2023 IEEE Symposium Series on Computational Intelligence (SSCI)
Mexico City, Mexico. December 5-8, 2023

Parameter Optimisation for Context-Adaptive Deep Layered Network for Semantic Segmentation

Ranju Mandal
Architecture Art & Design
Torrens University
Brisbane, Australia
ranju.mandal@torrens.edu.au

Brijesh Verma
Institute for Integrated and Intelligent Systems
Griffith University
Brisbane, Australia
b.verma@griffith.edu.au

***Abstract*— Evolutionary optimization methods have been utilized to optimize a wide range of models, including many complex neural network models. Manual parameter selection requires substantial trial and error and specialist domain knowledge of the inherent structure, which does not guarantee the best outcomes. We propose a three-layered novel architecture for semantic segmentation and optimize it using two distinct evolutionary algorithm-based optimization processes namely genetic algorithm and particle swarm optimization. To fully optimize an end-to-end image segmentation framework, the network design is tested using various combinations of a few parameters. An automatic search is conducted for the optimal parameter values to maximize the performance of the image segmentation framework. Evolutionary Algorithm (EA)-based optimization of the three-layered semantic segmentation network optimizes parameter values holistically, which produces the best performance. We evaluated our proposed architecture and optimization on two benchmark datasets. The evaluation results show that the proposed optimization can achieve better accuracy than the existing approaches.**

Keywords—image segmentation, semantic segmentation, scene parsing, genetic algorithm, deep learning

I. INTRODUCTION

The objective of semantic segmentation is to assign pre-determined semantic categories like objects (e.g., motor vehicle, animal, pedestrian) and backgrounds (e.g., roadway, tree, building) to each pixel in the image. The outcome of semantic segmentation is a dense pixel-wise annotation of an image that has an identical resolution as the input image. Semantic segmentation enables us to obtain a more precise and rich representation of the image content, which makes it a fundamental step in computer vision techniques with numerous applications such as image understanding and editing.

Semantic segmentation involves dividing an image into meaningful classes and identifying relationships between them, which presents challenges. Crucial systems, such as hazard detection, image compression, augmented reality, AI-based smart monitoring, robot vision, and autonomous vehicle navigation rely on accurately detecting the presence of predefined objects in an environment. The diverse appearance of objects, in both structured and unstructured complex natural scenes make it challenging to label pixels precisely with an object category. Fig.1 displays a selection of images from our experimental datasets with their annotated labels, demonstrating various challenges. The class objects exhibit variations in appearance and are influenced by factors like occlusion, illumination, angle, and size. The robust segmentation network must perform three tasks (classification, localization, and edge delineation) on each object to achieve high accuracy. The following is a brief overview of our key contributions to this paper:

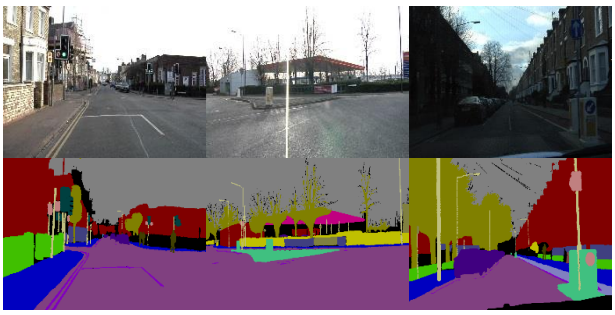

Fig. 1. Semantic segmentation densely predicts semantic categories from natural images taken in an unrestricted setting. Annotated ground truth images are presented in the bottom row, while original images are presented in the top row.

a) Our proposed architecture consists of three layers, holistically optimized through the evolutionary algorithm. These layers (i.e., visual feature layer, context feature layer, and integration layer) capture optimal parameter values. We conducted two sets of experiments. The evolutionary algorithms based on Particle Swarm Optimization and Genetic Algorithms play a crucial role in improving accuracy by selecting the optimal parameters.

b) The co-occurrence probability information between object classes from the training samples is extracted effectively by our architectures' contextual layer. It leverages the patterns in the object's spatial data, both locally and globally (block-wise), to obtain the desired information.

c) As evidenced by robust evaluation using the widely used Stanford Background Dataset [1] and CamVid [2] benchmark datasets, the proposed method exhibits satisfactory performance on both image segmentation datasets. Our results are comparable to previous methods on both SBD and the CamVid street scene dataset.

978-0-7381-4408-5/23/$31.00 ©2023 IEEE

II. PREVIOUS WORKS

A. Evolutionary Computation-based Optimisation

The application of evolutionary algorithms to address a variety of hyperparameter optimization and search problems has been widely reported. Two Evolutionary Algorithms (EA) such as Particle Swarm Optimization (PSO) [5, 6] and Genetic Algorithm (GA) [3, 4] remain at the forefront in solving parameter optimization problems. While a solution based on the PSO algorithm is efficient for continuous optimization, GA-based solutions are more suited for optimization with binary data. Previous research has employed Genetic Algorithms for optimizing the weights of artificial neural networks [7] and for determining the structure of neural network models [8, 9].

B. Semantic Segmentation

A comprehensive evaluation was conducted on the publicly available semantic segmentation datasets and the pros and cons of different ConvNets-based semantic segmentation architectures. Many CNN-based network models outperformed advanced semantic segmentation methods. Recently developed image or scene parsing models utilized implicit global [10] or local context [11] along with convolutional visual features. However, these methods are ineffective in integrating relative and absolute contextual information with the visual characteristics effectively. The proposed method, on the other hand, is based on explicit context integration. Two main techniques have been used in published semantic segmentation articles: multi-scale context-based methods and deep ConvNets variations.

Earlier techniques for classifying pixel labels involved either the individual extraction of visual features around each pixel in an image [12] or the extraction of features in patches [13]. Image segmentation architectures that utilized feature hierarchies employed region proposals to obtain class labels in some of the earliest scene segmentation tasks [14]. However, features based on global context perform better than those based on individual pixels, as the latter cannot effectively capture the statistics of adjoining regions, while feature extraction using patches is susceptible to background noise from objects. Spatial pyramid-based pooling in PSPNet [15] improves scene parsing accuracy by incorporating global contextual statistics. Nguyen et al. [14] introduced a network based on a hybrid Deep Network-Gaussian Process (GP) for the segmentation of scene images into lane and background regions. This architecture differs from existing deep learning approaches by combining a dense network of limited parameters with a robust nonparametric GP classifier. In both visual and quantitative evaluations, this Gaussian Process-based classifier outperforms SegNet [16] and Bayesian SegNet [17]. However, the study only evaluated and compared with other techniques for a single class (pedestrian lane).

Yu et al. [18] proposed BiSeNet as a two-branche architecture for real-time semantic segmentation. The Detail Branch captures the spatial details with wide channels and shallow layers. The Semantic Branch, in contrast, extracts categorical semantics with fewer channels and deep layers. These two features are merged to obtain a comprehensive feature representation. ParseNet [19] leverages image-level information by utilizing each pixel's global features. PSPNet [15] improves accuracy by employing a pyramid pooling module that gathers useful multi-scale contextual information. Zhang et al. [20] proposed EncNet introduces a Context Encoding Module to emphasize class-dependent feature maps and record semantic context. Deeplabv2 [21] and Deeplabv3 [22] integrate contextual information through the use of atrous spatial pyramid pooling, which involves dilated convolutions parallelly with varying rates. Despite the importance of context in real-world image parsing, the integration of both global and local context in a single network architecture has not yet been thoroughly explored. Our goal is to utilize the statistical characteristics to infer class labels by capturing essential neighboring class information as context.

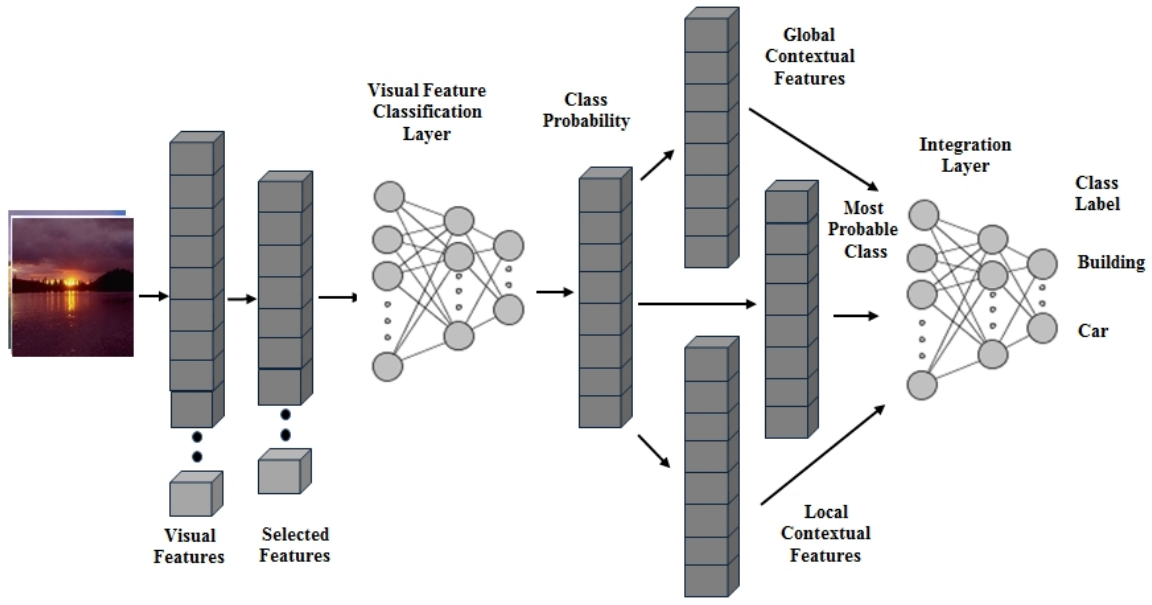

Fig. 2. A cutting-edge optimization approach is used to optimize the proposed new multi-layered image segmentation framework. In the first layer, we compute class probability using superpixel visual features. In order to estimate the probability for classes, the second layer computes context characteristics utilising superpixel blocks probability and votes from neighbouring superpixel. The probability vector acquired from the visual feature classification (layer 1) and the contextual attributes are ideally fused to get the final class label for each superpixel in the final integration layer (layer 2).

III. PROPOSED METHOD

We propose a new framework that optimises our deep semantic segmentation architecture using evolutionary algorithms. Two different optimization frameworks using Genetic Algorithm (GA) and Particle Swarm Optimisation (PSO) are utilized for a limited number of iterations to find the optimal parameter values for the segmentation architecture. Fig. 2 presents the building blocks of the architecture to achieve our goal of high-resolution semantic segmentation. The process starts with the extraction of the visual features from the segmented superpixel patches, which are then fed into the visual feature classification layer. This class-semantic supervised classifier produces class-wise probability matrices for all superpixels, indicating the probabilities of each superpixel belonging to one of the object classes. The contextual layer is divided into two parts, each of which extracts different types of contextual information (local and global region-based) from each superpixel in the image. Contextual features are generated by combining the class probability matrices obtained from the visual classification layer and the Object Co-occurrence Priors (OCPs) obtained during the training process. The visual and contextual features are integrated through a multi-layer perceptron (MLP) network, which assigns the final class marker for each superpixel.

A. Features based on visual properties

The proposed feature extraction algorithm uses superpixels, which are over-segmented clusters of pixels created from an input image, to extract visual features. In the feature selection process, the most impactful features are selected from the visual information. The visual feature classifier is trained using the visual features that are extracted from the superpixel patches of training image samples from the dataset. The visual feature classifier uses the superpixel-level feature vector as input, and a fully connected network with a single hidden layer learns various patterns for classification. The classifier's efficiency is enhanced throughout training by selecting appropriate feature subsets.

B. Features based on contextual properties

The contextual features of the classes are extremely important in the proposed framework. By calculating the probability of adjacent superpixels in both the close surroundings and the entire image, we precisely construct the

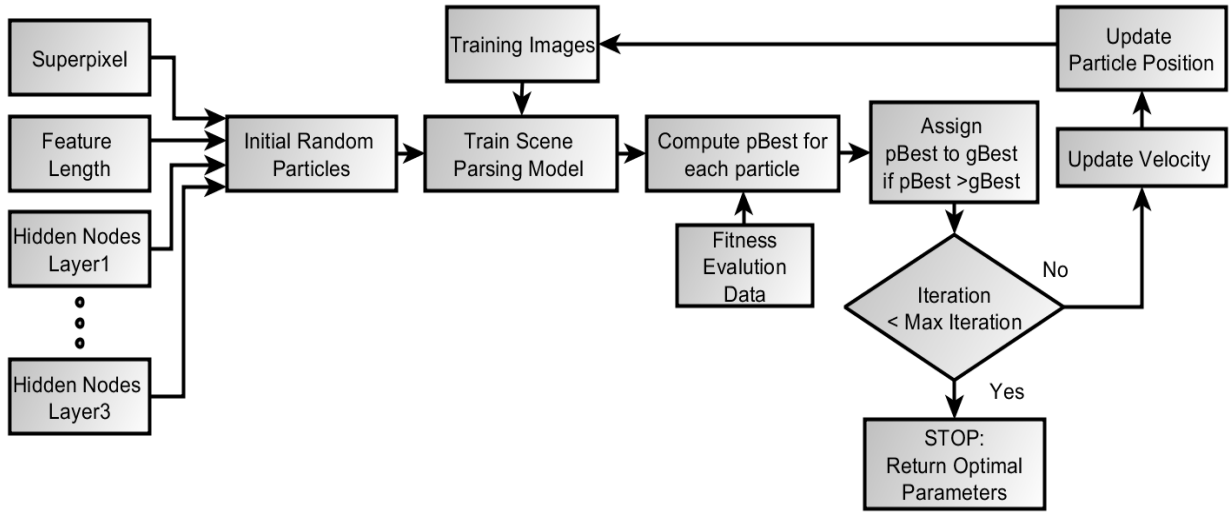

Fig. 3. The proposed flow diagram of PSO-based optimisation of the context-based neural networks for image segmentation.

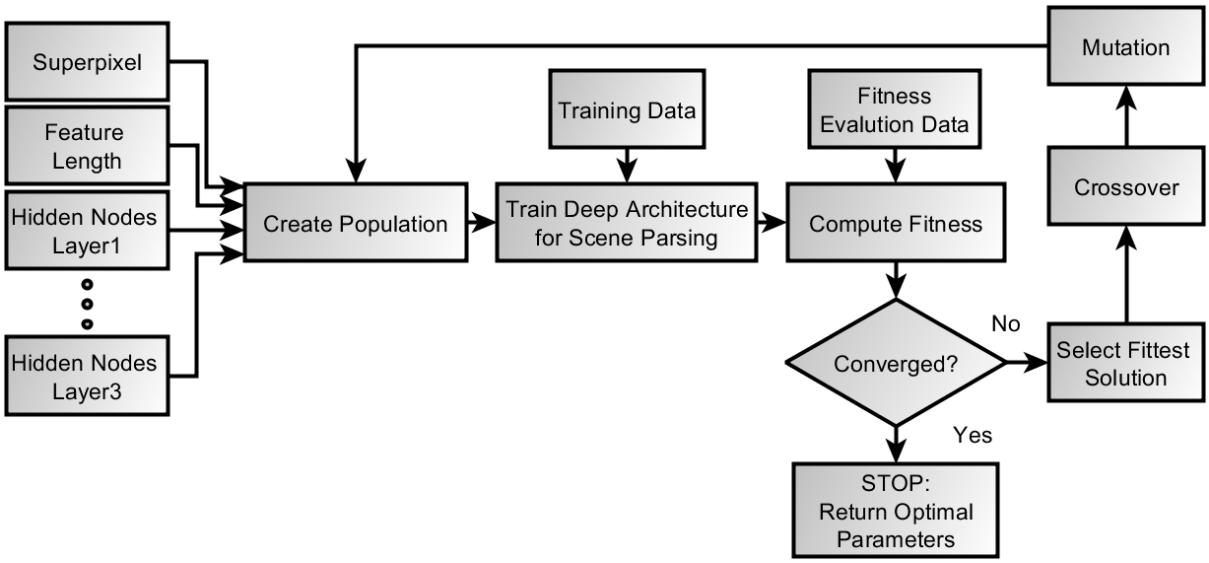

Fig. 4. The proposed architecture of the GA-based optimization system for image parsing

relationships between superpixels. Local probabilities are estimated using the immediately adjacent superpixels while calculating global probability, superpixels within a spatial block are assessed. The probability values take into account both local and global information.

Modeling local context using statistics of adjacent superpixels: The localized contextual information incorporates valuable information from the surrounding superpixels. The probability vector of the adjacent superpixels for each superpixel is estimated. Each superpixel casts a vote for the surrounding superpixels based on a pre-determined class label, which leads to the computation of an object co-occurrence priors matrix. This matrix provides class probability statistics for the surrounding superpixels for each superpixel and contains probability matrices for neighboring superpixels that were estimated from the training sample.

Modeling global context using statistics of block-wise superpixels: The reliance on long-range superpixel correlations is heavily influenced by block-wise voting for each superpixel. Both the absolute position of objects and their spatial arrangement offsets are considered using block-wise estimation. The block-wise encoding creates a balance of absolute and relative regions, encoding the absolute location of each block while encoding the relative location offsets of objects in the spatial relations between blocks. The spatial distributions encoding method of all blocks also encodes directional spatial relationships between blocks (E.g., interactions between left and right blocks).

C. Optimal integration of three features

The integration layer seeks to give every superpixel a class label through our final integration layer's weight optimisation. The most likely visual characteristic from visual feature-based prediction and two types of contextual characteristics from context-sensitive voting are integrated using optimized weights. A weighted majority technique is utilized to assign a class label to each superpixel. The integration layer (a single hidden layer MLP network) estimates the correlation of visual and contextual characteristics during integration.

D. Optimization

Particle Swarm Optimization (PSO): Our proposed PSO-based architecture optimisation method considers the feature dimension, the count of nodes in the visual feature classifier, the count of nodes in the integration layer, and other hyperparameters for optimisation. Gradient-based optimization methods are not effective for parameter optimization due to their non-differentiable and non-convex nature. The PSO technique creates multiple models from the initial model, initializing the population chromosomes or vectors with parameter range values (as presented in Fig. 3). The fitness evaluation determines which model performs the best by computing the cost. Due to the random initialization value of parameters during the model's optimisation process, some models perform better than others.

The segmentation network was optimised with a population size of 25, 0.01% mutation rate, and 0.05–0.1% crossover rate. Over the iterations, the accuracy improved, and its best value was attained when the crossover rate was 0.1. In our experiment, the convergence was accelerated during the high crossover rate. The proposed network learns parameters in each layer to generate an approximation.

Genetic Algorithm (GA): The proposed network architecture considers various parameters, including the feature dimension, count of nodes in the visual classification layer, and count of nodes in the integration layer (as shown in Fig. 4). Traditional gradient-based optimization techniques are not effective due to the non-differentiable and non-convex nature of the parameter optimization problem [23]. We use GA to create several offspring from the initial model, which is initialized with parameter range values for the population chromosomes or vector. The fitness evaluation tests compute the cost to determine the best-performing model. The models' performance can vary due to random initialization. The network selects the optimal set of parameters for each layer. The GA parameters for the network optimisation were set to a population size of 25, a mutation rate of 0.01, and a crossover rate between a range of 0.05 to 0.1. During the optimization process, a higher crossover rate was found to speed up convergence, with the best accuracy obtained when the crossover rate was 0.1. The proposed context-based architecture uses a GA-based framework to automatically determine the optimal value of parameters by searching the solution space. The value encoding approach was used in the genetic algorithm-based optimization, where the chromosome is represented by real, integer, or character values. This encoding scheme is well-suited for our continuous search problems, commonly used in neural networks for finding optimal weights.

IV. RESULTS AND DISCUSSIONS

This section presents the achieved accuracy from the proposed optimized architectures and discusses them. Our experimentation was carried out on two benchmark datasets, and a thorough performance review of recently published works in image segmentation techniques was performed to compare the performance.

A. Datasets

Segmentation evaluation dataset: The proposed methodology was validated using the Stanford Background Dataset (SBD) [1] and the CamVid [2] dataset. The SBD [1] consists of 715 outdoor scenery photos sourced from public databases and was annotated using an online platform operated by Amazon (Amazon Mechanical Turk), categorizing image pixels into eight predefined categories or undefined categories. For the experiments, the SBD is partitioned into three sets, namely training data (70% samples), fitness evaluation data (15%), and final test data (15%). The data samples are chosen at random for training, fitness evaluation, and testing.

The CamVid [2] dataset features manually annotated labels for 32 object classes and valuable experimental data for evaluating image segmentation models. The images in this dataset are extracted from video footage captured from a car's perspective, which adds diversity to the object classes represented. This dataset is well-regarded among computer vision researchers for its significant contributions, and its original resolution of 960 × 720 was down-sampled to 480 × 360 for our experiments, following previous studies. Like SBD, the CamVid dataset is also partitioned into 3 sets (i.e., 70% of samples are used for training models, and the remaining samples are equally allocated for fitness evaluation and testing. The training and fitness evaluation data sets are used for the stochastic gradient descent search and genetic beam search to induce the neural network models, while the test data set was used for final image segmentation to evaluate

the performance of the trained/learned neural networks. Note that the test was not used at any stage of the training/evolutionary process.

B. Implementation details

Our model was developed using the Python and MATLAB environments with the support of the Global Optimization Toolbox, Image Processing Toolbox, and Neural Network Toolbox. All evaluations were conducted on a High-Performance Computing (HPC) cluster facility with dedicated 16 processing nodes and 160 GB of RAM.

C. Evaluation on Stanford Dataset

Table I displays the results obtained from the Stanford dataset. The GA-optimised model scored 92.42% accuracy, and the PSO-optimised model achieved 92.35% accuracy with a population size set to 25. Table I also compares the accuracies reported by previous works with the accuracy of our proposed GA-based and PSO-based optimised models. Table I shows 92.18% class accuracy obtained using the GA-optimised model, while PSO optimised model obtained 92.15% accuracy. Fig. 5 showcases the qualitative results obtained from the Stanford Background dataset, indicating that the proposed solution predicted object pixels with high precision.

TABLE I. COMPARISON OF SEGMENTATION RESULTS (%) WITH BENCHMARK APPROACHES ON THE STANFORD BACKGROUND DATASET

Method	Pixel Acc.	Class Acc.
Gould et al. [1]	76.4	-
Lempitsky et al. [24]	81.9	72.4
Farabet et al. [25]	81.4	76.0
Sharma et al. [13]	82.3	79.1
Luc et al. [26]	75.2	68.7
Chen et al. [22]	87.0	75.9
Zhu et al. [27]	87.7	79.0
Proposed Method (GA)	**92.42**	**92.18**
Proposed Method (PSO)	**92.35**	**92.15**

D. Segmentation Results on CamVid dataset

On the CamVid test dataset, we used the broadly used Intersection-Over-Union (IoU) metric (i.e., the Jaccard Index) to assess how well our optimised models perform. The weighted Jaccard Index was used as the mean IoU metric to analyze the model's performance. It's a popular highly effective metric method used by the state-of-the art algorithms in semantic segmentation performance evaluation.

The proposed GA-optimized three-layered network model achieved 78.68% mIoU on the Stanford dataset without using any pre-trained weights when the population was set to 25. The PSO-optimised model outperformed the GA-based optimised model with an accuracy of 81.77% mIoU when the population was set to 25. Table II compares the mean IoU with previously published techniques on the CamVid dataset and shows that the network model outperformed existing techniques using the best parameter choices. Fig. 6 showcases the qualitative results of the proposed approach on the CamVid dataset, with the top row displaying the test samples, the middle row displaying the corresponding annotation of test samples, and the bottom row interpreting the model-generated results from our PSO-optimised segmentation model.

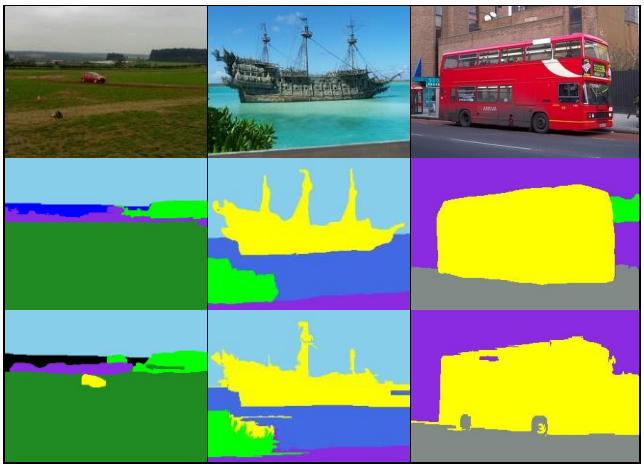

Fig. 5. Qualitative accuracy on the Stanford Background dataset is achieved by the PSO-optimised model, displayed column-wise with original, ground truth, and network-labeled images arranged from top to bottom.

TABLE II. COMPARISON OF SEGMENTATION RESULTS (%) WITH BENCHMARK APPROACHES ON THE CAMVID DATASET

Method	Pre-trained	mIoU (%)
Badrinarayanan et al. [16]	ImageNet	60.1
Huang et al. [28]	ImageNet	62.5
Yu and Koltun [10]	ImageNet	65.3
Zhao et al. [14]	ImageNet	69.1
Bilinski and Prisacariu [29]	ImageNet	70.9
Chandra et al. [30]	Cityscapes	75.2
Yu et al. [18]	ImageNet	78.5
Proposed Method (GA)	-	**78.6**
Proposed Method (PSO)	-	**81.7**

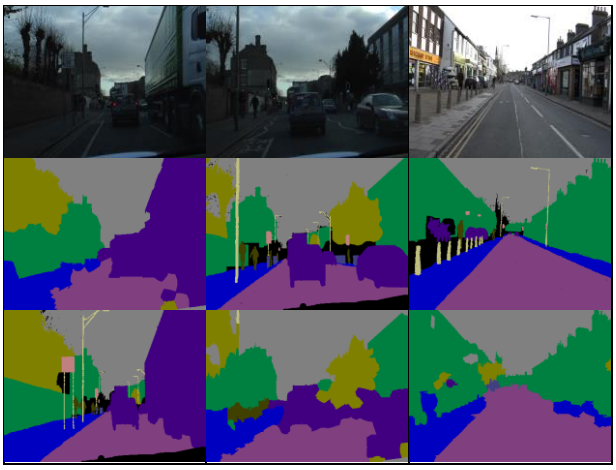

Fig. 6. Qualitative accuracy on the CamVid dataset is achieved by the PSO-optimised model. We presented a top-down, columnar strategy to show sample test images, ground truth annotation, and the segmentation model annotated images.

V. CONCLUSIONS

In this study, we propose a novel deep context-adaptive architecture for the segmentation of images semantically, and the performance was improved using evolutionary algorithms. The network utilizes the object co-occurrence priors matrices evaluated during the training phase to derive the contextual features. The context-based feature is intended to record label correlations of objects in a scene image over close-range and far-off distances. Our architecture optimization aims to determine the optimal combination of our network parameters that would produce the best accuracy. Two benchmark datasets were used in the experiments on the proposed image parsing model, which showed improved performance. Using the genetic algorithm as an optimizer on the SBD, we achieved an accuracy of 92.42 percent, while on the CamVid dataset, we secured an accuracy of 78.68 percent mIoU. The PSO algorithm achieved almost equal performance (92.35 percent accuracy) on the SBD and, however, outperformed the genetic algorithm with 81.77 percent accuracy on the CamVid dataset. The comparative study presented in Tables I and II shows that our proposed model performs better, as evidenced by the results. Future studies will concentrate on further performance-improving optimization of the proposed model.

ACKNOWLEDGMENT

This research project was supported under the Australian Research Council's Discovery Projects funding scheme (ARC-DP200102252).

REFERENCES

[1] S. Gould, R. Fulton, and D. Koller, "Decomposing a Scene into Geometric and Semantically Consistent Regions," *in Proc. IEEE Conference on Computer Vision and Pattern Recognition.*, pp. 1-8, 2009

[2] G. J. Brostow, J. Fauqueur, and R. Cipolla, "Semantic Object Classes in Video: A High-definition Ground Truth Database," *Pattern Recognition Letters*, vol. 30, no. 2, pp. 88-97, 2009

[3] Y. Sun, B. Xue, M. Zhang, and G. G. Yen, "Evolving Deep Convolutional Neural Networks for Image Classification," *IEEE Transactions on Evolutionary Computation*, vol. 24, no. 2, pp. 394-407, 2020.

[4] Y. Sun, B. Xue, M. Zhang, G. G. Yen, and J. Lv, "Automatically Designing CNN Architectures Using the Genetic Algorithm for Image Classification," *IEEE Transactions on Cybernetics*, vol. 50, no. 9, pp. 3840-3854, 2020.

[5] B. Wang, Y. Sun, B. Xue, and M. Zhang, "Evolving deep convolutional neural networks by variable-length particle swarm optimization for image classification," *in Proc. IEEE Congress on Evolutionary Computation (CEC)*, pp. 1-8, 2018.

[6] F. E. F. Junior and G. G. Yen, "Particle swarm optimization of deep neural networks architectures for image classification," *Swarm and Evolutionary Computation*, vol. 49, pp. 62–74, 2019.

[7] S. Ding, H. Li, C. Su, J. Yu, and F. Jin, "Evolutionary artificial neural networks: a review," *Artificial Intelligence Review*., vol. 39, no. 3, pp. 251–260, 2013.

[8] J. Bayer, D. Wierstra, J. Togelius, and J. Schmidhuber, "Evolving memory cell structures for sequence learning," *in Proc. International Conference on Artificial Neural Networks*, 2009, pp. 755–764.

[9] K. O. Stanley and R. Miikkulainen, "Evolving neural networks through augmenting topologies," *Evolutionary Computation*., vol. 10, no. 2, pp. 99–127, 2002.

[10] F. Yu and V. Koltun, "Multi-Scale Context Aggregation by Dilated Convolutions," *arXiv:*1511.07122, 2015.

[11] S. Choi, J. T. Kim, and J. Choo, "Cars Can't Fly Up in the Sky: Improving Urban-Scene Segmentation via Height-Driven Attention Networks*," in Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9373–9383, 2020.

[12] C. Farabet, C. Couprie, L. Najman, and Y. LeCun, "Learning Hierarchical Features for Scene Labeling," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1915–1929, 2013.

[13] A. Sharma, O. Tuzel, and D. W. Jacobs, "Deep Hierarchical Parsing for Semantic Segmentation," *in Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp. 530–538, 2015.

[14] T. N. A. Nguyen, S. L. Phung, and A. Bouzerdoum, "Hybrid Deep Learning-Gaussian Process Network for Pedestrian Lane Detection in Unstructured Scenes," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no. 12, pp. 5324–5338, 2020.

[15] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid Scene Parsing Network," *in Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2881–2890, 2017.

[16] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

[17] A. Kendall, V. Badrinarayanan, and R. Cipolla, "Bayesian SegNet: Model Uncertainty in Deep Convolutional Encoder-Decoder Architectures for Scene Understanding," *arXiv:*1511.02680, 2015.

[18] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "BiSeNet V2: Bilateral Network with Guided Aggregation for Real-Time Semantic Segmentation," *International Journal of Computer Vision*, pp. 3051-3068, 2021.

[19] W. Liu, A. Rabinovich, and A. C. Berg, "ParseNet: Looking wider to see better". *arXiv*:1506.04579, 2015.

[20] H. Zhang, K. Dana, J. Shi, Z. Zhang, X. Wang, A. Tyagi, and A. Agrawal, "Context Encoding for Semantic Segmentation," *in Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7151–7160, 2018.

[21] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, 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 (4), pp. 834-848, 2018.

[22] L. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *arXiv:*1706.05587, 2017.

[23] Y. Sun, B. Xue, M. Zhang, and G. G. Yen, "An Experimental Study on Hyper-parameter Optimization for Stacked Auto-Encoders*," in Proc. IEEE Congress on Evolutionary Computation (CEC)*, pp. 1-8, 2018.

[24] V. Lempitsky, A. Vedaldi, and A. Zisserman, "Pylon Model for Semantic Segmentation," *in Advances in Neural Information Processing System*, pp. 1485–1493, 2011.

[25] C. Farabet, C. Couprie, L. Najman, and Y. Lecun, "Learning hierarchical features for scene labeling," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1915-1929, 2013.

[26] P. Luc, C. Couprie, S. Chintala, and J. Verbeek, "Semantic Segmentation Using Adversarial Networks," *NIPS Workshop on Adversarial Training, Barcelona, Spain,* 2016.

[27] X. Zhu, X. Zhang, X.-Y. Zhang, Z. Xue, and L. Wang, "A Novel Framework for Semantic Segmentation with Generative Adversarial Network," Journal of Visual Communication and Image Representation, vol. 58, pp. 532–543, 2019.

[28] P.-Y. Huang, W.-T. Hsu, C.-Y. Chiu, T.-F. Wu, and M. Sun, "Efficient Uncertainty Estimation for Semantic Segmentation in Videos," *in Proc. of the European Conference on Computer Vision*, pp. 520–535, 2018.

[29] P. Bilinski and V. Prisacariu, "Dense Decoder Shortcut Connections for Single-Pass Semantic Segmentation," *in Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6596–6605, 2018.

[30] S. Chandra, C. Couprie, and I. Kokkinos, "Deep Spatio-Temporal Random Fields for Efficient Video Segmentation," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8915–8924, 2018.