IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Traitement du Signal
Vol. 39, No. 4, August, 2022, pp. 1235-1245
Journal homepage: http://iieta.org/journals/ts

A Semantic Segmentation Method for Road Environment Images Based on Hybrid Convolutional Auto-Encoder

Xiaona Song[1], Haichao Liu[1], Lijun Wang[1*], Song Wang[1], Yunyu Cao[1], Donglai Xu[2], Shenfeng Zhang[3]

[1] School of Mechanical Engineering, North China University of Water Resources and Electric Power, Zhengzhou 450011, China
[2] School of Computing, Engineering and Digital Technologies, Teesside University, Middlesbrough, TS1 3BX, UK
[3] Suzhou Tianshuo Intelligent Software Co., Ltd., Suzhou 215011, China

Corresponding Author Email: wanglijun@ncwu.edu.cn

https://doi.org/10.18280/ts.390416

Received: 26 April 2022
Accepted: 28 June 2022

Keywords:
semantic segmentation, autonomous vehicles, convolutional auto-encoder, deep learning

ABSTRACT

Deep convolutional neural networks (CNNs) have presented amazing performance in the task of semantic segmentation. However, the network model is complex, the training time is prolonged, the semantic segmentation accuracy is not high and the real-time performance is not good, so it is difficult to be directly used in the semantic segmentation of road environment images of autonomous vehicles. As one of the three models of deep learning, the auto-encoder (AE) has powerful data learning and feature extracting capabilities from the raw data itself. In this study, the network architecture of auto-encoder and convolutional auto-encoder (CAE) is improved, supervised learning auto-encoder and improved convolutional auto-encoder are proposed, and a hybrid convolutional auto-encoder model is constructed by combining them. It can extract low-dimensional abstract features of road environment images by using convolution layers and pooling layers in front of the network, and then supervised learning auto-encoder are used to enhance and express semantic segmentation features, and finally de-convolution layers and un-pooling layers are used to generate semantic segmentation results. The hybrid convolutional auto-encoder model proposed in this paper not only contains encoding and decoding parts which are used in the common semantic segmentation models, but also adds semantic feature enhancing and representing parts, so that the network which has fewer convolutional and pooling layers can still achieve better semantic segmentation effects. Compared to the semantic segmentation based on convolutional neural networks, the hybrid convolutional auto-encoder has fewer network layers, fewer network parameters, and simpler network training. We evaluated our proposed method on Camvid and Cityscapes, which are standard benchmarks for semantic segmentation, and it proved to have a better semantic segmentation effect and good real-time performance.

1. INTRODUCTION

With the rise of artificial intelligence, the research of autonomous vehicles has become a hot research topic of major scientific research institutions and automobile manufacturers in recent years. Visual perception, as an extremely important perception method in the environment perception of autonomous vehicles, uses on-board cameras to obtain the image information of the surrounding environment of the vehicle, and then perceives the driving environment of the vehicle through image analysis and recognition technology. In the application of autonomous vehicles, the semantic segmentation of road environment images is mainly used to achieve the basic semantic description of road scenes, providing necessary environmental information for the understanding of the scene of autonomous vehicles, so as to ensure the safety of autonomous vehicles.

The deep convolution neural networks are constructed to complete semantic segmentation in the existing road environment image semantic segmentation framework [1-6]. The layer number of the networks is as high as dozens or even hundreds of layers, a few days or even weeks are still needed to train the networks even on the high performance GPU, and the training is complex and difficult. In addition, the deep convolution and pooling operations in the full convolutional neural network will lose the location information and the relevant information between different regions in the road environment image, which is not conducive to solving the dual tasks of accurate classification and accurate positioning that must be completed in semantic segmentation, and affects the accuracy of semantic segmentation. For this purpose, researchers [7-10] improved the feature extraction capability of the network by improving convolution and pooling methods in deep architecture, and generated more accurate semantic segmentation results through multi-scale convolution, feature fusion, pyramid pooling and other methods. These improved methods not only improve the performance of network semantic segmentation, but also make the semantic segmentation model more complex, increase the difficulty of network training and affect the real-time performance of model running. Although researchers improve the real-time performance of semantic segmentation by streamlining the network or designing some lightweight network models [11-15], it is difficult to guarantee the accuracy of semantic

segmentation at the same time.

As one of the three models of deep learning, auto-encoder [16] has outstanding feature extraction and data reconstruction capabilities. Auto-encoders contain rich structural forms, such as sparse auto-encoders [17], denoising auto-encoders [18, 19], convolutional auto-encoders [20], stacked auto-encoders, etc. These different structural forms have different characteristics, which make them have different feature learning capabilities. Using auto-encoder to extract effective semantic features of road environment image can provide more information for semantic segmentation of road environment image.

A semantic segmentation method for road environment image based on hybrid convolutional auto-encoder is put forward in this paper. Due to the introduction of semantic feature enhancement and expression layer, the network can still have better semantic feature extraction ability and achieve better semantic segmentation accuracy even with fewer convolution and pooling layers.

The rest of this paper is organized as follows. Section 2 reviews recent semantic segmentation methods. Section 3 introduces our methods of semantic segmentation. Experiments are discussed and evaluated in Section 4. A summary in Section 5 concludes this paper.

2. RELATED WORK

Since deep learning has made great achievements in multiple tasks in the field of computer vision, deep neural networks are used to complete image semantic segmentation in most semantic segmentation methods.

The FCN [3] network proposed in 2015 was successfully applied to image semantic segmentation by replacing the full connection layers of CNN with the convolution layers, and realized pixel-level prediction through end-to-end training, establishing the subsequent semantic segmentation framework. Segnet [4] network constructed a deep convolution encoder-decoder architecture, and verified that the network improved the segmentation accuracy and achieved pixel-level semantic segmentation in CamVid [21] and KITTI [22] semantic segmentation data sets of road environment. Since then, many semantic segmentation networks, such as PSPnet [8] and RefineNet [9], are mostly based on encoder-decoder structure. Current approaches for semantic segmentation focus on optimizing the encoders that extract features and the decoders that outputs semantic segmentation results to improve segmentation accuracy. Such large models cannot be directly used in the autonomous vehicles due to the limited GPU memory. Further, most of these approaches are relatively complex and therefore do not meet the real-time requirements of autonomous vehicles. Therefore, improving the existing semantic segmentation model to have good real-time has become a research focus of semantic segmentation of road environment images. RTSEG [13] designed two different networks, Skipnet-Mobilenet and UNet-Mobilenet, to reduce the computing cost of network operation and built a real-time semantic segmentation network for road environment images. Similarly, E-net [14] and ERFnet [15] have also simplified and modified existing semantic segmentation networks to meet real-time requirements.

The deep neural networks are trained in the above methods with massive data and the network models are tremendous. Even if they run on high-performance GPUs, the training time is very prolonged, which affects the application of road images semantic segmentation.

Auto-encoder is a method of data dimension compression and feature expression based on unsupervised learning. In essence, it is a neural network to reconstruct input data and express features. Hamza et al. [23-25] shows that auto-encoder has strong feature extraction capability and rich structural forms, and it consists of encoding and decoding parts, which is similar to the semantic segmentation method based on full convolutional neural network framework. However, the auto-encoder continuously abstracts features through identity mapping to obtain a simple and effective representation of image features. Obviously, this unsupervised self-learning method cannot be used directly to complete semantic segmentation. However, by adding a supervised layer to the auto-encoder, the supervised learning AE can be forced to learn semantic features which are beneficial to semantic segmentation to complete semantic segmentation task. In Song et al. [26], the single-layer supervised learning sparse auto-encoder and supervised learning denoising auto-encoder models are stacked to form a supervised learning deep auto-encoder model to complete the semantic segmentation task of road environment images. This method shows powerful feature extraction and reorganization ability of the supervised learning AE, but on account of the connection structure, the computing data is too big, the memory needed in the model is also too big. So the images are down-sampled to reduce the data dimension in the training process. However, the down-sampling could cause the loss of original data and affect the accuracy of segmentation.

In order to solve the problem, a hybrid convolutional auto-encoder model is proposed to complete semantic segmentation. In this method, the supervised learning auto-encoder is combined with the convolutional auto-encoder model, and a hybrid convolutional auto-encoder model is constructed, forming a semantic segmentation method for high-precision images. It can extract low-dimensional abstract features of road environment images by using convolutional layer and pooling layer in front of network, and supervised learning auto-encoder are used to enhance and express semantic segmentation features, and finally de-convolutional layers and un-pooling layers are used to generate semantic segmentation results. This network architecture can avoid the information loss caused by excessive convolutional layers and pooling layers in the original full convolutional neural network framework. Because the semantic feature enhancement and expression layer are introduced into the network, the semantic feature extraction ability can be enhanced when there are fewer convolutional layers and pooling layers in the network, and better semantic segmentation accuracy is achieved.

3. THE PROPOSED METHOD

3.1 Supervised learning auto-encoder

In order to utilize auto-encoder directly for semantic segmentation, a supervised learning auto-encoder model was constructed by adding a supervised layer to the classical auto-encoder model. The objective function of the auto-encoder is to minimize the average reconstruction error between input data X and output data Z, while the objective function of the supervised learning auto-encoder model is to minimize the average error between supervised label X_{label} and output data Z, shown as follows.

$$J_{supervised}(W,b)=\frac{1}{m}\sum_{i=1}^{m}(\frac{1}{2}\|Z^i-X_{label}^i\|^2) \quad (1)$$

where, m is the number of training samples, Z^i represents the ith reconstruction sample, X_{label}^i is the road segmentation result of the ith training sample, W is the connection weight and b is the bias term.

The supervised learning AE model learns the useful features for segmentation and completes the semantic segmentation for road environments by minimizing the mean reconstruction error between the supervised label X_{label} and reconstructed sample Z. The architecture of the supervised learning AE model is shown in Figure 1.

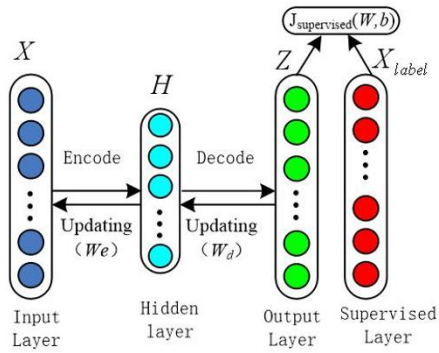

Figure 1. Architecture of supervised learning auto-encoder

The feature learning ability of classical unsupervised AE is restricted, because it can only learn the inherent features of the input samples. Nevertheless, by adding supervised learning layer, under the guidance of the supervision layer, the model can learn more relevant features conducive to semantic segmentation, so as to directly complete the semantic segmentation of the image. Auto-encoder has many variants corresponding to various applications, such as denoising auto-encoder, sparse auto-encoder. Different forms of auto-encoder have different characteristics, and the features extracted from the original data have different characteristics. By adding supervised layer to different auto-encoders, different models of supervised learning auto-encoders can be attained.

(1) Supervised learning sparse auto-encoder

The structure of the supervised learning sparse auto-encoder is similar to the basic model structure of the supervised learning auto-encoder. The sparse regular term is added to the objective function of the supervised learning auto-encoder, shown as follows.

$$J_{supervised\_sparse}(W,b)=\frac{1}{m}\sum_{i=1}^{m}(\frac{1}{2}\|Z^i-X_{label}^i\|^2)+\beta\sum_{j=1}^{S2}KL(\rho\|\hat{\rho}_j) \quad (2)$$

The first term of the objective function is similar to the objective function of the supervised learning auto-encoder, the second term is sparse regularization term which makes most neurons in the hidden layer of the network in a suppressed state. Therefore, it can extract sparse features of the image, which are sensitive to the contour or edge of the object in the image, and these features are helpful to the semantic segmentation. Therefore, supervised learning sparse auto-encoder can be used for semantic segmentation of road environment image to optimize the precision of semantic segmentation.

(2) Supervised learning denoising auto-encoder

Supervised learning denoising auto-encoder introduces supervised layer based on the model of denoising auto-encoder, and uses "clean" semantic segmentation label to guide the model to learn the most essential features of the input data, so as to obtain higher-level and more abstract feature expression of the input data. By minimizing the average reconstruction error between the supervised label Xlabel and the reconstructed sample Z, the feature extraction ability of the data in the noise environment is enhanced and the robustness of the network is improved. The objective function is the same as formula (1).

For unmanned vehicles, the on-board camera images will be affected by various factors such as weather, illumination, etc. To a certain degree, the supervised denoising auto-encoder can eliminate these interference factors, extract more robust image feature extracting and enhance the anti-interference ability of semantic segmentation and the accuracy of semantic segmentation.

In general, when various forms of supervised learning auto-encoders are used for semantic segmentation, the processes are divided into encoding and decoding. The main task of encoding is to extract image features which are favorable to semantic segmentation, while the task of decoding is to re-express the extracted features to generate semantic segmentation results. Therefore, its network structure is similar to the semantic segmentation method based on the full convolutional neural network framework. Nevertheless, in the semantic segmentation method based on the full convolutional neural network framework, the inherent convolution, pooling, de-convolution and un-pooling operations in the full convolutional neural network limit the form of network extraction feature and recombination feature. Under the operation of alternating convolution, pooling, de-convolution and un-pooling, the location information and the relevant information between different regions of the image are highly depleted. The supervised learning auto-encoder model with multiple structures can extract richer semantic segmentation features and automatically recombine these features to generate semantic segmentation maps. The loss of location information and inter-region correlation information is less and the semantic segmentation can be better completed.

3.2 Convolutional auto-encoder

The network structure of the Convolutional Auto-Encoder (CAE) is illustrated in Figure 2, which mainly comprises convolution, pooling, de-convolution and un-pooling. The first two parts are equivalent to the coding process used mainly for feature extraction, which is the same as the CNN feature extraction process. The latter two parts are equivalent to the decoding process through de-convolution and un-pooling to reconstruct the image. The training of CAE is similar to that of classical auto-encoder. The input samples are compared with the final reconstructed results, and the back propagation algorithm is used to minimize the error function to modify the training parameters. The error function of CAE is similar to that of traditional auto-encoder, which can be expressed as:

$$J_{CAE}(\theta)=\sum_{x\in S}L(x,y)+\lambda\|W\|_2^2 \quad (3)$$

where, S is training sample set, parameter λ is coefficient of regularization term, W is the weight.

3.3 Improved convolutional auto-encoder

The original convolutional auto-encoder structure only contains the convolutional layer, pooling layer, de-convolution layer and un-pooling layer, but does not contain the full connection layer. The supervised learning auto-encoder network designed in section 3.1 is a fully connected structure. In order to combine with the supervised learning auto-encoder, this paper improves the structure of the convolutional auto-encoder by adding a fully connected layer between the encoding part and decoding part of the network, and its basic structure is shown in Figure 3.

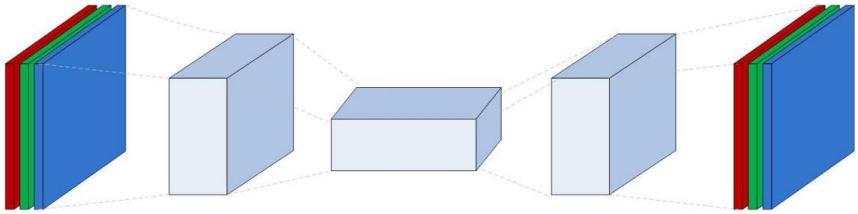

Figure 2. Architecture of convolutional auto-encoder

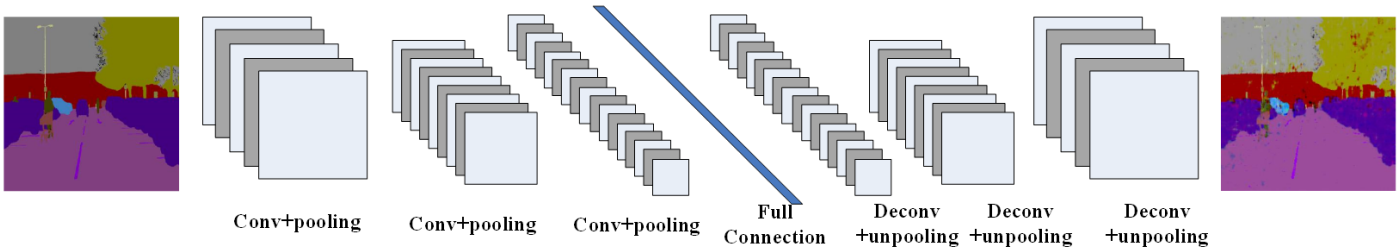

Figure 3. Architecture of improved convolutional auto-encoder

The original convolutional auto-encoder network extracts the features that represent the nature of the data through the alternation of convolution and pooling, but the operation of convolution and pooling also causes the loss of useful information in the image. By adding a full connection layer between the encoding part and the decoding part of the network, the features extracted from the encoding part can be further enhanced and effectively integrated through nonlinear mapping relationship, and then the enhanced and integrated features are added to the decoding part, so it has better image reconstruction capability. Obviously, the basic structure of the improved convolutional auto-encoder can only be used to reconstruct the image, but not for semantic segmentation of the image.

3.4 Convolutional hybrid auto-encoder

In order to complete semantic segmentation of road environment images, supervised learning auto-encoder is introduced into convolutional auto-encoder training, and a hybrid convolutional auto-encoder model is established. The design idea of this model is as follows.

Firstly, road environment images and its semantic segmentation labels are used to train two improved convolution auto-encoder models which can extract the characteristics of road environmental images and semantic segmentation labels respectively. Secondly, a supervised learning auto-encoder is built, in which the characteristic of road environment image is the input layer and the characteristics of the semantic segmentation tags is the supervised layer. Abstract features that can be used for image semantic segmentation are extracted by this supervised learning auto-encoder. Finally, two improved convolutional auto-encoder models and supervised learning auto-encoder are combined to generate a convolutional hybrid auto-encoder model.

Figure 4 shows the structure and training method of the hybrid convolutional auto-encoder model. The training process of the whole model is divided into four parts.

Firstly, the sample images are used to build a sample improved convolutional auto-encoder model, and its main purpose is to compress the road image sample, extract the representative data of the road images and obtain the encoding parameters of road images; Second, semantic segmentation labels are used to construct a label convolutional auto-encoder model. The main purpose of this model is to extract the data features of semantic segmentation tags and obtain the decoding parameters of semantic segmentation tags; Third, a supervised learning model automatic encoder is built to enhance and express the characteristics of semantic segmentation. In this model, the extracted road image characteristics in the first part is used as the input layer, the extracted semantic segmentation tags in the second part are used as the supervised layer and the better characteristics of the segmentation information can be obtained in this model.

Finally, the sample improved convolutional auto-encoder, the supervised learning auto-encoder and the label reconstruction convolutional auto-encoder model are stacked to form the hybrid convolutional auto-encoder model. The encoding part of the sample improved convolutional auto-encoder, the full connection layer of the sample improved convolutional auto-encoder, the middle layer of the supervised learning auto-encoder, the full connection layer of the label improved convolutional auto-encoder, and the decoding part of the label improved convolutional auto-encoder are stacked up in sequence. In the training process of this stage, a supervised layer is introduced into the convolutional hybrid auto-encoder model, and semantic segmentation labels are used as supervised information to fine-tune and optimize the parameters of the network.

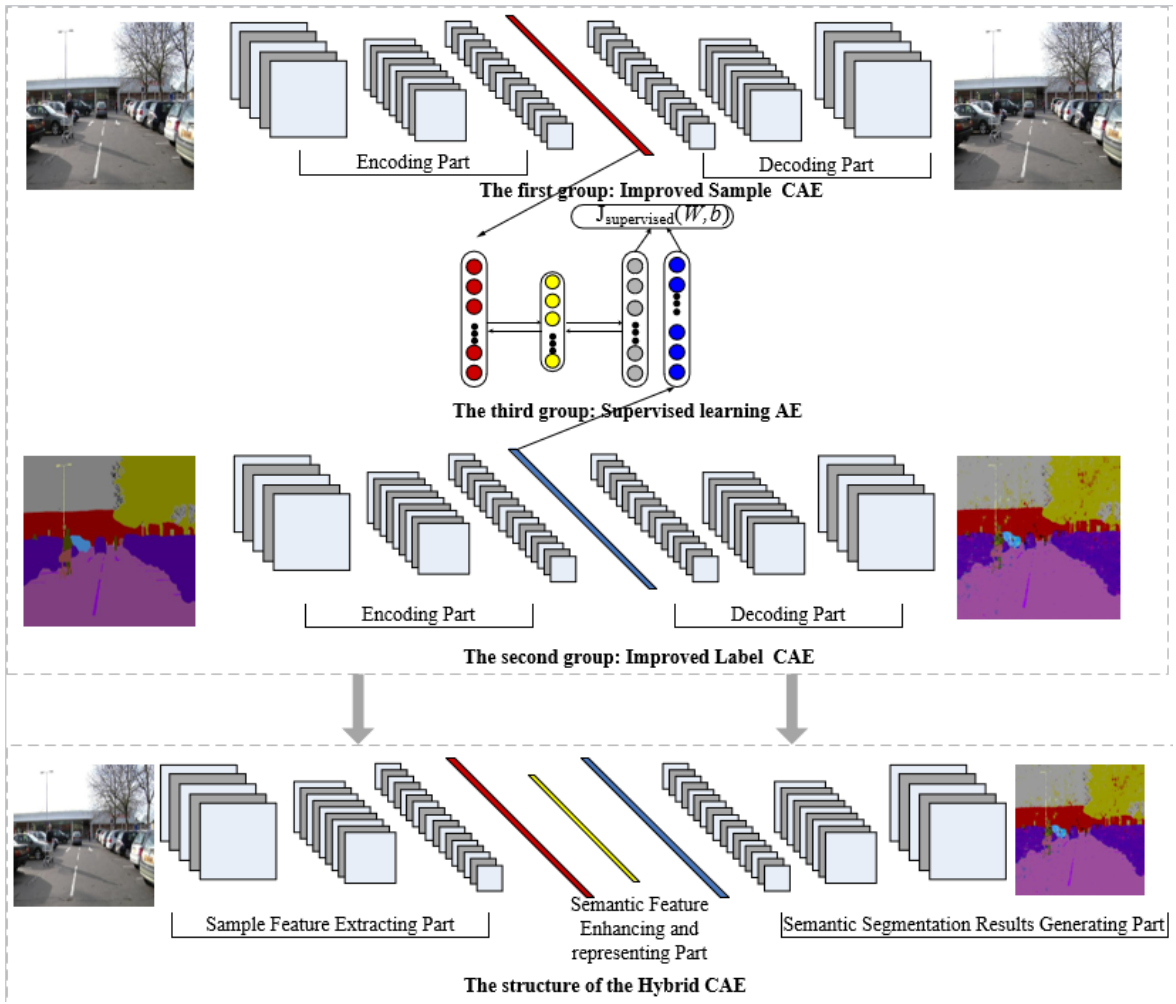

Figure 4. The structure and training method of the hybrid convolutional auto-encoder

As shown in Figure 4, the hybrid convolutional auto-encoder model is divided into sample feature encoding part, semantic feature enhancing and representing part, and semantic segmentation results generating part. Compared with the traditional semantic segmentation networks which only have encoding and decoding parts, the hybrid convolutional auto-encoder we established can extract features of the road samples to complete data reduction by convolution and pooling operation in the beginning of the model, enhance and reorganize the features with the supervised learning auto-encoder in the middle of the model, and generate the semantic segmentation results by de-convolution and un-pooling operations at the end of the model. The sample feature encoding part effectively compresses the original data to get a low-dimensional full connection vector, so the parameters of the supervised learning auto-encoder are greatly reduced. Even if the full connection structure is used in the model, the network parameters will not be exploded. Due to the addition of semantic feature enhancing and representing part, the network can obtain better semantic segmentation accuracy with fewer convolution and pooling layers, thus the network model is reduced and the real-time performance are improved.

4. EXPERIMENTS AND ANALYSIS

In order to evaluate the semantic segmentation performance of the hybrid convolutional auto-encoder model proposed in this paper, a series of experiments are conducted using Camvid and Cityscapes data sets. The image size is adjusted to 360×360 for the convenience of the experiment. Since the semantic segmentation object is the road environment image of unmanned vehicles in which the on-board memory is limited and the requirement of real-time is relatively high, the parameters of the hybrid convolutional auto-encoder model are reasonably set as follows.

In the sample feature encoding part, we set three convolutional layers and three pooling layers. The number of convolutional kernels at each layer of the convolutional layer is 64, 128 and 64 respectively, and the size of the convolutional kernels is 5×5. The pooling layer adopts max-pooling and the size is 2×2. The structure of semantic segmentation results generating part is symmetric with the sample feature encoding part, and the number of convolution kernels of the corresponding de-convolution layer is 64, 128 and 64. Other parameters are also the same as the encoding part. The semantic feature enhancing and representing part, namely the fully connected part in the model, has 500, 400 and 500 nodes respectively. The activation function of the hybrid convolutional auto-encoder is ReLu function except for the last layer, and the last layer is Sigmoid function.

Compared to the semantic segmentation model based on the Fully Convolutional Networks framework, the model proposed in this paper contains only three convolution layers, three full connection layers and three de-convolution layers. It is markedly smaller than the fully convolution framework of semantic segmentation model and the network parameters are highly reduced. It not only improves the real-time of the

semantic segmentation, but also reduces the memory requirement of the GPU in unmanned vehicles.

Since the hybrid convolutional auto-encoder model is constructed by stacking two improved convolutional auto-encoder models and one supervised learning auto-encoder, the accuracy of each sub-model will directly affect the performance of the hybrid convolutional auto-encoder model. Therefore, we evaluate the feature extraction ability of convolutional auto-encoder model, semantic transformation method of supervised learning auto-encoder model and semantic segmentation performance of hybrid convolutional auto-encoder.

4.1 Feature extraction ability of improved convolutional auto-encoder model

Table 1. Correlation coefficients of some sample images

Model	R Chanel	G Chanel	B Chanel	Average
1	0.9810	0.9964	0.9964	0.9912
2	0.9974	0.9990	0.9982	0.9982
3	0.9845	0.9959	0.9968	0.9924
4	0.9954	0.9976	0.9981	0.9970
5	0.9942	0.9975	0.9973	0.9963
6	0.9856	0.9920	0.9923	0.9900
7	0.9955	0.9981	0.9977	0.9971
8	0.9809	0.9969	0.9968	0.9915
9	0.9952	0.9979	0.9972	0.9968
10	0.9933	0.9976	0.9973	0.9961

The two groups of improved convolutional auto-encoder models extract the features of sample images and semantic segmentation labels respectively through unsupervised learning, and reconstruct the sample images and semantic segmentation labels using the extracted features. Obviously, if the errors between original images and reconstructed images are big, it shows that the extraction ability of the two improved convolution encoder automatically model of feature is not good, or the features extracted by the network are not representative and cannot reflect the essence of the original samples and labels. Only when the difference between the reconstructed images and the original input images is small, it indicates that the features extracted from the network are representative and can represent the original image. Therefore, we can evaluate the feature extraction ability of the two improved convolutional auto-encoder by assessing the performance of the reconstructed images. In this paper, the correlation coefficient is selected to measure the correlation between the original image and the reconstructed image, and the formula is as follows.

$$r = \frac{\sum_{m}\sum_{n}(A_{mn}-\bar{A})(B_{mn}-\bar{B})}{\sqrt{(\sum_{m}\sum_{n}(A_{mn}-\bar{A})^2)((\sum_{m}\sum_{n}(B_{mn}-\bar{B})^2)}} \tag{4}$$

where, m and n are rows and columns of image data respectively, $\bar{A}$ and $\bar{B}$ represent the average pixel value of two image matrices, and r is the correlation coefficient of images. The closer r is to 1, the more similar the two images are, and the closer r is to 0, the less similar the two images are.

As the images in the data set are color images and have RGB three channels, the evaluation of its correlation is divided into three channels respectively, and its values are averaged to obtain the final correlation coefficient. In this paper, the correlation of all training samples are measured, and the average correlation coefficient between sample reconstructed image and sample image is 0.9923, and that between label reconstructed image and label image is 0.9951. This indicates that the similarity of images is very high and it further indicates that the feature extraction ability of the improved sample convolutional auto-encoder and the improved label convolutional auto-encoder is very strong, and the models can obtain the most representative features of the data itself, which can be used to represent the data sample image and semantic label. Table 1 offers the correlation coefficients of some sample images.

When the training of the two improved convolutional auto-encoders is accomplished, we can extract road image features and semantic segmentation label features respectively using the encoding parts of the two models, shown as Figure 5. The training sample images and their semantic segmentation labels are added to the front end of the trained network, the features of road images and semantic segmentation labels can be extracted and compressed by the forward propagation of the network, which can be used for subsequent supervised learning AE.

4.2 Semantic transformation of the supervised learning auto-encoder model

The road image feature and semantic segmentation label feature extracted by the two improved convolutional auto-encoders are used to construct supervised learning auto-encoder in which road images feature is used as input layer and semantic segmentation label feature is used as supervised layer. Its main purpose is to extract the image semantic segmentation features and realize the transformation from road image features to semantic segmentation features. In this paper, we design three types of supervised learning auto-encoders, which are the supervised learning auto-encoder, supervised learning sparse auto-encoder and supervised learning denoising auto-encoder. By studying the training process of the three models, the final semantic transformation model is determined.

Figure 6 shows the error function curves of the training process of the three models. It can be seen from the figure that the error of the supervised learning sparse auto-encoder reduces fastest, and the final error is the smallest with the same number of iterations. Therefore, in the process of constructing a convolutional hybrid auto-encoder network, supervised learning sparse auto-encoder is selected to realize the transformation from road image features to semantic segmentation features.

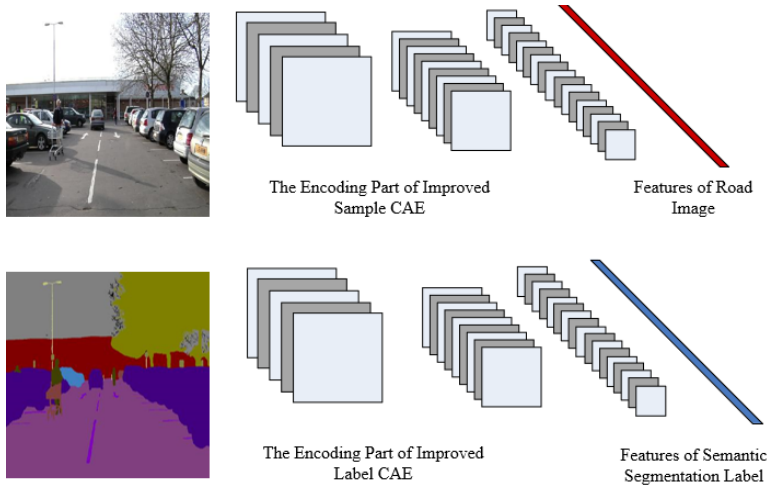

Figure 5. The features extracting process of improved convolutional auto-encoders

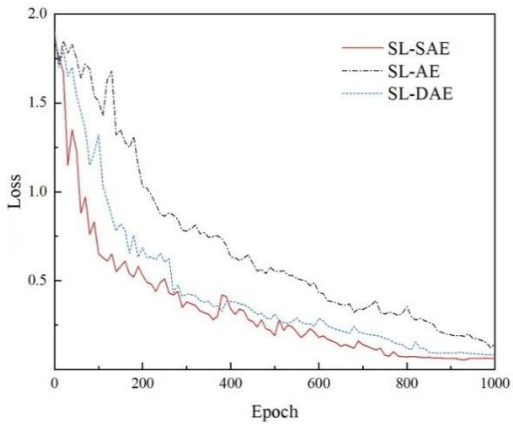

Figure 6. Error curves of different types of supervised learning auto-encoders

4.3 Semantic segmentation performance of hybrid convolutional auto-encoder

When the training of two improved convolutional auto-encoders and supervised learning sparse auto-encoder is accomplished, the hybrid convolutional auto-encoders are created according to the stacking method specified in section 3.4. Subsequently, the hybrid convolutional auto-encoder network is optimized again by adding a supervised layer (semantic segmentation label is used as the supervised layer) to generate the final model.

We divide the testing process into two steps. First, the test samples are added to the improved label convolutional auto-encoder model to assess its image reconstruction ability, and then the test samples are added to the final hybrid convolutional auto-encoder model to assess its semantic segmentation performance.

Figure 7 shows the results of some test samples. As can be seen from the comparison between the second and third lines in Figure 7, the reconstructed images of semantic labels in the first three images are nearly the same as the original label, while the reconstructed images of the last image are not similar enough to the original label image. The average similarity between the whole test sample and its reconstructed image is 0.9843. In general, the improved label convolutional auto-encoder can well complete the extraction of semantic label features.

According to the comparison between the second and fourth lines in Figure 7, it can be seen that the semantic segmentation results have clear object boundaries and the classification of each object is accurate. But some regions in the fourth image in Figure 7 are misclassified, and the reason is that when the sample labels are reconstructed ineffectively, the corresponding segmentation results of hybrid convolutional auto-encoder would also be affected. It shows that the hybrid convolutional auto-encoder will be affected by the accuracy of each early training model. In order to construct a more accurate semantic segmentation model, each sub-model must be carefully trained to improve the accuracy.

Then, we compare the hybrid convolutional auto-encoder proposed with SegNet based on the full convolutional network framework. Figure 8 shows the results of semantic segmentation.

As can be seen from Figure 8, compared with the results of SegNet, object boundaries in the results of hybrid convolutional auto-encoder are clearer and more accurate, misclassified regions are fewer and classification accuracy is higher. It shows that the hybrid convolutional auto-encoder can reasonably extract more accurate semantic segmentation features and generate semantic segmentation results. The main reason is that hybrid convolutional auto-encoder not only contains convolution, de-convolution, pooling and un-pooling layers, but also contains the semantic feature enhancing and representing layer, which use supervised learning to enhance and express road environment image features extracted by early convolution and pooling layers.

In this paper, the performance of hybrid Convolutional Auto-Encoder (H-CAE for convenience) on data set CamVid is quantitatively evaluated. The evaluation indexes are Precision Accuracy (PA) and Intersection-over-Union (IoU). The comparison results are shown in Table 2. Compared with SegNet, H-CAE improves its average PA and IoU by 18.3% and 17.1% respectively. Compared with E-Net, they are 12.9% and 26.9% higher, respectively. Therefore, H-CAE has higher semantic segmentation accuracy compared to other road image semantic segmentation methods. The results show that the H-CAE can extract semantic segmentation features better and complete the classification of different objects more accurately. It can be seen from the IoU that the H-CAE can achieve both accurate classification and accurate localization, which can better complete the dual tasks of object classification and accurate localization in semantic segmentation.

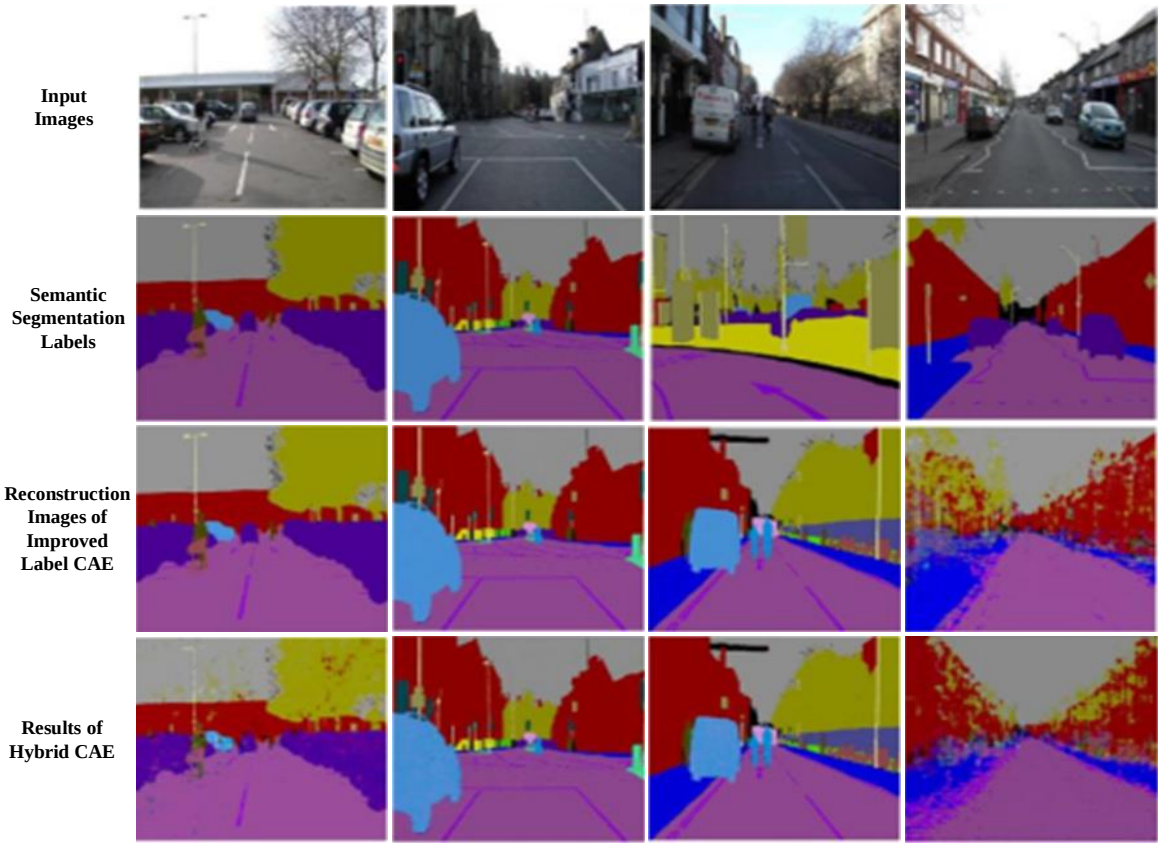

Figure 7. Results of some test samples on CamVid dataset

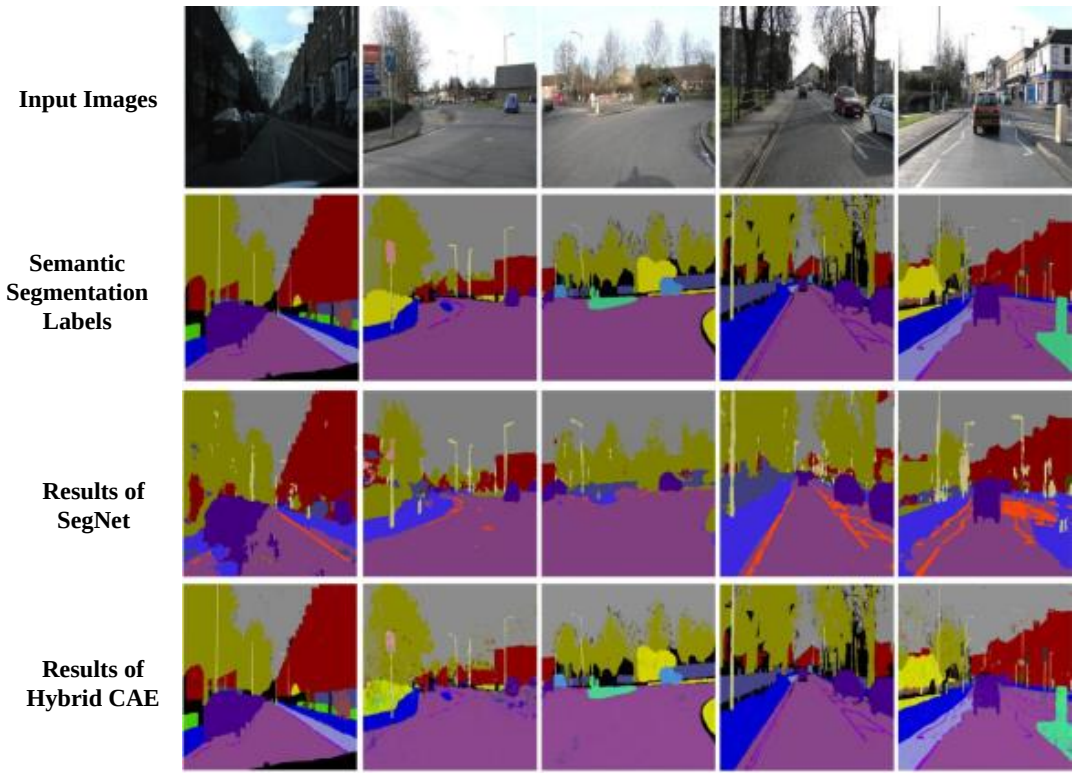

Figure 8. Results of different methods on CamVid dataset

Table 2. Comparison with other methods on CamVid dataset

Method	Bui.	Tre.	Sky	Car	Sig.	Roa.	Ped.	Fen.	Pol.	Sid.	Bic.	Class Avg.	mIoU
SegNet	88.8	87.3	92.4	82.1	20.5	97.2	57.1	49.3	27.5	84.4	30.7	65.2	55.6
E-Net	74.7	77.8	95.1	82.4	51.0	95.1	67.2	51.7	35.4	86.7	34.1	68.3	51.3
H-CAE	90.4	89.5	96.7	94.4	53.5	99.1	72.3	55.4	42.6	90.4	63.6	77.1	65.1

Table 3. Comparison with other methods on Cityscapes dataset

Method	Roa	Sid	Bui	Wal	Fen	Pol	TLi	TSi	Veg	Ter	Sky	Ped	Rid	Car	Tru	Bus	Tra	Mot	Bic	Class	Cat
Segnet	96.4	73.2	84.0	28.4	29.0	35.7	39.8	45.1	87.0	63.8	91.8	62.8	42.8	89.3	38.1	43.1	44.1	35.8	51.9	57.0	79.1
E-net	96.3	74.2	75.0	32.2	33.2	43.4	34.1	44.0	88.6	61.4	90.6	65.5	38.4	90.6	36.9	50.5	48.1	38.8	55.4	58.3	80.4
H-CAE	98.7	89.4	88.5	59.7	55.6	48.7	46.8	50.3	94.8	78.6	96.8	66.8	54.9	97.2	48.6	63.3	58.6	55.5	60.8	69.1	89.5

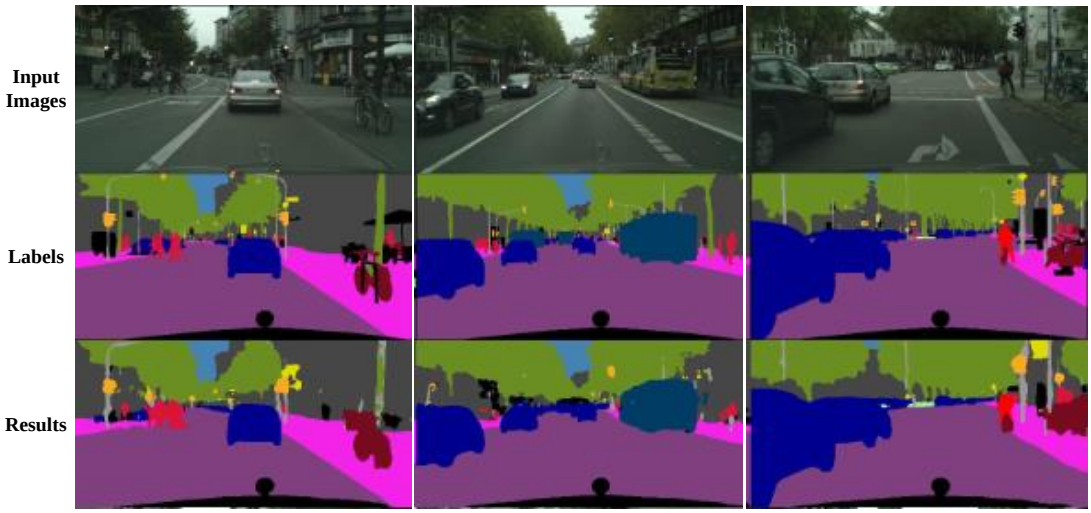

Figure 9. Semantic segmentation results on Cityscapes dataset

Table 4. Performances of different methods on Cityscapes dataset

Method	Conv-H AE	SegNet	E-net	PSPNet	DeepLabV3
Class IoU	69.1	57.0	58.3	81.2	81.3
Category IoU	89.5	79.1	80.4	91.2	91.6
Inference time(s)	0.008	0.054	0.014	0.33	0.41

In order to prove the effectiveness of the model, we also evaluate our model on Cityscapes. Figure 9 shows the results of some samples. We can see that we still achieve good semantic segmentation results on Cityscapes data sets and we can detect and classify the image objects such as road, cars, sidewalks, pedestrians accurately. Small targets in the images of the road environment, such as traffic lights, light poles can also be detected, but the object boundary is not accurate. This is because such samples occupy a small area in the images and the number of training samples is fewer, and it conforms to the expected law.

Meanwhile, the semantic segmentation performance of the H-CAE in Cityscapes is also quantitatively evaluated, and comparison with other methods is shown in Table 3. Compared with SegNet, the average IoU and average category IoU are improved 21.4% and 10.6% respectively, while the corresponding values are improved 18.6% and 11.4% respectively, compared with E-Net. These data indicate that the H-CAE model greatly improves the accuracy of semantic segmentation. H-CAE achieves the best results on all classes of Cityscapes data set, indicating that compared with the other two methods, this method can extract various semantic features more precisely and accomplish semantic segmentation accurately.

Since the application scenario of this paper is unmanned vehicles, real-time performance is very important. Therefore, the running time of H-CAE is evaluated to determine its real-time performance. Based on the same hardware environment, we compare the semantic segmentation methods, namely, H-CAE, SegNet and E-Net, and the two classical semantic segmentation models, namely PSPNet and DeepLabV3. The results are shown in Table 4. In Table 4, the average IoUand average class IoU of various methods are also provided. As can be seen from Table 4, compared with SegNet and E-Net, the running time of H-CAE is shorter and the accuracy is higher. Comparing with the two classical semantic segmentation models PSPNet and DeepLabV3, the accuracy of H-CAE is slightly lower, but the real-time performance is better. The main reason is that compared with PSPNet and DeepLabV3, the model built in this paper is smaller. However, PSPNet and DeepLabV3 models with large scale have higher requirements on model operating environment, memory and performance of the on-board GPU. However, due to the small model and short running time, the proposed method is more suitable for the road environment images of unmanned vehicles.

5. CONCLUSION

In this paper, the convolutional auto-encoder and supervised learning auto-encoder are combined to construct a hybrid convolutional auto-encoder model. It can extract features of the road images to complete data dimension reduction by convolution and pooling operation in the beginning of the model, enhance and reorganize the features with the supervised learning auto-encoder in the middle of the model, and generate the semantic segmentation results by de-convolution and un-pooling operations at the end of the model. Compared with the existing semantic segmentation model, the model we constructed has fewer layers, fewer parameters and simpler training. Experimental results on CamVid and

Cityscapes datasets prove that our model not only has a good semantic segmentation effect, but also has good real-time performance.

ACKNOWLEDGEMENTS

This work was supported financially by National Natural Science Foundation of China (Grant No.: 61472444 and 61671470) and the National Key Research and Development Program of China (Grant No.: 2016YFC0802904). This work was supported by the Scientific and Technological Project of Henan Province (Grant No.: 212102210069 and 222102210043), the "ZHONGYUAN Talent Program" of Henan Province (Grant No.: ZYYCYU202012112), the Zhengzhou Measurement and Control Technology and Instrument Key Laboratory (NO. 121PYFZX181), the Scientific research launching project for high-level talents of North China University of Water Resources and Electric Power (Grant No.: 202006008) and Fund of Innovative Education Program for Graduate Students at North China University of Water Resources and Electric Power (Grant No.: YK-2021-90).

REFERENCES

[1] Fan, M., Lai, S., Huang, J., Wei, X., Chai, Z., Luo, J., Wei, X. (2021). Rethinking BiSeNet for real-time semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9716-9725. https://doi.org/10.1109/CVPR46437.2021.00959

[2] Hou, Q., Zhang, L., Cheng, M.M., Feng, J. (2020). Strip pooling: Rethinking spatial pooling for scene parsing. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4003-4012. https://doi.org/10.1109/CVPR42600.2020.00406

[3] Long, J., Shelhamer, E., Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431-3440. https://doi.org/10.48550/arXiv.1411.4038

[4] Badrinarayanan, V., Kendall, A., Cipolla, R. (2017). Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE transactions on Pattern Analysis and Machine Intelligence, 39(12): 2481-2495. https://doi.org/10.1109/TPAMI.2016.2644615

[5] Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L. (2017). Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE Transactions on Pattern Analysis and Machine Intelligence, 40(4): 834-848. https://doi.org/10.1109/TPAMI.2017.2699184

[6] Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H. (2018). Encoder-decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), pp. 801-818.

[7] Pinheiro, P.O., Lin, T.Y., Collobert, R., Dollár, P. (2016). Learning to refine object segments. In European Conference on Computer Vision, pp. 75-91. https://doi.org/10.1007/978-3-319-46448-0_5

[8] Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J. (2017). Pyramid scene parsing network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2881-2890.

[9] Lin, G., Milan, A., Shen, C., Reid, I. (2017). Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1925-1934. https://doi.org/10.1109/CVPR.2017.549

[10] Roy, A., Todorovic, S. (2016). A multi-scale CNN for affordance segmentation in RGB images. In European Conference on Computer Vision, pp. 186-201. https://doi.org/10.1007/978-3-319-46493-0_12

[11] Yu, C., Wang, J., Peng, C., Gao, C., Yu, G., Sang, N. (2018). Bisenet: Bilateral segmentation network for real-time semantic segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), pp. 325-341.

[12] Mehta, S., Rastegari, M., Caspi, A., Shapiro, L., Hajishirzi, H. (2018). Espnet: Efficient spatial pyramid of dilated convolutions for semantic segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), pp. 552-568. https://doi.org/10.48550/arXiv.1803.06815

[13] Siam, M., Gamal, M., Abdel-Razek, M., Yogamani, S., Jagersand, M. (2018). Rtseg: Real-time semantic segmentation comparative study. In 2018 25th IEEE International Conference on Image Processing (ICIP), 1603-1607. https://doi.org/10.1109/ICIP.2018.8451495

[14] Paszke, A., Chaurasia, A., Kim, S., & Culurciello, E. (2016). Enet: A deep neural network architecture for real-time semantic segmentation. arXiv preprint arXiv:1606.02147. https://doi.org/10.48550/arXiv.1606.02147

[15] Romera, E., Alvarez, J.M., Bergasa, L.M., Arroyo, R. (2017). Erfnet: Efficient residual factorized convnet for real-time semantic segmentation. IEEE Transactions on Intelligent Transportation Systems, 19(1): 263-272. https://doi.org/10.1109/TITS.2017.2750080

[16] Rumelhart, D.E., Hinton, G.E., Williams, R.J. (1986). Learning representations by back-propagating errors. nature, 323(6088): 533-536. https://doi.org/10.1038/323533a0

[17] Ranzato, M.A., Poultney, C., Chopra, S., Cun, Y. (2006). Efficient learning of sparse representations with an energy-based model. Advances in Neural Information Processing Systems, 19.

[18] Vincent, P., Larochelle, H., Bengio, Y., Manzagol, P.A. (2008). Extracting and composing robust features with denoising autoencoders. In Proceedings of the 25th International Conference on Machine Learning, pp. 1096-1103. https://doi.org/10.1145/1390156.1390294

[19] Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., Manzagol, P.A., Bottou, L. (2010). Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. Journal of Machine Learning Research, 11(12).

[20] Masci, J., Meier, U., Cireşan, D., Schmidhuber, J. (2011). Stacked convolutional auto-encoders for hierarchical feature extraction. In International Conference on Artificial Neural Networks, pp. 52-59. https://doi.org/10.1007/978-3-642-21735-7_7

[21] Brostow, G.J., Fauqueur, J., Cipolla, R. (2009). Semantic object classes in video: A high-definition ground truth

database. Pattern Recognition Letters, 30(2): 88-97. https://doi.org/10.1016/j.patrec.2008.04.005
[22] Geiger, A., Lenz, P., Stiller, C., Urtasun, R. (2013). Vision meets robotics: The kitti dataset. The International Journal of Robotics Research, 32(11): 1231-1237. https://doi.org/10.1177/0278364913491297
[23] Hamza, M.A., Hassine, S.B.H., Abunadi, I., Al-Wesabi, F.N., Alsolai, H., Hilal, A.M., Motwakel, A. (2022). Feature selection with optimal stacked sparse autoencoder for data mining. Cmc-Computers Materials & Continua, 72(2): 2581-2596. https://doi.org/10.32604/cmc.2022.024764
[24] Alshathri, S.I., Vincent, D.J., Hari, V.S. (2022). Denoising letter images from scanned invoices using stacked autoencoders. Cmc-Computers Materials & Continua, 71(1): 1371-1386. https://doi.org/10.32604/cmc.2022.022458
[25] El-Shafai, W., Abd El-Nabi, S., El-Rabaie, E., Ali, A., Soliman, F., Algarni, A.D., Abd El-Samie, F.E. (2022). Efficient deep-learning-based autoencoder denoising approach for medical image diagnosis. Cmc-Computers Materials & Continua, 70(3): 6107-6125. https://doi.org/10.32604/cmc.2022.020698
[26] Song, X., Rui, T., Zhang, S., Fei, J., Wang, X. (2018). A road segmentation method based on the deep auto-encoder with supervised learning. Computers & Electrical Engineering, 68: 381-388. https://doi.org/10.1016/j.compeleceng.2018.04.003