Hindawi
Wireless Communications and Mobile Computing
Volume 2021, Article ID 5563875, 16 pages
https://doi.org/10.1155/2021/5563875

Research Article

PPANet: Point-Wise Pyramid Attention Network for Semantic Segmentation

Mohammed A. M. Elhassan,[1] **YuXuan Chen**,[1] **Yunyi Chen**,[1] **Chenxi Huang**,[1] **Jane Yang**,[2] **Xingcong Yao**,[1] **Chenhui Yang**,[1] **and Yinuo Cheng**[3]

[1]*School of Informatics, Xiamen University, Xiamen, Fujian Province 361005, China*
[2]*Department of Cognitive Science, University of California, San Diego, USA*
[3]*Beijing Jingwei Hirain Technologies Co., Inc, China*

Correspondence should be addressed to Chenxi Huang; supermonkeyxi@xmu.edu.cn, Chenhui Yang; chyang@xmu.edu.cn, and Yinuo Cheng; yinuo.cheng@hirain.com

Received 7 January 2021; Revised 30 January 2021; Accepted 3 April 2021; Published 30 April 2021

Academic Editor: Khin wee Lai

In recent years, convolutional neural networks (CNNs) have been at the centre of the advances and progress of advanced driver assistance systems and autonomous driving. This paper presents a point-wise pyramid attention network, namely, PPANet, which employs an encoder-decoder approach for semantic segmentation. Specifically, the encoder adopts a novel squeeze nonbottleneck module as a base module to extract feature representations, where squeeze and expansion are utilized to obtain high segmentation accuracy. An upsampling module is designed to work as a decoder; its purpose is to recover the lost pixel-wise representations from the encoding part. The middle part consists of two parts point-wise pyramid attention (PPA) module and an attention-like module connected in parallel. The PPA module is proposed to utilize contextual information effectively. Furthermore, we developed a combined loss function from dice loss and binary cross-entropy to improve accuracy and get faster training convergence in KITTI road segmentation. The paper conducted the training and testing experiments on KITTI road segmentation and Camvid datasets, and the evaluation results show that the proposed method proved its effectiveness in road semantic segmentation.

1. Introduction

Advanced driver assistance systems (ADAS) have gained massive popularity in the past decades, with much attention given by big car companies such as Tesla, Google, and Uber. ADAS, including adaptive cruise control (ACC), lateral guidance assistance, collision avoidance, traffic sign recognition, and lane change assistance, can be considered crucial factors in developing autonomous driving systems [1–3]. Early studies have developed to detect lanes using mathematical models and traditional computer vision algorithms. For instance, many algorithms have been developed to work on supervised and unsupervised approaches [4–7]. The current paradigm of research has shifted towards nontraditional machine learning methods, namely, deep learning. Deep learning methods have notable performance improvement and have been the dominant solution for many academia and industry problems because convolutional neural networks (CNNs) extract robust and representative features. The significant improvement in ADAS and autonomous driving field has been driven by deep learning success, particularly deep convolutional neural networks (CNNs).

Road detection is an essential component of many ADAS and autonomous vehicles. There is much active research focusing on performing road detection [8–19] and wide-ranging algorithms of various representations proposed for this regard. Semantic segmentation has been at the centre of this development. There is a significant amount of research using convolution neural network-based segmentation. As region-based representation [20], encoder-decoder networks [21–26] and several supporting approaches along with these networks have been used, while other supporting techniques

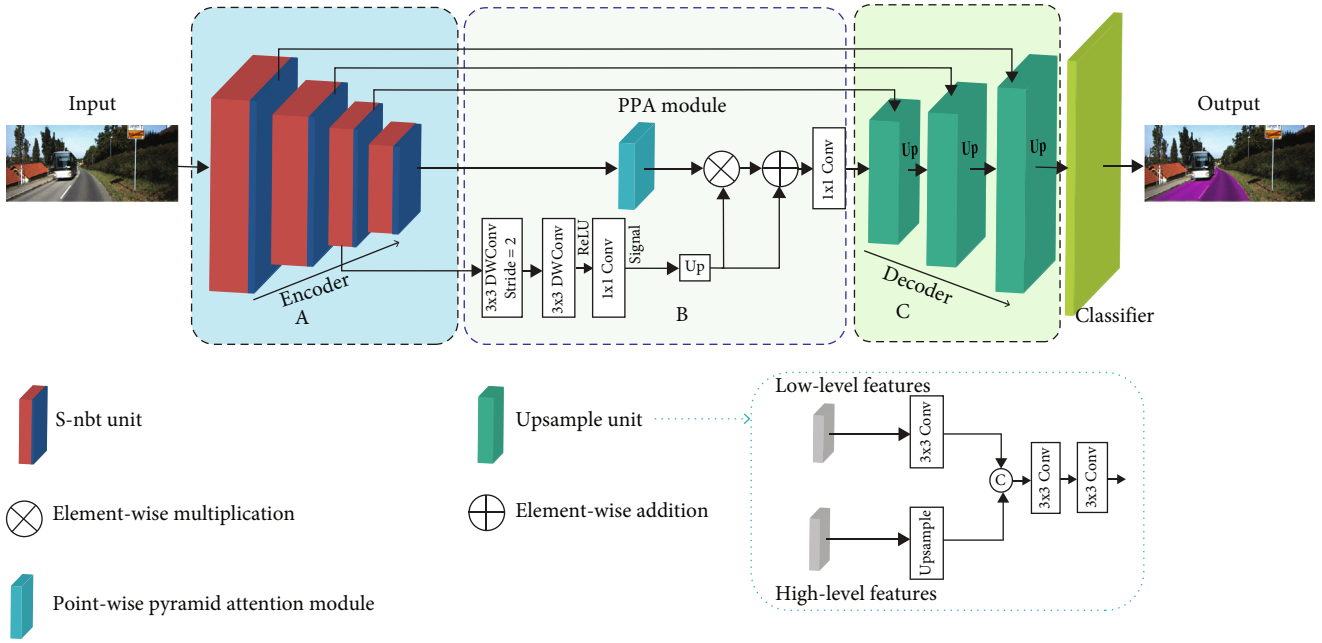


Figure 1: Overall architecture of the proposed PPANet. The encoder adopts squeeze-nbt in an FCN-like network, while PPA and an upsampling unit were employed in the decoder.

fused 3D LiDAR point cloud with 2D images, such as [27, 28]. In this paper, we focus on road segmentation using RGB images. Inspired by the seminal segmentation model U-Net [29], inception [30], SqueezeNet [31], and deep residual learning [32], we propose an architecture that takes the strengths of these well-established models and performs semantic segmentation more effectively. The proposed new architecture is named PPANet (point-wise pyramid attention network), which follows the encoder-decoder approach. In summary, our main contributions could be summarized as follows.

Firstly, we introduce a novel module named point-wise pyramid attention (PPA module) to acquire long-term dependency and multiscale features without much computation burden. Secondly, we design an upsampling module to help to recover the lost details in the encoder. Thirdly, based on the possibility for improvement, we propose a squeeze-nbt module to extract feature representations in the encoder. At last, we combine these modules in an encoder-decoder manner to construct our PPANet for semantic segmentation. The designed model was used to improve the performance of road understanding on KITTI road segmentation and Camvid datasets.

2. Related Works

2.1. Encoder-Decoder Method. In semantic segmentation, the main objective is to assign a categorical label to every pixel in an image, which plays a significant role in road scene understanding. The success and advances in deep convolutional neural network (CNN) models [30, 32–34] have a remarkable impact on pixel-wise semantic segmentation progress due to the rich hierarchical features [29, 35–38]. Usually, to obtain a more delicate result from such a deep network, it is essential to retain high-level semantic information when using low-level details. However, training such a deep neural network requires a large amount of data, but only a limited number of training examples are available in many practical cases. One way to overcome this problem is by employing transfer learning through a network that is pre-trained on a big dataset then fine-tuned on the targeted dataset, as done in [36, 39]. Another solution for such a problem is performing extensive data augmentation, as done in U-Net [29]. In addition to data augmentation, the model architecture should be designed to propagate the information from low levels to the corresponding high levels in a much easier way, such as U-Net.

2.2. Deep Neural Networks. Since the seminal AlexNet [33], model architecture with only eight layers, many studies have been proposed with new approaches for a classification task. Later on, these developed models were applied successfully to a different computer vision task, for example, to segmentation [36], object detection [34], video classification [40, 41], object tracking [42], human pose estimation [43], and super-resolution [44]. These successes spurred the design of a new model with a very large number of layers. However, these growing numbers of layers will need tedious hyperparameter tuning, and that can increase the difficulty of designing such kind of model. In 2014, VGGNet [34] was proposed, in which a significant improvement has been made by utilizing a wider and deeper network; their approach introduced a simple yet effective strategy for designing a very deep network. The quality of a deeper network has a significant impact on improving other computer vision tasks. ResNet [32] has come with an even very deeper model. However, increasing the depth of the network could cause a vanishing gradient problem [32]. Many techniques have been introduced to

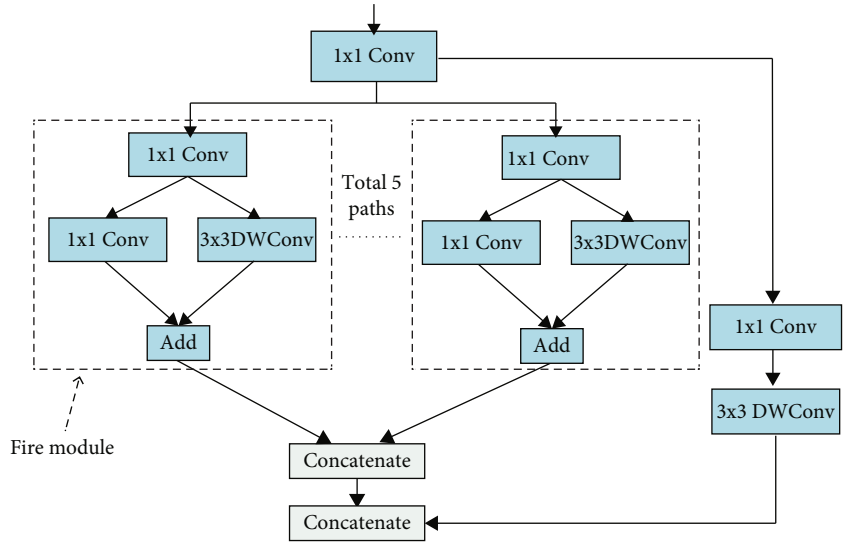


Figure 2: Squeeze-nbt module.

prevent vanishing gradients, for instance, using an initialization method MSR [45] and batch normalization [46].

Meanwhile, skip connection (identity mapping) was used to ease the training process of deep networks without vanishing gradient problems, although VGGNet has a simple architecture, which requires high computation capabilities. On the other hand, inception model families have been designed to perform well with constraint memory and low computation budget. In an Inception module, a split transform-merge strategy where the input feature maps are split into lower dimensions (using 1×1 convolutions) then transformed by a combination of specialized filters (7×7, 5×5, and 3×3) and merged in the end by concatenating branches is adopted.

2.3. Semantic Segmentation with CNN. Recent segmentation-based methods have a significant contribution to solving many computer vision problems, using a wide range of techniques such as a Fully Convolutional Network (FCN) [36], FCN with conditional random field CFD [46], region-based representation [20], encoder-decoder networks [21–23], and multidimensional recurrent networks [47]. Furthermore, pyramid pooling and its variance have a great impact on the recent advances in semantic segmentation [48–53].

2.4. Dilated Convolution-Based Architecture. Dilated convolution or atrous convolutions [53] are a powerful tool in the recent progress of semantic segmentation [52]. It is used to enlarge the receptive field while maintaining the same number of parameters. Recently, many approaches focus on multimodal fusion and contextual information aggregation to improve semantic segmentations [52, 54, 55]. ParseNet [56] applies average pooling to the full image to capture the global contextual information. Spatial pyramid pooling (SPP) [57] has inspired the use of pyramid pooling to aggregate multiscale contextual information such as pyramid pooling [51] module and atrous spatial pyramid pooling module (ASPP) [53, 58]. DenseASPP [59] is proposed to generate dense connections to acquire a larger receptive field. To empower the ASPP module, Xie et al. [60] introduced vortex pooling to utilize contextual information.

Table 1: The PPANet network architecture.

Input	Stage	Type	Stride	Output size
				$160 \times 600 \times 3$
Encoder	Stage 1	Squeeze-nbt unit	2	$80 \times 300 \times 32$
		Squeeze-nbt unit	1	$80 \times 300 \times 32$
	Stage 2	Squeeze-nbt unit	2	$40 \times 150 \times 64$
		Squeeze-nbt unit	1	$40 \times 150 \times 64$
	Stage 3	Squeeze-nbt unit	2	$20 \times 75 \times 128$
		Squeeze-nbt unit	1	$20 \times 75 \times 128$
	Stage 4	Squeeze-nbt unit	2	$10 \times 75 \times 256$
		Squeeze-nbt unit	1	$10 \times 75 \times 256$
Decoder	Centre	PPA module		$20 \times 75 \times 128$
	Dec 1	Upsampling unit	—	$20 \times 75 \times 128$
	Dec 2	Upsampling unit	—	$40 \times 150 \times 64$
	Dec 3	Upsampling unit	—	$80 \times 300 \times 32$
	Final	1×1 Conv	1	$160 \times 600 \times C$

3. Methodology

3.1. Architecture. In this work, we proposed a point-wise pyramid attention network (PPANet) for semantic segmentation, as shown in Figure 1. The network is constructed with an encoder-decoder framework. The encoder is similar to the classification networks; it extracts features and encodes the input data into compact representations. At the same time, the decoder is used to recover the corresponding representations. The squeeze-nbt unit in Figure 2 is used as the main building block for the different stages in the encoder part. Each stage in the encoder has two blocks of the squeeze-nbt unit and the feature map downsampled by half at each first block in each stage using stride convolution

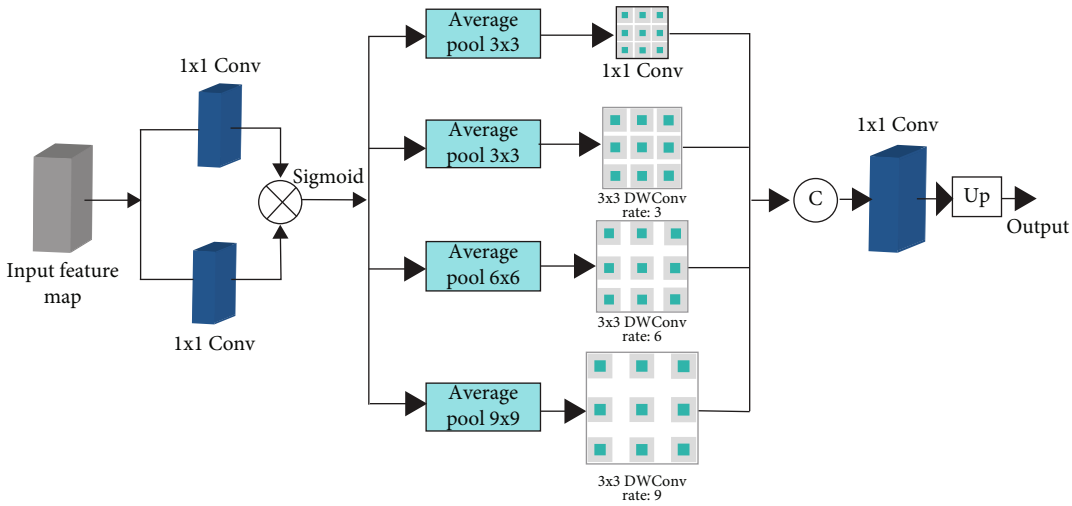


Figure 3: Point-wise pyramid attention module.

(for more details, see Table 1. The network has two other parts: the point-wise pyramid attention module and attention module inserted between the encoder and the decoder. These modules in the centre have been used to enrich the receptive field and provide sufficient context information. More details will be discussed in the following sections.

3.2. Basic Building Unit. This subsection elaborates the squeeze-nbt module architecture (as illustrated in Figure 2). It draws its inspiration from several concepts that have been introduced into recent state-of-the-art models in classification and segmentation, such as the fire module in SqueezeNet [31], depthwise separable convolution [61], and dilated convolution [58]. Figure 2 is the squeeze-nbt module and encoder-decoder framework. We introduce a new module named squeeze-nbt (squeeze nonbottleneck) module. It is based on a reduce-split-squeeze-merge strategy. The squeeze-nbt module first uses point-wise convolution to reduce the feature maps and then apply a parallel fire module to learn useful representations. To make squeeze-nbt computationally efficient, we adopted 3×3 dilated depthwise separable convolution instead of computationally expensive 3×3 convolution.

3.3. Upsampling Module. Several methods such as [21–23, 62], transpose convolution [63], or bilinear upsampling have been utilized broadly to gradually upsample encoded feature maps. In this work, we proposed the upsample module to work as a decoder and to refine the encoded feature maps by aggregating features of different resolutions. First, the low-level feature is processed with 3×3 convolution and in parallel the high-level features upsampled to match the features coming from the encoder; these different features are concatenated and refined with two consecutive 3×3 convolutions.

3.4. Point-Wise Pyramid Attention (PPA) Module. Segmentation requires both sizeable receptive field and rich spatial information. We proposed the point-wise pyramid attention (PPA) module, which is effective for aggregating global contextual information. As shown in Figure 3, the PPA module consists of two parts: the nonlocal part and vortex pooling. On the one hand, the nonlocal module will generate dense pixel-wise weight and extract long-range dependency. On the other hand, vortex atrous convolution is useful in detecting an object at multiple scales. By analysing the vortex pooling and nonlocal dependency, we fuse these two modules' advantages in one module named the point-wise pyramid attention (PPA) module. The PPA module consisted of three parallel vortex atrous convolution blocks with dilation rates of 3, 6, and 9 and one nonatrous convolution block.

The point-wise pyramid attention module is shown in Figure 3. Let X be the input feature map where $X \in R^{H \times W \times C}$ and W, H, and C are width, height, and channels, respectively. First, we apply two parallel convolution layers $F_1 \in R^{H \times W \times C}$ and $F_2 \in R^{H \times W \times C}$ to generate a feature map of $F_a \in R^{H \times W \times C'}$, where $C' = C/4$ indicates the number of channels of F_a:

$$\begin{aligned} F_1 &= \text{conv}1 \times 1(X), \\ F_2 &= \text{conv}1 \times 1(X). \end{aligned} \tag{1}$$

Then, we calculate the similarity matrix $S \in R^{HW \times HW}$ of F_1 and F_1 by a matrix multiplication $F_s = F_1 \times F_2^T$.

Lastly, softmax is applied to normalize the result and transform F_s to self-attention-like mechanism:

$$\text{Output} = \text{Softmax}(F_s). \tag{2}$$

4. Experimental Results and Analysis

In this section, comprehensive experiments on the KITTI road segmentation dataset [64] and Camvid dataset [65] are carried out. We evaluate the efficiency and effectiveness of

(a) Without mathematical morphology

(b) With mathematical morphology

Figure 4: Comparison between PPANet (a) without mathematical morphology and (b) with mathematical morphology. In the first picture, some pixels belong to nonroad classified as a road.

Table 2: Ablation study results on the KITTI dataset.

Method	AP (%)	Precision (%)	Recall	max F
PPANet-baseline	96.20	97.1	95.6	96.3
PPANet, $r = 3, 6, 13$	91.95	86.44	94.01	89.98
PPANet, $r = 2, 4, 8$	78.90	90.37	88.35	89.36
PPANet, $r = 2$	92.15	92.68	91.41	92.06
Decoder				
PPANet+upsampling unit	91.50	91.89	94.33	93.01
PPANet+PPA	93.54	95.77	95.48	95.16

Table 3: Comparison of our model with other networks on the KITTI road segmentation dataset, using average precision (AP), precision, recall, and F-measure.

Method	AP (%)	Pre (%)	Rec (%)	max F
SegNet [23]	89.40	90.90	89.50	90.10
ENet [21]	87.40	88.90	87.60	88.20
FastFCN [68]	95.10	96.10	94.70	95.30
LBN-AA [66]	94.70	96.00	94.10	95.00
DABNet [67]	93.40	94.90	93.50	94.10
AGLNet [69]	94.60	96.00	95.60	96.30
Ours (PPANet)	96.20	97.10	95.60	96.30

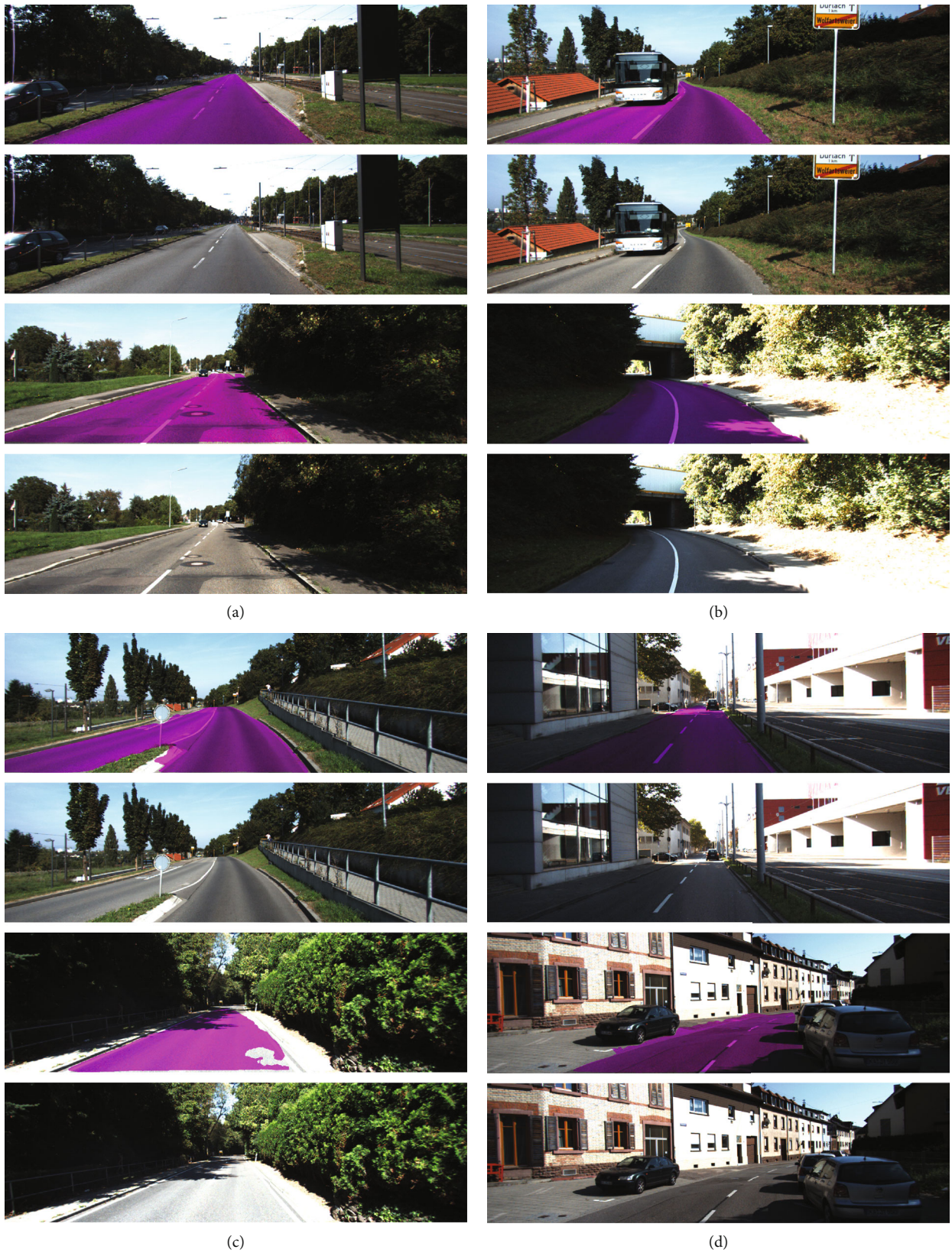
our proposed architecture. Firstly, an introduction to the datasets and the implementation protocols is given. We then elaborate on the loss function and the evaluation metrics used to train KITTI, followed by ablation studies and experiments with the SOTA models. Finally, we report a comparison on the Camvid dataset.

4.1. Datasets and Implementation Details

4.1.1. Datasets

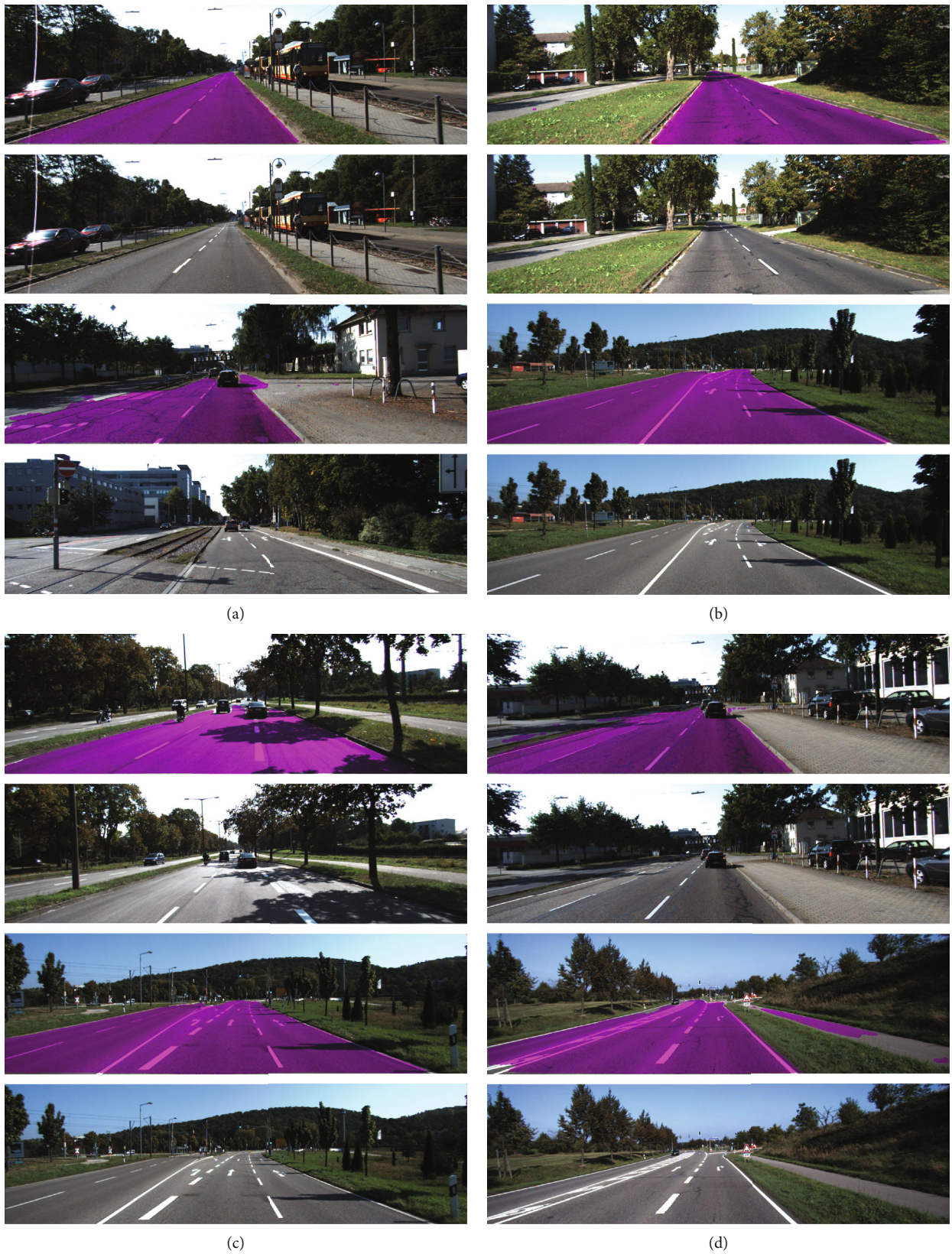
(1) KITTI Road Segmentation Dataset. It consists of 289 training images with their corresponding ground truth. The data in this benchmark is divided into 3 categories: urban marked (UM) with 95 frames, urban multiple marked lane (UMM) with 96 frames, and urban unmarked (UU). The dataset has a small number of frames and difficult lightning conditions, which make it very challenging. In total, it has 290 frames for testing (testing frames have no ground truth information). Training and testing frames were extracted from the KITTI dataset [64] at a minimum spatial distance of 20 m. Each image has a resolution of 375×1242. We split the dataset into three subsets: (a) training samples with 173 images, (b) validation samples with 58 images, and (c) testing samples with 58 images.

(2) Camvid Dataset. The Camvid dataset is an urban street scene understanding dataset in autonomous driving. It consists of 701 samples: 376 training samples, 101 validation samples, and 233 test samples, with 11 semantic categories such as building, road, sky, and bicycle, while class 12 contains unlabelled data that we ignore during training. The original image resolution for the Camvid dataset is 960×720. It has been downsampled into 360x before training for a fair comparison. A weighted categorical cross-entropy loss



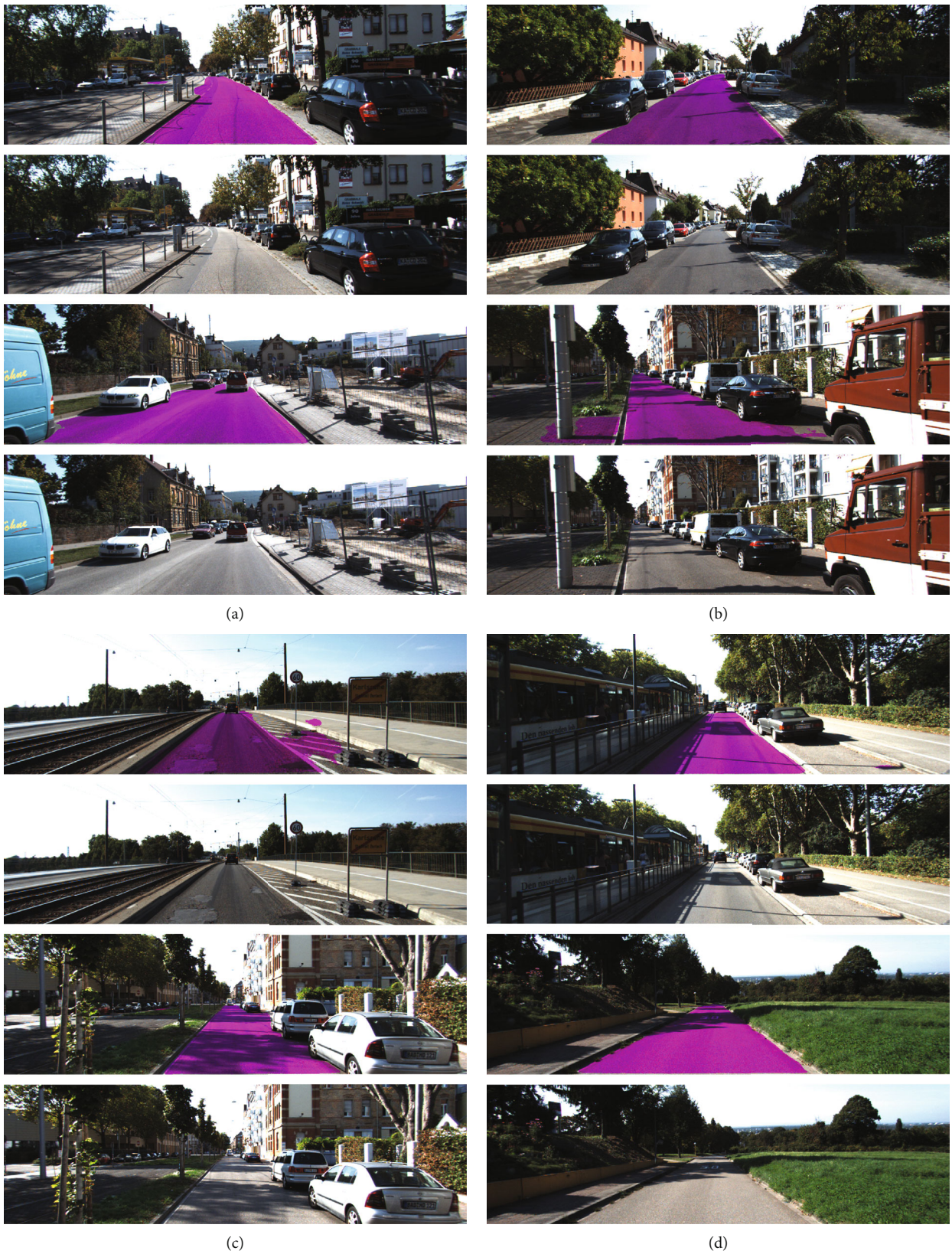
(a) (b) (c) (d)

Figure 5: Examples of road detection images for the UM test set obtained from the public benchmark suite in perspective view. (a, c) Show the segmentation results; (b, d) show the original images.



(a) (b) (c) (d)

Figure 6: Examples of road detection images for the UMM test set obtained from the public benchmark suite in perspective view. (a, c) Show the segmentation results; (b, d) show the original images.



(a) (b) (c) (d)

Figure 7: Examples of road detection images for the UU test set obtained from the public benchmark suite in perspective view. (a, c) Show the segmentation results; (b, d) show the original images.

Table 4: Results of the model on the Camvid dataset.

Method	Year	Params (M)	mIoU (%)
Deeplab-LFOV [58]	2017	262.1	61.6
PSPNet [51]	2017	—	69.1
DenseDecoder [73]	2018	—	70.9
SegNet [23]	2015	29.5	55.6
ENet [21]	2016	0.36	61.3
BiSeNet2 [70]	2018	5.8	68.7
CGNet [71]	2018	0.50	64.0
NDNet45-FCN8-LF [72]	2020	1.1	57.5
LBN-AA [66]	2020	6.2	68.0
DABNet [67]	2019	0.76	66.4
AGLNet [69]	2020	1.12	69.4
PPANet (ours)	—	3.01	70.10

was used to compensate for a small number of categories in the dataset.

4.1.2. Implementation Details. All experiments were implemented with one GTX1080Ti CUDA 10.2 and cuDNN 8.0 on Pytorch [58] deep learning framework. The Adam [59] optimizer is a stochastic-based optimizer used with an initial learning rate of $4e-4$ to train KITTI road segmentation and Camvid datasets. The learning rate is adjusted according to Equation (3), where α is the initial learning rate, F is a factor used to control the learning rate drop, D is the number of epochs to decrease the learning rate value, and i is the current epoch. In PPANet implementation, the learning rate is reduced by a factor of every 15 epochs. The proposed network is limited to run for a maximum of 300 epochs. Normal weight initialization [45] is used to initialize the model. Finally, $7e-3$ of l_2 regularization to deal with the model overfitting

$$\alpha_{i+1} = \alpha_1 \cdot F^{(1+i)/D}. \tag{3}$$

(1) Loss Function. There is a wide range of loss functions proposed over the years to perform semantic segmentation tasks. For instance, binary cross-entropy (BCE) has been applied to many research in classification and segmentation with remarkable success. Although it is convenient to train neural networks using BCE, it might not perform well in class unbalance. For instance, it does not perform well when it is used as the only loss function on KITTI road segmentation with PPANet. In this work, our total loss Equation (6) is a combination of dice loss Equation (5), which was proposed in Zhou et al. [62], and binary cross-entropy Equation (4). Let $p \in [0\ 1]$ be the prediction given by a sigmoid nonlinearity and let $p^{\wedge} \in [0\ 1]$ be the corresponding ground truth. Dice loss has been implemented in a different form in literature; for instance, in [62, 63], it has equivalent definitions, differing in the denominator value. Our experiment found that using the dice loss function that uses the summation of squared values of probabilities and ground truth in the denominator performs better. These functions are defined as follows.

The binary cross-entropy:

$$\mathrm{BCE}(p, \widehat{p}) = \sum_i p \log \widehat{p} + (1-p) \log (1-\widehat{p}). \tag{4}$$

Dice loss:

$$D_i(p, \widehat{p}) = \frac{\sum_i p_i \widehat{p_i}}{\sum_i \left(p_i^2 + \widehat{p}_i^2\right)}. \tag{5}$$

Total loss:

$$\mathrm{loss}_{\mathrm{total}} = \mathrm{BCE}(p, \widehat{p}) + D_i(p, \widehat{p}). \tag{6}$$

4.1.3. Evaluation Metrics on KITTI. Precision and recall evaluation metrics can be considered one of the most common metrics for evaluating a binary classification; following the methods used in [64, 66, 67], we evaluated our segmentation model using precision Equation (7), recall Equation (8), and F-measure Equation (11). The evaluation metrics are listed in the following equations:

$$\mathrm{PRE\ (precision)} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{TF}}, \tag{7}$$

$$\mathrm{REC\ (recall)} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \tag{8}$$

$$\mathrm{PFR} = \frac{\mathrm{FP}}{\mathrm{TP} + \mathrm{FP}}, \tag{9}$$

$$\mathrm{FNR} = \frac{\mathrm{FN}}{\mathrm{TP} + \mathrm{FN}}, \tag{10}$$

$$F\text{-measure} = \frac{2 \times \mathrm{PRE} \times \mathrm{REC}}{\mathrm{PRE} + \mathrm{REC}}, \tag{11}$$

$$\mathrm{Accuracy} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}. \tag{12}$$

4.1.4. KITTI Data Augmentation. Data augmentation comprises a wide range of techniques used to extend the training samples by applying random perturbations and jitters to the original data. In our model, an online data augmentation approach helps the model learn more robust features and increase the generalizability by preventing the model from seeing the same image twice, as slightly random modification to the input data is performed each time. Therefore, we perform a series of data transformation to deal with typical changes in road images, such as texture and colour changes and illumination. In particular, we implemented normalization, blurring, and changing the illumination. Data augmentation can lend itself naturally in the context of computer vision. For example, we can acquire additional training data from the original KITTI road segmentation images by applying the following transforms:

(1) Transformation that applied to both image and the ground truth

Table 5: Per-class results on the Camvid test set in terms of class mIoU scores.

Methods	Bui	Tree	Sky	Car	Sig	Roa	Ped	Fen	Pol	Side	Bic	mIoU (%)
SegNet [23]	88.8	87.3	92.4	82.1	20.5	97.2	57.1	49.3	27.5	84.4	30.7	55.6
ENet [21]	74.7	77.8	95.1	82.4	51.0	95.1	67.2	51.7	35.4	86.7	34.1	51.3
BiSeNet2 [70]	83.0	75.8	92.0	83.7	46.5	94.6	58.8	53.6	31.9	81.4	54.0	68.7
CGNet [71]	79.8	73.2	90.8	81.3	41.6	95.3	52.9	32.9	28.1	81.9	53.9	64.7
NDNet45-FCN8-LF [72]	85.5	84.6	94.8	82.6	39.2	97.4	60.1	37.3	17.6	86.8	53.7	57.5
LBN-AA [66]	83.2	70.5	92.5	81.7	51.6	93.0	55.6	53.2	36.3	82.1	47.9	68.0
DABNet	81.0	74.1	91.1	81.7	43.0	93.8	56.2	37.2	29.4	78.7	56.5	65.7
AGLNet [69]	82.6	76.1	91.8	87.0	45.3	95.4	61.5	39.5	39.0	83.1	62.7	69.4
PPANet (ours)	84.31	77.84	92.06	86.11	51.17	94.85	62.77	41.86	36.92	82.19	61.12	70.1

(i) Geometric transformations are used to alter the position of the point, such as translation, scaling, and rotations

(ii) Mirroring (horizontal flip)

(2) Transformations that applied to the image only since they affect only pixel values

(i) Normalize the input image by standardizing each pixel to be in $[-1, 1]$ range using Equation (13)

(ii) Random brightness adjustment

(iii) Gaussian blur

(iv) Random noise:

$$\text{input} \times \frac{2}{255} - 1.0 \tag{13}$$

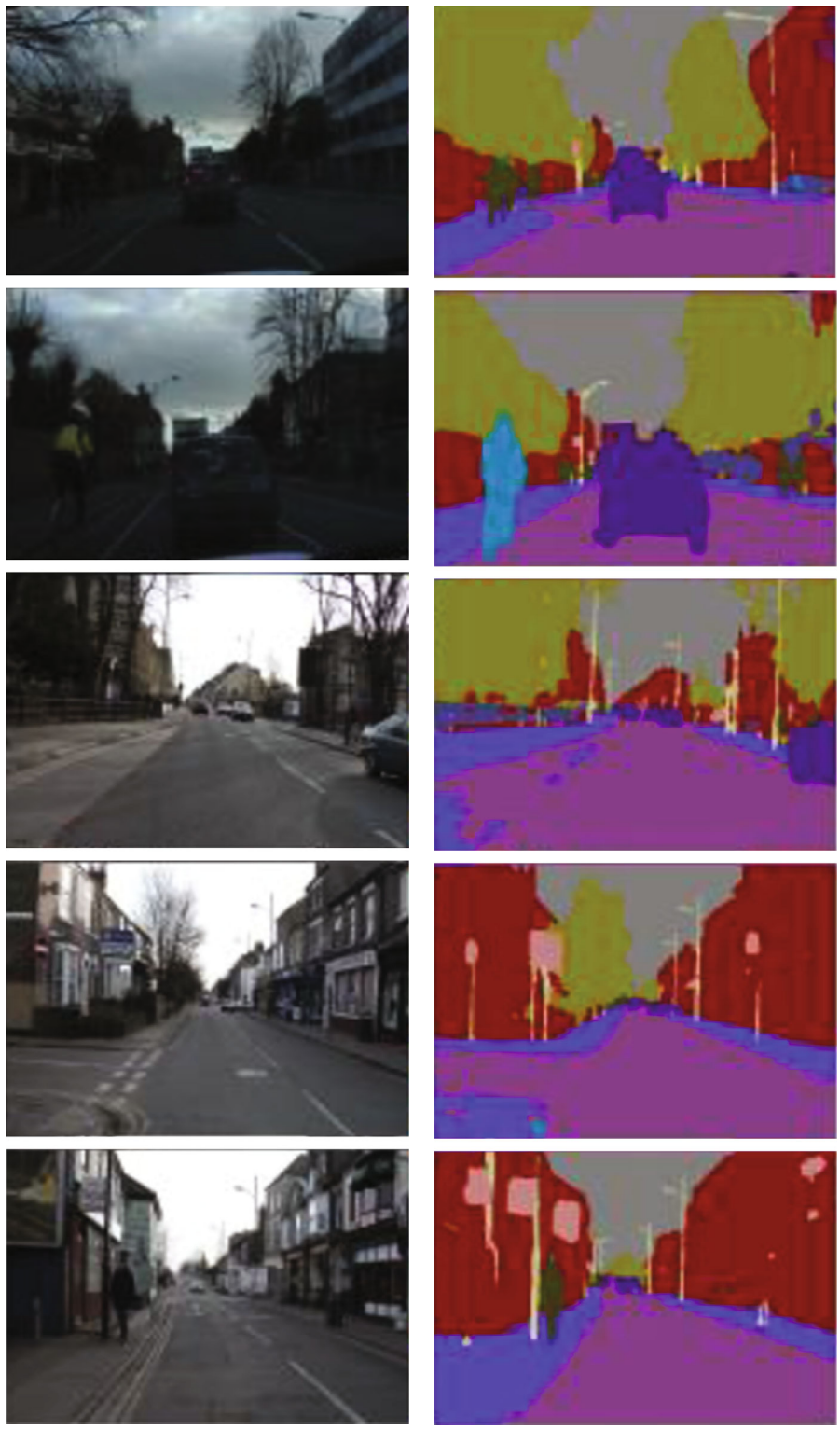
(1) Mathematical Morphology. Applying deep learning methods to the segmentation of a road sometimes results in some noise. Nonroad could be classified as a road and vice versa. Several mathematical morphology techniques can be used to remove the noise and improve the performance of the model in the testing time. An opening mathematical morphology process with square structuring elements of 15 $\times$ 15 sizes was used. It helped the network eliminate some of the nonroad classified as a road (false positive), as illustrated in Figure 4, where (a) represents the performance of the model without augmentation, and we can see the noise by the side of the road, and (b) shows the effect of removing this false positive when training the model.

4.2. Ablation Study

4.2.1. Encoder. We carried out some ablation studies to highlight the effectiveness of our proposed model structure. The proposed method baseline achieves 96.3% max F and 96.2% AP; then, we run experiments with different dilation rate settings. First, we gradually increased the dilation rates 3, 9, and 13 in the encoder at stages 2, 3, and 4, respectively, which result in a decrease of 6.32% F-measure and 4.25% average precision. To further examine and verify the effectiveness of our method with a range of dilation rates, we employed another combination dilation rates, 2, 4, and 8, which yield the lowest result in the ablation experiments with 89.36% F-measure and 78.9%. This sequence of dilation rates has given lower results. It seems that the combination of dilation rates is not effective for the PPANet encoder. Finally, we tested our model using a dilation rate of 2 in the three stages of the encoder and yield the best outcome for our model, as shown in Table 2. Therefore, we set the dilation rate of 2 for all three stages in the encoder.

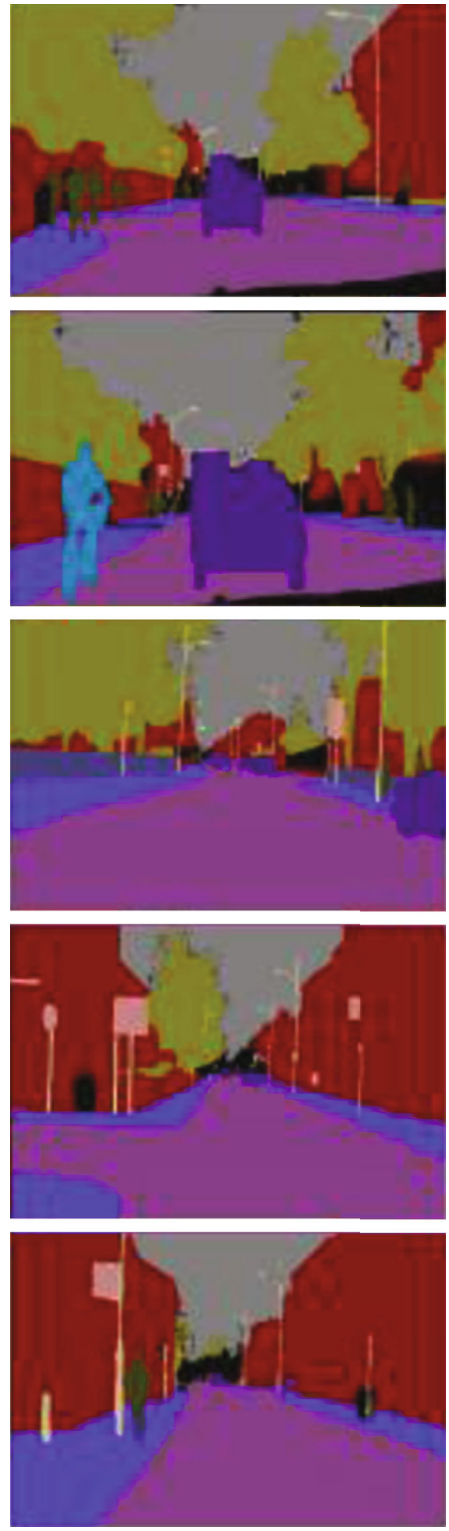
4.2.2. Decoder. We test two settings in the decoding part. First, using the upsampling unit comprises bilinear upsampling and convolution to restore the high-resolution feature from low-resolution features; this approach achieved good results. However, still, there is some information lost during downsampling the feature map in the encoding process. To maintain the highest possible global context feature, we designed a point-wise pyramid attention that is used to increase the model prediction performance with both PPA and upsampling unit; the decoder can aggregate information through a fusion of a multiscale feature. Therefore, it effectively captures local and global context features (see Table 2 for further comparison). It can be seen that the proposed upsample unit and the point-wise pyramid attention (PPA) improve the segmentation-based PPANet and helped achieve superior AP, precision, recall, and max F score compared to other models.

4.3. Comparing with the SOTA. In this subsection, we will present the overall qualitative and quantitative assessment of the trained model. The training and evaluation were conducted using the KITTI road segmentation and Camvid datasets. We then compare the model with a selected SOTA model. Table 3 shows the comparison of PPANet with other SOTA models on the KITTI road segmentation dataset. PPANet is designed for road scene understanding, and it is being trained end-to-end. As previously stated, the dataset has a limited amount of data divided into three categories: urban unmarked (UM road), urban multiple marked lanes (UMM road), and urban unmarked (UU road) as one category to help alongside data augmentation to overcome model overfitting. To rank the best performance among the chosen



(a) (b)

Figure 8: Continued.



(c)

Figure 8: Visual results of our method PPANet on the Camvid test set: (a) is the image, (b) is the prediction, and (c) is ground truth.

models for comparison and evaluation, we reported precision (PRE), recall (REC), and max F metrics, which are known metrics used to evaluate different approaches in binary semantic segmentation. We have chosen some state-of-the-art models to perform a comparison with our proposed PPA-Net model. These models include SegNet [23], ENet [21], FastFCN [68], LBN-AA [66], DABNet [67], and AGLNet [69]. The overall results of PPANet and other SOTA models

are illustrated in Figure 3. Our PPANet obtained the highest scores for all metrics, demonstrating the effectiveness of the proposed method for robust road detection; FastFCN ranked second in terms of precision and third for max F, while ALG-Net has better f-measure. ENet achieved the lowest results compared to all other models. It is designed for speed purposes.

For qualitative performance evaluation of our model in road segmentation, a visual representation of PPANet predictions in the KITTI dataset test set is presented in Figures 5–7 in perspective view for UM, UMM, and UU, respectively. We can see that the urban marked (Figure 5) road got the best prediction with almost no misclassification. For urban unmarked, there is little noisy prediction that can be improved using some postprocess optimization techniques such CRF or increasing the amount of data. When we move to the urban multiple unmarked, it has a higher misclassified road area; it has some area outside the road predicted as road. These false positive detections mainly occur in pole railway when it is close to the road, and also, the road detection is affected by shadow. So, our model with only 3.01 M parameters and without pretrained weights got quite excellent results in a small dataset such as KITTI road segmentation.

4.4. Comparison with SOTA Models on Camvid. In this subsection, we design an experiment to demonstrate our proposed network effectiveness and validity on the Camvid dataset. We train and evaluate the model in the training and validation images and validation set for 400 epochs. Then, the model was tested using the testing images and the results reported in Table 4 in terms of mean intersection over union (mIoU). From Table 4, we can see that the proposed PPANet method has superior performance in terms of mIoU. First, the model was compared with models that have been designed for real-time semantic segmentation such as ENet [21], BiSeNetv1 [70], CGNet [71], NDNet45-FCN8-LF [72], LBN-AA [66], DABNet [67], and AGLNet [69]. And also, we compared our proposed method with a non-real-time model such as DeebLabv2 [58], PSPNet [51], DenseDecoder [73], and SegNet [23]. Besides, we present the individual category results in the Camvid test set in Table 5. As can be seen, the proposed method obtained better accuracy in most of the classes. We also provide visual results in Figure 8.

5. Conclusion

This paper has presented an approach to scene understanding in monocular images. A novel encoder-decoder network for effective semantic segmentation is proposed, named PPANet. The encoder adopts split and squeeze operations in the residual layer to enhance information propagation and feature reuse. To effectively refine the encoded feature map, we design a decoder consisting of the upsampling unit and point-wise pyramid attention (PPA) module. The PPA module is inserted in the centre to enrich the receptive field and to aggregate global contextual information. The attention mechanism is utilized to refine the prediction using a sequence of depthwise convolution followed by sigmoid. This interaction between different features from the upsampling unit, PPA, and attention provides guidance for high-level and low-level features to improve the performance. The network is trained in an end-to-end manner on two popular datasets: KITTI road segmentation and Camvid. The experimental results showed that the proposed method improves the state of the art for road segmentation on small datasets such as the KITTI dataset and Camvid. Future works will include using pretrained weight as that has been the paradigm for most SOTA in this field. Also, we will investigate the potential of incorporating other sensors such as LiDAR into the architecture and test the effectiveness of our approach in dealing with data fusion and 3D road segmentation.

Data Availability

We have used the Camvid dataset and KITTI road segmentation dataset.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

The study was funded by the Fujian province Innovation Strategy Research Program (No. 2020R01020196) and Yongtai Artificial Intelligence Institute.

References

[1] H. Liu, S. E. Shladover, X.-Y. Lu, and X. Kan, "Freeway vehicle fuel efficiency improvement via cooperative adaptive cruise control," *Journal of Intelligent Transportation Systems*, pp. 1–13, 2020.

[2] I. Mahdinia, R. Arvin, A. J. Khattak, and A. Ghiasi, "Safety, energy, and emissions impacts of adaptive cruise control and cooperative adaptive cruise control," *Transportation Research Record*, vol. 2674, no. 6, pp. 253–267, 2020.

[3] Y. Jiang, "Modeling and simulation of adaptive cruise control system," 2020, https://arxiv.org/abs/2008.02103.

[4] E. Kurbatova, "Road detection based on color and geometry characteristics," in *2020 International Conference on Information Technology and Nanotechnology (ITNT)*, pp. 1–5, Samara, Russia, 2020.

[5] Y. Zhang, L. Wang, H. Wu, X. Geng, D. Yao, and J. Dong, "A clustering method based on fast exemplar finding and its application on brain magnetic resonance images segmentation," *Journal of Medical Imaging and Health Informatics*, vol. 6, no. 5, pp. 1337–1344, 2016.

[6] Y. Zhang, F.-l. Chung, and S. Wang, "Clustering by transmission learning from data density to label manifold with statistical diffusion," *Knowledge-Based Systems*, vol. 193, article 105330, 2020.

[7] Y. Zhang, F. Tian, H. Wu et al., "Brain MRI tissue classification based fuzzy clustering with competitive learning," *Journal of Medical Imaging and Health Informatics*, vol. 7, no. 7, pp. 1654–1659, 2017.

[8] B. Wang, V. Frémont, and S. A. Rodríguez, "Color-based road detection and its evaluation on the KITTI road benchmark," in

2014 IEEE Intelligent Vehicles Symposium Proceedings, pp. 31–36, Dearborn, MI, USA, 2014.

[9] L. Geng, J. Sun, Z. Xiao, F. Zhang, and J. Wu, "Combining CNN and MRF for road detection," *Computers & Electrical Engineering*, vol. 70, pp. 895–903, 2018.

[10] M. Passani, J. J. Yebes, and L. M. Bergasa, "CRF-based semantic labeling in miniaturized road scenes," in *17th International IEEE Conference on Intelligent Transportation Systems (ITSC)*, pp. 1902-1903, Qingdao, China, 2014.

[11] H. Liu, X. Han, X. Li, Y. Yao, P. Huang, and Z. Tang, "Deep representation learning for road detection through Siamese network," 2019, https://arxiv.org/abs/1905.13394.

[12] G. L. Oliveira, W. Burgard, and T. Brox, "Efficient deep models for monocular road segmentation," in *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 4885–4891, Daejeon, South Korea, 2016.

[13] F. Ren, X. He, Z. Wei et al., "Fusing appearance and prior cues for road detection," *Applied Sciences*, vol. 9, no. 5, p. 996, 2019.

[14] S. Gu, Y. Zhang, X. Yuan, J. Yang, T. Wu, and H. Kong, "Histograms of the normalized inverse depth and line scanning for urban road detection," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 8, pp. 3070–3080, 2018.

[15] L. Xiao, R. Wang, B. Dai, Y. Fang, D. Liu, and T. Wu, "Hybrid conditional random field based camera-LIDAR fusion for road detection," *Information Sciences*, vol. 432, pp. 543–558, 2018.

[16] L. Xiao, B. Dai, D. Liu, D. Zhao, and T. Wu, "Monocular road detection using structured random forest," *International Journal of Advanced Robotic Systems*, vol. 13, no. 3, p. 101, 2016.

[17] T. Rateke, K. A. Justen, V. F. Chiarella, A. C. Sobieranski, E. Comunello, and A. V. Wangenheim, "Passive vision region-based road detection," *ACM Computing Surveys*, vol. 52, no. 2, pp. 1–34, 2019.

[18] K. Goro and K. Onoguchi, "Road boundary detection using in-vehicle monocular camera," in *Proceedings of the 7th International Conference on Pattern Recognition Applications and Methods*, pp. 379–387, Funchal, Madeira, Portugal, 2018.

[19] Y. Lyu, L. Bai, and X. Huang, "Road segmentation using CNN and distributed LSTM," in *2019 IEEE International Symposium on Circuits and Systems (ISCAS)*, pp. 1–5, Sapporo, Japan, 2019.

[20] H. Caesar, J. Uijlings, and V. Ferrari, "Region-based semantic segmentation with end-to-end training," in *in European Conference on Computer Vision*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds., vol. 9905 of Lecture Notes in Computer Science, , pp. 381–397, Springer, 2016.

[21] A. Paszke, A. Chaurasia, S. Kim, and E. Culurciello, "Enet: a deep neural network architecture for real-time semantic segmentation," 2016, https://arxiv.org/abs/1606.02147.

[22] E. Romera, J. M. Alvarez, L. M. Bergasa, and R. Arroyo, "ERFNet: efficient residual factorized ConvNet for real-time semantic segmentation," *IEEE Transactions on Intelligent Transportation Systems*, vol. 19, pp. 263–272, 2018.

[23] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: a deep convolutional encoder-decoder architecture for image segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

[24] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "Unet ++: a nested U-net architecture for medical image segmentation," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pp. 3–11, Springer, 2018.

[25] J. Wang, H. Xiong, H. Wang, and X. Nian, "ADSCNet: asymmetric depthwise separable convolution for semantic segmentation in real-time," *Applied Intelligence*, vol. 50, no. 4, pp. 1045–1056, 2020.

[26] C. Hazirbas, L. Ma, C. Domokos, and D. Cremers, "Fusenet: incorporating depth into semantic segmentation via fusion-based CNN architecture," in *Asian conference on computer vision*, S. H. Lai, V. Lepetit, K. Nishino, and Y. Sato, Eds., vol. 10111 of Lecture Notes in Computer Science, , pp. 213–228, Springer, 2016.

[27] L. Caltagirone, M. Bellone, L. Svensson, and M. Wahde, "LIDAR–camera fusion for road detection using fully convolutional neural networks," *Robotics and Autonomous Systems*, vol. 111, pp. 125–131, 2019.

[28] H. Liu, Y. Yao, Z. Sun, X. Li, K. Jia, and Z. Tang, "Road segmentation with image-LiDAR data fusion in deep neural network," *Multimedia Tools and Applications*, vol. 79, no. 47, pp. 35503–35518, 2020.

[29] O. Ronneberger, P. Fischer, and T. Brox, "U-net: convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. MICCAI 2015*, N. Navab, J. Hornegger, W. Wells, and A. Frangi, Eds., vol. 9351 of Lecture Notes in Computer Science, pp. 234–241, Springer, Cham, 2015.

[30] C. Szegedy, W. Liu, Y. Jia et al., "Going deeper with convolutions," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–9, Boston, MA, USA, 2015.

[31] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5 MB model size," 2016, https://arxiv.org/abs/1602.07360.

[32] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, Las Vegas, NV, USA, 2016.

[33] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 60, no. 6, pp. 84–90, 2017.

[34] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, https://arxiv.org/abs/1409.1556.

[35] S. Zheng, S. Jayasumana, B. Romera-Paredes et al., "Conditional random fields as recurrent neural networks," in *Proceedings of the IEEE international conference on computer vision*, pp. 1529–1537, Santiago, Chile, 2015.

[36] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3431–3440, Boston, MA, USA, 2015.

[37] Z. Liu, X. Li, P. Luo, C.-C. Loy, and X. Tang, "Semantic image segmentation via deep parsing network," in *Proceedings of the IEEE international conference on computer vision*, pp. 1377–1385, Santiago, Chile, 2015.

[38] G. Lin, C. Shen, A. Van Den Hengel, and I. Reid, "Efficient piecewise training of deep structured models for semantic segmentation," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3194–3203, Las Vegas, NV, USA, 2016.

[39] L. Zhou, C. Zhang, and M. Wu, "D-LinkNet: LinkNet with pretrained encoder and dilated convolution for high resolution

satellite imagery road extraction," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2018, pp. 182–186, Salt Lake City, UT, USA, 2018.

[40] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1725–1732, Columbus, OH, USA, 2014.

[41] H. Tian, Y. Tao, S. Pouyanfar, S.-C. Chen, and M.-L. Shyu, "Multimodal deep representation learning for video classification," *World Wide Web*, vol. 22, no. 3, pp. 1325–1341, 2019.

[42] Y. Wu, J. Lim, and M.-H. Yang, "Online object tracking: a benchmark," in *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2411–2418, Portland, OR, USA, 2013.

[43] A. Toshev and C. Szegedy, "Deeppose: human pose estimation via deep neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1653–1660, Columbus, Ohio, USA, 2014.

[44] C. Dong, C. C. Loy, K. He, and X. Tang, "Learning a deep convolutional network for image super-resolution," in *Computer Vision – ECCV 2014. ECCV 2014*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds., vol. 8692 of Lecture Notes in Computer Science, pp. 184–199, Springer, Cham, 2014.

[45] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: surpassing human-level performance on ImageNet classification," in *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 1026–1034, Santiago, Chile, 2015.

[46] S. Ioffe and C. Szegedy, "Batch normalization: accelerating deep network training by reducing internal covariate shift," in *Proceedings of the 32nd International Conference on Machine Learning*, Lille, France, 2015.

[47] A. Graves, S. Fernández, and J. Schmidhuber, "Multi-dimensional recurrent neural networks," in *Artificial Neural Networks – ICANN 2007. ICANN 2007*, vol. 4668 of Lecture Notes in Computer Science, pp. 549–558, Springer, Berlin, Heidelberg, 2007.

[48] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Computer Vision – ECCV 2018. ECCV 2018*, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds., vol. 11211 of Lecture Notes in Computer Science, pp. 801–818, Springer, Cham, 2018.

[49] S. Mehta, M. Rastegari, A. Caspi, L. Shapiro, and H. Hajishirzi, "ESPNet: efficient spatial pyramid of dilated convolutions for semantic segmentation," in *Computer Vision – ECCV 2018. ECCV 2018*, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds., vol. 11214 of Lecture Notes in Computer Science, pp. 552–568, Springer, Cham, 2018.

[50] S. Mehta, M. Rastegari, L. Shapiro, and H. Hajishirzi, "ESPNetv2: a light-weight, power efficient, and general purpose convolutional neural network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 9190–9200, Long Beach, CA, 2019.

[51] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2881–2890, Honolulu, HI, USA, 2017.

[52] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," 2015, https://arxiv.org/abs/1511.07122.

[53] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, https://arxiv.org/abs/1706.05587.

[54] X. Lian, Y. Pang, J. Han, and J. Pan, "Cascaded hierarchical atrous spatial pyramid pooling module for semantic segmentation," *Pattern Recognition*, vol. 110, article 107622, 2021.

[55] Y. Zhang, S. Wang, K. Xia, Y. Jiang, P. Qian, and For the Alzheimer's Disease Neuroimaging Initiative, "Alzheimer's disease multiclass diagnosis via multimodal neuroimaging embedding feature selection and fusion," *Information Fusion*, vol. 66, pp. 170–183, 2021.

[56] W. Liu, A. Rabinovich, and A. C. Berg, "Parsenet: looking wider to see better," 2015, https://arxiv.org/abs/1506.04579.

[57] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 9, pp. 1904–1916, 2015.

[58] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. 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, 2018.

[59] M. Yang, K. Yu, C. Zhang, Z. Li, and K. Yang, "DenseASPP for semantic segmentation in street scenes," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3684–3692, Salt Lake City, UT, USA, 2018.

[60] C.-W. Xie, H.-Y. Zhou, and J. Wu, "Vortex pooling: improving context representation in semantic segmentation," 2018, https://arxiv.org/abs/1804.06242.

[61] F. Chollet, "Xception: deep learning with depthwise separable convolutions," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1251–1258, Honolulu, HI, USA, 2017.

[62] Q. Zhou, W. Yang, G. Gao et al., "Multi-scale deep context convolutional neural networks for semantic segmentation," *World Wide Web*, vol. 22, no. 2, pp. 555–570, 2019.

[63] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 1520–1528, Santiago, Chile, 2015.

[64] J. Fritsch, T. Kuehnl, and A. Geiger, "A new performance measure and evaluation benchmark for road detection algorithms," in *16th International IEEE Conference on Intelligent Transportation Systems (ITSC 2013)*, pp. 1693–1700, The Hague, Netherlands, 2013.

[65] G. J. Brostow, J. Fauqueur, and R. Cipolla, "Semantic object classes in video: a high-definition ground truth database," *Pattern Recognition Letters*, vol. 30, no. 2, pp. 88–97, 2009.

[66] G. Dong, Y. Yan, C. Shen, and H. Wang, "Real-time high-performance semantic image segmentation of urban street scenes," *IEEE Transactions on Intelligent Transportation Systems*, pp. 1–17, 2020.

[67] G. Li, I. Yun, J. Kim, and J. Kim, "DABNet: depth-wise asymmetric bottleneck for real-time semantic segmentation," 2019, https://arxiv.org/abs/1907.11357.

[68] H. Wu, J. Zhang, K. Huang, K. Liang, and Y. Yu, "FastFCN: rethinking dilated convolution in the backbone for semantic segmentation," 2019, https://arxiv.org/abs/1903.11816.

[69] Q. Zhou, Y. Wang, Y. Fan et al., "AGLNet: towards real-time semantic segmentation of self-driving images via attention-

guided lightweight network," *Applied Soft Computing*, vol. 96, p. 106682, 2020.

[70] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "BiSeNet: bilateral segmentation network for real-time semantic segmentation," in *Computer Vision – ECCV 2018. ECCV 2018*, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds., vol. 11217 of Lecture Notes in Computer Science, pp. 325–341, Springer, Cham, 2018.

[71] T. Wu, S. Tang, R. Zhang, and Y. Zhang, "CGNet: a light-weight context guided network for semantic segmentation," 2018, https://arxiv.org/abs/1811.08201.

[72] Z. Yang, H. Yu, M. Feng et al., "Small object augmentation of urban scenes for real-time semantic segmentation," *IEEE Transactions on Image Processing*, vol. 29, pp. 5175–5190, 2020.

[73] P. Bilinski and V. Prisacariu, "Dense decoder shortcut connections for single-pass semantic segmentation," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6596–6605, Salt Lake City, UT, USA, 2018.