Journal of Information Systems and Telecommunication
Vol.11, No.4, October-December 2023, 359-367
http://jist.acecr.org
ISSN 2322-1437 / EISSN:2345-2773

Comparing the Semantic Segmentation of High-Resolution Images Using Deep Convolutional Networks: SegNet, HRNet, CSE-HRNet and RCA-FCN

Nafiseh Sadeghi[1,2], Homayoun Mahdavi-Nasab[*1,2], Mansoor Zeinali[1,2], Hossein Pourghassem[1,2]

[1].Department of Electrical Engineering, Najafabad Branch, Islamic Azad University, Najafabad, Iran
[2].Digital Processing and Machine Vision Research Center, Najafabad Branch, Islamic Azad University, Najafabad, Iran

Received: 11 Oct 2022/ Revised: 04 Mar 2023/ Accepted: 25 Apr 2023

Abstract

Semantic segmentation is a branch of computer vision, used extensively in image search engines, automated driving, intelligent agriculture, disaster management, and other machine-human interactions. Semantic segmentation aims to predict a label for each pixel from a given label set, according to semantic information. Among the proposed methods and architectures, researchers have focused on deep learning algorithms due to their good feature learning results. Thus, many studies have explored the structure of deep neural networks, especially convolutional neural networks. Most of the modern semantic segmentation models are based on fully convolutional networks (FCN), which first replace the fully connected layers in common classification networks with convolutional layers, getting pixel-level prediction results. After that, a lot of methods are proposed to improve the basic FCN methods results. With the increasing complexity and variety of existing data structures, more powerful neural networks and the development of existing networks are needed. This study aims to segment a high-resolution (HR) image dataset into six separate classes. Here, an overview of some important deep learning architectures will be presented with a focus on methods producing remarkable scores in segmentation metrics such as accuracy and F1-score. Finally, their segmentation results will be discussed and we would see that the methods, which are superior in the overall accuracy and overall F1-score, are not necessarily the best in all classes. Therefore, the results of this paper lead to the point to choose the segmentation algorithm according to the application of segmentation and the importance degree of each class.

Keywords: Semantic Segmentation; Convolutional Neural Network; Deep Neural Network; High-Resolution Image Processing.

1- Introduction

Segmentation is the task process of assigning a label to every pixel in the image, based on features such as pixel intensity, color, texture, etc [1]. Nowadays, the subject of interest is semantic segmentation, predict the semantic category of each pixel from a given label set.

There are learning and anti-learning methods frequently used for segmentation [2]. Anti-learning methods, typically include graph cuts, level set, region growing, etc. and learning methods include fuzzy, neural, genetic algorithms and derivations [3]. Various learning methods have been created and developed in recent years, due to their considerable success in learning a hierarchy of features from high to low [2,4]. These methods were inspired by human brain's ability to receive, learn, and organize input information, especially visual data [5].

Convolutional neural network (CNN) is a form of learning techniques, in which local neighborhood pooling operations and trainable filters are alternatingly applied on the input images, resulting in a hierarchy of increasingly complex features [6-8]. Convolution layers in CNN try to find patterns in an image by convolving over it. So CNN may detect nonlinear mappings between the inputs and outputs [9].

LeCun et al. (1998) introduced the first structure of convolutional neural networks named LenNet. In the same year, they received an award for simulating their proposed network on the ImageNet dataset. LeNet had six

✉ **Homayoun Mahdavi Nasab**
mahdavinasab@iaun.ac.ir

convolutional layers, a pooling layer, and two FC[1] layers [10].

After introducing basic convolutional neural network architectures, researches in this field were continued in two directions: some studies focused on designing new convolutional network architectures and others focused on implementing techniques and strategies to optimize existing architectures [11,12]. In the following, we will introduce some of the most important architectures that were proposed after the creation of convolutional neural network.

In 2012, Krizhevsky et al. proposed a CNN structure called AlexNet with five convolutional layers, three pooling layers, two normalization layers, and three FC layers [13]. The innovation of AlexNet was its use of ReLU to reduce training time. Some have criticized this structure for implementing very heavy data augmentation.

Then, Simonyan et al. (2014) designed a deeper network named VGG with smaller filter sizes [14]. They design two structures of VGG with 16 and 19 layers. The proposed structure showed promising performance and could also be generalized to other datasets. Although the architectures introduced so far focused on window size and smaller steps in the first convolutional layer, VGG focused on an important aspect of convolution neural networks, called depth [15].

The VGG architecture included 1×1 convolutional filters, acting as linear transformation of the input. All hidden layers of the VGG had a ReLU unit for reducing training time. To have no change in spatial resolution after convolving, this architecture kept the convolution step fixed on one pixel. VGG had three FC layers follow a stack of convolutional layers: the first two have 4096 channels each, the third performs 1000-way ILSVRC classification and thus contains 1000 channels (means a channel for each class). The final last layer is soft-max.

Since its developers believed that localized response normalization (LRN) increased memory consumption and training time with no significant resolution improvement, VGG did not use LRN.

In the same year, Szegedy et al. designed a deeper and more computationally-efficient network called GoogleNet [16]. It had twenty-two layers, no FC layer, and a new module called Inception to increase efficiency. The developers claimed that it was twelve times faster than AlexNet with increased depth and width at the same computational cost. Although there were benefits to the increased size, it also increased the number of parameters. This made the network more susceptible to overfitting, especially with a limited number of labeled samples in the dataset.

SegNet was an important architecture proposed in 2015 on a set of camera images [17]. The SegNet architecture was based on encoder-decoder network with thirteen convolutional layers in the VGG16. Since the decoder part of SegNet is identical to the VGG, it was possible to achieve the pre-training benefits in this architecture. The decoder block consisted of five sub-blocks, each with convolutional layers and a downsample layer. Likewise, the decoding block had five sub-blocks with deconvolutional layers and an upsample layer. In fact, the innovation of the SegNet architecture was its use of upsampling layers (reconstructing the image in the original dimensions). In terms of memory use, accuracy, and reducing network parameters, the SegNet architecture demonstrated excellent performance compared to other architectures [18].

HRNet was a successful network that used a parallel integration strategy [19]. The first stage in this network was a high-resolution subnetwork, then high-to-low resolution subnetworks one after another adding to form other stages. The multi-resolution subnetworks were connected in parallel. Each high-to-low resolution representation gets information from other parallel representations, again and again, resulting into a rich high-resolution representation. The convolutional layers in this network were placed in parallel from high to low accuracy [20]. The network had a main subnetwork that produced feature maps with the same accuracy and a series of step-by-step convolutions that reduced accuracy. This network has a multiresolution composition.

After HRNet, Wang proposed its enhanced model built on the backbone network of HRNet. It used NDRB as the generic extractor for multi-scale contextual features. So CSE-HRNet could resolve intra-class heterogeneity and inter-class homogeneity.

Another recent architecture is RCA-FCN consisted of two network units, namely the spatial relation module and the channel relation module [21]. These two modules learn and reason about global relationships between any two spatial positions or feature maps. So they produce RA[2] feature representations. In other word, this model convolutions combine spatial information and channel relation information to record both spatial and channel relations.

Here we aimed to compare segmentation performance with four recent successful methods, SegNet, HRNet, CSE-HRNet, and RCA-FCN. We tried to reference important prior contributions that claim to be particularly successful despite being simple and convolutional-based. This paper is useful to choose the best algorithm for semantic segmentation in our desired application. Thus, section 2 will discuss semantic segmentation and introduce the four structures in more detail. Section 3 will present the implementation and the dataset in more details, and the

[1] Fully-Connected

[2] Relation-Augmented

segmentation results will be compared. Finally, section 4 will present a summary of results.

2- Semantic Segmentation of ISPRS Images

The Vaihingen 2-D semantic segmentation dataset of ISPRS includes 33 images with 3-10 million pixels. These are True Ortho Photos (TOP) taken from Vaihingen, Germany. The ground sampling distance is 9 cm and all pixels in these images are labeled in six classes: building, car, road, tree, low veg, and clutter. The images are eight-bit .tiff files with three bands corresponding to green, red, and near-infrared.

In the following, four new and important semantic segmentation methods for high-resolution remote sensing images will be evaluated. All of these methods lead to six class segmentation, as shown in Fig. 1.

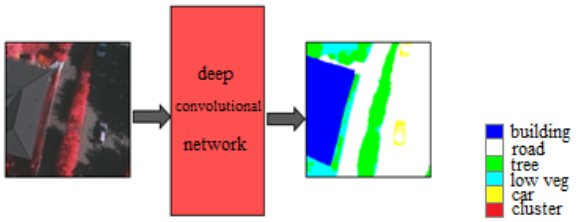

Fig. 1 Six Class Segmentation of Input Dataset

2-1- SegNet

The SegNet architecture was first introduced in 2015 for semantic segmentation on a set of camera images. Its topology was based on a decoder network with 13 convolutional layers in the VGG16 network. As with VGG16, this architecture can also achieve the benefits of pre-learning.

The SegNet topology is comprised of two parts: encoder and decoder. Each encoder performs convolution with a filter bank to produce a set of feature maps, and then they are batch normalized [22,23]. After that, an element-wise rectified linear activation function (ReLU), max(0, x), is applied. Next, a non-overlapping max-pooling, with the window size 2×2 and stride 2, is performed. The output is sub-sampled by a factor of 2. The purpose of using max-pooling is to achieve translation invariance over small spatial shifts in the input image.

Then the decoder upsamples the input feature maps using the memorized max-pooling indices from the corresponding encoder feature map. So sparse feature maps are produced, and then they are convolved with a trainable decoder filter bank. In this way dense feature maps will be produced. So the decoder upsamples and normalizes the stored feature maps. The softmax layer classifies each pixel independently and its output is an image with k channels, where k represents the number of classes.

Fig. 2 shows the schematic representation of the SegNet structure. Determination of boundaries was a success in SegNet architecture. It also showed great performance in terms of the number of network parameters [24], and the most important feature was its memory requirement, which was significantly lower than previous architectures. Therefore, due to its ability to quickly process a large area, SegNet matters when large-scale processing is necessary.

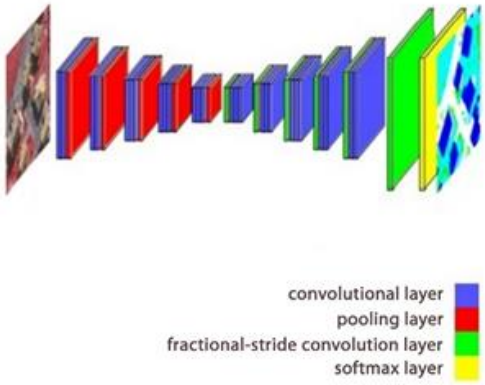

Fig. 2 SegNet Architecture [17]

2-2- HRNet

The main part of HRNet contains four stages with four parallel subnetworks. Each subnetwork is composed of a sequence of convolutions, also there is a down-sample layer across adjacent subnetworks to decrease the resolution to half. So the resolution is step by step decreased and the width (or the number of channels) is proportionally increased. The first stage contains four residual units. Each unit, the same as ResNet-50, is formed by a bottleneck with a width of 64, followed by one 3×3 convolution reducing the width of feature maps.

$$N_{11} \rightarrow N_{22} \rightarrow N_{33} \rightarrow N_{44}$$

In fact, Nsr represents the subnetwork in stage s, and r is the resolution index (Its resolution is $\frac{1}{2^{r-1}}$ of the resolution of the first subnetwork). It is obvious that the precision of each stage is $\frac{1}{2^{r-1}}$ of the first subnetwork's precision.

$$\begin{array}{ccccccc} \mathcal{N}_{11} & \rightarrow & \mathcal{N}_{21} & \rightarrow & \mathcal{N}_{31} & \rightarrow & \mathcal{N}_{41} \\ & \searrow & \mathcal{N}_{22} & \rightarrow & \mathcal{N}_{32} & \rightarrow & \mathcal{N}_{42} \\ & & & \searrow & \mathcal{N}_{33} & \rightarrow & \mathcal{N}_{43} \\ & & & & & \searrow & \mathcal{N}_{44}. \end{array}$$

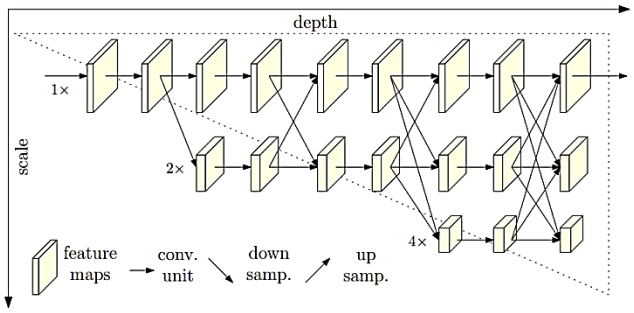

Fig. 3 HRNet architecture [20]

The multi-resolution subnetworks are in parallel. The resolution of parallel subnetworks in each stage will include the resolution of the previous stages and one stage below. An example of a network structure with four subnetworks is shown below:

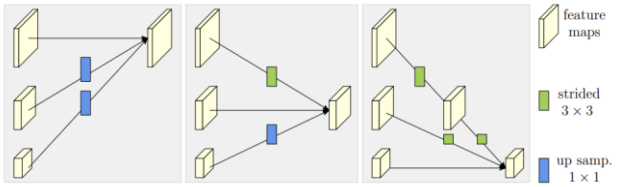

Fig. 4 From left to Right, the Exchange unit Aggregates the Information for High, Medium, and low Resolutions [20]

There are two versions of this network, HRNet-W32 and HRNet-W48. Here, 32 and 48 respectively represent the width (C) of the high-resolution subnetworks in the last three stages. The other three parallel subnetworks have widths of 64, 128, 256 for HRNet-W32 and 96, 192, and 384 for HRNet-W48. HRNet keeps high-resolution representation on the main stem throughout the network and lower-resolution parallel stems are produced via downsampling operations.

2-3- CSE_HRNet

Sometimes objects of the same class in aerial images acquired with high spatial resolutions show various shapes, scales, colors, and structures. Fig. 5 demonstrates some examples of this issue namely intra-class heterogeneity. In Fig. 5(a) cars have different colors, although they all belong to the car class. Similarly, in Fig. 5(b) buildings of the same category vary in texture and shape.

Meanwhile, we may face objects of the different classes having the same colors or interacted with cast shadows that present similar visual characteristics. This would lead to inter-class homogeneity problem.

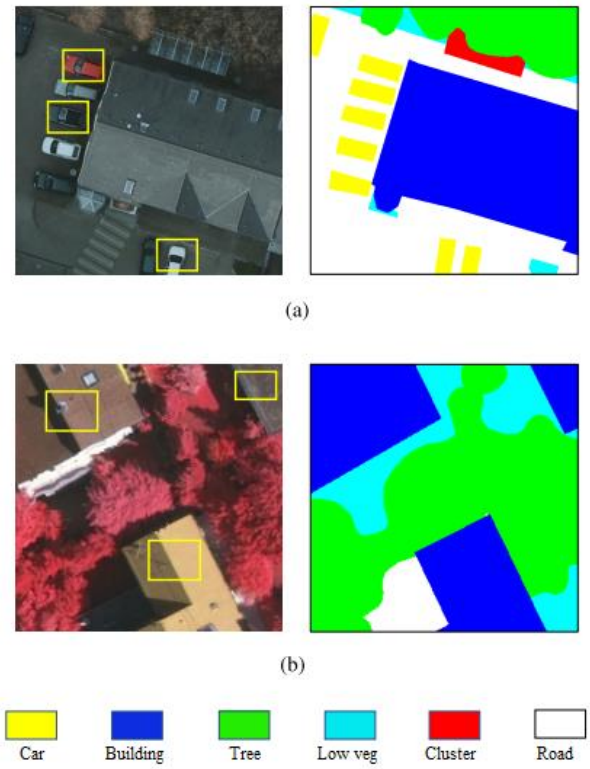

Fig. 5 Intra-class Heterogeneity (a) Cars have Different Colors (b) Buildings are Different in Appearance [25]

Fig. 6 shows objects which are similar in appearance while they should be categorized into separate semantic classes [26,27]. This issue named inter-class homogeneity is shown in Fig. 6. In Fig. 6(a) there are some areas of low veg and trees, which belong to two separate classes, have similar appearances [26,27]. Also, in Fig. 6(b) there are buildings and impervious surfaces are quite similar in appearance. These confusing objects pose extreme challenges for accurate and coherent segmentation [25,28].

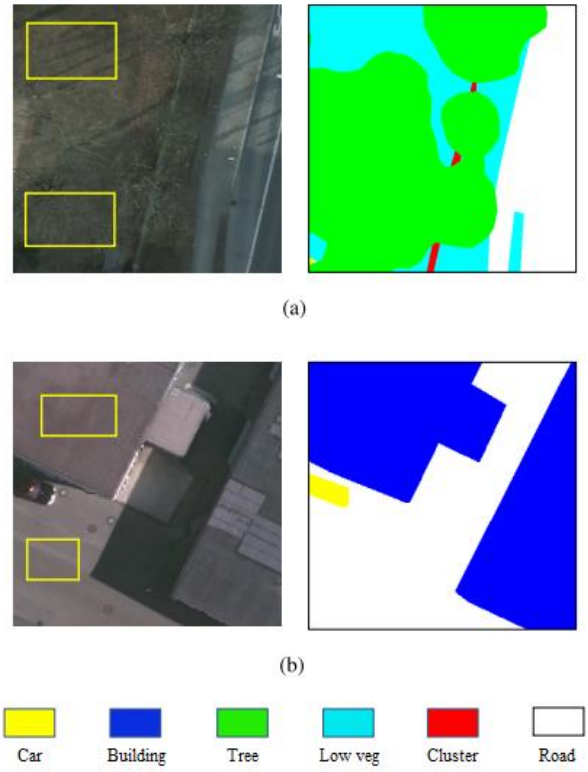

Fig. 6 Inter-Class Homogeneity (a) Trees and low veg are Similar (b) Buildings and Impervious Surface are Analogous [25]

The CSE-HRNet architecture was designed based on the backbone network of HRNet-W32. As with HRNet-W32, "W32" in CSE_HRNet32 represents the feature dimensions of high-resolution or the number of channels representations in the main sub-branch, and the number of other parallel channels will be 64, 128, and 256.

The pyramid structure can exploit the inherent multi-level features, and provide adequate semantic knowledge at all levels. So, the pyramidal feature hierarchy was introduced in this architecture to enhance multi-level semantic representations of the model [29-33]. CSE-HRNet can resolve intra-class heterogeneity and inter-class homogeneity simultaneously by using NDRB combined with the pyramidal multi-level feature hierarchy.

The hierarchy adopts a four-level top-down architecture where the strided convolution as the downsampling method is applied (the stride is set to 2). Widths and heights of feature maps (spatial resolutions) are then reduced by half after each downsampling, whereas the numbers of channels (feature dimensions) are doubled.

The first-level feature map of the pyramid is directly fed into the main high-resolution branch of the network. The second-, third-, and fourth- level feature maps are fused with the counterparts from the multi-resolution branches via the element-wise addition.

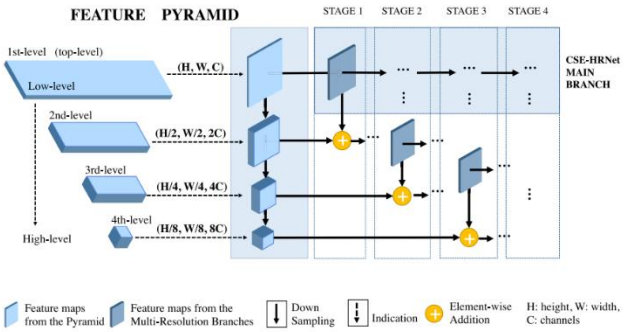

Fig. 7 CSE_HRNet Architecture [25]

2-4- RCA_FCN

Although it has been recognized that contextual relation can offer important cues for semantic segmentation tasks, but using convolution operations in prior convolutional neural networks leads to failure in modeling contextual spatial relations, due to their local valid receptive field [34-39]. However, some convolutional algorithms tried to address this problem using spatial propagation modules or graphical models, but they seek to capture global spatial relations implicitly with a chain propagation way. The effectiveness of these methods depends highly on the learning impact of long-term memorization [40]. Consequently, such models don't work well when long-range spatial relations exist. So, these models most of the time fail to capture long-range spatial relationships between entities, which leads to spatial fragmented prediction [26].

The most important goal of designing the RCA-FCN architecture is to solve spatial relation problems and access channel information. This structure introduces simple effective network units, namely, the spatial relationships module and the channel relationships module. So it can learn and reason about global relationships between every two feature maps or spatial positions, and produce RA feature representations. This network takes VGG16 as a backbone for multilevel feature extraction. Fig. 8 shows a representation of this architecture.

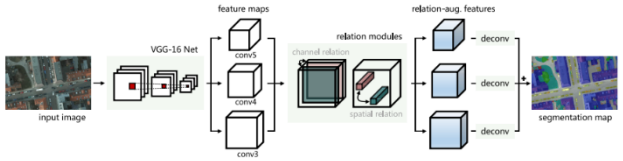

Fig. 8 Overview of the Relation Module [26]

As shown in Fig. 9, outputs of convolve3, convolve4, and convolve5 were fed into the channel relationships module and the spatial relationships module for generating RA features. Then these features were fed into convolutional

layers with 1×1 filters to squash the number of channels to the number of categories.

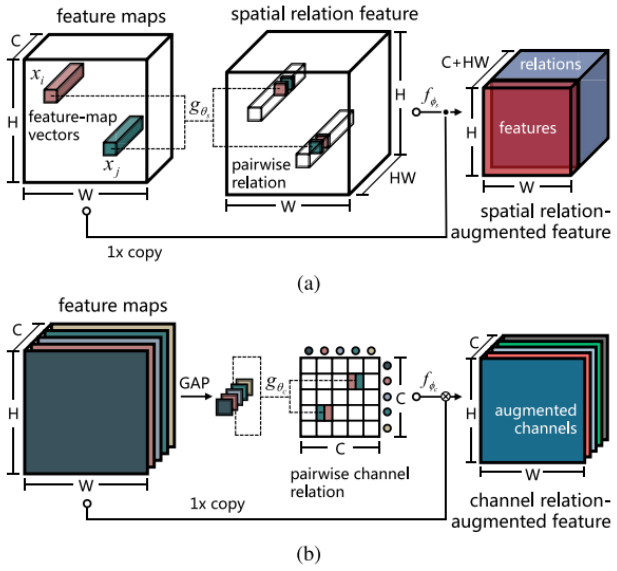

Fig. 9 (a) spatial relation module (b) channel relation module [26]

The convolved feature maps were finally upsampled to desired full resolution and element-wise added to generate final segmentation maps.

3- Implementation

The ISPRS segmentation dataset in Vaihingen was used in this implementation. This dataset is consist of 33 images collected over a 1.38 km^2 area and the average image size of 2494 × 2064 pixels. The spatial resolution is 9 cm. They have green (G), red (R), and near infrared (NIR) bands. Vaihingen dataset was provided by ISPRS-Commission III [26]. Images were captured using digital aerial cameras and mosaicked with Trimble INPHO OrthoVista [41].
Due to the large and diverse dimensions of images in the dataset, five random 240×240 crop were created from each image. Thus, a dataset of images with the same 240×240 resolution was obtained. This dataset was divided into the training dataset and the test dataset. So 60% of images were randomly selected and allocated to the training dataset and remaining 40% allocated to the test dataset. We train the four studied networks architectures in MATLAB 2021.

4- Comparison

The following metrics will be necessary for comparing and evaluating the segmentation performance of these models. They will be explained as follows:

4-1- Accuracy

Accuracy is the percentage of correctly classified instances, or in other words, the ratio of the true results to the total number of cases examined [42,43]. This factor cannot differentiate between FN and FP error and considers them the same.

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+FN+FP+TN}} \qquad (1)$$

4-2- Precision

This factor can determine how many of the correctly predicted cases really turned out to be positive [44]. Precision usually uses when the False Positive is a higher concern than the False Negatives.

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \qquad (2)$$

4-3- Recall

Recall determines how many True Positive cases can be predicted correctly with our model. This factor is also called sensitivity and it is a good choice for the unbalanced classes.

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \qquad (3)$$

4-4- F1

F1 can give a combined idea about two metrics, Precision and Recall. It is maximum when the Precision becomes equal to the Recall.

$$F1 = 2 \times \frac{\text{recall} \times \text{precision}}{\text{recall} + \text{precision}} \qquad (4)$$

In the following, the segmentation results of each model is presented according to the aforementioned metrics. Table (1) shows that no model is conclusively superior to others in terms of segmentation accuracy. A model may have high accuracy in some classes and wouldn't be so in some other classes. For example, although SegNet shows high accuracy in the car class, but it's not very good in classes such as tree and low veg. Compared to the other models, the RCA-FCN was also the most accurate in the building, tree, and low veg classes.

Table 1: Segmentation Accuracy for Each Class

Method	***classes accuracy (%)***				
	building	***car***	***road***	***tree***	***low veg***
SegNet	76.74	**95.93**	81.26	64.52	63.59
HRNet	93.34	63.32	87.35	84.97	74.96
CSE-HRNetsi	94.07	63.34	**88.03**	85.47	73.45
RCA-FCN	**94.12**	70.81	87.22	**89.25**	**87.67**

After observing the segmentation accuracy of each model in different classes, Table (2) shows the overall accuracy (OA) of the networks. The table shows that the CSE-

HRNet algorithm's overall accuracy is superior to the others, followed closely by RCA-FCN in second position.

Table 2: Segmentation Overall Accuracy

method	*Overall accuracy (%)*
SegNet	76.41
HRNet	85.06
CSE-HRNet	**89.23**
RCA-FCN	89.03

F1 is another metric for evaluating the performance of segmentation algorithms, and Table (3) shows its values for different algorithms. As what is said about accuracy, Table (3) clearly shows that no model is definitively superior in all classes. In terms of this metric, SegNet has not performed well in any class. HRNet has the highest F1 (88.94% and 83.19%) in the tree and low veg classes, and CSE-HRNet has the highest F1 (95.41% and 91.92%) in the building and road classes. However, the F1 score of these two architecture for other classes has minor differences with the maximum value. RCA-FCN achieved the best performance with the car class (87.16%). Therefore, one of these networks can be selected for segmentation based on the importance classes.

Table 3: Segmentation F1 for Each Class

Method	*Classes F1-score (%)*				
	building	*car*	*road*	*Tree*	*low veg*
SegNet	71.12	57.89	68.20	34.70	34.98
HRNet	92.91	84.28	91.68	**88.94**	**83.19**
CSE-HRNet	**95.41**	86.79	**91.92**	88.53	80.18
RCA-FCN	94.86	**87.16**	91.01	88.74	80.01

Table (4) shows an overall comparison of these models in terms of F1 and suggests that with an overall F1 score of 89.36%, HRNet can be considered the best network architecture.

Table 4: Segmentation Overall F1

Method	*Overall F1-score (%)*
SegNet	65.79
HRNet	**89.36**
CSE-HRNet	88.57
RCA-FCN	88.36

5- Conclusion

A network can be excellent for distinguishing a specific class of an image dataset and perform poorly in detecting the other classes from the dataset.

Unlike SegNet, the HRNet and CSE-HRNet architectures had generally acceptable results in the F1 and accuracy factors. The RCA-FCN structure can also be considered important not only for its near-ideal evaluation with the general factors, but for properly distinguishing some small classes, such as car and tree.

Finally, selecting the best structure for segmentation is fully dependent on image type and the class' importance for different applications. For studying the state of regional roads and traffic, a model with good accuracy for distinguishing the car and road classes is preferable. However, if the goal is to study the regional vegetation, the segmentation performance of the tree and low veg classes becomes more important.

References

[1] K. Farajzadeh, E. Zarezadeh, J. Mansouri, "Concept detection in images using SVD features and multi-granularity partitioning and classification", Journal of Information Systems & Telecommunication (JIST), 2017, pp. 172.

[2] M.J. Hasan, M. Sohaib, J.M. Kim, "An explainable ai-based fault diagnosis model for bearings", Sensors, 2021, Vol. 21, No. 12, pp. 4070.

[3] M. Ahmad, S. F. Qadri, S. Qadri, I. A. Saeed, S. S. Zareen, Z. Iqbal, A. Alabrah, H. M. Alaghbari, M. Rahman, S. A. Md, "A lightweight convolutional neural network model for liver segmentation in medical diagnosis", Computational Intelligence and Neuroscience, 2022.

[4] M. S. Al-Rakhami, M. M. Islam, M. Z. Islam, A. Asraf, A. H. Sodhro, and W. Ding, "Diagnosis of COVID-19 from X-rays using combined CNN-RNN architecture with transfer learning", MedRxiv, 2020, pp. 20181339.

[5] M. Islam, "An efficient human computer interaction through hand gesture using deep convolutional neural network", SN Computer Science, 2020, Vol. 1, No. 4, pp. 1-9.

[6] W. Li. R. Zhang, H. Deng, L. Wang, W. Lin, S. Ji, and D. Shen, "Deep convolutional neural networks for multi-modality isointense infant brain image segmentation", NeuroImage, 2015, Vol. 108, pp. 214-224.

[7] A. Sandooghdar, F. Yaghmaee, "Deep Learning Approach for Cardiac MRI Images", Journal of Information Systems and Telecommunication (JIST), 2022, Vol. 1, No. 37, pp. 61.

[8] E. Gholam, S.R. Kamel Tabbakh, "Diagnosis of Gastric Cancer via Classification of the Tongue Images using Deep Convolutional Networks", Journal of Information Systems and Telecommunication (JIST), 2021, Vol. 3, No. 35, pp. 191.

[9] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradientbased learning applied to document recognition", Proceedings of the IEEE, 1998, Vol. 86, No. 11, pp. 2278-2324.

[10] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition", Proceedings of the IEEE, 1998, VOL. 86, No. 11, pp. 2278-2324.

[11] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Region-based convolutional networks for accurate object detection

and segmentation", IEEE transactions on pattern analysis and machine intelligence, 2015, Vol. 38, No. 1, pp. 142-158.

[12] N. Audebert, B. Le Saux, and S. Lef`evre, "Beyond RGB: Very high resolution urban remote sensing with multimodal deep networks", ISPRS Journal of Photogrammetry and Remote Sensing, 2018, Vol. 140, pp. 20-32.

[13] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks", Advances in neural information processing systems, 2012, Vol. 25.

[14] K. Simonyan, and A. Zisserman, "Very deep convolutional networks for large-scale image recognition", arXiv preprint arXiv:1409.1556, 2014.

[15] Y. Mo, Y. Wu, X. Yang, F. Liu, and Y. Liao, "Review the state-of-the-art technologies of semantic segmentation based on deep learning", Neurocomputing, 2022, Vol. 493, pp. 626-646.

[16] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning", in Thirty-first AAAI conference on artificial intelligence, 2017.

[17] V. Badrinarayanan, A. Handa, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for robust semantic pixel-wise labelling", arXiv preprint arXiv: 1505.07293, 2015.

[18] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation", IEEE transactions on pattern analysis and machine intelligence, 2017, Vol. 39, No.12, pp. 2481-2495.

[19] K. Sun, Y. Zhao, B. Jiang, T. Cheng, B. Xiao, D. Liu, Y. Mu, X. Wang, W. Liu, and J. Wang, "High-resolution representations for labeling pixels and regions", arXiv preprint arXiv:1904.04514, 2019.

[20] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation", in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 5693-5703.

[21] D. Marmanis, J. D. Wegner, S. Galliani, K. Schindler, M. Datcu, and U. Stilla, "Semantic segmentation of aerial images with an ensemble of CNSS. ISPRS Annals of the Photogrammetry", Remote Sensing and Spatial Information Sciences, 2016, Vol. 3, pp. 473-480.

[22] S. Ioffe, and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift", In International conference on machine learning, 2015, pp. 448-456.

[23] V. Badrinarayanan, B. Mishra, and R. Cipolla, "Understanding symmetries in deep networks", arXiv preprint arXiv:1511.01029, 2015.

[24] H. Zamanian, H. Farsi, S. Mohamadzadeh, "Improvement in accuracy and speed of image semantic segmentation via convolution neural network encoder-decoder", Information Systems & Telecommunication (JIST), 2018, Vol. 6, No. 3, pp. 128-135.

[25] F. Wang, S. Piao, and J. Xie, "CSE-HRNet: A context and semantic enhanced high-resolution network for semantic segmentation of aerial imagery", IEEE Access, 2020, Vol. 8, No. 2, pp. 182475-182489.

[26] L. Mou, Y. Hua, and X. X. Zhu, "Relation matters: Relational context-aware fully convolutional network for semantic segmentation of high-resolution aerial images", IEEE Transactions on Geoscience and Remote Sensing, 2020, Vol. 58, No. 11, pp. 7557-7569.

[27] H. Luo, C. Chen, L. Fang, X. Zhu, and L. Lu, "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, 2019, Vol. 12, No. 9, pp. 3492-3507.

[28] N. Mboga, S. Georganos, T. Grippa, M. Lennert, S. Vanhuysse, and E. Wolff, "Fully convolutional networks and geographic object-based image analysis for the classification of VHR imagery", Remote Sensing, 2019, Vol. 11, No. 5, pp. 597.

[29] G. Zhang, T. Lei, Y. Cui, and P. Jiang, "A dual-path and lightweight convolutional neural network for high-resolution aerial image segmentation", ISPRS International Journal of Geo-Information, 2019, Vol. 8, No. 12, pp. 582.

[30] Z. Tu, X. Chen, A. L. Yuille, and S. C. Zhu, "Image parsing: Unifying segmentation, detection, and recognition", International Journal of computer vision, 2005, Vol. 63, No. 2, pp. 113-140.

[31] B. C. Russell, W. T. Freeman, A. A. Efros, J. Sivic, and A. Zisserman, "Using multiple segmentations to discover objects and their extent in image collections", In IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), 2006, Vol. 2, pp. 1605-1614.

[32] E. Borenstein, and S. Ullman, "Combined top-down/bottom-up segmentation", IEEE Transactions on pattern analysis and machine intelligence, 2008, Vol. 30, No. 12, pp. 2109-2125.

[33] J. Wu, J. Zhu, and Z. Tu, "Reverse Image Segmentation: A High-Level Solution to a Low-Level Task", In BMVC, 2014.

[34] Q. Zhao, and L. D. Griffin, "Better image segmentation by exploiting dense semantic predictions", arXiv preprint arXiv:1606.01481, 2016.

[35] R. Socher, C. C. Lin, A. Y. Ng, and C. D. Manning, "Parsing natural scenes and natural language with recursive neural networks", In Proc. IEEE Int. Conf. Mach. Learn. (ICML), 2011, pp. 129-136.

[36] J. Yao, S. Fidler, and R. Urtasun, "Describing the scene as a whole: Joint object detection, scene classification and semantic segmentation", In IEEE conference on computer vision and pattern recognition, 2012, pp. 702-709.

[37] A. Kae, K. Sohn, H. Lee, and E. Learned-Miller, "Augmenting CRFs with Boltzmann machine shape priors for image labeling", In Proceedings of the IEEE conference on computer vision and pattern recognition, 2013, pp. 2019-2026.

[38] H. Myeong, and K. M. Lee, "Tensor-based high-order semantic relation transfer for semantic scene segmentation", In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 3073-3080.

[39] J. J. Corso, "Toward parts-based scene understanding with pixel-support parts-sparse pictorial structures", Pattern Recognition Letters, 2013, Vol. 34, No. 7, pp. 762-769.

[40] Q. Li, Y. Shi, and X. Huang, "Building footprint generation by integrating convolution neural network with feature pairwise conditional random field (FPCRF)", IEEE Transactions on Geoscience and Remote Sensing, 2020, Vol. 58, No. 11, pp. 7502-7519.

[41] M. Cramer, "The DGPF-test on digital airborne camera evaluation overview and test design", Photogrammetrie-Fernerkundung-Geoinformation, 2010, pp. 73-82.
[42] M.J. Hasan, J.M. Kim, "Bearing fault diagnosis under variable rotational speeds using stockwell transform-based vibration imaging and transfer learning", Applied Sciences, Vol. 8, No. 12, pp. 2357.
[43] M.J. Hasan, J. Uddin, S.N. Pinku, "A novel modified SFTA approach for feature extraction", In 3rd International Conference on Electrical Engineering and Information Communication Technology (ICEEICT), 2016, pp. 1-5.
[44] M. Ghasemi, M. Kelarestaghi, F. Eshghi, A. Sharifi, "D 3 FC: deep feature-extractor discriminative dictionary-learning fuzzy classifier for medical imaging", Applied Intelligence, 2022, pp. 1-17.