Multi-level spatial attention network for image data segmentation

Jun Guo

School of Software,
Quanzhou University of Information Engineering,
Quanzhou, Fujian, China
Email: guojun20121203@163.com

Zhixiong Jiang*

School of Business,
Shanghai Dianji University,
Shanghai, China
Email: jiangzx@sdju.edu.cn
*Corresponding author

Dingchao Jiang

School of Electronic and Information Engineering,
Xi'an Jiaotong University,
Xi'an, Shaanxi, China
Email: jiangdingchao.18@gmail.com

Abstract: Deep learning models for semantic image segmentation are limited in their hierarchical architectures to extract features, which results in losing contextual and spatial information. In this paper, a new attention-based network, MSANet, which applies an encoder-decoder structure, is proposed for image data segmentation to aggregate contextual features from different levels and reconstruct spatial characteristics efficiently. To model long-range spatial dependencies among features, the multi-level spatial attention module (MSAM) is presented to process multi-level features in the encoder network and capture global contextual information. In this way, our model learns multi-level spatial dependencies between features by the MSAM and hierarchical representations of the input image by the stacked convolutional layers, which means the model is more capable of producing accurate segmentation results. The proposed network is evaluated on the PASCAL VOC 2012 and Cityscapes datasets. Results show that our model achieves excellent performance compared with U-net, FCNs, and DeepLabv3.

Keywords: deep learning; semantic segmentation; big data.

Reference to this paper should be made as follows: Guo, J., Jiang, Z. and Jiang, D. (2021) 'Multi-level spatial attention network for image data segmentation', *Int. J. Embedded Systems*, Vol. 14, No. 3, pp.289–299.

Biographical notes: Jun Guo is a Lecturer in the School of Software, Quanzhou University of Information Engineering, China. She received her MS from the Xiamen University Software School in 2011. She has engaged in teaching for more than ten years, presided over three provincial-level topics, published more than ten papers of various types, and participated in the compilation of one textbook. Her research interests include software engineering, blockchain, and information system.

Zhixiong Jiang received his BS in Computer Science from the Wuhan University of Technology, China, in 2001, his MS in Computer Science from the Huazhong University of Science and Technology, China, in 2004 and his PhD in Sociology from the Wuhan University, China, 2014. He is currently a Lecturer in the School of Business at the Shanghai DianJi University.

Dingchao Jiang received his BEng in Software Engineering from the Xi'an Jiaotong University, Xi'an, China, in 2012. He is currently pursuing his PhD in the School of Electronic and Information Engineering of Xi'an Jiaotong University, Xi'an, China. His current research interests include machine learning, deep learning, semantic image segmentation, probabilistic graphical model, big data, smart city, etc.

1 Introduction

Semantic image segmentation attempts to solve the issue of labelling every pixel with a semantic category in the image. The image segmentation can be applied to potential applications, like autonomous driving, robotic sensing, intelligent medical analysis, and image editing. With the advancement of deep learning, technologies of the internet of things (IoT) (Liang et al., 2011, 2016; Li et al., 2018a, 2020b), big data (Liang et al., 2020b; Li et al., 2020a) and artificial intelligence (AI) (LeCun et al., 2015; Goodfellow et al., 2016; Liang et al., 2019, 2020a) rapidly spread to almost all areas of our life, which greatly improves human health, safety, and productivity. Especially, convolutional neural networks (CNNs) have demonstrated outstanding results in various computer vision tasks such as image classification (Lee and Kwon, 2017; Chollet, 2017; Chan et al., 2015; He et al., 2016), object detection (Szegedy et al., 2013; Cai et al., 2016; Li et al., 2019) and natural language process (NLP) (Chiu and Nichols, 2016; Lopez and Kalita, 2017). Some techniques like dropout have been used as a method of regularisation to improve the generalisation ability of deep learning networks. The pooling operator is applied to summarise feature maps, introduce feature invariance, and reduce the spatial dimensions. Batch normalisation alleviates the problem of the vanishing gradient in deep networks. The residual blocks ease the training of very deep CNNs.

However, the size of feature maps extracted by the standard CNNs is decreased, which leads to the loss of spatial information. Some details of objects in the image, such as locations, structures, and boundaries, are abandoned through the pooling operator. In addition, the fixed size of the receptive field by employing multiple convolutional and pooling layers results in focusing on short-range contextual information and local features. These limitations of CNNs cause inconsistency and misclassification in the segmentation results.

To use feature representations extracted by the CNNs efficiently, some works have been proposed to aggregate information from multiple levels of the model. High-level features extracted by CNNs are commonly powerful in making coarse category classification while weak in reconstructing original resolution pixel labelling. To obtain spatial information, many models (Long et al., 2015; Ronneberger et al., 2015; Mostajabi et al., 2015) fuse low- and high-level features to combine coarse and fine representations of the input image by using skip connections. Long et al. (2015) upsampled the high-level feature that contains rich semantic information and fused it with the low-level feature by element-wise summing. Ronneberger et al. (2015) utilised a U-shape architecture and concatenated low-level feature maps from the encoder with feature maps in the decoder. Mostajabi et al. (2015) divided the input image into different levels of area and extracted features from these areas to make the pixel-wise classification. The selected feature maps in these methods are natural multi-scale due to the increasingly expanded receptive field. The problem with these approaches is the fixed size of the receptive field, which limits the model's ability to extract long-range contextual information.

To obtain contextual information in a broad receptive field, there are other works focusing on increasing the size of the receptive field effectively. Different shapes and sizes of objects in the image make it difficult for pixel-wise classification. To deal with these multi-scale objects, there are other networks (Chen et al., 2017a, 2017b) exploit multi-scale contextual information by using dilated convolution layers that have different dilation rates. But the dilated convolution may result in the loss of localisation information and the inconsistency in the final results due to its sampling mechanism. Zhao et al. (2017) presented PSPNet that applied pyramid pooling module to extract both local and global context information. However, the pooling-based approaches (Zhang et al., 2018; Zhao et al., 2017) extract contextual information in a non-adaptive way and treat all pixels equally, which contradicts the fact that pixels of different locations and features have different importance to each other.

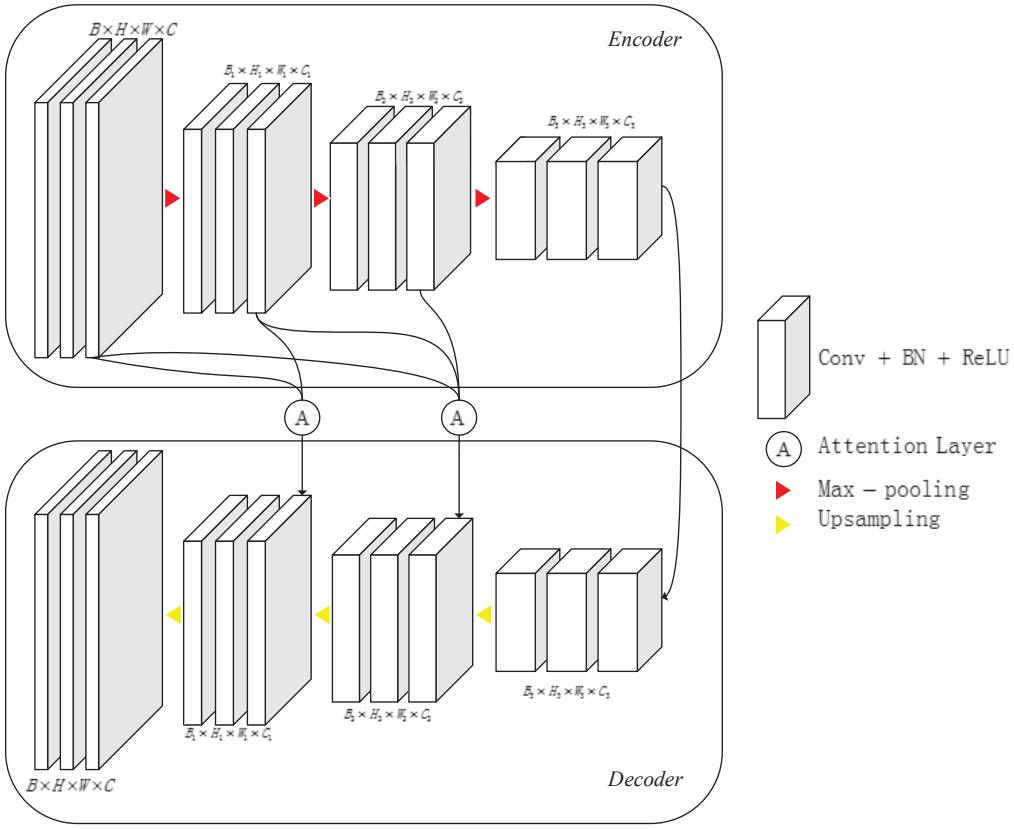
To capture long-range dependencies among pixels, Chen et al. (2017b) and Zheng et al. (2015) employed a fully connected conditional random field to model the relationships between all pixels directly. Visin et al. (2016) and Byeon et al. (2015) utilised recurrent neural networks to learn spatial dependencies between pixels. However, the parameters of these approaches are difficult to train.

To make up for the aforementioned deficiencies of CNNs applied in semantic image segmentation, this paper proposes a multi-level spatial attention network (MSANet). The proposed network utilises an encoder-decoder structure that has an encoder to extract multi-level features and a decoder to reconstruct spatial features and semantic information. In the network, a multi-level spatial attention module (MSAM) introduces the self-attention mechanism to aggregate multi-level contextual information. Specifically, two MSAM are applied in the network, which receives multi-level feature maps from the encoder and embeds long-range spatial dependencies between any two locations of the feature maps. To update the features of each location, MSAM fuses feature maps selected from the encoder by dilated convolution and aggregates features of all locations by using weighted summation. The weights are computed based on the feature similarities between two locations and are normalised by a softmax layer. Therefore, two locations with similar features are likely to be assigned the same label.

The main contributions of our work can be summarised as follows:

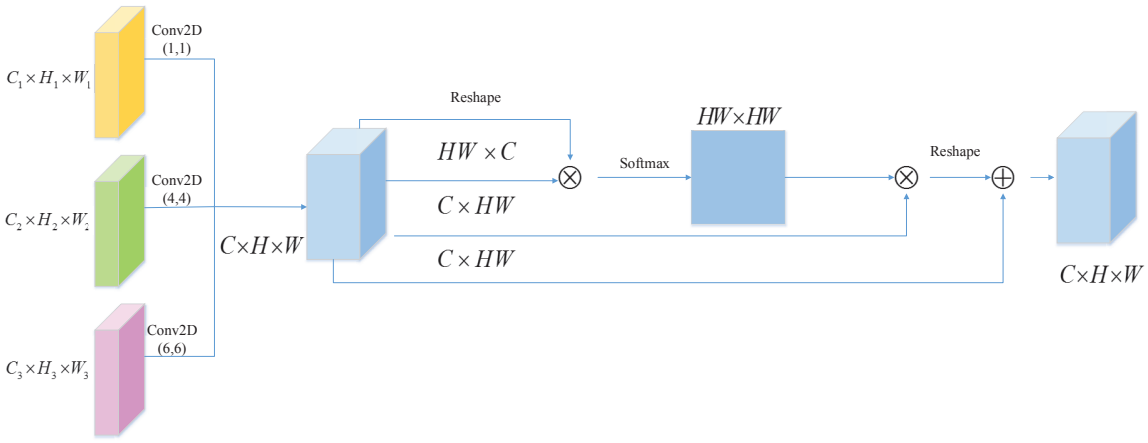
- A new attention-based network, MSANet, is proposed with a multi-level attention mechanism to utilise multi-level features and capture global contextual features.
- The MSAM is present to apply a self-attention mechanism among different level features for learning long-range spatial relationships between features.

Figure 1 The architecture of MASNet (see online version for colours)



Notes: We employ CNNs in the encoder and decoder, with max-pooling layers to downsample and unpooling layers to upsample. Between encoder and decoder, we design a MSAM to aggregate multi-level features.

Figure 2 The MSAM (see online version for colours)



- The proposed network is evaluated on the PASCAL VOC 2012 (Everingham et al., 2015) and Cityscape (Cordts et al., 2016) datasets. The results demonstrate that our attention-based model achieves competitive performance compared with state-of-the-art methods.

2 Related work

2.1 Semantic segmentation

Starting with the groundbreaking work of FCNs (Long et al., 2015), many models based on the deep CNN have been developed for semantic image segmentation. FCNs replaces the last fully connected layers of traditional CNNs with the convolutional layers to make a dense prediction. To get richer context information in a broad receptive field,

DeepLab (Chen et al., 2017b) applies the atrous convolution in the model to enlarge the size of the receptive fields of the networks. Compared with traditional convolution, atrous convolution can learn a set of denser feature maps without increasing the computation and the number of parameters. Considering different objects at various scales, Atrous Spatial Pyramid Pooling (ASPP) has been present. There are multiple parallel atrous convolution layers in ASPP with different dilation rates to learn richer feature maps at multiple scales. Peng et al. (2017) proposed a global convolutional network, which employs a fully convolutional structure to obtain local features and applies a large filter size in the network to make a dense classification. Differently, Zhao et al. (2017) presented a pyramid pooling module to aggregate the region-based context by inputting the extracted feature maps into multiple parallel pooling layer with various kernel sizes. These methods focus on expanding the size of the models' receptive field, which results in losing local information and long-range contextual features.

To capture long-range relationships between pixels, many methods have been proposed in semantic image segmentation. Chen et al. (2017b) utilised a fully connected conditional random field to post-process the classification outputs of the deep convolutional networks. Zheng et al. (2015) ran a fully connected conditional random field as a recurrent neural network and trained all parameters of both CNNs and the conditional random field in an end-to-end manner. Ji et al. (2020) designed a cascaded conditional random field, which receives feature maps from multiple layers in the decoder to refine the object boundaries. Qiu et al. (2020) fused multiple segmentation results of different deep CNNs and used a fully connected conditional random field to infer the label of each pixel. Some works turn to the recurrent neural network for long-range contextual information aggregation. Qiu et al. (2020) applied a VGG16 network to extract local features and stacked multiple recurrent neural network layers to capture global contextual dependencies all over the input image. Yan et al. (2016) utilised multiple spatial recurrent LSTM layers to build spatially relationships across different local areas among the input image. Luo et al. (2019) combined the fully convolutional network with channel attention mechanism (CAM) and applied multiple CAMs between different layers to obtain semantic and spatial location information.

2.2 Encoder-decoder

The encoder-decoder is widely used in image segmentation frameworks, like U-Net (Ronneberger et al., 2015) and DeepLab (Chen et al., 2017a). In this architecture, the encoder usually employs sophisticated CNNs that are powerful in extracting multiple level features by applying multiple convolutional and pooling layers. The decoder is used to reconstruct feature maps with original resolution from low-resolution feature maps that are outputted by the encoder. To utilise low-level features for getting more context information, most decoders utilise skip-connection. Ronneberger et al. (2015) directly concatenated the feature maps in the decoder with the selected feature maps from the encoder. Noh et al. (2015) applied multiple unpooling layers in the decoder to reuse the maximum value location recorded by the corresponding pooling layers in the encoder for resolution recovering.

2.3 Attention mechanism

To model implicit dependencies among the input data, attention mechanism is successfully used in image analysis and NLP for image caption (Xu et al., 2015), language translation (McCann et al., 2017) and classification (Wang et al., 2017) problems. In Wang et al. (2017), generated attention-aware features made it easier to train deep networks with many layers for image classification. In Lee and Kwon (2017), attention mechanism was used to characterise the spectral-spatial information for hyperspectral image classification. You et al. (2016) present a semantic attention model to provide a detailed, coherent description of objects in the image.

For semantic image segmentation, Hu et al. (2019) proposed an attention complementary network (ACNet) to gather features from RBG and depth branches for RGBD semantic segmentation. In the model, the attention complementary module is designed to fuse features selected from previous layers. Hu et al. (2020) used fast spatial attention to get rich spatial context at a small fraction of the computational cost. Zhang et al. (2019) used a decoupled spatial neural attention network for weakly supervised semantic segmentation, which applies the decoupled attention model to identify object regions and find the discriminative parts. Huang et al. (2019) presented criss-cross attention that captures the contextual dependencies for each pixel in its vertical and horizontal direction and builds recurrent criss-cross attention to aggregate global contextual information in the whole image. Li et al. (2018b) designed a feature pyramid attention module to fuse features with different scales and embed spatial context information. Then, global attention upsample module is presented to use global pooling features as guidance for selecting localisation characteristics from low-level features. Chen et al. (2016) directly input multi-size images to generate multi-scale feature maps and applied an attention mechanism to embed multi-scale spatial context for pixel-wise classification. Huang et al. (2017) provided the reverse attention network to capture the information of reverse-category in the confusing area. Liu et al. (2020) presented a CANet that introduces a covariance attention mechanism to model global dependencies for each location and channel by using the covariance matrix.

Different from these previous works, our proposed model, MSANet, exploits CNNs and two MSAMs to aggregate multi-scale spatial contextual information from multiple selected low-level features in the encoder for generating accurate segmentation results. Moreover, the proposed MSAM gives a global view of the input feature maps by equipping a large size of the receptive field.

Figure 3 Example of semantic segmentation results on PASCAL VOC 2012 (see online version for colours)

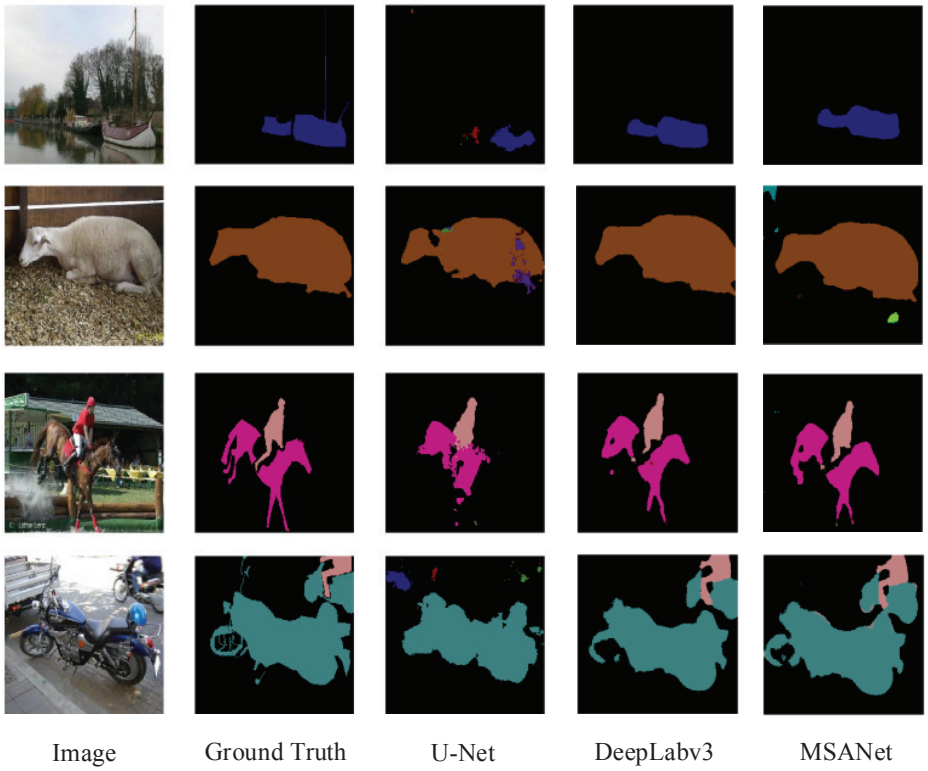
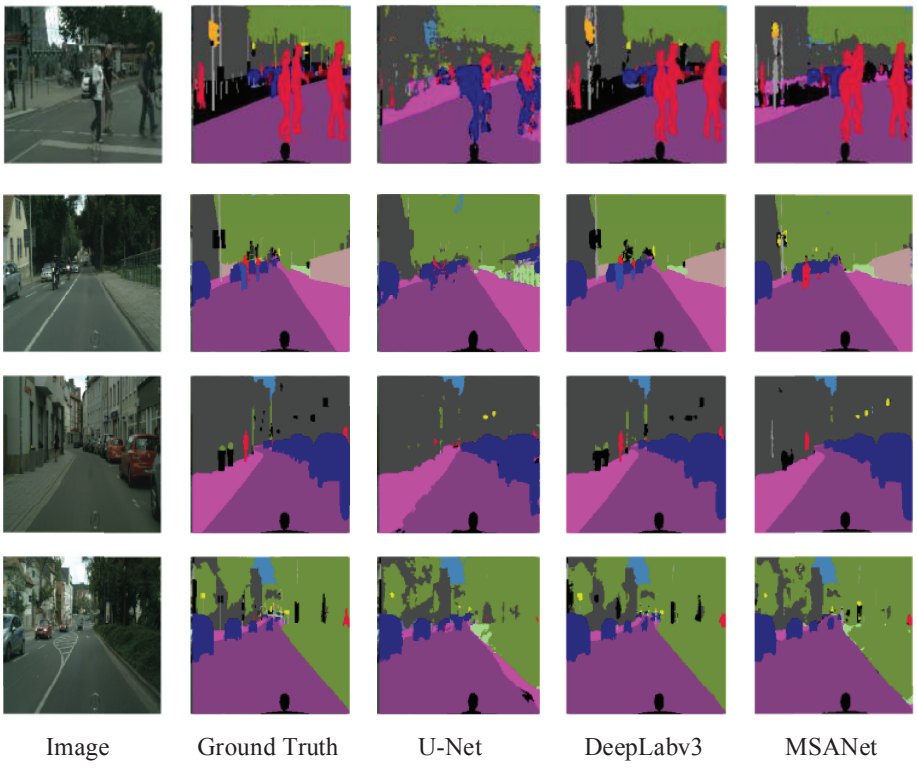


Figure 4 Example of semantic segmentation results on Cityscapes (see online version for colours)



3 Methods

This section describes the overall architecture of the proposed network for semantic image segmentation. The MSAM is then introduced.

3.1 Multi-level spatial attention network

The proposed network employs an encoder-decoder architecture, as shown in Figure 1. The proposed encoder

is composed of four convolutional blocks, and each block contains three convolutional layers with 3 × 3 filter sizes and a pooling layer. The first three pooling layers have a 2 × 2 filter size to learn representations, increase the size of the receptive field, introduce invariant, and reduce the number of parameters. The last pooling layer of the encoder has a 3 × 3 filter size to get a large size of the receptive field and sets the stride to 1 to keep the size of feature maps for preserving spatial information. Given an input image with resolution $H \times W$, each of the first three convolutional blocks in the encoder downsamples the feature maps by a factor of 2. The final feature maps have a size of $\frac{H}{8} \times \frac{W}{8}$.

Table 1 Detailed configuration of the proposed encoder network

Blocks	*Layers*	*Output size*	*Kernel*	*Stride*	*Pad*
Block 1	Conv2D	64 × 224 × 224	3 × 3	1	1
	Conv2D	64 × 224 × 224	3 × 3	1	1
	Conv2D	64 × 224 × 224	3 × 3	1	1
	Max-pooling	64 × 112 × 112	2 × 2	2	0
Block 2	Conv2D	128 × 112 × 112	3 × 3	1	1
	Conv2D	128 × 112 × 112	3 × 3	1	1
	Conv2D	64 × 224 × 224	3 × 3	1	1
	Max-pooling	128 × 56 × 56	2 × 2	2	0
Block 3	Conv2D	256 × 56 × 56	3 × 3	1	1
	Conv2D	256 × 56 × 56	3 × 3	1	1
	Conv2D	256 × 56 × 56	3 × 3	1	1
	Max-pooling	256 × 28 × 28	2 × 2	2	0
Block 4	Conv2D	512 × 28 × 28	3 × 3	1	1
	Conv2D	512 × 28 × 28	3 × 3	1	1
	Conv2D	512 × 28 × 28	3 × 3	1	1
	Max-pooling	512 × 28 × 28	3 × 3	1	1

Notes: 'Conv2D' denotes the convolutional layer. ReLU layers and batch normalisation layers are omitted from the table for brevity.

The proposed corresponding decoder also has four convolutional blocks, and each block contains three convolutional layers with 3 × 3, 1 × 1, and 3 × 3 filter size. Three unpooling layers are inserted in the decoder. Each unpooling layer can upsample the feature maps by a factor of 2 to reconstruct spatial characteristics, which puts values in the feature maps to the locations recorded in the corresponding pooling layer. To make the final pixel-level classification, the output feature maps of the decoder are input into a softmax layer to generate probabilistic maps that indicate the probability of each pixel belongs to each semantic category.

There are two MSAMs applied in the network to fuse the multi-scale feature maps from different layers in the encoder, embed low-level details, and capture long-range dependencies between any two positions. The first module selects two sets of low-level feature maps from the encoder as input, while the second module selects three sets of low-level feature maps. Each set of feature maps is the output of the activation function of the last convolutional layer of the corresponding convolution block. The outputs of each MSAM are cascaded with the feature maps in the decoder to enhance the feature representation and make a dense classification.

3.2 Multi-level spatial attention module

Due to multiple stacked convolutional and pooling layers, fully convolutional networks are limited in a fixed size of the receptive field, which results in focusing on local features and losing spatial and long-range contextual information. This inherent problem causes the inconsistency and misclassification in the segmentation result and the blurry of the boundaries.

To capture long-range spatial relationships between any two positions in the feature map without losing location information, this paper presents a MSAM. The proposed MSAM fuses multi-scale feature maps selected from the encoder and captures long-range contextual information from these feature maps by modelling dependencies between any two positions in the feature map.

As shown in Figure 2, the second MSAM is described. The sizes of three input feature maps are $C_1 \times H_1 \times W_1$, $C_2 \times H_2 \times W_2$, and $C_3 \times H_3 \times W_3$, respectively. These feature maps are then input into three convolutional layer with different size of filter and stride to have a same size. The outputs are cascaded with the size of $C \times H \times W$. These feature maps are feed into four parallel convolutional layers to produce four new feature maps $A \in R^{C\times H\times W}$, $B \in R^{C\times H\times W}$, $C \in R^{C\times H\times W}$, and $D \in R^{C\times H\times W}$. A and B are reshaped to $R^{C\times N}$, where $N = H \times W$. We then multiply the transpose of A by B, and use a softmax layer to compute the multi-level spatial attention map $S \in R^{N\times N}$:

$$s_{ij} = \frac{\exp(A_i \cdot B_j)}{\sum_{i=1}^{N} \exp(A_i \cdot B_j)} \tag{1}$$

where s_{ij} describes the dependency between the i^{th} position and j^{th}. The greater the value of s_{ij}, the more similar feature representations of the two location are, and the more likely the two location has the same label.

The feature map C is also reshaped to $R^{C\times N}$. We then multiply C by the transpose of S, and reshape the outputs to $R^{C\times H\times W}$. Afterward, we perform a element-wise sum between these outputs and the feature maps D to get the final outputs $E \in R^{C\times H\times W}$.

$$E_j = \sum_{i=1}^{N} C_i s_{ij} + D_j \tag{2}$$

It can be known that feature at each location in E is the weighted sum of the features at all locations. Therefore, the network has a global contextual view of the feature maps and capture contextual information according to the spatial dependencies between features.

4 Experiments

The architecture of our proposed model is shown in Figure 1. To evaluate the proposed model, experiments

are carried out on the PASCAL VOC 2012 (Everingham et al., 2015) and Cityscapes dataset (Cordts et al., 2016). The results have shown that the proposed model achieves outstanding segmentation performance in qualitative and quantitative on two datasets. The next subsections introduce the datasets and implementation details. Then experiments on the PASCAL VOC 2012 and Cityscapes datasets are present. Finally, the results of the two datasets are reported.

Table 2 Detailed configuration of the proposed decoder network

Blocks	*Layers*	*Output size*	*Kernel*	*Stride*	*Pad*
Block 1	Conv2D	1,024 × 28 × 28	3 × 3	1	1
	Conv2D	512 × 28 × 28	1 × 1	1	0
	Conv2D	256 × 28 × 28	3 × 3	1	1
	Unpooling	256 × 56 × 56	2 × 2	2	0
Block 2	Conv2D	512 × 56 × 56	3 × 3	1	1
	Conv2D	256 × 56 × 56	1 × 1	1	0
	Conv2D	128 × 56 × 56	3 × 3	1	1
	Unpooling	128 × 112 × 112	2 × 2	2	0
Block 3	Conv2D	256 × 112 × 112	3 × 3	1	1
	Conv2D	128 × 112 × 112	1 × 1	1	0
	Conv2D	64 × 112 × 112	3 × 3	1	1
	Unpooling	64 × 224 × 224	2 × 2	2	0
Block 4	Conv2D	128 × 224 × 224	3 × 3	1	1
	Conv2D	64 × 224 × 224	1 × 1	1	0
	Conv2D	21 × 224 × 224	3 × 3	1	1

Notes: 'Conv2D' denotes the convolutional layer. ReLU layers and batch normalisation layers are omitted from the table for brevity.

4.1 Datasets

- *PASCAL VOC 2012:* The original dataset contains 1,464 images for training and 1,449 for validation. 90% of the training and validation images are selected and augmented by flipping the image horizontally and vertically for training. 10% of the training and validation images are selected for testing. The dataset has one background category and 20 foreground categories, such as airplane, bicycle, boat, bus, car, motorbike, train, bottle, chair, dining table, potted plant, sofa, TV/monitor, bird, cat, cow, dog, horse, sheep, and person, that are pixel-level labelled.
- *Cityscapes:* The dataset contains 5,000 images captured from 50 cities for semantic understanding of urban street scenes. There are 2,979 training images, 500 validating images, and testing 1,525 images. Each image has 2,048 × 1,024 pixels. These pixels are annotated with one background class and 19 semantic classes, such as flat surfaces, humans, vehicles, constructions, objects, nature, sky, etc. This dataset is also augmented by flipping images horizontally and vertically for training.

4.2 Implementation details

The proposed model is implemented in the Pytorch framework. The architecture is illustrated in Figure 1.

Table 3 Detailed configuration of the proposed MSAM

Modules	*Layers*	*Output size*	*Kernel*	*Stride*	*Pad*
Module 1	Conv2D	64 × 56 × 56	12 × 12	4	4
	Conv2D	64 × 56 × 56	6 × 6	2	2
	Conv2D	128 × 56 × 56	3 × 3	1	1
	Conv2D	256 × 56 × 56	1 × 1	0	1
	Conv2D	256 × 56 × 56	1 × 1	0	1
	Conv2D	256 × 56 × 56	1 × 1	0	1
	Conv2D	256 × 56 × 56	1 × 1	0	1
Module 1	Conv2D	64 × 112 × 112	6 × 6	2	2
	Conv2D	64 × 112 × 112	3 × 3	1	1
	Conv2D	128 × 112 × 112	1 × 1	0	1
	Conv2D	128 × 112 × 112	1 × 1	0	1
	Conv2D	128 × 112 × 112	1 × 1	0	1
	Conv2D	128 × 112 × 112	1 × 1	0	1

Notes: 'Conv2D' denotes the convolutional layer. ReLU layers and batch normalisation layers are omitted from the table for brevity.

Table 4 Segmentation results on the PASCAL VOC 2012 dataset

Method	*mIoU*	*Mean accuracy*
FCNs	62.7%	90.3%
U-Net	67.1%	92.6%
DeepLabv1	71.6%	93.2%
DeepLabv3	77.5%	95.1%
MSANet without attention	72.3%	92.8%
MSANet	78.1%	94.6%

Table 5 Segmentation results on the Cityscapes dataset

Method	*mIoU*	*Mean accuracy*
FCNs	59.3%	87.5%
U-Net	63.4%	88.6%
DeepLabv1	70.4%	91.3%
DeepLabv3	76.8%	94.5%
MSANet without attention	73.0%	91.4%
MSANet	76.3%	93.1%

The detailed configurations of the encoder and the decoder are shown in Tables 1 and 2, respectively. Both the proposed encoder and decoder are initialised with zero-mean Gaussian. The input image is firstly fed into the encoder to learning hierarchical feature representations that are downsampled to $\frac{1}{8}$ resolution of the original input image. Therefore, the low, fine feature maps are transferred to high, coarse feature maps by performing a lot of convolution and pooling operators. These final local feature maps produced by the encoder go through the decoder to reconstruct semantic and spatial information. In the decoder, three unpooling layers are utilised to recover the resolution of feature maps by upsampling feature maps to sparse feature maps. Then, a group of

convolutional layers is applied to produce a set of dense feature maps. The proposed MSAM is used to embed global contextual information into the local feature maps that are concatenated with the feature maps in the decoder. The output feature maps are input into a softmax layer to generate probability maps that represent the probability of each pixel assigned to every predefined category.

Two MSAM are applied in the network. The detailed configurations of these two modules are illustrated in Table 3. To obtain the same resolution of the feature maps, the selected feature maps from the encoder, which have different sizes, are feed to two or three parallel convolutional layers with different filter sizes and strides. These convolutional layers are equipped with larger filter sizes for low-level feature maps to increase the size of the receptive field. Therefore, the module captures multi-scale feature maps from multiple levels in the encoder. These multi-scale feature maps are also inputted into four parallel convolutional layers and reshaped. To generate the multi-level spatial attention map, matrix multiplication is performed between two feature matrices, and the result is feed into a softmax layer to be normalised. The normalised multi-level spatial attention map multiply by another feature matrix to utilise the spatial dependencies between features. To output the final feature maps, these feature maps are element-wise summed with the fourth feature maps. At last, these final feature maps are cascaded with the feature maps from the corresponding layer in the decoder.

We conduct all our experiments in python, using Pytorch. An Nvidia GeForce TITAN X is used for all experiments. The Adam optimiser is used to fine-tune the network on the pixel-classification task on the cross-entropy loss function. We use an initial learning rate of 0.001, multiplying by $\left(1-\frac{iter}{total\_iter}\right)^{0.9}$ after each iteration. The batch size of training and testing is 4. We reshape all the images to the size of 250 $\times$ 250. During training, we randomly crop the input images to the size 224 $\times$ 224. The maximum training time is set to 200 epochs.

The performance of our network is measured by two metrics:

- *Mean accuracy:* This metric outputs the average prediction accuracy over all classes. It can be defined as

$$MeanAccuracy = \frac{1}{N_{cls}} \sum\nolimits_{i=1}^{N_{cls}} \frac{N_{ii}}{\sum_{j=1}^{N_{cls}} N_{ij}} \tag{3}$$

  where N_{cls} is the total number of categories. N_{ii} is the number of pixels belonging to class i and are labelled as class i. N_{ij} is the number of pixels belonging to class i and are labelled as class j.

- *Mean IoU:* This metric, also referred to as the Jaccard index, is essentially a method to quantify the percent overlap between the target mask and the prediction output. It can be defined as

$$mIoU = \frac{1}{N_{cls}} \sum\nolimits_{i=1}^{N_{cls}} \frac{N_{ii}}{\sum_{j=1}^{N_{cls}} N_{ij} + \sum_{j=1}^{N_{cls}} N_{ji} - N_{ii}} \tag{4}$$

  where N_{ii} is the number of true positives, N_{ij} is the number of false positives, and N_{ji} it the number of false negatives. mIoU is the most widely used evaluation metric.

We also compare our model with other approaches such as FCNs (Long et al., 2015), U-Net (Ronneberger et al., 2015), DeepLabv1 (Chen et al., 2017b) and DeepLabv3 (Chen et al., 2017a). The performance has shown that our model gets better results, as illustrated in Tables 4 and 5. To qualitative evaluate our model, some prediction of our model is shown in Figures 3 and 4. From the prediction of our model, we can see that our model well reconstruct the shapes and boundaries of objects.

4.3 Evaluation on the PASCAL VOC 2012 dataset

The proposed MSANet is compared with other approaches such as FCNs (Long et al., 2015), U-Net (Ronneberger et al., 2015), DeepLabv1 (Chen et al., 2017b), and DeepLabv3 (Chen et al., 2017a), where FCNs, U-Net, and DeepLabv1 are baseline methods, and DeepLabv3 is one of the state-of-the-art methods. These three approaches are initialised by their pre-trained models and fine-tuned on our preprocessed image datasets. As shown in Table 4, MSANet obtains 93.6% mean pixel-accuracy, where FCNs get 94.6%, U-Net gets 92.6%, DeepLabv1 gets 93.2%, and DeepLabv3 gets 95.1%. These models achieve similar performance in mean pixel-accuracy. However, MSANet obtains 78.1% mean pixel-accuracy, where FCNs get 63.7%, U-Net gets 67.1%, DeepLabv1 gets 71.6%, and DeepLabv3 gets 77.5%. MSANet performs better than other models evaluated in mIoU. MSAM is also evaluated by comparing the MSANet with the network that doesn't apply the attention mechanism. The MSANet without attention mechanism obtains 92.8% in mean pixel-accuracy and 72.3% in mIoU. Therefore, MASM improves 5.8% in mIoU and 1.8% in mean pixel-accuracy on the PASCAL VOC 2012 dataset.

To qualitatively evaluate these models, some segmentation results of MSANet and other methods are shown in Figure 3. In the first row of Figure 3, U-Net only locates and recognises a small part of the area of two boats. Moreover, U-Net makes wrong classifications of the area among the area of the small boat. However, MSANet and DeepLabv3 segment the semantic area of two boats with complete shape, clear boundary, and correct label. In the second row, U-Net outlines the area of the sheep. But there are some areas labelled with incorrect labels. Although making some misclassification of the areas in the upper left and lower right corner, MSANet maintains consistency in the semantic area of the sheep with smooth boundary and complete structure. In the third row, U-Net mixed the area of the human and the horse, and the segmented area of the horse is fully deformed. The segmentation result of the MSANet has separated the two semantic areas, and the shape and structure of the human and horse are complete. In the fourth row, U-Net completely regardless of the semantic area in the

upper right corner. MSANet has segmented the human and two motorcycles with clear boundaries. Therefore, these examples of segmentation results validate that MSANet can segment the semantic areas with complete shapes, smooth boundaries, and consistency.

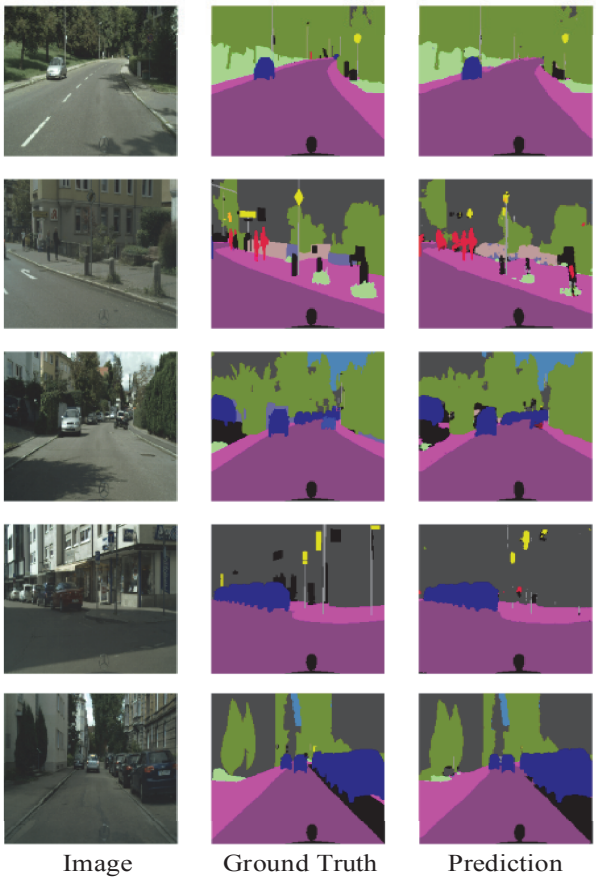
Figure 5 Segmentation results produced by MSANet on the PASCAL VOC 2012 dataset (see online version for colours)

4.4 *Evaluation on the Cityscapes dataset*

Experiments are conducted on the Cityscapes dataset to further evaluate the performance of our model. As shown in Table 5, MSANet obtains 93.1% mean pixel-accuracy and 76.3% mIoU that outperform all baseline methods and can be competitive with DeepLabv3. In addition, these quantitative results also verify that MASM improves 3.3% in mIoU and 1.7% in mean pixel-accuracy on the Cityscapes dataset.

Some segmentation results produced by U-Net, DeepLabv3, and MSANet are visualised in Figure 4. In the first row of Figure 4, U-Net classifies the area of the human as a car and regardless of the traffic light. However, MSANet accurately locates and recognises the human and the light. In the third row, U-Net still ignores human on the left. MSANet has segmented the small semantic area of the human. These segmentation results prove that MSANet can precisely locate and classify objects, including small objects such as humans and traffic lights.

Figure 6 Segmentation results produced by MSANet on the Cityscapes dataset (see online version for colours)



4.5 *Discussion*

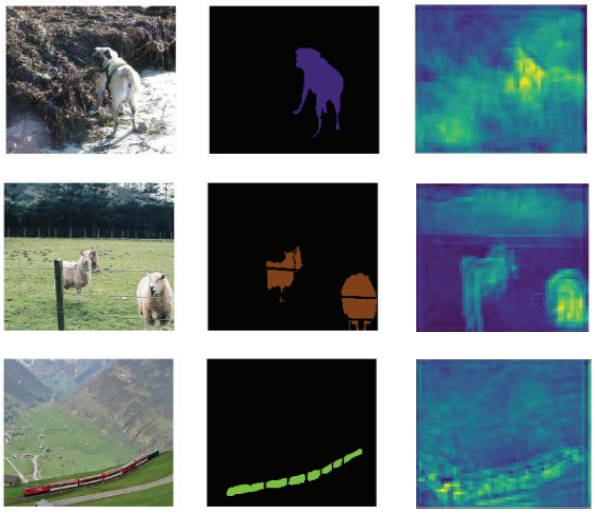
To further evaluate the effectiveness of the MSANet, more examples of segmentation results produced by MSANet are provided. As seen in Figure 5, MSANet can segment persons, chairs, buses, sofas, and motorcycles. But some areas of the sofa under the person are classified as chair in the fourth row of Figure 5. And some separated areas of the sofa are connected in the results. These results verify that MSANet can precisely segment the semantic areas and always maintain consistency in the results. But MASNet also makes mistakes when encountering complex structures such as the legs of the sofa and areas around the table in the fourth row of Figure 5.

The segmentation results in Figure 6 validate that our model can efficiently deal with small objects. For example, the traffic lights in the fourth row, the little car in the first row, and persons in the second row are all recognised. But some thin and long objects like the poles in the fourth row of the Figure 6 and the mast in the first row of the Figure 3. The reason for the phenomenon is that the number of pixels in the areas of these objects is insufficient for learning local feature representations for pixel-wise classification.

As shown in Figure 7, some activation maps randomly selected from the MSAM are visualised to verify the capability of the MSAM. As seen in Figure 7, the highlight areas precisely locate and roughly outline semantic objects

in the image, such as the dog, train, and sheep. The widely distributed highlight areas have validated MASM's ability to capture long-range contextual information.

Figure 7 Visualisation of MSAM (see online version for colours)



5 Conclusions

This paper proposes an attention-based network, MSANet, for semantic image segmentation, which applies an encoder-decoder structure to extract multi-level local features and reconstruct spatial information. Especially, the model utilises the MSAM to fuse multi-level features selected from different layers in the encoder and capture global spatial dependencies between local features. The experiments have demonstrated that the MSAM can deal with long-range contextual information efficiently, keep consistency in the segmentation results, and improve the segmentation performance. Our model has achieved outstanding performance on the PASCAL VOC 2012 and the Cityscapes datasets.

Acknowledgements

This work is supported by Fujian Province Educational Research Projects of Young and Middle-aged Teachers, under Grant 2018J01570 and JAT200818.

References

Byeon, W., Breuel, T.M., Raue, F. and Liwicki, M. (2015) 'Scene labeling with lstm recurrent neural networks', *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.3547–3555.

Cai, Z., Fan, Q., Feris, R.S. and Vasconcelos, N. (2016) 'A unified multi-scale deep convolutional neural network for fast object detection', *European Conference on Computer Vision*, pp.354–370.

Chan, T-H., Jia, K., Gao, S., Lu, J., Zeng, Z. and Ma, Y. (2015) 'PCANet: a simple deep learning baseline for image classification?', *IEEE Transactions on Image Processing*, Vol. 24, No. 12, pp.5017–5032.

Chen, L-C., Yang, Y., Wang, J., Xu, W. and Yuille, A.L. (2016) 'Attention to scale: scale-aware semantic image segmentation', *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.3640–3649.

Chen, L-C., Papandreou, G., Schroff, F. and Adam, H. (2017a) *Rethinking Atrous Convolution for Semantic Image Segmentation*, arXiv [online] https://arxiv.org/pdf/1706.05587.

Chen, L-C., Papandreou, G., Kokkinos, I., Murphy, K. and Yuille, A.L. (2017b) '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.

Chiu, J.P. and Nichols, E. (2016) 'Named entity recognition with bidirectional LSTM-CNNs', *Transactions of the Association for Computational Linguistics*, Vol. 4, No. 2016, pp.357–370.

Chollet, F. (2017) 'Xception: deep learning with depthwise separable convolutions', *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.1251–1258.

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R. and Schiele, B. (2016) 'The cityscapes dataset for semantic urban scene understanding', *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.3213–3223.

Everingham, M., Eslami, S.A., Van Gool, L., Williams, C.K., Winn, J. and Zisserman, A. (2015) 'The pascal visual object classes challenge: a retrospective', *International Journal of Computer Vision*, Vol. 111, No. 1, pp.98–136.

Goodfellow, I., Bengio, Y., Courville, A. and Bengio, Y. (2016) *Deep Learning*, MIT Press, Cambridge.

He, K., Zhang, X., Ren, S. and Sun, J. (2016) 'Deep residual learning for image recognition', *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.770–778.

Hu, X., Yang, K., Fei, L. and Wang, K. (2019) 'AcNet: attention based network to exploit complementary features for RGBD semantic segmentation', *2019 IEEE International Conference on Image Processing (ICIP)*, pp.1440–1444.

Hu, P., Perazzi, F., Heilbron, F.C., Wang, O., Lin, Z., Saenko, K. and Sclaroff, S. (2020) *Real-time Semantic Segmentation with Fast Attention*, arXiv [online] https://arxiv.org/pdf/2007.03815.

Huang, Q., Xia, C., Wu, C., Li, S., Wang, Y., Song, Y. and Kuo, C-C.J. (2017) *Semantic Segmentation with Reverse Attention*, arXiv [online] https://arxiv.org/pdf/1707.06426.

Huang, Z., Wang, X., Huang, L., Huang, C., Wei, Y. and Liu, W. (2019) 'CCCNet: criss-cross attention for semantic segmentation', *Proceedings of the IEEE International Conference on Computer Vision*, pp.603–612.

Ji, J., Shi, R., Li, S., Chen, P. and Miao, Q. (2020) 'Encoder-decoder with cascaded CRFs for semantic segmentation', *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 31, No. 5, pp.1926–1938.

LeCun, Y., Bengio, Y. and Hinton, G. (2015) 'Deep learning', *Nature*, Vol. 521, No. 7553, pp.436–444.

Lee, H. and Kwon, H. (2017) 'Going deeper with contextual CNN for hyperspectral image classification', *IEEE Transactions on Image Processing*, Vol. 26, No. 10, pp.4843–4855.

Li, X., Liu, S., Wu, F., Kumari, S. and Rodrigues, J.J. (2018a) 'Privacy preserving data aggregation scheme for mobile edge computing assisted IoT applications', *IEEE Internet of Things Journal*, Vol. 6, No. 3, pp.4755–4763.

Li, H., Xiong, P., An, J. and Wang, L. (2018b) *Pyramid Attention Network for Semantic Segmentation*, arXiv [online] https://arxiv.org/pdf/1805.10180.

Li, Y., Chen, Y., Wang, N. and Zhang, Z. (2019) 'Scale-aware trident networks for object detection', *Proceedings of the IEEE International Conference on Computer Vision*, pp.6054–6063.

Li, X., Tan, J., Liu, A., Vijayakumar, P., Kumar, N. and Alazab, M. (2020a) 'A novel UAV-enabled data collection scheme for intelligent transportation system through UAV speed control', *IEEE Transactions on Intelligent Transportation Systems*, Vol. 22, No. 4, pp.2100–2110.

Li, X., Liu, T., Obaidat, M.S., Wu, F., Vijayakumar, P. and Kumar, N. (2020b) 'A lightweight privacy-preserving authentication protocol for VANETs', *IEEE Systems Journal*, Vol. 14, No. 3, pp.3547–3557.

Liang, W., Xingming, S., Zhiqiang, R., Jing, L. and Chengtao, W. (2011) 'A sequential circuit-based IP watermarking algorithm for multiple scan chains in design-for-test', *Radioengineering*, Vol. 20, No. 2, pp.533–539.

Liang, W., Liao, B., Long, J., Jiang, Y. and Peng, L. (2016) 'Study on PUF based secure protection for IC design', *Microprocessors and Microsystems*, Vol. 45, No. PA, pp.56–66.

Liang, W., Li, K-C., Long, J., Kui, X. and Zomaya, A.Y. (2019) 'An industrial network intrusion detection algorithm based on multifeature data clustering optimization model', *IEEE Transactions on Industrial Informatics*, Vol. 16, No. 3, pp.2063–2071.

Liang, W., Huang, W., Long, J., Zhang, K., Li, K-C. and Zhang, D. (2020a) 'Deep reinforcement learning for resource protection and real-time detection in IoT environment', *IEEE Internet of Things Journal*, Vol. 7, No. 7, pp.6392–6401.

Liang, W., Fan, Y., Li, K-C., Zhang, D. and Gaudiot, J-L. (2020b) 'Secure data storage and recovery in industrial blockchain network environments', *IEEE Transactions on Industrial Informatics*, Vol. 16, No. 3, pp.2063–2071.

Liu, Y., Chen, Y., Lasang, P. and Sun, Q. (2020) 'Covariance attention for semantic segmentation', *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Long, J., Shelhamer, E. and Darrell, T. (2015) 'Fully convolutional networks for semantic segmentation', *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.3431–3440.

Lopez, M.M. and Kalita, J. (2017) *Deep Learning Applied to NLP*, arXiv [online] https://arxiv.org/pdf/1703.03091.

Luo, H., Chen, C., Fang, L., Zhu, X. and Lu, L. (2019) 'High-resolution aerial images semantic segmentation using deep fully convolutional network with channel attention mechanism', *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, Vol. 12, No. 9, pp.3492–3507.

McCann, B., Bradbury, J., Xiong, C. and Socher, R. (2017) 'Learned in translation: contextualized word vectors', *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp.6297–6308.

Mostajabi, M., Yadollahpour, P. and Shakhnarovich, G. (2015) 'Feedforward semantic segmentation with zoom-out features', *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.3376–3385.

Noh, H., Hong, S. and Han, B. (2015) 'Learning deconvolution network for semantic segmentation', *Proceedings of the IEEE International Conference on Computer Vision*, pp.1520–1528.

Peng, C., Zhang, X., Yu, G., Luo, G. and Sun, J. (2017) 'Large kernel matters – improve semantic segmentation by global convolutional network', *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.4353–4361.

Qiu, Y., Cai, J., Qin, X. and Zhang, J. (2020) 'Inferring skin lesion segmentation with fully connected CRFs based on multiple deep convolutional neural networks', *IEEE Access*, Vol. 8, pp.144246–144258.

Ronneberger, O., Fischer, P., Brox, T. (2015) 'U-net: Convolutional networks for biomedical image segmentation', *International Conference on Medical image computing and computer-assisted intervention*, pp.234-241.

Szegedy, C., Toshev, A. and Erhan, D. (2013) 'Deep neural networks for object detection', *Proceedings of the 26th International Conference on Neural Information Processing Systems*, Vol. 2, pp.2553–2561.

Visin, F., Ciccone, M., Romero, A., Kastner, K., Cho, K., Bengio, Y. and Courville, A. (2016) 'ReSeg: a recurrent neural network-based model for semantic segmentation', *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp.41–48.

Wang, F., Jiang, M., Qian, C., Yang, S., Li, C., Zhang, H. and Tang, X. (2017) 'Residual attention network for image classification', *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.3156–3164.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R. and Bengio, Y. (2015) 'Show, attend and tell: neural image caption generation with visual attention', *International Conference on Machine Learning*, pp.2048–2057.

Yan, Z., Zhang, H., Jia, Y., Breuel, T. and Yu, Y. (2016) *Combining the Best of Convolutional Layers and Recurrent Layers: A Hybrid Network for Semantic Segmentation*, arXiv [online] https://arxiv.org/pdf/1603.04871.

You, Q., Jin, H., Wang, Z., Fang, C. and Luo, J. (2016) 'Image captioning with semantic attention', *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.4651–4659.

Zhang, H., Dana, K., Shi, J., Zhang, Z., Wang, X., Tyagi, A. and Agrawal, A. (2018) 'Context encoding for semantic segmentation', *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.7151–7160.

Zhang, T., Lin, G., Cai, J., Shen, T., Shen, C. and Kot, A.C. (2019) 'Decoupled spatial neural attention for weakly supervised semantic segmentation', *IEEE Transactions on Multimedia*, Vol. 21, No. 11, pp.2930–2941.

Zhao, H., Shi, J., Qi, X., Wang, X. and Jia, J. (2017) 'Pyramid scene parsing network', *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.2881–2890.

Zheng, S., Jayasumana, S., Romera-Paredes, B., Vineet, V., Su, Z., Du, D. and Torr, P.H. (2015) 'Conditional random fields as recurrent neural networks', *Proceedings of the IEEE International Conference on Computer Vision*, pp.1529–1537.