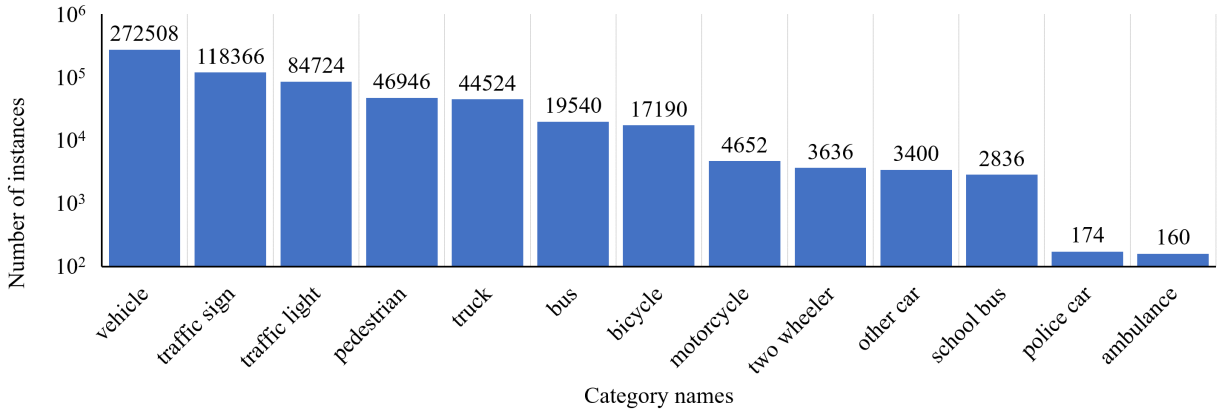

FIGURE 3. Instance statistics of 13 object classes of our proposed ICV22 dataset. Number of instances of each category follows a long-tail distribution. Vehicle class includes the highest number of instances and the ambulance class has the lowest.

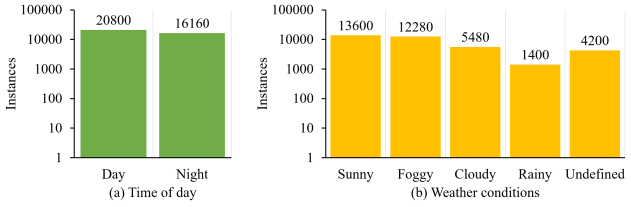

FIGURE 4. Frame statistics of (a) Time of day and (b) Weather conditions.

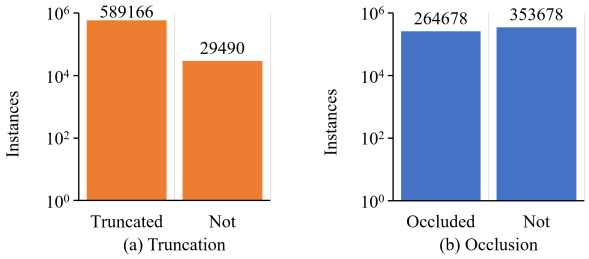

FIGURE 5. Instance statistics of (a) Truncations and (b) Occlusions of the proposed ICV22 dataset.

seasons in various South Korean cities(e.g., Busan, Ulsan, Seoul, Incheon, etc.). The introduced ICV22 dataset covers five types of weather conditions (i.e., sunny, foggy, cloudy, rainy and undefined) during both day and night as shown in Figure 4. All images were manually annotated for dense pixel-level annotation, aiming for a high diversity of foreground objects and overall scene layout.

B. CLASSES AND ANNOTATIONS

1) OBJECT DETECTION CLASSES

Object detection is a fundamental task not only for self-driving systems but also for all other areas of computer vision. The newly introduced ICV22 dataset provided box annotations of 13 classes for each of the 36,960 frames. The instance statistics for each category of the object detection task are shown in Figure 3.The vehicle class had the highest number of instances with 272,508 instances throughout the dataset, and the police car and ambulance classes had the least number of object instances with 174 and 160 instances respectively. Additionally, the statistics of visibility attributes such as ''occluded'' and ''truncated'' are shown in Figure 5.

2) SEMANTIC SEGMENTATION CLASSES

The ICV22 dataset provided pixel-level fine annotations for objects on the road for 33 semantic segmentation classes. Annotation and quality control required an average of 6 min for a single image. Annotators were asked to label the image from back to front, such that no object boundary was marked more than once. Given the label scheme, annotations can easily be extended to cover additional or more fine-grained classes. The ICV22 dataset defines 33 visual classes for semantic segmentation annotation, which are classified into four categories: moving-objects, road-objects, static-objects, and others. These four categories were further divided into twelve sub-categories, as shown in Figure 6. Classes are selected based on their relevance from an application standpoint, practical considerations regarding the annotation effort, and facilitation of compatibility with existing self-driving-based datasets(e.g., [32], [33], [34]).

Because the ICV22 dataset is specifically tailored for joint object detection and semantic segmentation tasks and is mainly aimed at developing a real-life panoptic driving system, comparisons in terms of annotation and volume density with other datasets(i.e., Cityscapes [32], CamVid [39], and KITTI [40]) were conducted. The Cityscapes [32] dataset contains 20K coarse and 5K fine annotations of images from 50 different cities, CamVid consists of 10 minutes of video footage with pixel-wise annotations for over 700 frames, and DUS consists of 5K images from which 500 have been annotated. The comparative statistics of the semantic pixel-wise annotations are listed in Table 1.

Although the statistics indicate that the annotation density of the ICV22 dataset is considerably lower than that of other datasets, the number of annotated images for the semantic

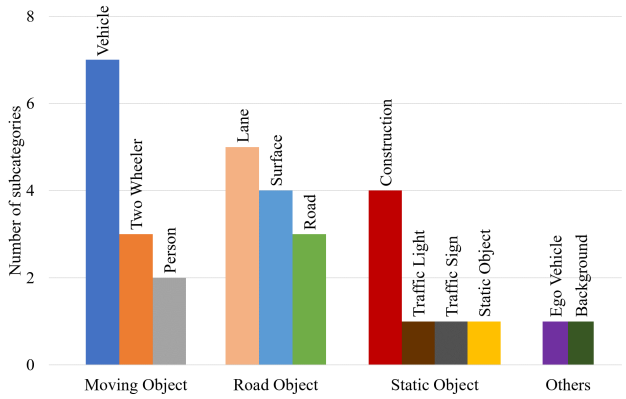

FIGURE 6. Proposed ICV22 dataset semantic segmentation classes which are divided in four categories (Moving Object, Road Object, Static Object, and Others).

TABLE 1. Comparison of proposed dataset(ICV22) for absolute number and density of annotated pixels with Cityscapes, DUS, KITTI(upscaled to 1280 × 720 pixels to maintain the original aspect ratio).

	#pixel [billion]	annot. density(%)
Cityscapes(coarse)	12.35	67.5
CamVid(fine)	0.62	(96.2)
DUS	0.14	63.0
KITTI(fine)	0.23	88.9
ICV22(**ours**)	**18.43**	29.7

segmentation task(which is 36,960) is considerably higher with more accurate annotations for each class. The main reason for the lower annotation density is the background zero pixels of certain classes (e.g. sky, buildings and trees). While annotating the ICV22 dataset, precise annotation of each annotated class feature (e.g., visibility of the hands and, legs of an annotated pedestrian) was prioritized. Therefore, the models can learn more about each of the predicted classes of semantic segmentation tasks by utilizing the ICV22 dataset for real-life applications.

V. EXPERIMENTS

A. DATA PREPROCESSING

Experiments were first performed on the ICV22 dataset. All image frames contained accurately annotated objects for both the tasks. In contrast to previous studies (e.g., Peng et al. [41] and DsPNet [30]), all 36,960 image frames were efficiently used for joint training and testing, rather than partially utilizing the newly introduced dataset. The dataset was further divided into 70% training, 15% validation and 15% test images. To validate the generalization and robustness of the proposed model, additional experiments were conducted on the widely used Cityscapes [32] dataset. The Cityscapes [32] dataset is mainly annotated for segmentation related tasks(i.e., semantic segmentation, instance segmentation, and panoptic segmentation). The dataset contains 25,000 images of traffic scenes, 20,000 of which are coarsely annotated, and the other 5,000 images have fine-grained annotations of high quality. However, 5,000 images with quality annotations were used for the experiments following the settings of existing research works (DsPNet [30]). Because the annotations of the images in the testing set were not publicly available, 3,475 images in the training set were further divided into 2,975 images for training and 500 images for testing. However, the Cityscapes [32] dataset is annotated mainly for segmentation type tasks, and the bounding boxes for objects are not available on the public website. To qualify the dataset for the object detection task, four boundary values (leftmost, rightmost, uppermost and nethermost) of a semantic segment were computed to form a bounding box for an object in a traffic scene. Therefore, for fair evaluation and comparison with existing methods(for joint object detection and semantic segmentation) only eight classes were chosen for each task. Because the dataset contains 33 classes, the other classes were set as background classes.

B. IMPLEMENTATION DETAILS

Data augmentation techniques, such as scaling the images, rotation, translation and resizing were used to process images to handle geometric distortions. During training and testing, the original size of the images was resized to 640 × 384 to maintain fairness when compare with other methods(for joint object detection and semantic segmentation).

1) EXPERIMENTED CNN BACKBONE NETWORKS

Joint semantic segmentation and object detection network efficiently utilize two recently introduced CNN backbone networks as encoders: CSPResNet50 [42] and Xception65 [36]. CSPResNet50 [42] is a convolutional neural network which is a recent version of ResNet with a cross stage partial network(CSPNet [42]). CSPNet partitions the feature map of the base layer into two parts and merges them through a cross-stage hierarchy, significantly improving the performance of the detection task.

The Xception65 [36] network was selected for its state-of-the-art performance in various studies that handle semantic segmentation tasks(e.g., EfficientNet [43] and DeepLabV3+ [44]). The network architecture uses depthwise separable convolutions which are more computationally efficient than traditional convolutions. Computational efficiency is essential for the task, as the goal of a semantic segmentation task is to assign a label to each pixel in an image that can be computationally expensive.

2) TRAINING SETTINGS

On an NVIDIA RTX 3090 24GB GPU with an initial learning rate of 0.0001, 16 batch size and Adam optimizer, the network was trained for 300 epochs, with the segmentation task frozen in the first 200 epochs, detection task frozen for the next 50 epochs and 50 epochs of end-to-end training at the final stages. The implemented joint multi-class semantic segmentation and object detection network architecture uses

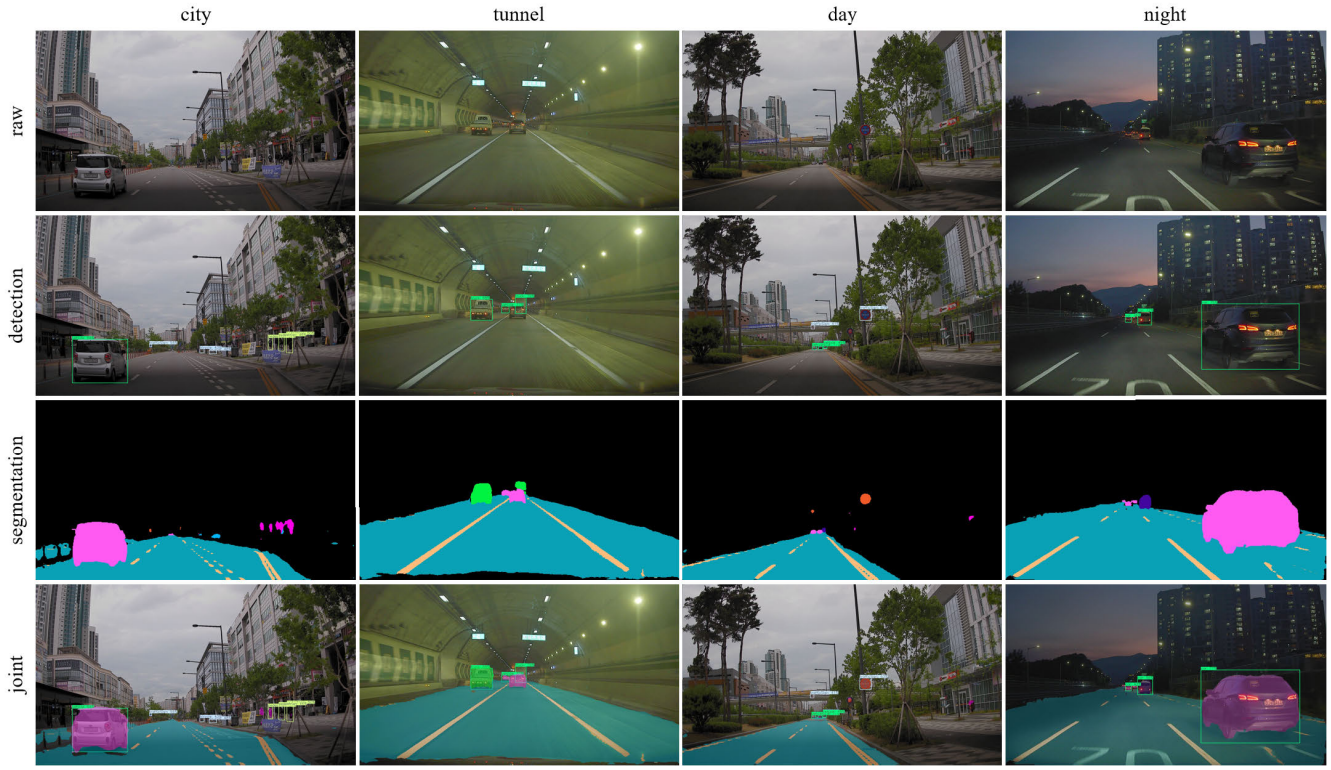

FIGURE 7. Qualitative results of joint object detection and semantic segmentation network on proposed ICV22 dataset.

TABLE 2. Comparison of object detection results on the proposed ICV22 dataset.

Method	vehicle	truck	s.Bus	otherCar	t.Light	t.Sign	bicycle
YoloV5 [16]	**71.4**	**63.4**	-	**63.2**	**58.6**	**80.7**	**54.3**
Ours (Xception-65)	63.9	55.0	38.6	45.0	50.7	70.4	39.7
Ours (CSPResnet-50)	65.1	58.9	36.3	48.1	55.6	72.3	40.8
	pedestrian	**t.Wheeler**	**bus**	**motorcycle**	**ambulance**	**policeCar**	**mAP**
YoloV5 [16]	**31.0**	**32.3**	**71.0**	**51.7**	43.4	56.5	**51.9**
Ours (Xception-65)	29.6	21.0	57.6	39.2	41.8	**57.5**	46.9
Ours (CSPResnet-50)	29.4	19.8	59.9	44.6	**51.6**	51.2	48.7

mAP and recall to evaluate the performance of a traffic object detection task. To provide fairness in comparisons with other models, mAP50 was computed by the average of the average precision calculated for all classes at a single intersection over union(IoU) threshold of 0.5.

$$mAP = \frac{1}{C}\sum_{c=1}^{C} AP_c \tag{5}$$

The recall refers to the percentage of the total number of relevant results that are correctly classified by the algorithm.

$$Recall = \frac{TP}{TP + FN} \tag{6}$$

mIoU was utilized to evaluate the performance of the semantic segmentation task. In segmentation, IoU is a metric that is represented by dividing the area of intersection of the predicted and ground truth mask areas of the union of two masks.

$$mIoU = \frac{1}{C}\sum_{c=1}^{C} \frac{AreaofOverlap_c}{AreaofUnion_c} \tag{7}$$

C. EXPERIMENTAL RESULTS

1) PERFORMANCE ANALYSIS ON ICV22 DATASET

Tables 2 and 3 list the results of the proposed method compared with state-of-the-art single-tasking object detection and semantic segmentation methods on the ICV22 dataset. ICV22 dataset differs from other earlier mentioned autonomous driving based datasets in terms of road architecture, traffic complexity and diversity of location(e.g. includes frames in both cities and remote road areas). For the semantic

TABLE 3. Comparison of semantic segmentation results on the proposed ICV22 dataset.

Method	road	lane	bus	car	person	t.Light	t.Sign	truck	t.Wheeler	mIoU	mAcc
DeeplabV3++	**94.8**	73.6	85.3	**90.6**	72.4	75.9	78.1	72.5	71.9	79.5	88.1
Ours (Xception-65)	93.2	**77.7**	90.5	85.1	87.1	88.4	**81.9**	87.3	**89.9**	86.8	**92.1**
Ours (CSPResnet-50)	93.7	77.6	**92.1**	85.3	**87.4**	**88.7**	81.7	**88.4**	89.9	**87.2**	91.8

TABLE 4. Comparison of object detection results on the Cityscapes dataset.

Method	person	rider	car	truck	bus	train	mbike	bike	mAP
DspNet [30]	23.0	27.5	52.8	30.8	48.1	40.5	19.8	25.1	33.4
BlitzNet [2]	28.7	31.8	63.9	34.1	57.2	45.1	20.6	26.6	38.5
PairNet [3]	21.6	28.8	48.8	33.2	53.4	49.3	14.2	22.4	34.0
TripleNet [3]	21.1	27.4	49.6	33.3	52.5	42.6	19.4	21.4	33.4
Peng et al. [41]	28.7	32.8	63.9	**35.7**	**58.6**	**50.5**	23.7	26.5	40.0
Ours (CSPResnet-50)	**32.1**	**34.9**	**68.9**	28.4	52.2	49.5	**24.0**	**31.4**	**40.2**

TABLE 5. Comparison of semantic segmentation results on the Cityscapes dataset.

Method	road	swalk	build	wall	fence	pole	t.light	t.sign	veg.	terrain
DspNet [30]	89.8	63.2	80.1	38.4	28.0	11.6	22.3	36.1	81.6	49.2
BlitzNet [2]	88.4	58.2	78.1	30.7	31.6	10.5	11.4	24.4	80.6	41.5
PairNet [3]	87.4	58.9	77.1	39.2	29.8	8.4	13.9	25.7	75.5	44.5
TripleNet [3]	87.7	60.6	77.7	38.3	30.1	9.1	12.0	29.5	80.7	45.8
Peng et al. [41]	90.7	65.0	81.0	**45.3**	33.6	17.8	26.4	38.5	83.2	48.1
Ours (CSPResnet-50)	80.8	52.8	68.4	44.6	**37.9**	**33.7**	**37.2**	36.5	73.4	47.0
Method	**sky**	**person**	**rider**	**car**	**truck**	**bus**	**train**	**mbike**	**bike**	**mIoU**
DspNet [30]	82.0	51.0	32.0	86.0	63.2	71.0	62.2	36.8	51.0	56.6
BlitzNet [2]	82.9	50.3	26.1	85.2	56.7	67.3	60.3	28.3	47.1	52.6
PairNet [3]	79.3	48.1	29.8	83.5	57.5	65.0	51.5	32.2	46.8	51.8
TripleNet [3]	80.5	49.1	27.8	84.8	63.0	68.9	49.9	30.5	48.4	51.3
Peng et al. [41]	**83.8**	52.5	33.1	**86.4**	60.2	65.6	56.3	35.0	50.9	55.4
Ours (CSPResnet-50)	76.1	**52.7**	**42.8**	72.9	**71.5**	**74.7**	**66.9**	**46.2**	**54.7**	56.4

segmentation task on ICV22 dataset, our method outperforms the NVIDIA's semantic segmentation network on majority of the trained segmentation classes(e.g. road, lane, car, trafficLight, trafficSign, truck and twoWheeler). Additionally, the performance of all of the segmentation types are between the IoU range of 77% and 89% which is exceptional for further developments of real-life self-driving application for South Korean roads. For object detection task, our method is compared with YoloV5 [16] model which is a popular single task network for object detection. The overall on Table 2 indicate the close performance accuracies on most of the object detection classes and better performance on two of the 13 trained classes.

Figure 7 shows the qualitative analysis of the joint object detection and semantic segmentation model on the newly introduced ICV22 dataset. To test the raw images, two raw images captured during daytime on the city road and one raw image on the tunnel with comparatively poor light conditions are selected. First, end-to-end trained model is tested for detection task only, detecting cars and pedestrians located on some distance from the camera's location point. Second, the raw images are tested for segmentation task only, segmenting nine selected semantic segmentation classes of ICV22 dataset. Finally, joint object detection and semantic segmentation qualitative performance analysis is performed on the given raw images as shown in the Figure 7.

2) COMPARISON WITH BASELINES ON CITYSCAPES DATASET

Tables 4 and 5 present the object detection and semantic segmentation results of the proposed model with the CSPResNet50 [42] backbone. The comparison tables are based on the results released on Peng et al. [41] as the source code of most of the existing models is not publicly available. Figure 8 shows comparison of the qualitative performance results of object detection and semantic segmentation task results trained on the introduced joint object detection and

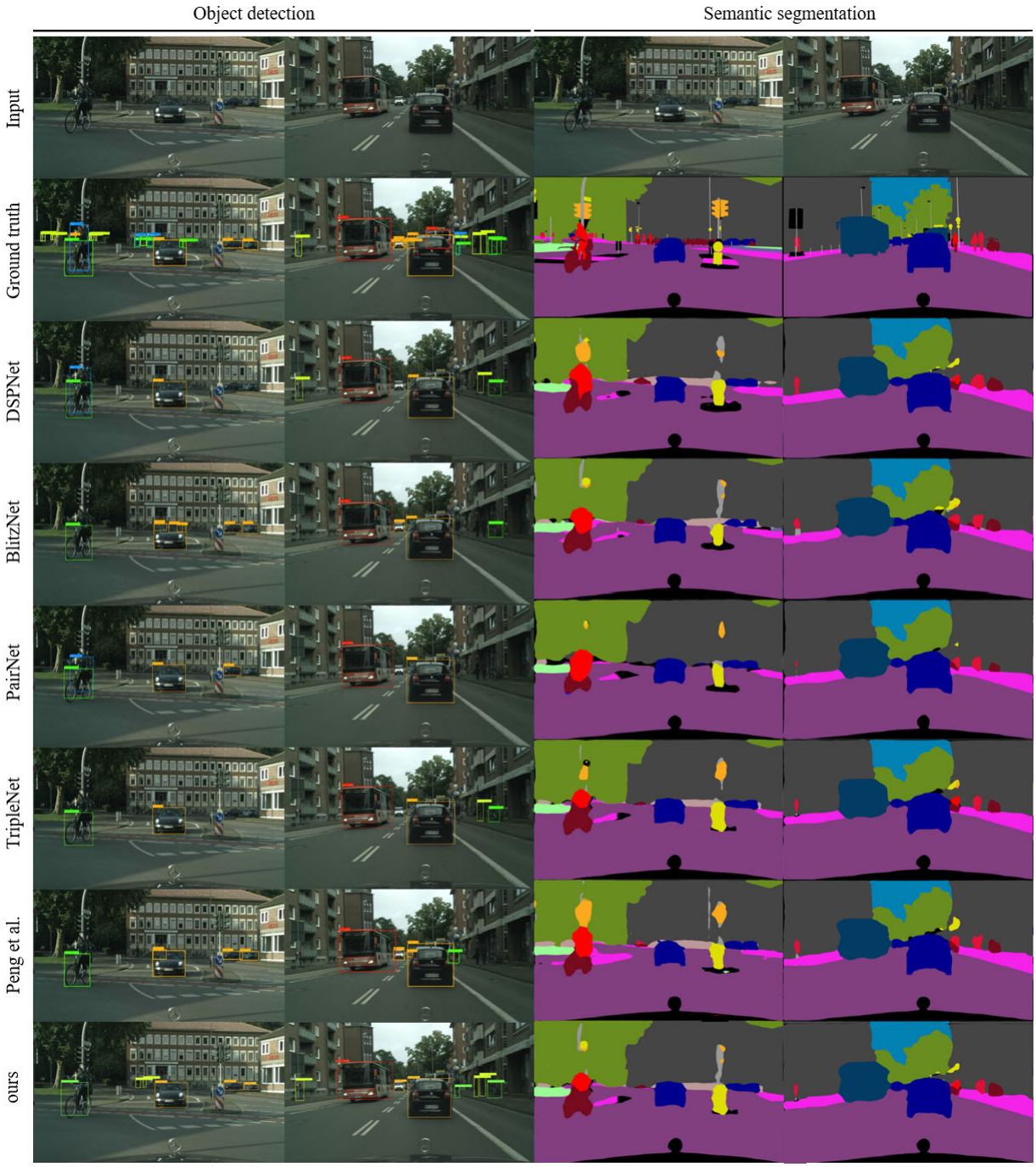

FIGURE 8. Qualitative comparison results with previously developed models on Cityscapes dataset.

semantic segmentation network with existing networks on the Cityscapes dataset.

The first row shows the raw test images, and the second row shows the ground truth of the given raw input images for the object detection and semantic segmentation tasks. Because the network was trained for only 2,975 images of the Cityscapes dataset, the performance results for both tasks were relatively lower than those for the ICV22 dataset.

a: OBJECT DETECTION

Table 4 compares object detection results of the jointly trained object detection and semantic segmentation network with those of other previously developed models. The object detection results indicate that the overall 34.8 mAP is better than three of the existing joint object detection and semantic segmentation models. Additionally, the results of three classes(i.e. person, car and bike) jointly trained

TABLE 6. Comparison of different training combinations with three loss functions.

Method	mAP50	mIoU	Accuracy
Baseline(Cross entropy)	43.5	83.2	88.4
Cross entropy + Dice	45.9(+2.4)	86.6(+3.4)	89.6(+1.2)
Cross entropy + Focal	44.7(-0.2)	84.4(-1.2)	90.7(+1.1)
Focal + Dice	48.7(+4.0)	86.3(+1.9)	92.1(+1.4)

TABLE 7. Joint training vs single task training.

Method	Recall	mAP50	mIoU	Accuracy	FPS
Detection only	77.6	49.4	-	-	62.6
Segmentation only	-	-	85.6	91.0	65.7
Joint training	79.8(+2.2)	48.7(-0.7)	87.2(+1.6)	91.8(+0.8)	42.2

for object detection are superior to those of other existing networks.

b: SEMANTIC SEGMENTATION

Table 5 compares the semantic segmentation results of the proposed method with previously developed joint object detection and semantic segmentation networks(e.g. DspNet [30], BlitzNet [2], PairNet [3], TripleNet [3] and Peng et al. [41] on the Cityscapes dataset. The joint object detection and semantic segmentation system was trained with the CSPResNet50 [42] backbone network as an encoder because, the network illustrated better results on the ICV22 dataset. Our method exhibited better performance on six of the 18 semantic segmentation classes(i.e. fence, pole, traffic light, rider, train and mbike). The reason for effectiveness of the proposed method is the multi-scale feature fusion network with the BiFPN mechanism utilized in the segmentation head. Combined segmentation loss consisting of weighted focal loss and dice loss by addressing the problem of incorrectly classified voxels and improved class distribution learning.

VI. ABLATION STUDIES

Ablation experiments were designed to illustrate the effectiveness of utilizing combined losses in the segmentation task. In the experiments, three combinations of cross-entropy loss, dice loss and focal loss were implemented and trained for performance evaluation of the network. For the baseline result, a single use of cross-entropy loss was selected and compared with other two loss combinations. From the results listed in Table 6, the highest mAP results for the detection task and the highest accuracy results for the segmentation task can be observed in the combination of focal loss and dice loss.

The combination of cross-entropy loss and dice losses yielded the highest mIoU results for the semantic segmentation task.

Additionally, to verify effectiveness of the joint object detection and semantic segmentation model, the performance results of multitask training and single-task training are compared. The model was trained on (1) object detection task only (2) semantic segmentation task only and (3) joint training and testing for both tasks on ICV22 dataset.

Table 7 illustrate that jointly training both tasks have a positive effect on the overall performance of the model rather than training and testing the model for a single task.

VII. CONCLUSION

This study proposed a novel joint object detection and semantic segmentation network that can efficiently output object detection and semantic segmentation results. The method utilizes an encoder-decoder mechanism with an additional feature fusion network. The model was evaluated and compared with single-tasking models using proposed ICV22 dataset. Additionally, the performance of the method on widely used public Cityscapes was compared with existing multitasking models. The results showed that the addition of the feature fusion mechanism to the encoder-decoder structure improved the performance of both tasks when compared with existing architectures.

The limitations of this research are that the proposed ICV22 dataset can be further annotated for additional classes of both object detection and semantic segmentation tasks. Many of the surrounding environments(e.g., buildings, trees) are provided with a zero pixel and the density of semantic segmentation annotations in a single frame is lower than existing related datasets. However, the scale of the ICV22 dataset was significantly larger. As the annotation process is time consuming and requires manual labor, only the most essential classes on the road are tested for both tasks to evaluate performance of the proposed method. Additionally, memory efficiency of the joint object detection and semantic segmentation of the model can be further improved while maintaining state-of-the-art performance accuracy for both tasks.

ACKNOWLEDGMENT

Data collection is funded by the National Information Society Agency(NIA, South Korea) and the dataset introduced in this research will soon be released as a part of "The Open AI Dataset Project (AI-Hub, South Korea)." All the relevant data information can be accessed through "AI-Hub (www.aihub.or.kr)."

REFERENCES

[1] Y. Qian, J. M. Dolan, and M. Yang, "DLT-Net: Joint detection of drivable areas, lane lines, and traffic objects," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 11, pp. 4670–4679, Dec. 2019.

[2] N. Dvornik, K. Shmelkov, J. Mairal, and C. Schmid, "BlitzNet: A real-time deep network for scene understanding," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4154–4162.

[3] J. Cao, Y. Pang, and X. Li, "Triply supervised decoder networks for joint detection and segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7392–7401.

[4] M. Tan, R. Pang, and Q. V. Le, "EfficientDet: Scalable and efficient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10781–10790.

[5] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.

[6] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, 2015, pp. 1–15.

[7] Z. Cai and N. Vasconcelos, "Cascade R-CNN: Delving into high quality object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6154–6162.
[8] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2117–2125.
[9] G. Song, Y. Liu, and X. Wang, "Revisiting the sibling head in object detector," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11563–11572.
[10] T. Liang, H. Bao, W. Pan, and F. Pan, "Traffic sign detection via improved sparse R-CNN for autonomous vehicles," *J. Adv. Transp.*, vol. 2022, pp. 1–16, Mar. 2022.
[11] J. Cao, J. Zhang, and X. Jin, "A traffic-sign detection algorithm based on improved sparse R-CNN," *IEEE Access*, vol. 9, pp. 122774–122788, 2021.
[12] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.
[13] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 7263–7271.
[14] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*.
[15] A. Bochkovskiy, C.-Y. Wang, and H.-Y. Mark Liao, "YOLOv4: Optimal speed and accuracy of object detection," 2020, *arXiv:2004.10934*.
[16] G. Jocher, A. Stoken, J. Borovec, A. Chaurasia, L. Changyu, A. Hogan, J. Hajek, L. Diaconu, Y. Kwon, Y. Defretin, A. Lohia, B. Milanko, B. Fineran, D. Khromov, and D. Yiwei, "Ultralytics/YOLOv5: V5.0—YOLOv5-P6 1280 models, AWS, supervise.ly and YouTube integrations," Tech. Rep., Apr. 2021. [Online]. Available: https://zenodo.org/record/4679653
[17] C. Li, L. Li, H. Jiang, K. Weng, Y. Geng, L. Li, Z. Ke, Q. Li, M. Cheng, W. Nie, Y. Li, B. Zhang, Y. Liang, L. Zhou, X. Xu, X. Chu, X. Wei, and X. Wei, "YOLOv6: A single-stage object detection framework for industrial applications," 2022, *arXiv:2209.02976*.
[18] C.-Y. Wang, A. Bochkovskiy, and H.-Y. Mark Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," 2022, *arXiv:2207.02696*.
[19] C.-Y. Fu, W. Liu, A. Ranga, A. Tyagi, and A. C. Berg, "DSSD: Deconvolutional single shot detector," 2017, *arXiv:1701.06659*.
[20] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.
[21] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.
[22] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2016, pp. 770–778.
[23] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," 2019, *arXiv:1902.09212*.
[24] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder–decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.
[25] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Munich, Germany: Springer, Oct. 2015, pp. 234–241.
[26] L. C. Chen, G. Papandreou, and I. Kokkinos, "DeepLab: Semantic image segmentation with deep convolutional nets, Atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Jun. 2018.
[27] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2881–2890.
[28] L. Fan, W. Wang, F. Zha, and J. Yan, "Exploring new backbone and attention module for semantic segmentation in street scenes," *IEEE Access*, vol. 6, pp. 71566–71580, 2018.
[29] H.-H. Han and L. Fan, "A new semantic segmentation model for supplementing more spatial information," *IEEE Access*, vol. 7, pp. 86979–86988, 2019.
[30] L. Chen, Z. Yang, J. Ma, and Z. Luo, "Driving scene perception network: Real-time joint detection, depth estimation and semantic segmentation," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2018, pp. 1283–1291.
[31] I. Kokkinos, "UberNet: Training a universal convolutional neural network for low-, mid-, and high-level vision using diverse datasets and limited memory," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6129–6138.
[32] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3213–3223.
[33] G. Neuhold, T. Ollmann, S. R. Bulo, and P. Kontschieder, "The Mapillary Vistas dataset for semantic understanding of street scenes," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4990–4999.
[34] F. Yu, H. Chen, X. Wang, W. Xian, Y. Chen, F. Liu, V. Madhavan, and T. A. Darrell, "BDD100K: A diverse driving dataset for heterogeneous multitask learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 2636–2645.
[35] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2009, pp. 248–255.
[36] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1251–1258.
[37] S. Miraliev, S. Abdigapporov, J. Alikhanov, V. Kakani, and H. Kim, "Edge device deployment of multi-tasking network self-driving operations," in *Proc. Int. Conf. Next Gen. Comput.*, Oct. 2022, pp. 83–85.
[38] S. Abdigapporov, S. Miraliev, J. Alikhanov, V. Kakani, and H. Kim, "Performance comparison of backbone networks for multi-tasking in self-driving operations," in *Proc. 22nd Int. Conf. Control, Autom. Syst. (ICCAS)*, Nov. 2022, pp. 819–824.
[39] G. J. Brostow, J. Fauqueur, and R. Cipolla, "Semantic object classes in video: A high-definition ground truth database," *Pattern Recognit. Lett.*, vol. 30, no. 2, p. 8897, 2009.
[40] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The KITTI vision benchmark suite," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 3354–3361.
[41] J. Peng, Z. Nan, L. Xu, J. Xin, and N. Zheng, "A deep model for joint object detection and semantic segmentation in traffic scenes," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2020, p. 18.
[42] C.-Y. Wang, H.-Y. Mark Liao, Y.-H. Wu, P.-Y. Chen, J.-W. Hsieh, and I.-H. Yeh, "CSPNet: A new backbone that can enhance learning capability of CNN," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 390–391.
[43] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn.*, May 2019, pp. 6105–6114.
[44] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking Atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*.
[45] T. Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. DollBr, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, Sep. 2014, pp. 740–755.
[46] J. Jain, J. Li, M. Chiu, A. Hassani, N. Orlov, and H. Shi, "OneFormer: One transformer to rule universal image segmentation," 2022, *arXiv:2211.06220*.
[47] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, "Masked-attention mask transformer for universal image segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 1290–1299.

SHAKHBOZ ABDIGAPPOROV (Student Member, IEEE) received the B.B.A. degree from Inha University, in 2021, where he is currently pursuing the M.Sc. degree in electrical and computer engineering.

He is a dedicated and ambitious researcher in the field of electrical and computer engineering. Throughout his academic career, he has developed a strong interest in the field of deep learning for computer vision and autonomous vehicles. His research interest includes developing novel approaches that can improve the accuracy and efficiency of computer vision applications in challenging real-world scenarios.

SHOKHRUKH MIRALIEV (Student Member, IEEE) received the B.B.A. degree from Inha University, in 2021, where he is currently pursuing the M.Sc. degree.

He is a dedicated and passionate researcher in the field of electrical and computer engineering. His research interests include the application of deep learning to computer vision and autonomous vehicles. Specifically, he is exploring innovative techniques that enable machines to perceive and analyze visual information in a manner similar to that of human beings, with the goal of enhancing the accuracy and efficiency of computer vision systems.

VIJAY KAKANI (Member, IEEE) received the B.Sc. degree in electronics and communication engineering from Jawaharlal Nehru Technological University, Kakinada, India, in 2012, the M.Sc. degree in computers and communication systems from the University of Limerick, Ireland, in 2014, and the Ph.D. degree in information and communication engineering and future vehicle engineering from Inha University, South Korea, in 2020.

Currently, he is an Assistant Professor with the Department of Integrated System Engineering, School of Global Convergence Studies, Inha University. His research interests include autonomous vehicles, sensor signal processing, applied computer vision, deep learning, systems engineering, and machine vision applications.

HAKIL KIM (Member, IEEE) received the M.Sc. and Ph.D. degrees in electrical and computer engineering from Purdue University, in 1985 and 1990, respectively.

In 1990, he joined the College of Engineering, Inha University, Incheon, South Korea, where he is currently a Full Professor with the Department of Information and Communication Engineering. In order to retain the balance between academic research and commercial development, he founded Vision Inc., in 2014, where he is also the CEO. His research interests include biometrics, intelligent video surveillance, and embedded vision for autonomous vehicles. Since 2003, he has been actively involved as the Project Editor of the International Standardization of Biometrics, ISO/IEC JTC1/SC37.