Pedestrian Detection for Autonomous Cars: Inference Fusion of Deep Neural Networks

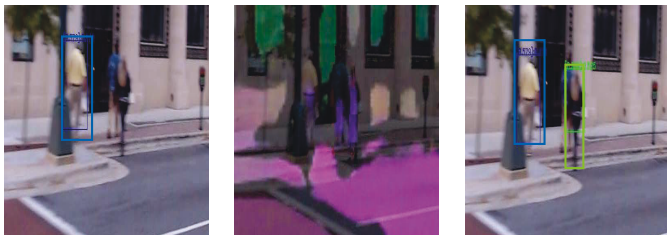
Muhammad Mobaidul Islam, Abdullah Al Redwan Newaz, and Ali Karimoddini

***Abstract*— Network fusion has been recently explored as an approach for improving pedestrian detection performance. However, most existing fusion methods suffer from runtime efficiency, modularity, scalability, and maintainability due to the complex structure of the entire fused models, their end-to-end training requirements, and sequential fusion process. Addressing these challenges, this paper proposes a novel fusion framework that combines asymmetric inferences from object detectors and semantic segmentation networks for jointly detecting multiple pedestrians. This is achieved by introducing a consensus-based scoring method that fuses pair-wise pixel-relevant information from the object detector and the semantic segmentation network to boost the final confidence scores. The parallel implementation of the object detection and semantic segmentation networks in the proposed framework entails a low runtime overhead. The efficiency and robustness of the proposed fusion framework are extensively evaluated by fusing different state-of-the-art pedestrian detectors and semantic segmentation networks on a public dataset. The generalization of fused models is also examined on new cross pedestrian data collected through an autonomous car. Results show that the proposed fusion method significantly improves detection performance while achieving competitive runtime efficiency.**

***Index Terms*— Pedestrian detection, autonomous vehicles, object detection, semantic segmentation, fusion, deep learning.**

I. Introduction

THE advent of autonomous and semi-autonomous driving technologies has led to the development of vehicles with different levels of autonomy equipped with Advanced Driver Assistance Systems (ADAS), in which automatically identifying pedestrians is a crucial safety requirement. Deep learning-based pedestrian detection methods have gained much attention in recent years due to their superior detection accuracy. A vast majority of deep learning-based pedestrian detection methods rely on object detectors, typically by detecting bounding boxes and their associated class confidence scores. Pedestrian detection, however, is more challenging than generic object detection problems since the image space variability of this class is very large as pedestrians appear in various poses, clothing, and various articulations of body parts. Furthermore, different sizes, aspect ratios, and partial occlusion of pedestrians make it challenging to detect pedestrians in an image frame.



(a) (b) (c)

Fig. 1. Fusion of network inferences can improve detection accuracy: (a) using a state-of-the-art object detection model can predict only one pedestrian, (b) our semantic segmentation network provides useful information though it is noisy and difficult to interpret and accurately localize the pedestrians, and (c) the developed fusion mechanism in this paper can accurately detect both pedestrians.

Deep learning approaches for pedestrian detection either rely on the configurations in anchor boxes or high level semantic information [1]. The former approaches utilize a set of anchor boxes combined with an image classification network to localize pedestrians [2], [3], [4]. The later approaches label each pixel of an image with its corresponding class. Each of these methods has its own pros and cons. For instance, consider the scenario in Fig. 1 in which one can see that an anchor box-based detection method is unable to detect all pedestrians. On the other hand, given the same scene as an input, the semantic segmentation network yields noisy but useful information that can be used for enhancing pedestrian detection.

Therefore, in this paper, we hypothesize that jointly predicting pedestrians using both object detection and semantic segmentation networks enables efficient and robust pedestrian detection in challenging environments. The crux of the challenge is that it is not straightforward to combine the inferences of multiple asymmetric networks for the following reasons. First, pedestrian inferences appear to take place in different domains. Second, since semantic segmentation cannot segregate individual pedestrians, it is challenging to obtain one-to-one correspondence between the inferences of

Manuscript received 28 May 2021; revised 14 January 2022, 7 May 2022, and 30 July 2022; accepted 13 September 2022. Date of publication 30 September 2022; date of current version 5 December 2022. This work was supported in part by the North Carolina Department of Transportation (NCDOT) under Award RP2019-28 and Award TCE2020-03, in part by the National Science Foundation under Award 2018879 and Award 2000320, and in part by Ford Motor Company. The Associate Editor for this article was H. Dong. *(Corresponding author: Ali Karimoddini.)*

Muhammad Mobaidul Islam and Ali Karimoddini are with the Department of Electrical and Computer Engineering, North Carolina Agricultural and Technical State University, Greensboro, NC 27411 USA (e-mail: mmislam@aggies.ncat.edu; akarimod@ncat.edu).

Abdullah Al Redwan Newaz is with the Department of Computer Science, University of New Orleans, New Orleans, LA 70148 USA (e-mail: aredwann@uno.edu).

Digital Object Identifier 10.1109/TITS.2022.3210186

semantic segmentation and an object detection network. Third, the input-output of each network might have different sizes. Fourth, a sequential operation of fusing inferences from multiple networks requires a high computation time.

To address the above problems, we propose a novel fusion method that combines asymmetric inferences of multiple networks for enhanced pedestrian detection. Our fusion method is designed in such a way that parallelizes the image computations in terms of pedestrian detection and semantic segmentation from different networks. We first map overlapping areas of asymmetric inferences of two networks into the pixel domain. As semantic segmentation cannot segregate individual pedestrians, we adopt an anchor box-based detection method to calculate the overlapping area for individual pedestrians. We then propose a novel scoring metric to calculate the confidence score for pedestrians by correlating the outputs of different networks. The contributions of this paper, therefore, are as follows:

1) We propose a novel approach for fusing inferences of an object detection network and a semantic segmentation network to improve the robustness and performance of the overall pedestrian detection.
2) We introduce a pixel-wise consensus-based approach to address the challenge of combining asymmetric inferences of multiple networks.
3) The proposed inference fusion framework is generic, agnostic, and modular in the sense that it can be employed for any choices of object detection and semantic segmentation networks.
4) We create a pedestrian dataset with 1746 annotations of 867 images using an autonomous car platform for cross dataset evaluation purposes.
5) Extensive validation experiments are carried out including the fusion of 8 object detectors and 4 semantic segmentation networks (overall 32 fusion models) by benchmarking over two different datasets to assess the performance and the run-time efficiency of the developed fused models for pedestrian detection.

In contrast to existing pedestrian detection methods, our framework offers the following advantages:

- **Easy development:** our architecture is naturally well suited for iterative development and testing by leveraging the existing state-of-the-art networks.
- **Modularity:** our framework offers a modular structure where any object detection and semantic segmentation networks can be adopted.
- **Parallelism:** our framework can handle the computations of anchor box detection and semantic segmentation networks in parallel, resulting in better runtime efficiency.
- **Scalability:** our framework can also obtain higher runtime efficiency by combining relatively less complex networks.
- **Maintainability:** Due to the modular structure of our framework, one can easily add or replace a network.

The remainder of the paper is organized as follows. Section II describes object detection and semantic segmentation network architectures and reviews the existing network fusion mechanisms. Section III presents the proposed methodology. Section IV demonstrates the experimental results for the implementation of different fused models along with their detection performance and runtime analysis. Finally, Section V provides the concluding remarks.

II. Related Work

In this section, we will briefly introduce the state-of-the-art pedestrian detection and masking methods.

A. Classical Pedestrian Detection

Before the deep learning era, hand-crafted features were a popular choice to capture localization signals in an image pervasively [5], [6], [7], [8], [9], [10], [11], [12], [13]. For instance, combining gradient direction and edge orientation on small spatial regions of an image, a Histogram of Oriented Gradients (HOG) feature descriptor is created to detect pedestrians [5]. Different versions of Local Binary Pattern (LBP) based pedestrian feature descriptors, including Semantic LBP and Fourier LBP are also used to detect pedestrians [13]. Semantic LBP and Fourier LBP exploit the idea of a geometrical interpretation and a Fourier boundary descriptor, respectively. Training weak classifiers in the Euclidean space for faster computation, AdaBoost is designed for faster pedestrian detection [8]. Exploiting motion information, covariance descriptors can be used to detect pedestrians in multi-camera settings [10]. Moreover, filtered channel features and low-level visual features along with spatial pooling demonstrate a significant improvement in pedestrian detection [14], [15]. However, a major drawback of these classical methods is that they generally perform poorly when pedestrians are partially occluded, pedestrians have a different articulation of body parts, and environmental conditions (e.g., illumination level, background, etc.) change.

B. Deep Learning-Based Pedestrian Detection

Deep learning-based algorithms [3], [16], [17], [18], [19], [20] have created a breakthrough in pedestrian detection and taken the leading position in solving the pedestrian detection problem. For example, [4], [21], [22] proposed a pedestrian detection method based on Faster R-CNN that utilizes a Region Proposal Network (RPN) to generate pedestrian candidates, resulting in both improved detection performance and runtime efficiency compared to its predecessors [23], [24]. Nevertheless, the computation cost of Faster R-CNN still is high for realtime applications such as autonomous driving. One way to achieve a better runtime efficiency is to use single stage pedestrian detectors due to the fact that by leveraging the power of CNN, they combine feature extraction, location regression, and region classification. However, these methods often suffer from low accuracy [25]. Asymptotic Localization Fitting (ALF) was introduced in [20] by employing the Single Shot Detector (SSD) which was stacked together with multiple predictors for localization of bounding boxes. In [26], the authors used YOLO for pedestrian detection, which is a remarkably fast single stage anchor free detector. Another

anchor-free detector is Center and Scale Prediction (CSP) [16] that can generate bounding boxes without any requirements of extra post-processing schemes by utilizing concatenated feature maps for predicting pedestrians. In [27], the authors introduced the RetinaNet that overcomes the foreground-background class imbalance problem by incorporating an additional focal loss along with classification and localization losses in the loss function.

C. Pedestrian Masking Using Semantic Segmentation

Semantic segmentation is the process of making the dense prediction for inferring semantic labels for pixels in an image frame so that each pixel denotes the class information of its corresponding object. Most semantic segmentation networks are based on a Fully Convolutional Network (FCN). For instance, FCN is used in [28] and [29] to introduce a pixel-wise prediction for an end-to-end semantic segmentation. FCN along with a dilated convolution is also used in the dense prediction problem [30] that requires high resolution features. Another approach is using Generative Adversarial Network (GAN) [31] for semantic segmentation to improve labeling accuracy. In [32], authors proposed Pix2Pix GAN for a general-purpose solution to image-to-image translation problems. In [33], authors proposed SegGAN in which a GAN is adopted to refine the segmentation masks. To design a low convolutional input/output network for semantic segmentation, HarDNet [34] added a soft constraint on each layer and achieved high accuracy and low memory traffic. To reduce the computational cost, an encoder-decoder based semantic feature extraction method is introduced in [35]. Recently, SegFormer [36] introduced a hierarchically structured transformer encoder that provides multi-scale contextual information for enhancing accuracy in a semantic segmentation network. To tackle the problem of computational cost more efficiently, Deep Dual-resolution Networks (DDRNets) [37] introduces a composition of two deep branches with multiple bilateral fusions between them.

D. Pedestrian Masking Using Instance Segmentation

Instance segmentation combines the object detection and the semantic segmentation in a unified framework. The process of instance segmentation begins with identifying each object instance within an image utilizing a detection network, and then, predicting instance masks in a pixel-to-pixel manner by utilizing an extended subnetwork. Popular instance segmentation frameworks are based on R-CNN [38], [39]. For example, Mask R-CNN provides the state-of-the-art framework for object instance segmentation extending a parallel FCN branch of the Faster R-CNN for predicting masks. On the other hand, a single shot-based instance segmentation networks such as YOLACT [40] and Mask SSD [1] generate comparatively fast instances at the expense of lower detection accuracy compared to the Mask R-CNN. To accomplish the instance segmentation task, YOLACT avoids the re-pooling operation and produces instance masks by linearly combining the prototypes and mask-coefficient generated from two subtasks. Likewise, Mask SSD utilizes a subnetwork that outputs pixel-wise segmentation for each detection while providing the multi-scale and feedback features from different layers as input. Recently, a Point-based instance segmentation is introduced in [39] which can provide comparable performance to the region-based Mask R-CNN with faster inference.

E. Pedestrian Detection Using Fusion Methods

Deep fusion of multiple networks demonstrates the advantage of capturing a variety of complementary information for detecting pedestrians robustly. Deep fusion can also be applied at different layers of the networks for the detection reinforcement [35], [41], [42], [43], [44], [45], [46], [47], [48] In [41], the authors introduced a Hyper-learner that integrates different kinds of channel features into CNN-based detectors in a multi-tasking manner. In [43], the authors proposed an unsupervised learning algorithm to learn a non-linear mapping from the RGB channels to the thermal channel. In [44], the authors proposed an infusion network where semantic segmentation provides additional supervision that helps in guiding features in shared layers along with the pedestrian detection network. In [42], the authors introduced MGAN that uses an additional Mask-Guided Attention branch to produce a pixel-wise attention map to guide the network attention, resulting in improved performance on occluded pedestrians. In [35], the authors introduced panoptic segmentation that merges semantic segmentation and instance segmentation by combining the output of PSPNet [49] and Mask R-CNN. Hence, the first network is a Region Reconstruction Network and the second network is a Multi Scale Detection Network. In [47], a Scale-Aware Fast R-CNN framework is introduced that adaptively combines the output of multiple sub-networks to detect pedestrians of different scales. In [48], the authors introduced a halfway fusion method that fuses feature information of the color image and thermal image by utilizing a deconvolutional single shot multi-box detector.

Though the aforementioned fusion methods exhibited the improvement of the detection accuracy, it is often found that such fusion architectures are complex, resulting in a high computation cost and significant reduction of the run-time efficiency due to the sequential operation of subnetworks at each step. In addition, the training of the entire fused architecture challenges the development of such models. For instance, a monolithic deep fusion mechanism cannot directly utilize pre-trained weights of backbone networks — either the entire network needs to be trained from the scratch or a layer-wise pre-training strategy can be used. The closest work to this paper is [45] which fuses multiple deep neural networks with a soft-rejection method to adjust the confidence in the detection results, and optionally uses semantic segmentation network to improve detection accuracy. More specifically, the method in [45] is a two-stage fusion approach in which initially multiple deep neural networks are fused to generate candidate pedestrian proposals, followed by an optional semantic segmentation network which rejects false proposals. This is proven to be very computationally expensive due to the large input size and complex architecture of the network. In contrast, our proposed fusion method has the

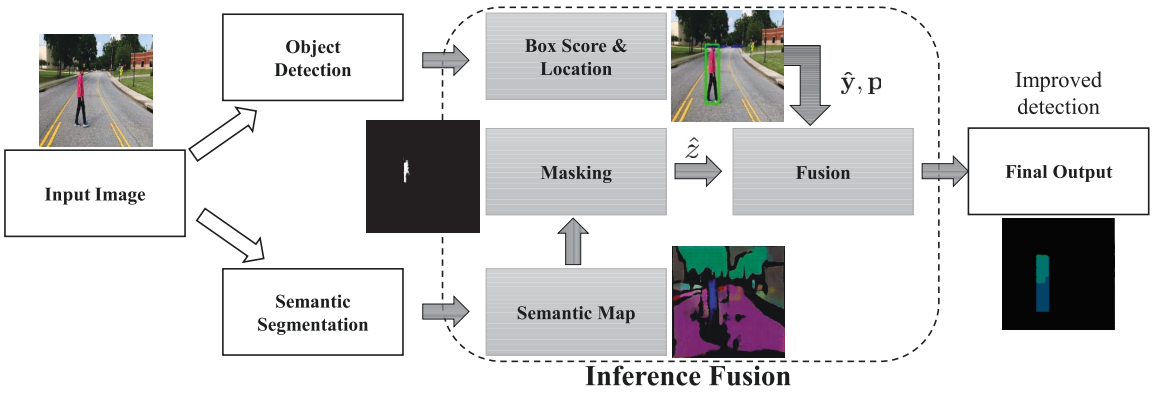


Fig. 2. **Proposed inference fusion architecture:** An input image is fed into the object detection branch and semantic segmentation branch simultaneously. The output of the semantic map is masked for pedestrian instances and mapped into the detected bounding boxes from the object detection network. The inferences of both networks are fused based on the consensus on detected pedestrians.

following merits: Firstly, our framework provides a general-purpose solution to combine pedestrian detection networks with semantic segmentation networks. Secondly, by fusing two asymmetric inferences, our method improves the detection accuracy. Finally, the parallel processing of the two networks allows us to achieve a competitive runtime efficiency.

III. Methodology

In this section, we propose an asymmetric inference fusion architecture for pedestrian detection. As shown in Fig. 2, the proposed framework has two branches. The first branch utilizes an object detection method to predict the bounding box score and location of pedestrians, and the second branch utilizes semantic segmentation and masking techniques to generate pedestrian inferences. Finally, we develop a fusion mechanism to combine these two different inferences to predict pedestrians jointly. Next, we will discuss the components of the proposed framework in detail.

A. Inferring Pedestrians From an Object Detector

To detect pedestrians, first, we need to classify the pedestrians, and then, localize their corresponding positions within a given input image. More specifically, consider a generic pedestrian detector that takes an input image $x \in \mathcal{X}$ according to a data distribution $\mathcal{D}$. We use supervised learning to detect pedestrians which requires training pedestrian detectors with the label data, $y \in \mathcal{Y}$. Canonically, the ground truth labels are given in terms of bounding boxes that represent the true pedestrians' positions on images. In the training context, the detector is tasked to classify pedestrians and provide their precise bounding boxes within the known scenes. The output of the detector for a given input image $x \in \mathcal{X}$ includes the vector of bounding boxes $\hat{\mathbf{y}}$ and their corresponding confidence scores vector $\mathbf{p}$, such that each predicted (or estimated) bounding box $\hat{y} \in \hat{\mathbf{y}}$ associated with a confidence score $p \in [0, 1]$. In the course of training, the goal of a good detector is to make its prediction $\hat{y}$ as close as possible to the ground truth y. To capture this relationship, we can define a loss function $\mathcal{L}(y, \hat{y}|x)$ that measures a distance between the predicted labels $\hat{y}$ and their corresponding ground truth y for a given input image x. Different types of loss functions are studied in the literature to improve pedestrian detection performance. A pedestrian detector leverages at least two types of loss functions, i.e., the classification loss and the localization loss. A popular choice of classification loss is Cross-Entropy loss. On the other hand, a common choice of localization/regression loss is Smooth-L1 loss. In the testing context, we evaluate the detector performance in unseen data. Evaluation metrics for pedestrian detection include Average Precision and Log Average Miss Rate.

Standard deep learning-based pedestrian detectors come in two flavors: single-stage detectors and two-stage detectors. A single-stage based pedestrian detector, e.g., SSD and RetinaNet, requires only a single pass through the neural network and predicts all the bounding boxes directly. This leads to a simpler and faster model architecture, thereby suitable for edge devices. On the other hand, a two-stage pedestrian detector, e.g., Faster R-CNN, Cascade Mask R-CNN, and MGAN, requires a pre-trained ImageNet model such as VGG-16 or MobileNet, followed by a region proposal network (RPN) to detect pedestrians [42]. Although most two-stage pedestrian detectors are known to be relatively slow, they are very accurate and provide state-of-the-art performance.

B. Inferring Pedestrians From a Semantic Map

Pedestrian masking is a byproduct of the semantic segmentation which has a wide range of applications in autonomous driving. Unlike an object detector that uses bounding boxes to estimate the position of objects, the semantic segmentation focuses on labeling each pixel of an image with its corresponding class. The output of semantic segmentation is called semantic map which is crucial for a scene understanding, inferring support-relationships among objects. In this work, from a semantic map, we are only interested in those dense predictions that represent pedestrians. For this purpose, a masking method (based on color thresholds) is applied to separate and group the pixels related to the pedestrian class from a semantic map.

Now consider a generic semantic segmentation system that takes an input image $x \in \mathcal{X}$ according to a data distribution $\mathcal{D}$. Unlike the ground truth label data of a pedestrian detector, the ground truth data $z \in \mathcal{Z}$ for a semantic segmentation network is given in terms of polygons with different filled colors to represent different objects in a scene. During training, a deep neural network learns the feature information using different feature descriptors comparing each pixel of the output of the network with the corresponding pixel in the ground truth segmentation image. At the training time, the task of the segmentation classifier is to group pixels based on their object category within the known scenes. The output of the classifier includes the groups of colored pixels denoted by $\hat{z} \in \mathcal{Z}$ for each class for a given input image $x \in \mathcal{X}$. In the course of training, the goal of a good classifier is to make its prediction $\hat{z}$ as close as possible to the ground truth z. To capture this relationship, we can define a generic loss function $\mathcal{L}(z, \hat{z}|x)$ that measures matching information between the predicted group of pixel $\hat{z}$ and the ground truth z for a given input image x. During testing with a semantic segmentation model, each pixel of an input image x is classified as a predefined class $\hat{z}$ based on the learned information. For semantic segmentation, Cross-Entropy loss is also a popular choice. Standard semantic segmentation networks predominantly come in the form of three architectures: Fully Convolutional Network (FCN), Encoder-decoder, and Two-pathway architectures. In the FCN, fully convolution layers generate a spatial map for pixel-wise semantic segmentation [28], [50]. In the Encoder-decoder architecture such as SegFormer and FCHarDNet, the encoder generates feature maps containing several feature information (e.g., shape, size) and the decoder takes this information to produce the segmentation maps. Some lightweight backbone networks, e.g., MobileNet [51] and ShuffleNet [52], often are used as an encoder to reduce the computational burden of this architecture. Finally, in Two-pathway architecture, a shallow path along with the main path are used to extract semantic information. Thus, this architecture avoids information loss during repeated downsampling of feature maps in the Encoder-decoder architecture.

C. Inference Fusion Architecture

We propose an inference fusion architecture that combines unalike predictions of multiple networks. More specifically, we fuse two networks: (i) the first one is an object detection network architecture that is used for detecting objects in the form of bounding boxes along with their confidence scores over the entire image, and (ii) the second one is a semantic segmentation network architecture that is used for semantic map generation based on the same input image that the object detector has used. The challenge is that it is not straightforward to combine the inferences of these two models for the following reasons. First, the output of the object detector is in the form of bounding boxes with corresponding confidence scores whereas the output of semantic segmentation is a group of pixels dedicated for the pedestrian class. Second, since the semantic segmentation cannot segregate individual pedestrians, it is challenging to obtain one-to-one correspondence from their inferences. Third, the outputs of each branch do not share the same image size so that they can be easily combined without any modification, e.g., the Pix2Pix GAN semantic segmentation branch predicts over an image size of 256 × 256 whereas the SSD MobileNet object detection branch predicts over an image size of 300 × 300.

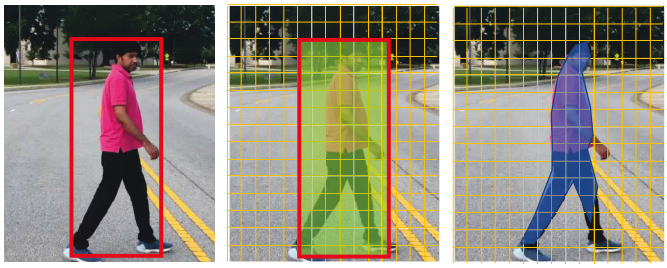


Fig. 3. We utilize pixel-level information from two deep neural network models to compute joint confidence scores: (a) the output of the pedestrian detector, (b) the pixels within the pedestrians' bounding box, and (c) the pixels for the pedestrian class from the semantic map within the box.

To overcome these challenges, we apply masking operation on the semantic map to carry out inferences over the pedestrians, and then, scale those masks to overlay the output image of the object detection network inference. Fig. 2 shows our proposed fusion framework with two branches: the object detection branch and the semantic segmentation branch. The key idea is to let the two branches with two different models reach a common consensus on pixel-level information. For this purpose, we use the object detection branch that takes images as input to generate bounding boxes and corresponding confidence scores. In parallel, the semantic segmentation branch and a masking process are applied to translate the same input image to a gray-scale mask-based image to predict pedestrians from a semantic map. This will yield pixel-wise information about pedestrians that needs to be mapped into bounding boxes to compute joint confidence scores. For instance, consider the output image of an object detector in Fig. 3.(a). We identify the pixels within the pedestrians' bounding boxes as shown in Fig. 3.(b). On the other hand, we identify the pixels for the pedestrian class from the semantic map as shown in Fig. 3.(c). These steps allow us to calculate the number of overlapping pixels by combining two asymmetric sources of inferences. Let $A \in \mathbb{N}$ be the total number of pixels within a predicted bounding box $\hat{y}$ shown in Fig. 3.(b), which formally can be calculated as follows:

$$A = |\{\ell \in x | \ell \in \hat{y}\}|, \tag{1}$$

where ℓ is image pixel and $|.|$ represents the cardinality of a set. Let $N \in \mathbb{N}$ be the number of masked pixels for the pedestrian class $\hat{z}$ within the bounding box $\hat{y}$ as shown in Fig. 3.(c). Formally, we can compute N as follows:

$$N = |\{\ell \in x | \ell \in \hat{z} \wedge \ell \in \hat{y}\}|. \tag{2}$$

We then compute the amount of non-overlapped pixels between the two inferences which are denoted by $M \in \mathbb{N}$ such that $M = A - N$. The joint confidence score for a given

bounding box can be calculated using Eqn. (3).

$$Score := \begin{cases} clip\left(\frac{\frac{N}{1-p} + M.p}{A}, 1\right) & \text{if } p < 1, \\ p & \text{otherwise,} \end{cases} \quad (3)$$

where $p \in \mathbb{R}^+$ is the confidence score of pedestrian class from the object detector. As it is obvious from the Eqn. (3), when there is no consensus between the two models, i.e., $N = 0$, or the pedestrian detector has full confidence, i.e., $p = 1$, we then rely on the confidence score of pedestrian detector to classify pedestrians. Otherwise, we compute the weighted average of overlapped pixels, N, and non-overlapped pixels, M, in the bounding box with the weights of $\frac{1}{1-p}$ and p, respectively. Clearly, with this method, we are taking into account the overlapping pixels with a much higher weight, which complements the lack of confidence in the object detection module based on the consensus (overlapped pixel) of the two models. Finally, while boosting the confidence scores, we make sure that the joint confidence score does not exceed 1 by clipping the scoring function. The summary of the fusion process is provided in Algorithm 1.

Algorithm 1 Network Fusion

Require: image x

$\hat{\mathbf{y}}, \mathbf{p} \Leftarrow$ `object_detection_branch(x)`
$\hat{z} \Leftarrow$ `semantic_segmentation_branch(x)`
$detection \Leftarrow \{\}$
for all $\hat{y} \in \hat{\mathbf{y}}$ and $p \in \mathbf{p}$ **do**
  Compute A from x given $\hat{y}$ using Eqn. (1)
  Compute N from $\hat{z}$ given $\hat{y}$ using Eqn. (2)
  Compute M such that $M := A - N$
  Compute $Score$ from p, A, N, M using Eqn. (3)
  $detection \Leftarrow detection \cup \{\hat{y}, Score\}$
end for
return $detection$

IV. Experimental Results

In this section, we evaluate the proposed inference fusion framework with numerous pedestrian detectors and pedestrian mask segmentation models. We benchmark our results on a 64-bit Ubuntu 18.04 server that has an Intel(R) Core(TM) i9-7900X CPU @ 3.30GHz with 64GB memory, which also has two NVIDIA GeForce RTX 2080 GPUs with 8GB memory each. We used Python programming language with frontend Keras and Pytorch libraries and backend TensorFlow library for training, testing, and benchmarking of our models. We also use Docker to build image with different versions of Python and associated libraries. The developed codes and tools are publicly available at https://github.com/ACCESSLab/InferenceFusion, the model weights for our trained networks are available at https://bit.ly/3qoFvAt (the weights of other pre-trained networks that are used in this study are available online in provided references), and the developed Geensboro Dataset is accessible at https://bit.ly/3GpTtaS.

A. Datasets and Training

For evaluation purposes, we use CityPersons dataset [14] which is a very diverse and challenging pedestrian dataset, collected through autonomous cars in 27 different cities in Germany and neighboring countries. In total, CityPersons dataset contains 2,975 training images and 400 test images. The advantage of CityPersons dataset is that it has been widely used for training different pedestrian detectors. We, therefore, use four pre-trained object detector models including Cascade Mask R-CNN with HRNet backbone, Cascade Mask R-CNN with MobileNet backbone, CSP, and MGAN which are already pre-trained on CityPersons dataset. In addition, to include more representative of pedestrian detection networks, we train another four pedestrian detection networks on CityPersons dataset including SSD MobileNet v1, SSD MobileNet FPN, Faster R-CNN, and RetinaNet. To expedite the training time and loss convergence, we use pretrained weights for these networks from the MSCOCO dataset [53] and fine-tune them on the CityPersons dataset using the progressive training method [54]. Note that the reason to not solely rely on pre-trained weights from the MSCOCO dataset and fine-tune the models on CityPersons is that the MSCOCO dataset lacks in complex driving backgrounds, partial occlusions, difficult body articulation, small size pedestrians.

For the semantic segmentation models, we use Cityscapes dataset [55]. The Cityscapes dataset is in fact CityPersons dataset with semantic annotations. Similarly, Cityscapes dataset has been widely used for training semantic segmentation models. Therefore, we use three semantic segmentation models including SegFormer, DDRNet, and FCHarDNet which are already pre-trained on Cityscapes dataset. In addition, we train Pix2PixGAN on this dataset as the fastest semantic segmentation network. Similar to the detection networks, the same set of 2,975 training images are used for training all of these semantic segmentation networks and 400 images for testing purposes.

The CityPersons dataset is collected before 2017, often containing difficult examples of crowded scenarios captured under low illumination conditions. To address these concerns and also for the purpose of cross data evaluation [54] and assessing the generalizability of the developed framework for other driving scenes particularly at the US, we created a cross pedestrian dataset. For this purpose, we used Aggies Autonomous Auto driving passenger vehicle equipped with two front-facing cameras —one for a short distance and the other one for the long-distance views. The images are collected in downtown Greensboro city in North Carolina states, USA. This dataset consists of 867 test images with 1746 pedestrian annotations. This dataset is hereafter referred to as the "Greensboro Downtown dataset."

B. Parallel Implementation of Inference Fusion Mechanism

Since we use a pedestrian detector and a semantic segmentation network to predict pedestrians jointly, a parallel implementation of inference fusion speeds up the runtime of the overall pedestrian detection system. For this purpose, our inference fusion method is implemented using

TABLE I

BENCHMARK ON THE CITYPERSONS DATASET. CITYPERSONS → CITYPERSONS REFERS TO TRAINING ON CITYPERSONS DATASET AND TESTING ON CITYPERSONS DATASET. ALL PEDESTRIAN DETECTORS PERFORMANCE IMPROVE BY FUSING WITH A PEDESTRIAN MASKING NETWORK USING OUR FRAMEWORK. HERE, THE INFERENCE FUSION WITH SEGFORMER MODEL ACHIEVES THE HIGHEST AP AND THE LOWEST MR

CityPersons → CityPersons	Baseline		SegFormer		DDRNet		FCHarDNet		Pix2PixGAN	
	AP	MR	AP	MR	AP	MR	AP	MR	AP	MR
Cascade Mask R-CNN HRNet	65.94	34.05	**67.49**	**32.50**	66.88	33.11	66.85	33.14	66.50	33.49
Cascade Mask R-CNN MobileNet	64.80	35.19	**66.49**	**33.50**	65.56	34.43	65.81	34.18	65.21	34.78
CSP	54.28	45.71	**56.99**	**43.00**	55.50	44.49	55.96	44.03	55.31	44.68
MGAN	52.61	47.38	**54.77**	**45.22**	54.51	45.48	54.54	45.45	54.41	45.58
SSD MobileNet v1	15.99	84.00	**17.88**	**82.11**	16.78	83.21	16.92	83.07	16.68	83.31
SSD MobileNet FPN	12.54	87.45	**32.61**	**67.38**	20.88	79.11	21.43	78.56	18.86	81.13
Faster R-CNN Inception v2	46.30	53.69	**47.40**	**52.59**	46.70	53.29	47.10	52.89	46.67	53.32
RetinaNet	34.84	65.15	**42.38**	**57.61**	38.77	61.22	39.73	60.26	38.28	61.71

TABLE II

CROSS DATASET EVALUATION ON THE GREENSBORO DATASET. CITYPERSONS → GREENSBORO REFERS TO TRAINING ON CITYPERSONS DATASET AND TESTING ON GREENSBORO DATASET. ALL PEDESTRIAN DETECTORS PERFORMANCE ARE IMPROVED BY FUSING WITH A PEDESTRIAN MASKING NETWORK USING OUR FRAMEWORK. HERE, SEGFORMER AND FCHARDNET ARE TWO TOP PERFORMING PEDESTRIAN MASK SEGMENTATION BRANCHES

CityPersons → Greensboro	Baseline		SegFormer		DDRNet		FCHarDNet		Pix2PixGAN	
	AP	MR	AP	MR	AP	MR	AP	MR	AP	MR
Cascade Mask R-CNN HRNet	86.93	13.06	90.74	09.25	90.29	9.70	**91.26**	**8.73**	90.35	9.64
Cascade Mask R-CNN MobileNet	85.32	14.67	89.28	10.71	87.53	12.46	**89.44**	**10.55**	87.55	12.44
CSP	48.88	51.11	70.07	29.92	65.01	34.98	**75.26**	**24.73**	65.49	34.50
MGAN	50.48	49.51	**67.59**	**32.40**	66.06	33.93	67.56	32.43	66.16	33.83
SSD MobileNet v1	14.09	85.90	**28.93**	**71.06**	21.22	78.77	28.71	71.28	21.17	78.82
SSD MobileNet FPN	46.30	53.69	68.30	31.69	57.43	42.56	**74.56**	**25.43**	57.39	42.60
Faster R-CNN Inception v2	84.298	15.70	85.43	14.56	84.34	15.65	**86.01**	**13.98**	84.40	15.59
RetinaNet	68.15	31.84	79.73	20.26	78.68	21.31	**81.96**	**18.03**	78.97	21.02

a multi-processing approach for detecting pedestrians concurrently. In one process, we implement the pedestrian detection branch, while in another process, we implement the pedestrian mask segmentation branch. Each of these processes is implemented using a separate docker container and a discrete GPU. We, however, use the shared memory to fuse the outputs of these branches. The two parallel branches are then cascaded with the fusion inference. We observed that fusing inferences takes much less computational time in contrast to the detection or segmentation branches. As an evidence, we can see from Table IV, the runtime of the overall process is primarily governed by the slower branch.

C. Benchmarking State-of-the-Art Pedestrian Detectors

We thoroughly evaluated and compared our method against state-of-the-art pedestrian detection methods. In the pedestrian detection branch of our fusion framework, we used baseline pedestrian detection methods including bounding-box-based detection methods (i.e., Faster R-CNN, RetinaNet, CSP, SSD MobileNet, and SSD FPN) and instance segmentation networks (i.e., Cascade Mask-R-CNN with MobileNet and HRNet backbones, and MGAN). In the pedestrian mask segmentation brnach, however, we used pedestrian mask segmentation networks (i.e., SegFormer, DDRNet, FCHarDNet, and Pix2PixGAN).

The CSP, Cascade Mask R-CNN with MobileNet and HRNet backbones, and MGAN are already pre-trained by other researchers on the CityPersons dataset [54]. Further, we used SSD MobileNet, SSD FPN, Faster R-CNN, and RetinaNet which are already pre-trained on the MSCOCO dataset [53], and applied progressive training to fine-tune them on CityPersons dataset. The pedestrian mask segmentation networks including SegFormer, DDRNet, and FCHarDNet which are already trained on the CityScape dataset [55]. In addition, we trained the Pix2PixGAN from the scratch on the CityScape dataset.

Table I shows the detailed performance evaluation of the developed fusion inference framework trained and tested on CityPersons dataset with different choices of pedestrian detection and pedestrian mask segmentation networks, compared to the baseline networks. We can observe that the performance of pedestrian detectors is significantly improved when fused with a pedestrian mask segmentation network using the proposed fusion inference method. For instance, fusing Cascade Mask R-CNN HRNet with SegFormer, or Cascade Mask R-CNN MobileNet with SegFormer increases the average precision (AP) from 65.94 to 67.49 and from 64.80 to 66.49, respectively, and decreases the miss rate (MR) from 34.05 to 32.50 and from 35.19 to 33.50, respectively. Likewise, CSP, MGAN, SSD MobileNet v1, Faster R-CNN show about $1\% \sim 3\%$ higher AP and about $1\% \sim 3\%$ lower MR after fusing with Segformer.

We also conducted the cross-dataset evaluation for our benchmarking. For this purpose, we used our fusion framework whose detectors are trained on the CityPersons dataset, and tested them on Greensboro dataset. This provides an insight into whether the developed fused models are over-fitted on the CityPersons dataset and how well they can adapt to different datasets with new scenes. Table II shows the cross

Fig. 4. **Detection comparisons:** The top row represents the input images. Then, the middle row indicates the corresponding semantic map of input images. Finally, the bottom row shows the ground truth, baseline detection, and detection with the fusion mechanism. Green, blue, and red color bounding boxes represent ground truth, baseline detection, and improvement with fusion, respectively.

dataset evaluation and comparison of the fused models against all eight baseline detectors. As it can be seen in Table II, all baseline detection networks show improvement on the Greensboro dataset after fusing with mask segmentation networks. In particular, mask segmentation networks SegFormer and FCHarDNet performed significantly better in terms of enhancing the baseline detectors. This is due to the fact that these segmentation networks retain semantic maps at higher resolution, increasing the likelihood of important information being captured in bounding box selection, which is especially beneficial for computing consensus in our proposed inference fusion method. At their best, these two fusion methods achieve 13.82 ± 10.06 higher AP and 13.82 ± 10.061 lower miss rate than baseline detectors, respectively. Particularly, we observe a significant improvement (20.82 ± 7.55 lower miss rate) after applying the proposed inference fusion method on the lower performance detection networks, i.e., CSP, SSD Mobilenet FPN, and RetinaNet.

To show the robustness and accuracy of our model in challenging scenarios, we visualize the detection results of our method in Fig. 4 Fig. 4 showcases detection results of the SSD MobileNet FPN and its improved results after fusing with the SegFormer. In Fig. 4, the first and second rows represent the input images and the corresponding pixel-wise colored semantic map from the SegFormer, respectively. The third row in Fig. 4 shows the detection output of SSD MobileNet FPN and the improvement after applying the fusion mechanism. Here in the third row, green bounding boxes represent the ground truth labels while blue bounding boxes represent the detection results of the SSD mobileNet FPN. As it can be seen, there are some situations when the SSD MobileNet FPN cannot detect pedestrians but when fused with the SegFormer, it can predict them more accurately. We highlighted such pedestrians with red bounding boxes in Fig. 4. As we can observe from the first, second, third, and fifth columns, the proposed fusion method can detect multiple pedestrians more accurately than the baseline SSD MobileNet FPN, which fails to detect red boxes. Finally, in the fourth column, the SSD MobileNet FPN missed a partially occluded pedestrian. However, our proposed fusion method is able to detect this pedestrian by combining inferences from the SSD MobileNet FPN and the SegFormer.

TABLE III

AVERAGE PRECISION AND MISS RATE OF DIFFERENT FUSION METHODS ON THE CALTECH DATASET

Fusion Method	Caltech Dataset	
	AP	MR
F-DNN [45]	34.09	65.90
F-DNN2+SS [45]	35.75	64.29
SDS-RCNN [44]	31.90	68.09
Ours (Cascade Mask-RCNN+DDRNet)	**44.46**	**55.53**
Ours (Cascade Mask-RCNN+SegFormer)	**45.02**	**54.97**

D. Comparison With Other Fusion Methods

We compared our fusion method with state-of-the-art fusion methods including F-DNN [45], F-DNN2+SS [45], and SDS-RCNN [44]. An overview of these methods was provided under Section II. These models have been trained and tested over the Caltech dataset [56]. To illustrate performance improvements that can be achieved by our method over existing fusion methods, we also extended the evaluation of our method by testing it on the Caltech test dataset. Caltech pedestrian dataset has different types of instances including partially and heavily occluded pedestrians. Table III exhibits the average precision (AP) and miss-rate (MR) of

TABLE IV
RUNTIME EFFICIENCY OF FUSION METHODS

	Det. time	SegFormer (time = 1.400s)	DDRNet (time = 0.039s)	FCHarDNet (time = 0.115s)	Pix2PixGAN (time = 0.011s)
Cascade Mask R-CNN HRNet	0.450	1.401	**0.451**	**0.451**	**0.451**
Cascade Mask R-CNN MobileNet	0.340	1.401	**0.341**	**0.341**	**0.341**
MGAN	0.170	1.401	**0.171**	**0.171**	**0.171**
CSP	0.220	1.401	**0.221**	**0.221**	**0.221**
SSD MobileNet v1	0.011	1.401	0.040	0.116	**0.012**
SSD MobileNet FPN	0.026	1.401	0.040	0.116	**0.027**
Faster R-CNN Inception v2	0.040	1.401	0.041	0.116	**0.041**
RetinaNet	0.091	1.401	**0.092**	0.116	**0.092**

different fusion methods on the Caltech dataset, including our models. From Table III, it can be seen that our fusion method outperforms other fusion methods in terms of higher AP and lower MR. As we can observe from Table III, F-DNN, F-DNN2+SS, and SDS-RCNN have AP of 34.09, 35.75, and 31.90, respectively, whereas our two fusion models Cascade Mask-RCNN+DDRNet and Cascade Mask-RCNN+Segformer show higher AP of 44.46, and 45.02, respectively. Table III also shows that existing fusion methods, i.e., F-DNN, F-DNN2+SS, and SDS-RCNN have a miss rate of 65.90, 64.29, and 68.09, respectively, whereas our two fusion models Cascade Mask-RCNN+DDRNet and Cascade Mask-RCNN+Segformer show lower miss rate of 55.53 and 54.97, respectively. These results demonstrate that the proposed approach effectively fuses inferences from asymmetric networks and accurately detects pedestrians under a wide variety of challenging scenarios.

E. Runtime Efficiency

Typically, combining asymmetric inference information in a network fusion architecture comes at the expense of a significant loss in speed due to the large input size, complex network structures, and sequential processes [45]. However, instead of sequentially processing the information, our proposed fusion method executes the information from the two branches in parallel. Then, this information is passed through the fusion inference mechanism whose computation time is negligible. Therefore, the overall computation time of the proposed fusion framework is almost the maximum of the computation time of the slower branch. For instance, as we can observe from the Table IV, the Cascade Mask R-CNN with the backbone of HRNet takes around 0.45 second to detect pedestrians and the SegFormer spends 1.40 seconds to generate pedestrian masks, whereas the overall computation time for our fusion method is 1.401 seconds. Interestingly, the combinations of high performing detectors with low computationally demanding pedestrian mask generation networks such as DDRNet, FCHarDNet, and Pix2PixGAN improved detection performance with competitive runtime efficiency. Table IV shows a thorough runtime comparison among all fused models. As it can be seen in Table IV, with our proposed fusion approach, we can reach to 83 fps when fusing Pix2PixGan and SSDMobileNet, whereas the best runtime efficiency reported in [45] is around $7 \sim 8$ fps (note that [45] has used a more advanced GPU). For the sake of readability, we also highlighted fused models with competitive runtime efficiency in Table IV.

V. CONCLUSION

This study was carried out to assess the inference fusion between two different deep neural networks for detecting pedestrians. We proposed a novel, modular, scalable, and maintainable fusion framework that combines asymmetric inferences from object detectors and semantic segmentation networks for jointly detecting multiple pedestrians. Our key idea is to introduce a computationally efficient consensus-based scoring method to fuse pair-wise pixel relevant information from these two networks to boost the final confidence scores accordingly. Further attention was given to improve the runtime efficiency among fusion models. We demonstrated that a real-time and accurate pedestrian detection system can be developed by running an object detection and a semantic segmentation model in parallel. We thoroughly investigated the performance of the proposed fusion framework with several different object detectors and semantic segmentation networks. We also created a new cross pedestrian dataset utilizing an autonomous car platform. This dataset allowed us to evaluate the generalizability of our fused models. All datasets, the developed models under this study, the codes and evaluation tools are made publicly available on the project website. The results exhibited that fused models outperform current state-of-the-art pedestrian detectors in terms of (lower) miss rate and (higher) average precision values. Future work includes the integration of the multi-modal sensor information into our fusion framework to ultimately create a generalizable, robust, efficient pedestrian detection system that takes advantage of both network and sensor fusion. In addition, the developed framework with enhanced pedestrian detection performance paves the way toward tracking pedestrians in a driving scene for more informed decision-making by autonomous cars.

REFERENCES

[1] H. Zhang, Y. Tian, K. Wang, W. Zhang, and F.-Y. Wang, "Mask SSD: An effective single-stage approach to object instance segmentation," *IEEE Trans. Image Process.*, vol. 29, pp. 2078–2093, 2020.

[2] W. Liu *et al.*, "SSD: Single shot MultiBox detector," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 21–37.

[3] 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.

[4] X. Zhao, W. Li, Y. Zhang, T. A. Gulliver, S. Chang, and Z. Feng, "A faster RCNN-based pedestrian detection system," in *Proc. IEEE 84th Veh. Technol. Conf. (VTC-Fall)*, Sep. 2016, pp. 1–5.

[5] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2005, vol. 1, no. 1, pp. 886–893.

[6] N. Dalal, B. Triggs, and C. Schmid, "Human detection using oriented histograms of flow and appearance," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2006, pp. 428–441.

[7] J.-G. Wang, J. Li, W.-Y. Yau, and E. Sung, "Boosting dense SIFT descriptors and shape contexts of face images for gender recognition," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.-Workshops*, Jun. 2010, pp. 96–102.

[8] S. Paisitkriangkrai, C. Shen, and J. Zhang, "Fast pedestrian detection using a cascade of boosted covariance features," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 18, no. 8, pp. 1140–1151, Aug. 2008.

[9] O. Tuzel, F. Porikli, and P. Meer, "Region covariance: A fast descriptor for detection and classification," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2006, pp. 589–600.

[10] G. Gualdi, A. Prati, and R. Cucchiara, "Covariance descriptors on moving regions for human detection in very complex outdoor scenes," in *Proc. 3rd ACM/IEEE Int. Conf. Distrib. Smart Cameras (ICDSC)*, Aug. 2009, pp. 1–8.

[11] M. Oren, C. Papageorgiou, P. Sinha, E. Osuna, and T. Poggio, "Pedestrian detection using wavelet templates," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 97, Jun. 1997, pp. 193–199.

[12] W. Gu, C. Xiang, and H. Lin, "Modified HMAX models for facial expression recognition," in *Proc. IEEE Int. Conf. Control Autom.*, Dec. 2009, pp. 1509–1514.

[13] Y. Mu, S. Yan, Y. Liu, T. Huang, and B. Zhou, "Discriminative local binary patterns for human detection in personal album," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.

[14] S. Zhang, R. Benenson, and B. Schiele, "CityPersons: A diverse dataset for pedestrian detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3213–3221.

[15] S. Paisitkriangkrai, C. Shen, and A. Van Den Hengel, "Strengthening the effectiveness of pedestrian detection with spatially pooled features," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2014, pp. 546–561.

[16] W. Liu, S. Liao, W. Ren, W. Hu, and Y. Yu, "High-level semantic feature detection: A new perspective for pedestrian detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5187–5196.

[17] I. Hasan, S. Liao, J. Li, S. U. Akram, and L. Shao, "Generalizable pedestrian detection: The elephant in the room," 2020, *arXiv:2003.08799*.

[18] X. Wang, T. Xiao, Y. Jiang, S. Shao, J. Sun, and C. Shen, "Repulsion loss: Detecting pedestrians in a crowd," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7774–7783.

[19] X. Ouyang, Y. Cheng, Y. Jiang, C.-L. Li, and P. Zhou, "Pedestrian-synthesis-GAN: Generating pedestrian data in real scene and beyond," 2018, *arXiv:1804.02047*.

[20] W. Liu, S. Liao, W. Hu, X. Liang, and X. Chen, "Learning efficient single-stage pedestrian detectors by asymptotic localization fitting," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 618–634.

[21] M. M. Islam, A. A. R. Newaz, R. Tian, A. Homaifar, and A. Karimoddini, "A computationally effective pedestrian detection using constrained fusion with body parts for autonomous driving," in *Proc. 5th IEEE Int. Conf. Robot. Comput. (IRC)*, Nov. 2021, pp. 106–110.

[22] M. M. Islam, A. A. R. Newaz, B. Gokaraju, and A. Karimoddini, "Pedestrian detection for autonomous cars: Occlusion handling by classifying body parts," in *Proc. IEEE Int. Conf. Syst., Man, Cybern. (SMC)*, Oct. 2020, pp. 1433–1438.

[23] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 580–587.

[24] Q. Zhu, M.-C. Yeh, K.-T. Cheng, and S. Avidan, "Fast human detection using a cascade of histograms of oriented gradients," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2, Jun. 2006, pp. 1491–1498.

[25] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, "OverFeat: Integrated recognition, localization and detection using convolutional networks," 2013, *arXiv:1312.6229*.

[26] W. Lan, J. Dang, Y. Wang, and S. Wang, "Pedestrian detection based on YOLO network model," in *Proc. IEEE Int. Conf. Mechatronics Autom. (ICMA)*, Aug. 2018, pp. 1547–1551.

[27] 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.

[28] 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.

[29] A. Nurhadiyatna and S. Loncaric, "Semantic image segmentation for pedestrian detection," in *Proc. 10th Int. Symp. Image Signal Process. Anal.*, Sep. 2017, pp. 153–158.

[30] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder–decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 801–818.

[31] P. Luc, C. Couprie, S. Chintala, and J. Verbeek, "Semantic segmentation using adversarial networks," 2016, *arXiv:1611.08408*.

[32] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1125–1134.

[33] X. Zhang, X. Zhu, X.-Y. Zhang, N. Zhang, P. Li, and L. Wang, "SegGAN: Semantic segmentation with generative adversarial network," in *Proc. IEEE 4th Int. Conf. Multimedia Big Data (BigMM)*, Sep. 2018, pp. 1–5.

[34] P. Chao, C.-Y. Kao, Y. Ruan, C.-H. Huang, and Y.-L. Lin, "HarDNet: A low memory traffic network," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3552–3561.

[35] A. Kirillov, R. Girshick, K. He, and P. Dollar, "Panoptic feature pyramid networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6399–6408.

[36] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "SegFormer: Simple and efficient design for semantic segmentation with transformers," 2021, *arXiv:2105.15203*.

[37] Y. Hong, H. Pan, W. Sun, and Y. Jia, "Deep dual-resolution networks for real-time and accurate semantic segmentation of road scenes," 2021, *arXiv:2101.06085*.

[38] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2961–2969.

[39] L. Qi *et al.*, "PointINS: Point-based instance segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 10, pp. 6377–6392, Oct. 2022.

[40] D. Bolya, C. Zhou, F. Xiao, and Y. J. Lee, "YOLACT: Real-time instance segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9157–9166.

[41] J. Mao, T. Xiao, Y. Jiang, and Z. Cao, "What can help pedestrian detection?" in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3127–3136.

[42] Y. Pang, J. Xie, M. H. Khan, R. M. Anwer, F. S. Khan, and L. Shao, "Mask-guided attention network for occluded pedestrian detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 4967–4975.

[43] D. Xu, W. Ouyang, E. Ricci, X. Wang, and N. Sebe, "Learning cross-modal deep representations for robust pedestrian detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5363–5371.

[44] G. Brazil, X. Yin, and X. Liu, "Illuminating pedestrians via simultaneous detection and segmentation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4950–4959.

[45] X. Du, M. El-Khamy, J. Lee, and L. Davis, "Fused DNN: A deep neural network fusion approach to fast and robust pedestrian detection," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2017, pp. 953–961.

[46] M. M. Islam, A. A. R. Newaz, and A. Karimoddini, "A pedestrian detection and tracking framework for autonomous cars: Efficient fusion of camera and LiDAR data," in *Proc. IEEE Int. Conf. Syst., Man, Cybern. (SMC)*, Oct. 2021, pp. 1287–1292.

[47] J. Li, X. Liang, S. Shen, T. Xu, J. Feng, and S. Yan, "Scale-aware fast R-CNN for pedestrian detection," *IEEE Trans. Multimedia*, vol. 20, no. 4, pp. 985–996, Apr. 2018.

[48] Y. Lee, T. D. Bui, and J. Shin, "Pedestrian detection based on deep fusion network using feature correlation," in *Proc. Asia–Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA ASC)*, Nov. 2018, pp. 694–699.

[49] 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.

[50] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected CRFs," 2014, *arXiv:1412.7062*.

[51] A. G. Howard *et al.*, "MobileNets: Efficient convolutional neural networks for mobile vision applications," 2017, *arXiv:1704.04861*.

[52] X. Zhang, X. Zhou, M. Lin, and J. Sun, "ShuffleNet: An extremely efficient convolutional neural network for mobile devices," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6848–6856.
[53] T.-Y. Lin *et al.*, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2014, pp. 740–755.
[54] I. Hasan, S. Liao, J. Li, S. U. Akram, and L. Shao, "Generalizable pedestrian detection: The elephant in the room," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 11328–11337.
[55] M. Cordts *et al.*, "The cityscapes dataset for semantic urban scene understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3213–3223.
[56] P. Dollár, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: An evaluation of the state of the art," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 4, pp. 743–761, Apr. 2012.

Muhammad Mobaidul Islam received the B.Sc. degree in electrical and electronic engineering (EEE) from the Bangladesh University of Engineering and Technology (BUET), Bangladesh, in 2009. He is currently pursuing the Ph.D. degree with the Autonomous Cooperative Control of Emergent Systems of Systems (ACCESS) Laboratory, Department of Electrical and Computer Engineering, North Carolina A&T State University, USA. His current research focuses on computer vision applications, vision-based control, and drive-by-wire control for autonomous vehicles. In particular, his research focuses on developing deep learning-based models for pedestrian detection.

Abdullah Al Redwan Newaz received the M.S. and Ph.D. degrees in information science from the Japan Advanced Institute of Science and Technology, Ishikawa, Japan, in 2014 and 2017, respectively. He is an Assistant Professor of computer science at the University of New Orleans, New Orleans, LA, USA. He is also the Director at the Intelligent Robotics Laboratory, UNO. Prior to that, he was a Post-Doctoral Researcher at North Carolina A&T State University, from 2020 to 2022; Rice University, Houston, TX, USA, from 2018 to 2020; and Nagoya University, Nagoya, Japan, from 2017 to 2018. His research interests include environmental monitoring using unmanned systems, planning under uncertainties for safety-critical systems, vision-based perception systems for autonomous cars, integrated perception, and planning for multi-agent systems.

Ali Karimoddini received the Bachelor of Electrical and Electronics Engineering degree from the Amirkabir University of Technology, Iran, in 2003, the Master of Science degree in instrumentation and automation engineering from the Petroleum University of Technology in 2007, and the Ph.D. degree in electrical engineering from the National University of Singapore, Singapore, in 2013. He is currently an Associate Professor at the ECE Department, North Carolina A&T State University, NC, USA. He is also the Director at the ACCESS Laboratory and the NC-CAV Center of Excellence on Advanced Transportation, and the Deputy Director of the TECHLAV DoD Center of Excellence, North Carolina A&T State University. His research interests include control and robotics, autonomy, resilient control systems, smart transportation, connected and autonomous vehicles, human–machine interactions, cyber-physical systems, flight control systems, and multi-agent systems.