FAST SEMANTIC IMAGE SEGMENTATION FOR AUTONOMOUS SYSTEMS

Christos Papaioannidis *Ioannis Mademlis* *Ioannis Pitas*

Department of Informatics, Aristotle University of Thessaloniki, Greece

ABSTRACT

Fast semantic image segmentation is crucial for autonomous systems, as it allows an autonomous system (e.g., self-driving car, drone, etc.) to interpret its environment on-the-fly and decide on necessary actions by exploiting dense semantic maps. The speed of semantic segmentation on embedded computational hardware is as important as its accuracy. Thus, this paper proposes a novel framework for semantic image segmentation that is both fast and accurate. It augments existing real-time semantic image segmentation architectures by an auxiliary, parallel neural branch that is tasked to predict semantic maps in an alternative manner by utilizing Generative Adversarial Networks (GANs). Additional attention-based neural synapses linking the two branches allow information to flow between them during both the training and the inference stage. Extensive experiments on three public datasets for autonomous driving and for aerial-perspective image analysis indicate non-negligible gains in segmentation accuracy, without compromises on inference speed.

Index Terms— Semantic image segmentation, multi-task learning, Convolutional Neural Networks, autonomous systems.

1. INTRODUCTION

Semantic image segmentation consists in predicting dense semantic maps, where a class label is assigned to each pixel of the input image. It is one of the most important tasks for autonomous systems perception, as it allows them to understand their environment using simple RGB cameras and adjust their actions accordingly. Recent technological advances (e.g., self-driving cars, autonomous drones) have highlighted the importance of real-time semantic image segmentation, since laggy/delayed semantic map estimation may potentially have catastrophic results. Moreover, these algorithms are typically executed on embedded computing boards to avoid potential delays due to connectivity issues between the autonomous system and the cloud or a ground station. However, fast execution on embedded computers with limited computational capabilities often compromises semantic map estimation accuracy [1].

Several recent semantic image segmentation methods based on Deep Neural Networks (DNNs) have attempted to bridge the gap between fast execution and estimation accuracy [2, 3, 4, 5]. For example, [2] utilized a lightweight Convolutional Neural Network (CNN) architecture to process the input image in multiple resolutions and predict accurate semantic maps in real-time. In a similar manner, a two-stream CNN with lightweight architecture was proposed in [3] to capture both semantic context and spatial details in the input image. However, certain important input data patterns may be missed by these lightweight network architectures which, if successfully captured, could further increase semantic map estimation accuracy.

To this end, *this paper introduces a novel semantic image segmentation framework* that increases semantic map estimation accuracy without hurting execution speed. An *innovative deep neural architecture is proposed* that augments typical semantic image segmentation network architectures by adding an auxiliary, parallel neural branch. Essentially, this extracts from the input data semantic information that is partially complementary to that captured by the main image segmentation branch and passes it to the latter one, thus enriching its semantic features. To achieve this, the auxiliary neural branch is trained under the Generative Adversarial Network (GAN) [6] framework, so as to recreate RGB images that resemble the ground-truth segmentation maps. In contrast to the typical supervised objective used to train the segmentation branch, GAN training is facilitated by a Discriminator classifier that validates the outputs of the auxiliary neural branch and provides it with indirect supervision signals through the *adversarial loss*. Thus, the auxiliary branch (acting as a Generator in GAN terminology) learns to capture semantic information that may have been missed by the main segmentation branch. Information flow between the auxiliary and the segmentation branches is realized through attention-based neural synapses that are placed between them.

Overall, the novel contributions of the proposed architecture are: i) the auxiliary neural branch, and ii) the attention-based synapses linking it with the main semantic segmentation branch. They are both designed to be lightweight and, therefore, to have negligible effect to the overall execution speed during inference. This was experimentally confirmed on three public datasets: two for autonomous driving and one for aerial-perspective semantic image segmentation.

This work has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 871479 (AERIAL-CORE).

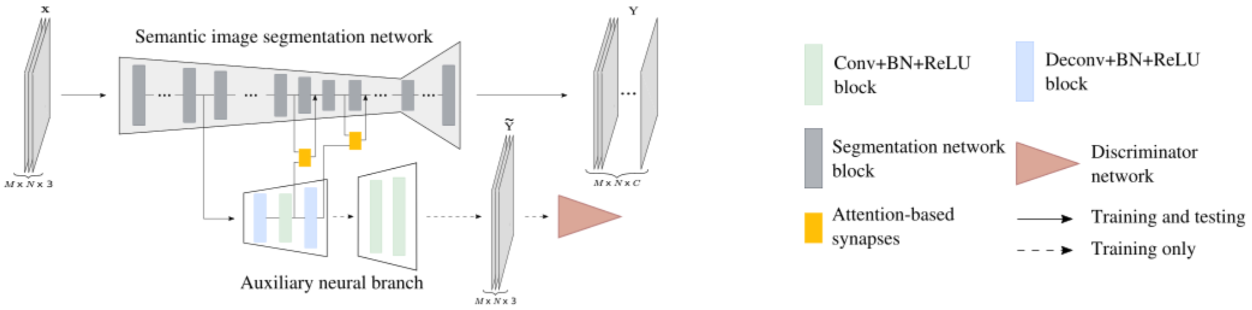

Fig. 1. The proposed deep neural architecture.

2. FAST SEMANTIC IMAGE SEGMENTATION

Semantic image segmentation is a heavily researched topic. Recent methods achieve remarkable performance with the help of complex deep neural architectures [7, 8]. In this work, however, only real-time image segmentation is considered, due to its importance in autonomous systems. Real-time image segmentation methods aim to accomplish both increased segmentation accuracy and fast inference, typically by utilizing lightweight neural architectures [2, 3, 4, 5, 9, 10].

Having this goal in mind, an image cascade network was proposed in [2], which processed the input image in multiple resolutions and effectively fused the encoded multi-resolution information to obtain accurate high-resolution semantic maps. Similarly, [10] employed lightweight variants of general purpose network architectures (e.g., ResNet18 [11]) to process the input image in two different resolutions. Then, by simply upsampling and fusing outputs of intermediate layers, the network was able to predict accurate semantic maps while retaining a high execution speed. Similarly, a lightweight two-branch neural architecture for image segmentation was proposed in [3]. The first branch consisted of a shallow CNN in order to encode the low-level information in the input image, while the second one captured high-level context. The feature maps of the two branches were subsequently combined using a fusing module to produce the final semantic maps. This approach was extended in [5] by introducing more efficient semantic image segmentation network modules for the two-branch network architecture, along with an improved feature aggregation layer for information fusion. Finally, [1] proposed a variation of the two-branch network architecture approaches, which greatly increased execution speed. However, this increase in inference speed came at the expense of segmentation map estimation accuracy.

The proposed framework follows an entirely different approach which is explained in the following Section. Thus, segmentation accuracy is increased without compromises in real-time execution speed.

3. AUGMENTING SEMANTIC IMAGE SEGMENTATION NETWORKS

This paper presents a novel deep neural architecture for fast and accurate semantic image segmentation. It introduces an auxiliary neural branch to augment the main semantic image segmentation branch and provide it with additional semantic information, in order to increase its estimation accuracy. This information is able to flow from the auxiliary branch to the main one, during both the training and the inference stage, through a novel type of neural synapses that are placed between the two branches. Since fast execution is essential, both novel components of the proposed method (i.e., auxiliary neural branch, neural synapses) are carefully designed to maintain fast inference. The overall architecture can be seen in Fig. 1.

Let $\mathbf{X} \in \mathbb{R}^{M\times N\times 3}$ be an RGB input image and $\mathbf{Y} \in \mathbb{R}^{M\times N\times C}$ be the corresponding ground-truth semantic map that must be estimated by the semantic image segmentation branch, where C is the total number of semantic classes. In order to ensure that the auxiliary branch captures information that is both useful and complementary to the one encoded by the main semantic image segmentation branch, it is tasked to reconstruct an RGB representation of the ground-truth semantic map, $\tilde{\mathbf{Y}} \in \mathbb{R}^{M\times N\times 3}$, where each semantic class is encoded with a different RGB color. This is achieved by training the auxiliary neural branch under the conditional GAN [12] framework, where an additional Discriminator network is employed to validate the outputs of the auxiliary branch. That is, the auxiliary branch aims to output "realistic" RGB representations of the ground-truth semantic maps that can not be identified by the Discriminator network, which is adversarially trained to detect the "fake" RGB semantic map representations that are produced by the auxiliary branch. Therefore, the auxiliary branch tries to minimize the conditional GAN objective $\mathcal{L}_{cGAN}$, while the Discriminator aims to maximize it, where:

$$\mathcal{L}_{cGAN} = \mathbb{E}_{(\mathbf{X},\tilde{\mathbf{Y}})}[\log D(\mathbf{X},\tilde{\mathbf{Y}})] + \mathbb{E}_{\mathbf{X}}[\log(1 - D(\mathbf{X}, G(\mathbf{X}))], \quad (1)$$

G, D are the auxiliary neural branch and the Discriminator network, respectively.

Along with the conditional GAN objective, two additional loss functions are utilized, in order to balance the strength of the two subnetworks during the adversarial training process and thus, ensure a smooth training of the auxiliary neural branch [13]. The utilized costs are the similarity loss function $\mathcal{L}_{sim} = \|\tilde{\mathbf{Y}} - G(\mathbf{X})\|_1$, which is used to help the auxiliary branch produce "realistic" outputs, and the multi-label classification loss function $\mathcal{L}_{occ}$ that is used to strengthen the Discriminator network by tasking it to predict semantic class occurrence in $\tilde{\mathbf{Y}}$. Therefore, the overall objective that is utilized to train the auxiliary neural branch is given by:

$$\mathcal{L}_{aux} = \min_{G} \max_{D} \mathcal{L}_{cGAN} + \lambda_1 \mathcal{L}_{sim} + \lambda_2 \mathcal{L}_{occ}, \tag{2}$$

where λ_1, λ_2 are hyperparameters used to scale the contribution of $\mathcal{L}_{sim}$ and $\mathcal{L}_{occ}$ in the total loss.

Information captured by the auxiliary neural branch must be able to flow to the main semantic image segmentation branch during inference, in order to enrich its features and enable it to predict accurate semantic maps. However, any information exchange of this nature requires special treatment, since in the majority of cases it will only partially be truly useful for semantic segmentation; a significant portion may essentially act as noise, thus eventually harming performance. Therefore, the two branches are interlinked in the proposed neural architecture using *attention-based synapses*. These allow the semantic image segmentation branch to actively query the auxiliary branch about the exchanged information. They rely on the dot-product attention mechanism [14], which enables the segmentation branch to select features from the auxiliary branch based on the cross-attention map, calculated between their feature maps. The exchanged information $\mathbf{CA}$ is then computed as follows:

$$\mathbf{CA} = softmax(\mathbf{M}_1 \mathbf{M}_2^T)\mathbf{M}_2, \tag{3}$$

where $\mathbf{M}_1$, $\mathbf{M}_2$ are feature maps of the semantic image segmentation and the auxiliary networks, respectively, after they have been transformed by a simple linear layer.

Our implementation of the proposed framework adopts the real-time network architecture of [3] for the main semantic image segmentation branch. However, *any real-time semantic image segmentation network architecture could be utilized in its place*. The employed semantic image segmentor calculates the final semantic maps by fusing the outputs of two parallel neural pathways, which act on the input image. It is trained in a fully supervised manner using the following objective [3]:

$$\mathcal{L}_{main} = \mathcal{L}_p + \alpha \sum_{i=1}^{2} \mathcal{L}_{a_i}, \tag{4}$$

where $\mathcal{L}_p$ is the principal loss used to supervise the entire segmentation branch and $\mathcal{L}_{a_i}$, $i = 1, 2$, are auxiliary loss terms for deep supervision [15]. Both $\mathcal{L}_p$ and $\mathcal{L}_{a_i}$ are typical Softmax loss functions, while α is used to scale the contribution of the deep supervision loss terms.

All building blocks of the proposed method are combined together to form an efficient network architecture for semantic image segmentation. The auxiliary neural branch consists of three convolution-BatchNorm-ReLu blocks [16] and two deconvolution-BatchNorm-ReLu ones, while the Discriminator network is a PatchGAN [17] classifier extended by a simple classification layer for the semantic class occurrence prediction task. The unified network architecture of the proposed method is trained using a multitask loss function that considers the objectives of both the auxiliary and the main segmentation branch:

$$\mathcal{L} = \beta \mathcal{L}_{main} + (1 - \beta)\mathcal{L}_{aux}, \tag{5}$$

where β is a hyperparameter for adjusting focus between the two tasks. The Discriminator network and the final layers of the auxiliary neural branch are only necessary during training and, therefore, are simply discarded afterwards, in order to avoid additional computational costs during inference. This is depicted in Fig. 1.

4. EXPERIMENTAL EVALUATION

The proposed framework was evaluated for a self-driving car scenario using the Cityscapes [18] and Cambridge-driving Labeled Video Database (CamVid) [19] datasets and for an autonomous Unmanned Aerial Vehicle (UAV)/drone flight scenario using the DroneCrowd [20] dataset. Cityscapes consists of 5000 finely annotated, high-resolution (1024 × 2048 pixels) images of urban street scenes, out of which 2975, 500 and 1525 are for training, validation and testing purposes, respectively. CamVid is a smaller dataset, consisting of 701 images with a resolution of 720 × 960 pixels. The total image set is split into three groups: 367 for training, 101 for validation and 233 for testing. DroneCrowd consists of 1790 images depicting human crowds in a wide range of scenes. These images are annotated with their ground-truth segmentation maps that represent crowd regions and are split in two sets, containing 1199 training and 591 test images. The evaluation metrics in all cases were the class Intersection over Union (IoU) and the inference speed in frames-per-second (FPS).

For all datasets, the training data were augmented online using random scaling in the range of $[0.5, 2]$ as well as random horizontal flipping. The unified network architecture of the proposed framework was trained for both objectives for 120 epochs using the SGD optimizer with a momentum of 0.9, weight decay of 0.0005, batch size equal to 8 and initial learning rate equal to 0.001, which is reduced in each epoch using the "poly" learning rate strategy with the power of 0.9. The Discriminator network was trained using the Adam optimizer [21] with a constant learning rate of 0.0002. β was

Table 1. Evaluation on Cityscapes [18] validation and test sets in terms of mIoU and inference time in FPS. *Backbone* indicates the backbone CNNs pretrained on ImageNet [22].

	Backbone	Input size	mIoU (%)		FPS
			val	*test*	
ESPNetV2 [23]	ESPNetV2	512 × 1024	66.40	66.20	—
ERFNet [24]	—	512 × 1024	70.00	68.00	41.7
Fast-SCNN [1]	—	1024 × 2048	68.60	68.00	**123.5**
DFANet A [4]	Xception A	1024 × 1024	—	71.30	100.0
DABNet [25]	—	1024 × 2048	—	70.10	27.7
GUN [26]	DRN-D-22	512 × 1024	69.60	70.40	33.3
SwiftNet [10]	ResNet-18	1024 × 2048	75.40	75.50	39.9
BiSeNet [3]	ResNet-18	768 × 1536	74.80	74.70	65.5
Proposed	ResNet-18	768 × 1536	**76.03**	**76.14**	52.9

Table 2. Evaluation on the CamVid [19] test set. Input size in all cases is 720×960. *Backbone* indicates whether a specific CNN is used as the backbone network and *Pretrain* denotes the dataset used for pretraining it.

	Backbone	Pretrain	mIoU (%)	FPS
DFANet A [4]	Xception A	ImageNet	64.70	**120.0**
ICNet [2]	PSPNet-50	ImageNet	67.10	27.8
SwiftNet [10]	ResNet-18	ImageNet	72.58	—
BiSeNet [3]	ResNet-18	ImageNet	68.70	89.4
Proposed	ResNet-18	ImageNet	**73.97**	68.2
BiSeNet [3]	ResNet-18	Cityscapes	75.09	89.4
Proposed	ResNet-18	Cityscapes	**76.92**	68.2

empirically set to 0.7, balancing the two objectives in favor of semantic image segmentation, while α was set to 1. λ_1 and λ_2 were set to 10 and 0.1, respectively, in order to ensure smooth training of the auxiliary neural branch.

First, the proposed framework was evaluated on the Cityscapes validation and test sets. Comparison results are depicted in Table 1, along with the backbone CNN, the input image resolution and inference speed in FPS. Note that inference speed measurements in this case are conducted using a single *Nvidia GTX 1080 Ti* GPU. The results show that the proposed framework outperforms all competing methods in terms of mIoU. Specifically, when compared to the best-performing competitors, the proposed framework outperforms both the baseline *BiSeNet* and *SwiftNet* by a margin up to 1.4% and 0.6%, respectively. In the latter case, this is despite the fact that *SwiftNet* uses higher resolution inputs, which also impacts its execution speed, rendering it slower that the proposed framework by 13 FPS. Moreover, the comparison between the proposed method and the baseline *BiSeNet* shows that the additional auxiliary neural branch and neural synapses do not hurt its real-time execution.

Evaluation results on the CamVid test set are reported in Table 2. As it can be seen, the proposed framework improves mIoU by 5.2% and 1.4% when compared to the best performing competing methods *BiSeNet* and *SwiftNet*, respectively. Furthermore, the impact of pretraining the entire model (instead of the backbone feature extraction CNN only) is also investigated, using the Cityscapes dataset during pretraining.

Table 3. Crowd detection performance in IoU of both *crowd* and *n-crowd* classes on the DroneCrowd [20] dataset.

	640 × 360			1280 × 720		
	crowd	*n-crowd*	FPS	*crowd*	*n-crowd*	FPS
FCN$_t$ [27]*	49.5	92.2	10.2	61.8	95.0	4.4
FCN$_p$ [27]†	50.6	92.6	7.9	64.9	95.3	3.2
CSRNet [28]*	78.6	97.9	5.8	79.4	97.9	1.5
BiSeNet [3]	80.6	98.0	**32.5**	83.5	98.1	**28.0**
Proposed	**86.4**	**98.8**	22.3	**86.9**	**99.0**	20.1

* Output thresholding was applied to obtain semantic maps.

† Output thresholding and Gaussian blur was applied to obtain semantic maps.

This further increases the accuracy of the proposed framework, which still outperforms the *BiSeNet* baseline.

Finally, comparison results on DroneCrowd dataset are presented in Table 3. The IoU for both *crowd* and *n-crowd* classes of all models are reported in two different input resolutions, along with inference speed in FPS. Inference speed is measured using a *Nvidia Jetson Xavier* computing board, in order to simulate an embedded execution scenario. When compared to *FCN*$_t$ and *FCN*$_p$, the proposed framework significantly increased *crowd* class IoU in both input resolutions. Comparison against *CSRNet* and *BiSeNet* also shows that the proposed framework increased *crowd* class IoU by 7.8% and 5.8%, respectively. In addition, the proposed framework manages to detect the *n-crowd* class more efficiently. Inference speed results show the the proposed framework is much faster that *FCN*$_t$, *FCN*$_p$ and *CSRNet*. When compared to the *BiSeNet* baseline, the proposed network architecture is slower by 10.2 FPS in the worst-case scenario. However, this is not critical, as inference speed remains near-real-time.

5. CONCLUSIONS

This paper presented a novel semantic image segmentation framework that achieves both increased test accuracy and fast inference. The proposed framework can augment any real-time semantic image segmentation network by an auxiliary, parallel neural branch, whose objective is to recreate RGB representations of the semantic maps. This alternative semantic map prediction approach enables it to capture complementary semantic information and pass it to the main segmentation branch, in order to enhance its features. Information exchange is realized through a set of novel attention-based neural synapses that are added between the two branches and allow the main one to actively select only the useful portion of the information encoded by the auxiliary branch. Experiments on public datasets (concerning autonomous driving and drone-perspective visual analysis) and extensive comparisons against competing methods showed that the proposed framework indeed increases semantic map estimation accuracy, without compromising fast execution. This was confirmed both on a desktop GPU and on an embedded computing board suitable for autonomous systems.

6. REFERENCES

[1] Rudra PK Poudel, Stephan Liwicki, and Roberto Cipolla, "Fast-scnn: fast semantic segmentation network," *arXiv preprint arXiv:1902.04502*, 2019.

[2] Hengshuang Zhao, Xiaojuan Qi, Xiaoyong Shen, Jianping Shi, and Jiaya Jia, "Icnet for real-time semantic segmentation on high-resolution images," in *European Conference on Computer Vision (ECCV)*, 2018.

[3] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang, "Bisenet: Bilateral segmentation network for real-time semantic segmentation," in *European Conference on Computer Vision (ECCV)*, 2018.

[4] Hanchao Li, Pengfei Xiong, Haoqiang Fan, and Jian Sun, "Dfanet: Deep feature aggregation for real-time semantic segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[5] Changqian Yu, Changxin Gao, Jingbo Wang, Gang Yu, Chunhua Shen, and Nong Sang, "Bisenet v2: Bilateral network with guided aggregation for real-time semantic segmentation," *arXiv preprint arXiv:2004.02147*, 2020.

[6] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative adversarial nets," *Advances in Neural Information Processing Systems (NIPS)*, 2014.

[7] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia, "Pyramid scene parsing network," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[8] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, pp. 834–848, 2017.

[9] Adam Paszke, Abhishek Chaurasia, Sangpil Kim, and Eugenio Culurciello, "Enet: A deep neural network architecture for real-time semantic segmentation," *arXiv preprint arXiv:1606.02147*, 2016.

[10] Marin Orsic, Ivan Kreso, Petra Bevandic, and Sinisa Segvic, "In defense of pre-trained imagenet architectures for real-time semantic segmentation of road-driving images," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[12] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014.

[13] Christos Papaioannidis, Vasileios Mygdalis, and Ioannis Pitas, "Domain-translated 3d object pose estimation," *IEEE Transactions on Image Processing*, vol. 29, pp. 9279–9291, 2020.

[14] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," *arXiv preprint arXiv:1706.03762*, 2017.

[15] Liwei Wang, Chen-Yu Lee, Zhuowen Tu, and Svetlana Lazebnik, "Training deeper convolutional networks with deep supervision," *arXiv preprint arXiv:1505.02496*, 2015.

[16] Sergey Ioffe and Christian Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International Conference on Machine Learning (ICML)*, 2015.

[17] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros, "Image-to-image translation with conditional adversarial networks," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[18] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele, "The cityscapes dataset for semantic urban scene understanding," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[19] Gabriel J Brostow, Jamie Shotton, Julien Fauqueur, and Roberto Cipolla, "Segmentation and recognition using structure from motion point clouds," in *European Conference on Computer Vision (ECCV)*, 2008.

[20] C. Papaioannidis, I. Mademlis, and I. Pitas, "Autonomous UAV safety by visual human crowd detection using multi-task deep neural networks," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2021.

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

[22] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems (NIPS)*, 2012.

[23] Sachin Mehta, Mohammad Rastegari, Linda Shapiro, and Hannaneh Hajishirzi, "Espnetv2: A light-weight, power efficient, and general purpose convolutional neural network," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[24] Eduardo Romera, José M Alvarez, Luis M Bergasa, and Roberto Arroyo, "Erfnet: Efficient residual factorized convnet for real-time semantic segmentation," *IEEE Transactions on Intelligent Transportation Systems*, vol. 19, no. 1, pp. 263–272, 2017.

[25] Gen Li, Inyoung Yun, Jonghyun Kim, and Joongkyu Kim, "Dabnet: Depth-wise asymmetric bottleneck for real-time semantic segmentation," *arXiv preprint arXiv:1907.11357*, 2019.

[26] Davide Mazzini, "Guided upsampling network for real-time semantic segmentation," *arXiv preprint arXiv:1807.07466*, 2018.

[27] M. Tzelepi and A. Tefas, "Graph embedded convolutional neural networks in human crowd detection for drone flight safety," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 5, no. 2, pp. 191–204, 2019.

[28] Y. Li, X. Zhang, and D. Chen, "Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.