Context-Adaptive Deep Learning for Efficient Image Parsing in Remote Sensing: An Automated Parameter Selection Approach

Basim Azam
IIIS
Griffith University)
Brisbane, Australia
basim.azam@griffith.edu.au

Brijesh Verma
IIIS
Griffith University
Brisbane, Australia
b.verma@griffith.edu.au

Mengjie Zhang
CDSAI
Victoria University of Wellington
Wellignton, New Zealand
mengjie.zhang@vuw.ac.nz

***Abstract*—Image parsing is among the core tasks in the field of image processing and computer vision having wide-ranging applications in the areas of autonomous driving, image interpretation, medical analysis, and remote sensing. The modern techniques despite performing the labeling tasks accurately face several challenges. Among these challenges, the computation of contextual information and the selection of optimized parameters are of prime importance in pixel-wise segmentation tasks. We propose a novel context adaptive image parsing framework that utilizes a unique parameter selection strategy to produce final pixel labels. The automatic parameter selection minimizes the computational overhead, reduces the time complexity, and improves upon the segmentation labels produced. The proposed framework is evaluated on Wuhan Dense Labelling Dataset (WHDLD). In addition, a comprehensive comparison with state-of-the-art image segmentation techniques is presented. Finally, the analysis supporting the dominance of proposed architecture is presented in collation with existing techniques.**

***Index Terms*—image parsing, deep learning, remote sensing, neural networks**

I. Introduction

The field of computer vision has observed an immense surge since the inception of deep convolutional neural networks. Numerous tasks, such as image classification, object recognition, object localization, pose estimation, and semantic segmentation have perceived state-of-the-art results [1]–[3]. Image parsing is among the core tasks in computer vision applications and assists the improvement in the areas of autonomous driving [4], interpretation of visual scenarios [5] and pixel-wise segmentation for remote sensing applications [6].

The advent of Convolutional Neural Networks (CNN) is the fundamental step towards the modern image segmentation architectures. The CNNs compute richer feature representations in contrast to the traditional hand-crafted features. Modern CNN architecture such as AlexNet [7], VGGNet [8], ResNet [9], MobileNet [10], and EfficientNet [11] adopt the path of decreasing spatial size of feature representations from a high resolution input image to low level feature maps. It is very difficult to leave the performances of these neural network unnoticed. On the other hand, the expensive convolutional computations and the lack of integrated contextual information within such architecture requires attention as well.

This research was supported under Australian Research Council's Discovery Projects funding scheme (project number DP200102252)

The assimilation of machine learning within computer vision application have assisted the revolution of object detection [12], image classification [13], and pixel-wise segmentation [14]. Over the years, research community has consistently been trying to investigate the fusion of contextual properties with such architecture. This research investigates the relation-aware context information and the automated optimization of parameters in the context of image parsing. The study presents a deep layered network enriched with context information, and the evaluation of proposed architecture on the remote sensing data. The contribution of the manuscript can be summarized as follows:

1) A unique parameter optimization algorithm that opts the most efficient set of parameters exploring particle swarm optimization to improve upon the pixel-wise labelling.
2) A deep layered architecture of image parsing, that maneuvers the adjacent and spatial context information along with proposed parameter optimization to produce state-of-the-art segmentation results.
3) A comprehensive comparison and extensive analysis with the state-of-the-art architecture, demonstrating the superior performance on remote sensing data.

The rest of the paper is organized as follows. The background of the pixel-wise segmentation approaches is presented in Section II. The proposed automated parameter optimization and image parsing architectures are described in section III. Section IV describes the experimental setup and results achieved in comparison to the previous approaches. The conclusion is presented in Section V.

II. Literature Review

The primary research using the deep CNNs paved the way for research conducted over the years. Many of the image parsing techniques were proposed exploring the convolutional neural network for such tasks.

Several classification networks make use of the traditional fully convolutional network techniques [15], [16] by having their fully connected layers abolished. These networks explore convolutional and pooling layers reducing size of feature maps. Ultimately leading towards fine spatial activations categorized as low-level feature representation. The researchers worked on ways to up-sample the feature representations. The pixel-wise segmentation frameworks were developed recovering the representations in dual ways, the initial being the symmetric reconstruction while the other as asymmetric recovering of feature maps.

The segmentation networks including ERFNet [17] and ESPNet [18] are known for computation of semantic information by employing convolutional layers, and these architectures produce dense feature predictions using the representation computed. On the other hand, the segmentation networks such as U-Net [19], SegNet [20], DeconvNet [21] Encoder-Decoder [22], are formed in shape of the encoder-decoder networks, the aim of these networks is just as similar, the computation of pixel-wise class labels for the input images. RefineNet [23] is very popular example segmentation architecture that is formed with asymmetric reconstructed architecture. Some of other research works also make us of the light up-sampling by employing dilated convolutions, recombinator architectures and the improvements in skip connections.

The segmentation architecture can generate high level feature representations to be computed and maintained throughout the network, e.g., interlinked CNNs [24] , GridNet [25], and the DenseNet [26]. The initial work lacks the capability to determine the initiation point for start of parallel stream and the ability to intelligently exchange information in the parallel streams. These architectures do not use the residual connection and normalization, ultimately resulting in poor performances. UNet [19] , and SegNet [20] diffuse the low-level features (high-to-low down-sampling process) and the high level features for same resolution (low-to-high up-sampling process). DeepLabv2-v3 [27] architecture also has the fusion process, that combines the pyramid features computed using pyramid pooling and atrous spatial pyramid pooling (ASPP).

The optimal selection of hyperparameters in neural networks is challenging issue [28]. While researchers have investigated the optimization problem over the years, the automatic and optimal selection of parameters for segmentation architecture is still an arduous problem . The research establishes that the optimum selection of hyperparameters lead to better performance of the segmentation architecture. The selection ultimately proves of significant importance in model building process. The idea of choosing relevant features have been explored in the literature over the years [9]–[11], [20]. The selection of features provides optimal number of features by removing redundancy among the set. The architecture proposed in the literature for segmentation tasks lack the ability to consider the context and neighbour information within an image. While, each pixel is correlated with other pixels in the images, so the pixel-wise segmentation tasks can be considered to adapt contextual information among pixels. . The primary research using the deep CNNs paved the way for research conducted over the years. Many of the image parsing techniques were proposed exploring the convolutional neural network for such tasks.

III. Proposed Framework

The image parsing framework is illustrated in figure 1. Initially, the superpixel representations are obtained from the input image by SLIC [29] algorithm. For each image in the dataset, the superpixels are also used to compute the contextual information in parallel to the visual representations. The architecture exploits the rich contextual information at multiple levels, the context representations are computed using the adjacent superpixel information and the block-wise information of superpixel occurrences. Furthermore, the dataset specific contextual information is used to compute the superpixel-wise probability of specific class label. Finally, the output of the visual classifier, adjacent superpixel occurrence, and spatially block-wise occurrence are fed to another multilayer perceptron to decide the final superpixel label.

A. The Particle Swarm Optimization

Particle swarm optimization is a population based stochastic optimization method, which was inspired by the social behavior of bird flocking. Such social behavior can be observed in a social group, the behavior of the individual is not only dependent the past experiences and cognition but it heavily relies on the pattern of overall social behavior.

So, the technique must choose best among a population of probable solutions also referred to as candidate solutions. The individual solutions exhibit movement based on the best individual standing position and the best position of the whole population. Consider the image parsing function

$$f_{ImageParsing}(X) = f(x_1, x_2, x_3, ..., x_n) \tag{1}$$

where x_i is the search variable in terms of parameters, representing the set of free variables of the given function. The aim is to find the value of such that the function $f(X)$is minimum in the search space.

In the global best PSO, the position of individual particle is affected by the best-fit particle in the entire swarm. The global best PSO makes use of the star social network topology, in which the social information is acquired from every particle in the whole swarm. In this method, the distinctive particles $i\varepsilon[1, ..., n]$ have a prevalent position in search space x_i, a prevalent velocity v_i, and a individual-level best position $P_{best,i}$ in search space. The individual-level best $P_{best,i}$ position represents the position in search space where particle i had the smallest value as estimated by the objective function f, considering a minimization problem.

Generally, In the minimization problems the individual-level best position $P_{best,i}$ at the next time step $t+1$, where $t\varepsilon[0, ..., N]$ can be estimated as

$$P_{best,i}^{t+1} = \begin{Bmatrix} P_{best,i}^{t} & iff(x_i^{t+1} > P_{best,i}^{t}) \\ x_i^{t+1} & iff(x_i t+1 \leq P_{best,i}^{t}) \end{Bmatrix} \tag{2}$$

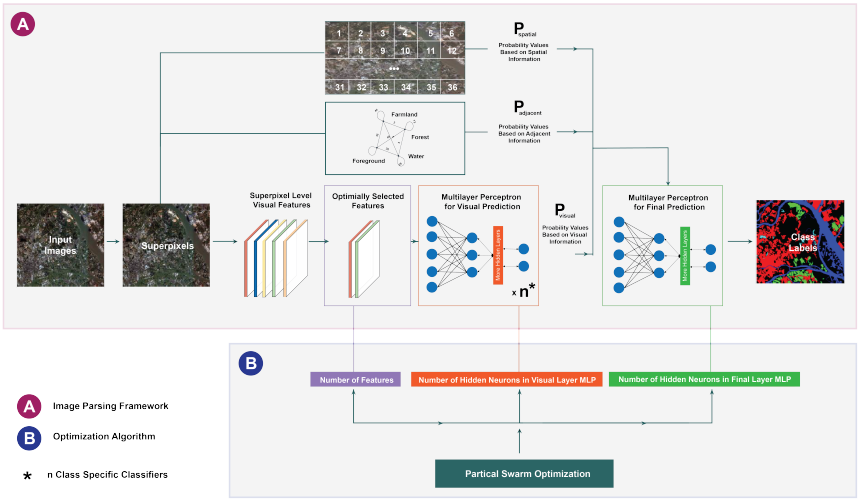

Fig. 1. **A - Image Parsing Framework**: *The Visual Probability Estimation*-this layer takes a subset of optimum super pixel visual features,*The Contextual Probability Estimation*-this layer estimates the probabilities for pixel labels using adjacent and spatial contextual modules. **B - Optimization Algorithm** - this helps find the most suitable number of parameters for the architecture. (Best Viewed in Color)

where $f : \mathrm{R}^{\mathrm{n}} \rightarrow \mathrm{R}$ is the fitness function, The global best position G_{best} at time step $t$is calculated as:

$$G_{best} = min\{P^t_{best,i}\} \tag{3}$$

where $i\epsilon[1, ..., n]$ and $n > 1$ This, it becomes essential to note that the values obtained from individual-level best is the best position that the individual particle has visited since beginning of time steps. While the global best position is the optimum position discovered by the particles in the entire swarm.

For the global PSO method the velocity of particle is calculated as

$$v^{t+1}_{ij} = v^t_{ij} + c_1 r^t_{1j}[P^i_{best,i} - x^t_{ij}] + c_2 r^t_{2j}[G^i_{best,i} - x^t_{ij}] \tag{4}$$

where v^t_{ij} is velocity vector, x^t_{ij} is position vector, $P^t_{best,i}$is personal best position,$G^t_{best,i}$ is the global best position,c_1 and c_2 are positive acceleration constants, r^t_{1j} and r^t_{2j} are random numbers at time t.

B. The Visual Probability Estimation

The raw images form the training dataset along with corresponding ground truth information are fed to the network. Initially, the superpixels are computed to extract the visual features from, these superpixels are computed using Simple Linear Iterative Clustering (SLIC) [30] algorithm. For the sake of these experiments, we generate 512 superpixels uniformly. These superpixels provide the baseline for the computation of visual features. Visual features include the color variation, geometric differences, and texture information. The standard deviation, mean, histograms of textons and masks are computed using these superpixels. Additionally, the computation of SIFT descriptors make the feature vector a bit lengthy. Towards this end, the proposed particle swarm optimization provides the optimal number of features to training the visual multilayer perceptron MLP_{vis}. The aim of training this classifier is to assign every pixel of the image to class label.

C. The Contextual Probability Estimation

In parallel to the visual probability estimation, the proposed network also makes use of the superpixel information to learn about occurrences of objects. The contextual probability estimation refers to the idea of using the information present in training dataset to suggest the probability of class labels occurring concurrently. The contextual information is estimated in two parts, the initial being the adjacent occurrence of the class labels, while the other being the occurrence of class labels in block of superpixels.

The adjacent contextual module and the spatial contextual modules vote towards the other class label. The superpixels vote towards the adjacent superpixel labels and towards the superpixel labels in a spatial block. So, the class label of each superpixel is produced by majority votes of labels. The adjacent superpixel information is referred to as local contextual information and the spatial superpixel information as global contextual information. The local and global votes are normalized to probability values as to match the probability of class labels using MLP_{vis}.

D. Final Class-wise Label Prediction

The probability values estimated from the contextual probability estimation layer and the visual probability estimation layer are combined to compute final class-wise labels for every superpixel.

IV. Experimental Evaluation

A. Benchmark and Evaluation Metrics

1) Wuhan Dense Labelling Dataset (WHDLD): The Wuhan Dense Labelling Dataset (WHDLD) contains 4940 RGB images cropped out of remote sensing images of Wuhan city. The images are of 256 x 256 pixels in size

with a resolution of 2m. These images are labelled into six categories as bare soil, road, building, vegetation, pavement and water. An overview of the dataset and the percentage breakdown of pixel labels in the dataset is provided in figure 2.

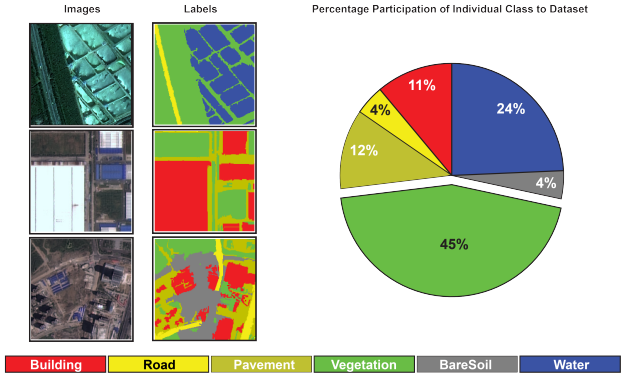

Fig. 2. The overview of the Wuhan Dense Labelling Dataset (WHDLD) used in the study. Left : Visual samples with labels, Right: Percentage breakdown by category.

2) Evaluation Metrics: The performance of the proposed network is evaluated on the basis of Pixel Accuracy (PA), Mean Pixel Accuracy (MPA), and the mean Intersection Over Union (mIoU). The pixel accuracy generally has higher tendency of being insensitive towards class with less occurrences, while the mIoU exhibits super-sensitivity towards minority object categories. To overcome this hurdle we make use of the F1 scores (F1). The PA, MPA, mIoU and F1 are computed as:

$$PixelAccuracy = \frac{\sum_{l=0}^{n} prob_{ll}}{\sum_{l=0}^{n}\sum_{m=0}^{n} prob_{lm}}$$

$$MeanPixelAccuracy = \frac{1}{l+1}\sum_{l=0}^{n}\frac{prob_{ll}}{\sum_{m=0}^{n} prob_{lm}}$$

$$mIoU = \frac{1}{l+1}\sum_{l=0}^{n}\frac{prob_{ll}}{\sum_{m=0}^{n} prob_{lm} + \sum_{m=0}^{n} prob_{ml} - prob'_{ll}}$$

$$F1Score = 2 \times \frac{precision \times recall}{precision + recall}$$

3) Implementation Details: The proposed architecture explores a range of parameters to opt for the best combination using PSO algorihtm.The details for the parametric specification are provided in the Table I. The three variable features include number of optimal visual-level features selected, number of hidden neurons in the visual probability estimation layer, and the number of neuron in the final layer. The maximum iteration, swarm size, ranges for the variables and initialization parameters for PSO are provided in the table as well.

[1] F_i, is number of selected features,N_{vis} is number of neurons to compute visual probability, and N_{cl} is number of neurons to compute final class label

TABLE I
THE PARAMETRIC SPECIFICATIONS TO INITIALIZE THE PARTICLE SWARM OPTIMIZATION TO FIND THE BEST SOLUTION FOR IMAGE PARSING FRAMEWORK (IN TERMS OF SELECTED FEATURES AND NUMBER OF HIDDEN NEURONS)

Parameter Specifications for PSO	
Variables	F_i, N_{vis}, N_{cl} [1]
Maximum Iterations	1000
Swarm Size	25
Rang F_i	50 - 100
Range N_{vis}	16 – 64
Range N_{cl}	16 – 32
Inertia Weight	1
Inertia Weight Damping Ratio	0.99
Personal Learning Coefficient	1.5
Global Learning Coefficient	2

4) Experimental Setup: The proposed architecture is implemented on MATLAB 2021a, various additional toolboxes have been used that include Machine Learning Toolbox and Deep Learning Toolbox. The experiments are conducted using the High-Performance Computing (HPC) clusters.

B. Comparison with Stat of the Art

The Table II presents the performance analysis of proposed architecture. The individual class-wise pixel accuracies are presented to describe the better and enhanced performance of the architecture. The per-class metric assists to verify the better performance of the architecture The proposed architecture achieved higher accuracies for classes building, bare soil, pavement, vegetation, and water. Table III presents the comparison of proposed architecture in terms of pixel accuracy, mean pixel accuracy, mean intersection over union, and f1-score. The proposed technique achieved slightly improved pixel accuracy value of 84.67%, better mean pixel accuracy value of 77.08%, mIoU value of 61.84% and F1 score of 73.91%. The results achieved are better in comparison to the segmentation approaches proposed over the years and evaluated on the WHDLD dataset.The WHDLD comprises of diverse scenarios representing combinations of pixels of water, vegetation, road, pavement, building, and bare soil. The adjacent and spatial context information assist the computation of final pixel labels.

C. Ablation Study

The ablation studies present the refinement using the context modules and the optimization process. The improvements in segmentation accuracies using the contextual exploration and optimization process are presented through the ablation studies. The comparison is carried out within the architecture on the WHDLD dataset. The global context information and the local context information are used to show the enhanced results. The results are also quantified based on various stages of the optimization, based on the optimal parameters the segmentation accuracies are presented. Table IV highlights the impact of adjacent context information and spatial context information on the overall architecture. The results are compared between the accuracy values obtained from the multilayer-perceptron (prior to the inclusion of

TABLE II
RESULTS ON THE WHDLD DATASET IN TERMS OF CLASS-WISE ACCURACY IN COMPARISON WITH THE STATE-OF-THE-ART SEGMENTATION TECHNIQUES.

Technique / Label	Building	BareSoil	Pavement	Vegetation	Road	Water
Proposed Technique	**84.72**	**76.36**	**80.62**	**93.42**	77.21	95.73
Segnet [20]	47.68	63.25	51.47	54.65	86.47	95.65
Tiramisu [31]	50.31	68.92	53.58	70.05	88.21	96.6
U-Net [19]	43.1	70.75	52.61	58.67	89.19	97.09
U-NetAtt [32]	47.97	72.74	48.94	60.58	90.00	**97.51**
FGC [33]	50.28	72.64	53.84	57.93	89.65	**97.29**
MSFCN [34]	52.18	**74.50**	55.18	68.8	**90.02**	**97.51**

TABLE III
RESULTS ON THE WHDLD DATASET IN TERMS OF PIXEL ACCURACY, MEAN PIXEL ACCURACY, MEAN INTERSECTION OVER UNION, AND F1 SCORE IN COMPARISON WITH THE STATE-OF-THE-ART SEGMENTATION TECHNIQUES.

	Pixel Accuracy	Mean Pixel Accuracy	mIoU	F1
Proposed Technique	**84.67**	**77.08**	**61.84**	**73.91**
Segnet [20]	80.29	63.78	52.94	66.53
Tiramisu [31]	82.19	70.71	58.17	71.28
U-Net [19]	81.83	67.72	55.7	68.57
U-NetAtt [32]	82.6	69.74	56.91	69.62
FGC [33]	82.98	68.86	57.36	70.27
MSFCN [34]	**84.16**	72.08	**60.36**	**73.03**

context adaptive modules) and the overall proposed optimized architecture. The class-wise scores improved 34%, 65%, 30%, 4%, 47%, 8% respectively for building, bare soil, pavement, vegetation, road, and water. The overall pixel accuracy improved from 53% to 84%, mean pixel accuracy from 50% 77%, mIoU 45% to 61%, and F1 score from 51% to 74%. The overall improvement are notable using the context adaptive modules.

The Table V presents the comparative analysis at three different stages of optimizations, the initial being first notable improvement, the second being the middle of overall iterations and the final stage is defined by final best solution of particle swarm optimization (PSO) algorithm. The architecture improved from 53%-pixel accuracy to 68%, 76% and 84% for all the three stages of optimization. The mean pixel accuracy scores were improved to 77% at the final optimized stage as compared to the initial 50%. The mIoU scores improved 27% using the proposed optimized architecture. The F1-scores were also improved from 51% to 74%. The results establish the efficacy of the proposed approach, and it also shows the improvements in scores using the optimization algorithm. It can be argued that the proposed architecture has established capability of adapting to datasets using the adjacent and spatial context information.

D. The Effectiveness of the Optimization

The proposed architecture exhibits notable enhancement for class-wise segmentation. The architecture explores various combination of features, hidden layer neurons in visual layer, hidden layer neuron in final layer, ultimately resulting in improved segmentation scores. The optimization improves the pixel accuracy from 53% in visual layer to 84% in final layer for WHDLD dataset. The comparison in Table V supports the argument that optimization module incorporated in the architecture improves the overall architecture.The improvements in pixel accuracy and evaluation scores establish the effectiveness of the proposed optimization algorithm.

TABLE IV
ABLATION STUDY ON WHDLD DATASET CONSIDERING THE ADDITIONAL CONTEXTUAL INFORMATION.

	MLP_{vis}	Adjacent Context Information	Spatial Context Information	Proposed Optimized Architecture
Building	50.22	✓	✓	*84.72*
Bare Soil	11.37	✓	✓	*76.36*
Pavement	50.31	✓	✓	*80.62*
Vegetation	89.9	✓	✓	*93.42*
Road	30.52	✓	✓	*77.21*
Water	87.69	✓	✓	*95.73*
PA	53.33	✓	✓	**84.67**
MPA	50.52	✓	✓	**77.08**
mIoU	45.90	✓	✓	**61.84**
F1 Score	51.31	✓	✓	**73.91**

TABLE V
COMPARATIVE ANALYSIS OF RESULTS IN TERMS OF ACCURACY USING THE PARTICLE SWARM OPTIMIZATION FOR PARAMETER SELECTION.

	MLPvis	Optimization Stage 1	Optimization Stage 2	Proposed Optimized Architecture
Building	50.22	74.41	73.24	*84.72*
Bare Soil	11.37	56.65	61.42	*76.36*
Pavement	50.31	57.67	71.67	*80.62*
Vegetation	89.90	84.23	91.23	*93.42*
Road	30.52	53.38	63.42	*77.21*
Water	87.69	86.17	91.28	*95.73*
PA	53.33	68.75	76.24	**84.67**
MPA	50.52	61.53	70.56	**77.08**
mIoU	45.90	55.43	58.32	**61.84**
F1 Score	51.31	58.91	61.38	**73.91**

V. Conclusion

The paper presents a novel parameter selection-based image parsing framework that explores additional contextual information to produce final labels. The notable novelties include the optimization of parameters, and the computation of contextual information. The paper demonstrates the improved pixel accuracy, mean pixel accuracy, mean intersection over union and f1-scores using the proposed image parsing architecture. The architecture achieves 84%-pixel accuracy, 77% mean pixel accuracy, mIoU 61% and F1-score of 73% on WHDLD dataset. In comparison to the state-of-the-art techniques the proposed approach achieves better scores. The incorporation of optimization algorithm and the additional context information improves the segmentation accuracies. In our future research, the aim will be to investigate the proposed architecture on a number of real-world image parsing datasets.

References

[1] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 38, no. 2, p. 295–307, 2016.

[2] R. Q. F. Liu, G. Lin and C. Shen, "Structured learning of tree potentials in crf for image segmentation," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 6, pp. 2631–2637, 2018.

[3] Y. Y. B. L. C. Li, W. Xia and J. Tang, "Segmenting objects in day and night: Edge-conditioned cnn for thermal image semantic segmentation," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 7, pp. 3069–3082, 2021.

[4] S. Y. Alaba and J. E. Ball, "Deep learning-based image 3-d object detection for autonomous driving: Review," *IEEE Sensors Journal*, vol. 23, no. 4, pp. 3378–3394, 2023.

[5] D. Elayaperumal and Y. H. Joo, "Robust visual object tracking using context-based spatial variation via multi-feature fusion," *Information Sciences*, vol. 577, pp. 467–482, 2021.

[6] W. Han, L. Wang, R. Feng, L. Gao, X. Chen, Z. Deng, J. Chen, and P. Liu, "Sample generation based on a supervised wasserstein generative adversarial network for high-resolution remote-sensing scene classification," *Information Sciences*, vol. 539, pp. 177–194, 2020.

[7] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 3431–3440.

[8] Z. Liu, J. Li, Z. Shen, G. Huang, S. Yan, and C. Zhang, "Learning efficient convolutional networks through network slimming," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2755–2763.

[9] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, "Residual attention network for image classification," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 6450–6458.

[10] D. Sinha and M. El-Sharkawy, "Thin mobilenet: An enhanced mobilenet architecture," in *2019 IEEE 10th Annual Ubiquitous Computing, Electronics Mobile Communication Conference (UEMCON)*, 2019, pp. 0280–0285.

[11] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proceedings of the 36th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, K. Chaudhuri and R. Salakhutdinov, Eds., vol. 97. PMLR, 09–15 Jun 2019, pp. 6105–6114.

[12] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 580–587.

[13] W. Rawat and Z. Wang, "Deep convolutional neural networks for image classification: A comprehensive review," *Neural Computation*, vol. 29, pp. 2352–2449, 2017.

[14] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 6230–6239.

[15] E. S. J. Long and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Transaction on Pattern Analaysis and Machine Inteligence*, vol. 39, no. 4, pp. 640–651, 2014.

[16] C. Gao, J. Yan, X. Peng, and H. Liu, "Signal structure information-based target detection with a fully convolutional network," *Information Sciences*, vol. 576, pp. 345–354, 2021.

[17] E. Romera, J. M. Álvarez, L. M. Bergasa, and R. Arroyo, "Erfnet: Efficient residual factorized convnet for real-time semantic segmentation," *IEEE Transactions on Intelligent Transportation Systems*, vol. 19, no. 1, pp. 263–272, 2018.

[18] J. Du, J. Song, K. Cheng, Z. Zhang, H.-X. Zhou, and H. Qin, "Efficient spatial pyramid of dilated convolution and bottleneck network for zero-shot super resolution," *IEEE Access*, vol. 8, pp. 117 961–117 971, 2020.

[19] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, Eds. Cham: Springer International Publishing, 2015, pp. 234–241.

[20] A. K. V. Badrinarayanan and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

[21] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1520–1528.

[22] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 801–818.

[23] G. Lin, A. Milan, C. Shen, and I. Reid, "Refinenet: Multi-path refinement networks for high-resolution semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1925–1934.

[24] Z. Yin, V. Yiu, X. Hu, and L. Tang, "End-to-end face parsing via interlinked convolutional neural networks," *Cognitive Neurodynamics*, vol. 15, no. 1, pp. 169–179, 2021.

[25] I. Bozcan, J. Le Fevre, H. X. Pham, and E. Kayacan, "Gridnet: Image-agnostic conditional anomaly detection for indoor surveillance," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 1638–1645, 2021.

[26] B. Lodhi and J. Kang, "Multipath-densenet: A supervised ensemble architecture of densely connected convolutional networks," *Information Sciences*, vol. 482, pp. 63–72, 2019.

[27] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, pp. 834–848, 2018.

[28] D. Gong, Z. Zhang, Q. Shi, A. van den Hengel, C. Shen, and Y. Zhang, "Learning deep gradient descent optimization for image deconvolution," *IEEE transactions on neural networks and learning systems*, vol. 31, no. 12, pp. 5468–5482, 2020.

[29] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "Slic superpixels compared to state-of-the-art superpixel methods," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 11, pp. 2274–2282, 2012.

[30] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "Slic superpixels compared to state-of-the-art superpixel methods," *IEEE transactions on pattern analysis and machine intelligence*, vol. 34, no. 11, pp. 2274–2282, 2012.

[31] D. V. A. R. S. Jégou, M. Drozdzal and Y. Bengio, "The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation," in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2017, pp. 1175–1183.

[32] J. S. et. al., "Attention gated networks: Learning to leverage salient regions in medical images," *Medical Image Analysis*, vol. 53, no. 12, pp. 197–207, 2019.

[33] S. Ji, Z. Zhang, C. Zhang, S. Wei, M. Lu, and Y. Duan, "Learning discriminative spatiotemporal features for precise crop classification from multi-temporal satellite images," *International Journal of Remote Sensing*, vol. 41, no. 8, pp. 3162–3174, 2020.

[34] R. Li, S. Zheng, C. Duan, L. Wang, and C. Zhang, "Land cover classification from remote sensing images based on multi-scale fully convolutional network," *Geo-spatial information science*, pp. 1–17, 2022.