A Graph-based Context Learning Technique for Image Parsing

Basim Azam
Institute for Integrated and Intelligent Systems
School of Infomration and Communication Technology
Griffith University, Brisbane, Australia
basim.azam@griffithuni.edu.au

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

***Abstract*—The modern deep learning-based architectures have performed well for pixel-wise segmentation tasks. The consideration of context is of vital importance for generation of accurate semantic information. In this research, a deep learning-based image parsing framework is proposed that utilizes novel relation-aware context learning technique. The proposed technique explores the graph constructs from the training data to learn the co-occurring context associations of object category labels using the graph edge connections. The proposed graph-based context learning technique defines the scene specific relation-awareness among semantic object categories, e.g., the probability of sky, road and building to co-exist in a scene is high. The proposed image parsing architecture (including the novel graph-based context learning technique) is evaluated on the benchmark datasets. In addition, a comprehensive comparison with existing image parsing techniques is presented to establish the efficacy of the scene-graph generation. The in-depth investigation of graph generation is presented to demonstrate the improvement in pixel-wise labeling.**

Keywords—Image Parsing, Relation Aware Graph, Pixel-Wise Segmentation, Semantic Segmentation, Deep Learning.

I. INTRODUCTION

An immense growing body of literature in the field of pixel-wise segmentation tasks helps recognize the significance of the topic in computer vision area [1],[2],[3],[4]. Several closely related tasks such as image classification [5], object localization/recognition [6], instance segmentation [7] and pose estimation [8] have also observed surged since the emergence of the convolutional neural networks. Central to the entire discipline of computer vision is the image parsing task, as it aids the processes of autonomous driving vehicles, remote sensing, autonomous surveillance, and interpretation of visual scenes.

In the recent years, deep learning and neural network-based architectures have revolutionized the computer vision applications. The literature establishes the fact that Convolutional Neural Networks (CNNs) have surpassed other traditional methods [9] by greater margins in terms efficiency and accuracy. The CNNs are known for estimating rich feature information as compared to the traditional feature computations, this improvement has helped surpass the human-level performance on various tasks. Interestingly, the networks have also deepened over the years in terms of architectural length and computational complexity.

Image parsing is a fundamental and powerful part of the computer vision [10]. While image parsing refers to individual pixel-level classification, the goal is to segment image into meaningful semantic parts such as sky, water, road and trees [11]. The ideal representation of image data into semantic categories can be observed in the fig. 1. The research community has produced state-of-the-art image parsing frameworks that achieve remarkable result on

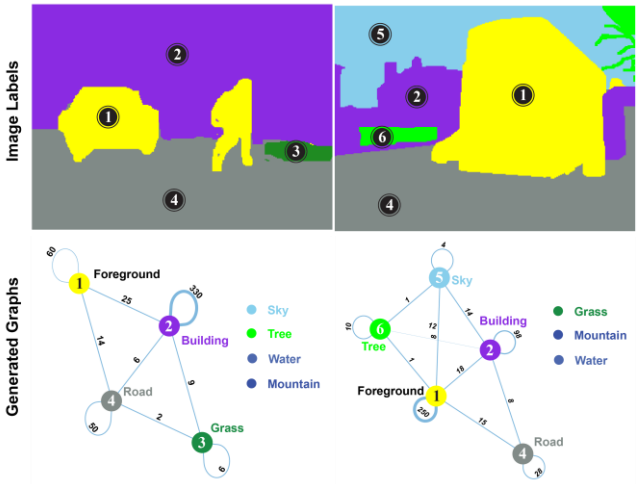

Fig. 1. Scene Aware Graph Generation: The image presents scene aware graphs generated on Stanford BG dataset. The relationship between different classes can be observed, sky and building pixels have direct noticeable relationship, the association of co-occurring pixels is also being represented using the graph edge connections. The weights of the edges represent the strength of relationship between two object categories.

benchmark datasets, the evolution of these architecture is based on increased depth of architecture, improved convolution computations, and consideration of contextual information. Many techniques have performed well with context adaptation [11], however further investigations are required to develop an architecture that considers relation-aware information in the dataset and improves upon the efficiency of segmentation architecture.

The objectives of this research include the investigations of context information using the generated relation-aware graphs in the context of image parsing, and to present the proposed architecture in comparison to the state-of-the-art approaches evaluated on benchmark datasets. The contribution of the manuscript can be summarized as follows:

1) Graph-based context-learning technique that extracts the context information from the image datasets to improve the image parsing.

2) Image parsing framework that utilizes the unique graph-based relation-aware context learning technique to produce better pixel-wise labels.

3) A detailed and comprehensive comparison on the benchmark datasets with extensive analysis-based discussion, outlining the improved performance of image parsing framework on the benchmark datasets.

The overall structure for the rest of the paper takes the form of four subsequent sections. The related works are

presented in section II, while the proposed relation-aware graph generation and overall image parsing architecture is described in section III. Section IV presents the benchmark datasets, evaluation of proposed architecture on the benchmarks, and comparison with state-of-the-art (SOTA) image parsing frameworks, this section also provides detailed discussion on both the results achieved and efficiency of proposed architecture. Final conclusions and observations are presented in Section V.

II. BACKGROUND

In this section, the research work related to the image parsing is presented. Specifically, insights to the image segmentation methods proposed to explore the context-information will be presented briefly.

The generation of pixel-wise labels have been a topic of research consideration over the years. The preliminary research work presented used tradition feature computations from the pixels [12], and the upgraded architectures made use of superpixel groups to extract features from [13]. Many approaches proposed afterwards generate superpixels based upon similarities, to provide alike features. While these approaches lacked the context information, for which Conditional Random Fields (CRFs) [14] were proposed. CRFs improved the segmentation results but proved to be computationally expensive and exhaustive.

The Fully Convolutional Networks (FCNs) have been used extensively for semantic labelling tasks [15]. The researchers presented diverse modifications to the baseline approach, improving the segmentation accuracies. The FCN architecture has also been used in concurrence with Conditional Random Fields (CRF), enhancing the computed feature representations and ultimately improving the segmentation results. However, the combination of FCN and CRF estimates the context information in a fashion similar to down-sampling, thus losing significant amount of context information.

The FCNs have also been used in conjunction with attention overheads to improve the segmentation accuracies, However, these attention mechanisms are insufficient to provide explanations for contextual information accumulated. Although if the architecture can be provided with relation aware object category information, not only would it be enough to improve the segmentation accuracy but also might enhance the transparency for context adaptation.

The encoder-decoder based SegNet [16] architecture produces dense pixel-wise labels utilizing the convolutional layers in encoder. The network encodes the feature maps using the convolutional layers, while the decoder up-samples the low-resolution feature maps with help of pooling indices. The up-sampled input resolution feature maps are fed to pixel-wise classification layer to produce final pixel labels. The SegNet does not store the full feature map in decoder path, hence is less efficient in comparison to other approaches considering entire feature maps. In addition, it lacks the consideration of any pixel-wise associations in terms of context information to produce final labels. The Parsnet [17] presents relatively simpler approach to add global context information to DCNNs. The network averages the features at individual layers and the resultant context features are appended to subsequent layer. The additional consideration of pooled features improve the results on benchmark data, however lacks the actual incorporation of global information. The Dual Relation-Aware Attention Network (DRANet) [18] integrates the contextual information with FCN using two additional attention modules. The attention mechanisms in DRANet are explored in spatial and channel dimensions. The pixels and channels aggregate the context features as per their association with other pixels or channels. However, the drawback of these attention modules is their limitation is few numbers of centers around which the attention information is acquired, as the consideration of whole image increases computational and memory overheads.

The Criss-Cross Net (CCNet) [19] proposes criss-cross attention module to capture the context information among all set of pixels. Although the CCNet computes information in horizontal and vertical manner, it not only lacks the overall connections among pixels present in the surrounding area but it can also be improved to estimate the context information on object category level. Whereas Object Context Network (OCNet) [20] pays attention individually to all the pixels belonging to object category. The OCNet explores a binary relationship matrix between the defined object category and all other category pixels. While the attention of binary object category is beneficial to the cause, the network still lacks co-occurrence information between multiple object categories at once.

The image parsing architecture proposed over the years have made significant attempts to incorporate context information. Such techniques consider variations in the conventional convolutional layers, dilated convolutional layers and their systematic fusion to deal with the contextual attributes. The context information estimated however, lacks information among relationship between adjacently occurring object categories. Many of recent segmentation networks consider the traditional features representations but lack ability to compute whole context of image i.e. the semantic object categories occurring besides the superpixels. While the object categories have strong correlation with co-occurring objects categories, it becomes essential to reformulate image parsing to consider relation information in image scenes

The modern segmentation architecture despite being very considerate for generation of context information provide room for investigations in terms of relation-awareness. The research establishes [21] the generation of adequate of contextual information leads to improved label segmentations. The scene-graph generation provides an additional context computation to traditional pipeline for image parsing architectures. And the computation of relation-awareness enhances the accuracies.

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

III. PROPOSED ARCHITECTURE

This section presents the detailed architecture of proposed image parsing framework and the scene-aware graph generation technique for contextual information. The fig. 2 represents the high-level algorithm of the proposed architecture and the main components involving the superpixel computation, feature extraction, contextual feature estimation using graphs, and the final pixel labelling. The implementation is divided into two components for presentation, initially the graph-based context learning technique is described, which is followed by the details on incorporation of the estimated graph-based context information into the image parsing architecture to produce final pixel labels.

A. *Graph-based Context Learning Technique*

Assuming $G = (V,E)$ is an undirected graph with vertices $v \in V$, the set of object categories, and edges $(v_j, v_k) \in E$ represent the relationship between pixel object categories. The weight $w(v_j, v_k)$ of edges $(v_j, v_k) \in E$ is a positive measure of adjacency and connectivity representing the strength of relationship. The weight $w(v_j, v_k)$ between two nodes can be estimated as:

$$w(v_j, v_k) = \begin{cases} \sum_{\alpha=1}^{M} \left(\gamma * \mu\left(r_i^j, r_i^k\right)\right)_\alpha + 1, & j \neq k \\ 0, & if\ \alpha = 0 \\ \sum_{\alpha=1}^{M} \left(\beta * \mu\left(r_i^j, r_i^k\right)\right)_\alpha + 1, & j = k \end{cases}$$

where γ and β are regulatory parameters that help capturing the adjacency of the superpixel regions r_i^j and r_i^k, and α represents number of connections (M adjacently connected superpixels) between class labels j and k. The equation represents three possibilities while estimating the weights of connectivity, which are as:

a) The superpixel regions r_i^j and r_i^k with corresponding class label j and k, segmented from image I_i are adjacent to each other and have α connections.

b) The superpixel regions r_i^j and r_i^k are not connected to each other.

ALGORITHM 1: GRAPH GENERATION FOR CONTEXT LEARNING TECHNIQUE

Input: Superpixel Segmented Image Labels
Initialize: Input Image Superpixels, Corresponding Labels
Output: Relationship Aware Graph

```
for each image
    construct the superpixels S from Image I
    construct label graph G = (V,E)
    pick a superpixel V_i
    for each superpixel
        calculate the adjacent superpixel R_i = {r_1, r_2, r_2 ..., r_n}
        for each V_{i,j} ∈ R_i
            if Adjacent Connectivity = True
                update the weight of edge
                update R_agg(r_i^j, r_i^k)
            endif
        end
    end
end
```

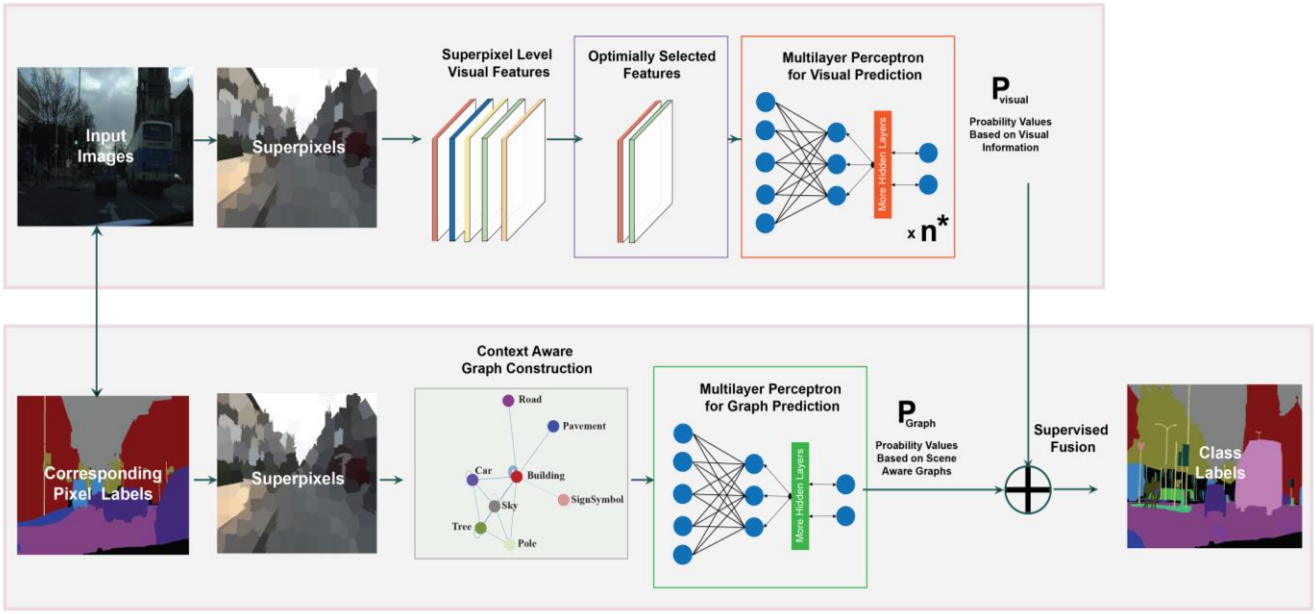

Fig. 2. The proposed scene specific graph-based context adaptation image parsing framework. (A) The top layer describes overview of visual probability estimation by initially computation of superpixels, estimation of visual attributes from the provided superpixels, the selection of features selection and classifier training. (B) The bottom layer describes the generation of scene aware context graphs, these constructed graphs are generalized for whole dataset and are used to estimate probabilities of pixel labels, finally these probabilities are fused assign each object category a pixel label.

c) The superpixel regions r_i^j and r_i^k segmented from image I_i are adjacent to each other in α appearances and have same class labels i.e., $j = k$.

The relation expression between superpixel regions r_i^j and r_i^k

$$R_{r_i^j, r_i^k} = \left\{ R^{adjacency}_{r_i^j, r_i^k}, R^{position}_{r_i^j}, R^{position}_{r_i^k} \right\}$$

where the adjacency metric may be represented as

$$R^{adjacency}_{r_i^j, r_i^k} = w(r_i^j, r_i^k)$$

$$R^{position}_{r_i^j} = [x_i^j, y_i^j, w_i^j, h_i^j],$$

and

$$R^{position}_{r_i^k} = [x_i^k, y_i^k, w_i^k, h_i^k]$$

The relation expressions between every superpixel region in the training dataset is aggregated to form relationship-aware matrix. The aggregated relational expression is given as:

$$R^{agg}(r_i^j, r_i^k) = \sum_{r_i^j, r_i^j \ \forall \ i=1,2,\ldots N} R_{r_i^j, r_i^k}$$

where the adjacency metric is aggregated for every image in the dataset. so, we normalize the relation aware expression estimated using graph generation to probability value as

$$P^{Graph}\left(l_{j,k \in class\ labels} \middle| r_i^j\right) = \frac{R^{agg}(r_i^j, r_i^k)}{\sum_{1<j<M} R^{agg}(r_i^j, r_i^k)}$$

where l_j in this expression represents the label for class j, while l_k refers to every label present in the dataset including j, while M is total number of class labels.

Thus, for image parsing problem the graph construction can help investigating the contextual relationship among pixel-wise label classes. The elements in V are superpixels and the weight of an edge is the measure of adjacency between superpixels ultimately providing the information on adjacency between semantic label categories. The weight of edges between two vertices of different object class proves extremely beneficial along with weights of edges for same class superpixels. This provides more likeliness of computing intrinsic information for contextual superpixel occurrences.

B. Image Parsing Architecture

The overall image parsing architecture utilizes the estimated superpixels to extract the visual features, and highly important and optimally selected feature vectors are used to train the class-specific classifiers. The graph-based context aware probabilities are used in conjunction with multilayer-perceptron based pixel label probabilities to produce final category-wise labels. The mathematical foundation and structure of the image parsing architecture is briefly presented in the subsection.

Let $I(v) \in \mathbb{R}^3$ be an image having set of pixels v, the purpose of image parsing is to assign class labels:

$$L = \{l_i | i = 1,2,3, \ldots, M\}$$

and M is the number of all classes. For superpixel based

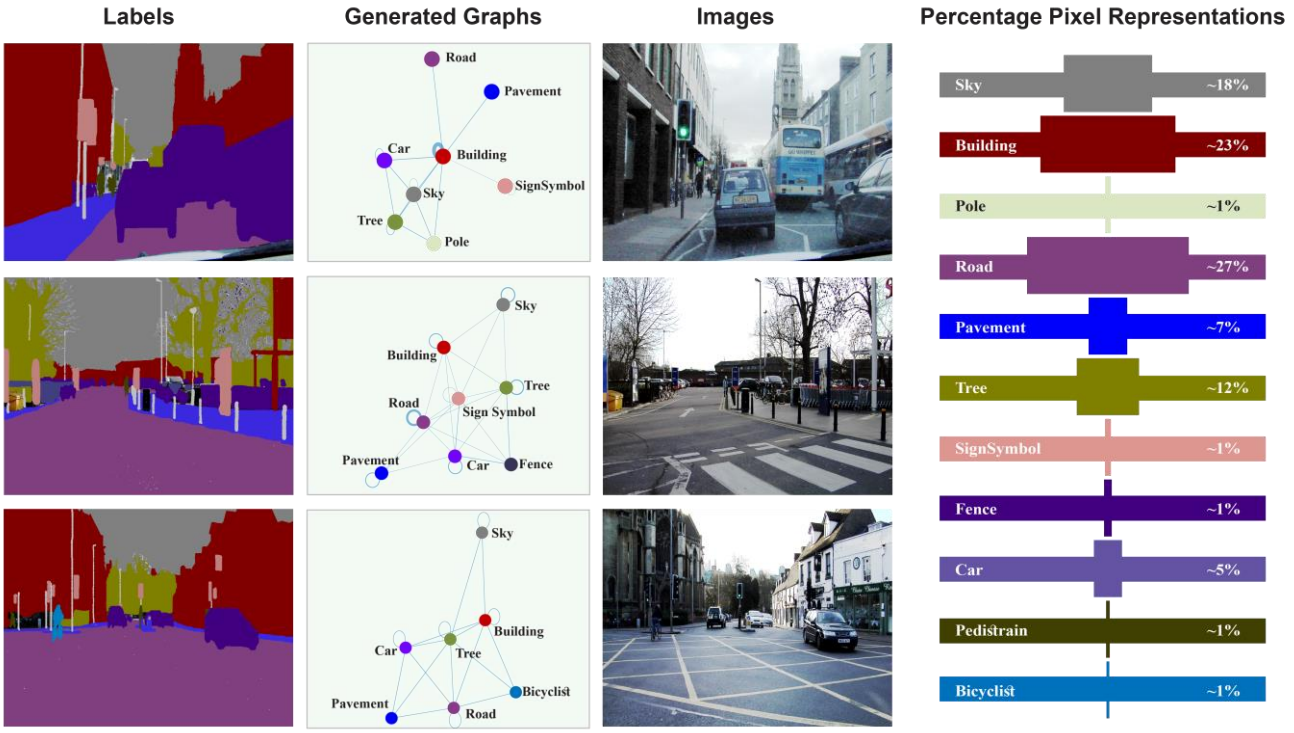

Fig. 3. The figure describes the generation of graphs on CamVid dataset and percentage wise pixel ratio to the overall dataset for each category. It can be observed that the CamVid dataset is dominated by the sky, building, road and tree object categories in terms of pixel representations. The remaining classes collectively constitute of merely about 17% of the data.

object classification, let

$$S(r) = \{r_j | j = 1,2,3, \dots, N\}$$

indicate the superpixel over segmented from I and N be the number of all superpixels. The corresponding visual features are

$$F^r = \{f_j^r | j = 1,2,3, \dots N\}$$

where F^r is the set of visual features extracted. The task is reformulated to label the pixels v in a superpixel r_j. While in similar pattern the relationship between the superpixel can be modeled using the graph superpixel information estimated from input images. The probability based on the visual features can be estimated as:

$$P^{Visual}(l_i|r_j, \aleph_\theta) = P(l_i|f_j^r)$$
$$\text{s.t } \sum_{1<i<M} P^{Visual}(l_i|r_j) = 1$$

where $\aleph_\theta$ is the binary trained classifier for class i. The probabilities obtained from the scene-aware graph generation and visual features based trained classifiers can be fused to obtained the final superpixel as:

$$P(l_i|r_j) = b_{1,i}^c + w_{1,i}^V * P^{Visual}(l_i|r_j) + w_{1,i}^G * P^{Graph}(l_{j,k \in class\ labels}|r_i^j)$$

For all M classes it can be presented as

$$P(L|r_j) = [P(l_1|r_j), \dots, P(l_i|r_j), \dots, P(l_M|r_j)]$$

Finally, the superpixel r_j will be assigned to the class label $\acute{l}$ which has the maximum probability across all classes using a majority voting strategy:

$$r_j \in \acute{l}\ if\ P(\acute{l}|r_j) = \max_{1<i<M} P(l_i|r_j)$$

IV. Experimental Evaluation.

A. Experimental Setup

MATLAB 2020a is utilized to implement the proposed graph-based context learning image parsing framework. Deep Learning Toolbox, Optimization, and Machine Learning Toolbox are some of the frequently used toolboxes to explore the functional implementations. Utilizing the High-Performance Computing (HPC) system made available by Griffith University, the experiments are carried out. 64 CPUs and 128 GB of RAM are two of the resources that have been frequently used.

B. Superpixel Computation

The initial step in the overall network is compute the superpixels from the provided image. In our method we use SLIC to segment the image into superpixels. The superpixel reduce the overall number of inputs by grouping the pixels and allow the computation of meaningful feature representations. For the sake of this study, the number of superpixels computed from each image is maintained at K = 512, which is well suited in terms of both the computational cost and accuracy for superpixel labels.

C. Benchmark Dataset

1) CamVid Dataset

The Cambridge-driving Labelled Video Database (CamVid) constitutes of driving scenarios captures in 5 sequences originally at the resolution 960 x 720. The footages captured were sample individually to provide 701 total frames with corresponding labelled object categories. The dataset provides driving perspective and heterogeneous occurrences of the object categories. The research community has been using 11 labelled object categories comprising majority of the labelled pixels in the dataset. These classes include sky, building, pole, road, pavement, tree, sign-symbol, fence, car, pedestrian, and bicyclist. The percentage representation of the pixels and some visual samples can be observed in fig. 3.

2) Stanford background Dataset

The Stanford background dataset is a collection of images featuring both urban and rural scenes. It includes a total of 715 images, each with dimensions of approximately 240 x 320 pixels. These images come from multiple datasets and contain a variety of objects, including sky, trees, roads, grass, water, mountains, and buildings. The foreground is also considered a separate class, making a total of 8 object classes in the dataset. Researchers often use this dataset for training and testing machine learning models, with the typical split being 80% for training and 20% for testing. The diverse range of objects and settings in this dataset make it a valuable resource for a variety of tasks related to image classification and object recognition.

D. Evaluation Metrics

The proposed image parsing framework is evaluated using the most widely used evaluation metrics that include pixel accuracy, mean pixel accuracy, and mean intersection over union. The pixel accuracy is perceived to be more biased towards the representation of classes with higher pixel-wise occurrences and is given as:

$$Pixel\ Accuracy = \frac{\sum_{i=0}^{m} p_{ii}}{\sum_{i=0}^{m} \sum_{j=0}^{m} p_{ij}}$$

where m is the total number of classes, p_{ii} represents total number of true positives, and p_{ij} represents false positives.

While the mIoU is estimated by taking the average of IoU values produced for each class, and the IoU is defined as the area of overlap between the predicted labels and ground truth labels divided by the area of union between these two. The mathematical formulation is given as:

$$mIoU = \frac{1}{m+1} \sum_{i=0}^{m} \frac{p_{ii}}{\sum_{j=0}^{m} p_{ij} + \sum_{j=0}^{m} p_{ji} - p_{ii}}$$

where m is number of classes, i represents actual class of pixel, j highlights predicted class of pixel, the true positive values are represented by p_{ii} , the number of false positives and false negatives are represented by p_{ij} , and p_{ji} respectively.

E. Evaluation Results

The proposed Graph-based context learning technique is evaluated on benchmark datasets. In this subsection, the results are presented using the evaluation metrics mostly used in the literature. Firstly, we present the overall mIoU scores for Camvid data and global pixel accuracy on Stanford background dataset. Next, detailed and comprehensive class-wise comparison is presented in contrast to the existing image parsing techniques.

Table 1 The proposed architecture with feature selection module achieves competitive global and average accuracies on the Camvid dataset. The (%) accuracies values are listed in comparison to the previous approaches.

Method	SegNet [16]	ENet [23]	SegNet v2 [16]	NDNet45-FCN8-LF [41]	JPANet-S [42]	DFANet-A [8]	BiSeNet1 [29]	WFDCNet [43]	LBN-AA [44]	BiSeNet2 [29]	AGLNet [32]	DSANet [35]	RELAXNet [36]	RegSeg [16]	**Proposed Approach**
IoU(%)	46.4	51.3	55.6	57.5	63.8	64.7	65.6	67.5	68	68.7	69.4	69.9	71.2	80.9	81.2

Table 2 The proposed scene-aware image parsing framework acquires better class-wise accuracies on the Camvid dataset. The (%) accuracies are presented in comparison with previous methods.

Method / Category	**Building**	**Tree**	**Sky**	**Car**	**SignSymbol**	**Road**	**Pedestrian**	**Fence**	**Pole**	**Pavement**	**Bicyclist**
SFM + App. [22]	46.2	61.9	89.7	68.6	42.9	89.5	53.6	46.6	0.7	60.5	22.5
ENet [23]	74.7	77.8	95.1	82.4	51	95.1	67.2	51.7	35.4	86.7	34.1
FCN+Comb [24]	79.7	77.2	85.7	86.1	45.3	94.9	45.9	69	25.2	86.2	57.9
L.L.D. [25]	80.7	61.5	88.8	16.4	n/a	98	1.09	0.05	4.13	12.4	0.07
B-Net-VGG-LCM [26]	81.4	75.3	92.8	82.5	42.8	89.2	60.8	47.8	36.3	66.4	54.8
Boosting [27]	81.5	76.6	96.2	78.7	40.2	93.9	43	47.6	14.3	81.5	33.9
DeepLab – LFOV [28]	81.5	74.6	89	82.2	42.3	92.2	48.4	27.2	14.3	75.4	50.1
BiSeNet1 [29]	82.2	74.4	91.9	80.8	42.8	93.3	53.8	49.7	25.4	77.3	50
G-FRNet [30]	82.5	76.8	92.1	81.8	43	94.5	54.6	47.1	33.4	82.3	59.4
Dilation 8 [31]	82.6	76.2	89	84	46.9	92.2	56.3	35.8	23.4	75.3	55.5
AGLNet [32]	82.6	76.1	91.8	87	45.3	95.4	61.5	39.5	39	83.1	62.7
BiSeNet2 [29]	83	75.8	92	83.7	46.5	94.6	58.8	53.6	31.9	81.4	54
Dilation + FSO-DF [33]	84	77.2	91.3	85.6	49.9	92.5	59.1	37.6	16.9	76	57.2
CNN-CRF [34]	84.3	65.3	95.6	74.6	0.4	93.5	25.6	32.3	13.8	85	54.3
DSANet [35]	84.3	77.8	92.1	86.1	51.2	94.8	62.8	41.9	35.9	82.2	60.1
RELAXNet [36]	84.8	78.2	93.2	84.5	50.3	94.7	64.3	47.8	45.1	83.4	56.9
ReSeg [37]	86.8	84.7	93	87.3	48.6	98	63.3	20.9	35.6	87.3	43.5
Super Parsing [38]	87	67.1	96.9	62.7	30.1	95.9	14.7	17.9	1.7	70	19.4
FPANet [39]	88.8	78.7	92.7	93.1	28.4	94.3	75.2	56	50.3	90.4	73.4
SegNet [16]	88.8	87.3	92.4	82.1	20.5	97.2	57.1	49.3	27.5	84.4	30.7
CAB-Net [40]	91.1	88.9	95.7	93	64.8	94.7	66.5	70.5	29.8	85.3	60.3
Proposed Approach	95.93	94.06	94.86	96.98	86.71	95.6	40.61	70.78	89.38	87.23	66.97

To evaluate our proposed technique against the state-of-the-art methods, firstly we present evaluation results on Camvid benchmark dataset. Graph-based context learning technique achieves 81.2% mIoU, which is noticeably higher than other image parsing methods. The widely cited RegSeg [16] achieved 80.9% mIoU value on the said data. While the proposed approach outperforms recently proposed image parsing techniques like BiSeNet [29], DSANet [35], and RelaxNet [36].

The proposed graph-based context learning technique achieves better class-wise accuracies on the Camvid dataset, outperforming existing image parsing methods for majority of the object categories. Our method achieves 95.9%, 94.1%, 94.9%, 96.9%, 86.7%, 96%, 40.6%, 70.8%, 89.4%, 87.2%, and 67% respectively for object categories building, tree, sky, car, sign-symbol, road, pedestrian, fence, pole, pavement and bicyclist classes. The graph-based technique improves more than 5% for building and tree classes, while the scores are comparably better for sky, car, fence pole and bicyclist in comparison to FPANet [39], SegNet [16], and CAB-Net [40]. The class-wise pixel accuracy value for pedestrian is lower in comparison to the other approaches, apparently the training pixels for pedestrian class are

Table 2 The proposed Graph-based context learning technique achieves better accuracy scores on the Stanford Background Dataset. The accuracy values listed in comparison with state-of-the-art architectures.

	Sky	**Tree**	**Road**	**Grass**	**Water**	**Building**	**Mountain**	**Foreground**
Gould [45]	92.6	61.4	89.6	82.4	47.9	82.4	13.8	53.7
Munoz [46]	91.6	66.3	86.7	83.0	59.8	78.4	5.0	63.5
Sharma [47]	94.8	71.6	90.6	88.0	73.5	82.2	10.2	59.9
DeepLab v3 [48]	89.3	72.2	87.2	77.4	72.7	80.0	48.6	66.9
DeepLab v3 2 [48]	89.3	72.5	87.3	77.5	72.7	80.0	49.1	66.6
DeepLab v3+LoAd [49]	89.4	72.8	87.7	77.9	73.8	80.8	50.1	67.5
Proposed Approach	95.3	86.8	94.9	93.7	93.3	93.8	89.6	81.8

Table 4 Global pixel accuracy values listed in comparison with state-of-the-art architectures. The proposed Graph-based context learning technique achieves better accuracy scores on the Stanford Background Dataset

Method	**Gould [45]**	**Kumar [50]**	**Lemprisky [51]**	**Farabet [52]**	**Sharma [47]**	**Luc [53]**	**DeepLab [48]**	**SegWgan [54]**	**Deeplab Load [49]**	**Munoz [46]**	**Proposed Approach**
Pixel Accuracy	76.4	79.4	81.9	81.4	82.3	75.2	87.0	87.7	75.0	70.7	**89.7**

relatively lower than majority of other classes. It leads to the argument that graph-based context learning requires adequate amount of pixel values to learn the contextual information.

The results presented in table 4 demonstrate the efficacy of the proposed graph-based context module in improving the accuracy of pixel labelling for object categories in Stanford background dataset. The proposed approach achieved an overall improvement in global pixel accuracy of 89.7% on the Stanford dataset when compared to state-of-the-art methods. This significant increase in global pixel accuracy value reveals that the proposed approach is capable of accurately producing pixel labels with corresponding object categories, which is a crucial aspect of image parsing.

The proposed architecture is evaluated on Stanford background dataset, the comparison is present with state-of-art-existing techniques. Table 3 presents the class-wise pixel accuracy values using the proposed technique compared with other techniques. The graph-based context learning technique achieves about 94%, 87%, 95%, 94%, 93%, 94%, 90% and 82% for object categories sky, tree, road grass, water, building, mountain, and foreground. The proposed technique achieves comparatively better scores object categories mountain and foreground in comparison to the other approaches. The global pixel accuracy obtained using the graph-based learning technique is 89.7% that is quite better in comparison to other techniques. The achieved scores are 7% better than Lemprisky [51], 8% better than Farabet [52], about 3% improved than DeepLab [48], and 2% higher than state-of-the-art SegGan [54]. Thus, proving the overall efficacy of the proposed graph-based context learning technique on the Stanford benchmark dataset.

The proposed approach, by including graph-based context information to the architecture, not only improves the overall efficacy of the proposed architecture but also improves class-wise accuracies for object-categories with least pixel representations. This is particularly important for image parsing tasks, as object categories with fewer pixels. The proposed graph-based context module addresses this issue by incorporating additional contextual information using graphs, which allows the architecture to label these object categories more accurately.

F. Enhancement using Graph-based Context Learning Technique

The graph-based context enhances the global pixel accuracy values from 85.6 % to 91.8% by including the said module for Stanford dataset. There is noticeable percentage increase in accuracy scores for object categories building sky, car, road, while the accuracy scores are way higher for other object categories in the Camvid dataset with inclusion of graph-based context information. Table 5 presents the pixel accuracy values for each category for the proposed technique with inclusion of graph context module. The pixel-wise accuracy for classes building, tree, sky, car, enhanced about 6%, 4%, 6%, while car, pedestrian, fence, pavement and bicyclist improved 17%, 15%, 10%, 19%, and 21% respectively.

The table 5 also shows the result of proposed approach on Stanford background dataset. It can be observed that the incorporation of graph-based context improves the pixel wise accuracy values for object categories sky, tree, road, grass, water, building, mountain, and foreground about 2%, 6%, 3%, 8%, 13%, 7%, 3%, and 8% respectively. The global pixel accuracy value improved from 86.2% to 89.7%. It can be argued that the proposed graph-based context incorporation not only improves the overall efficacy of the proposed architecture but also improves class-wise accuracies for object-categories with least pixel representations. Moreover, it is also worth noting that the proposed approach has a greater impact on object categories with fewer pixel representations.

Table 5 The accuracy value comparison for Stanford and Camvid Dataset describing the efficacy of proposed Graph-based Context inclusion.

Camvid Dataset			Stanford Dataset		
	(%) Accuracy Values Proposed Approach			(%) Accuracy Values Proposed Approach	
Graph Context Module		☑	Graph Context Module		☑
Building	89.93	95.93	Sky	93.3	95.3
Tree	80.64	94.06	Tree	80.2	86.8
Sky	88.62	94.86	Road	91.6	94.9
Car	79.04	96.98	Grass	85.1	93.7
Sign Symbol	31.85	86.71	Water	80.2	93.3
Road	80.60	88.07	Building	86.8	93.8
Pedestrian	26.83	40.61	Mountain	86.7	89.6
Fence	61.43	70.78	Foreground	73.9	81.8
Pole	79.01	89.38	Global Pixel Accuracy	86.2	89.7
Pavement	68.30	87.23			
Bicyclist	45.06	66.97			
Global Pixel Accuracy	85.63	91.80			

V. Conclusion

In conclusion, we have presented a novel image parsing framework that utilizes graph-based context learning technique to improve pixel labelling accuracy. Our proposed approach incorporates additional contextual information by generating graphs from object category pixel labels and computing contextual information using these graphs. Our experimental results on the Stanford Background Dataset and the Camvid benchmark dataset demonstrate that our proposed approach outperforms state-of-the-art techniques, achieving 89.7% global pixel accuracy on the Stanford Background Dataset and 81.2% mean intersection over union (mIoU) on the Camvid benchmark dataset. The incorporation of scene aware graph information proves beneficial in achieving these results and holds great potential for further improvements in image parsing. The proposed graph-based context learning technique improves the accuracy scores by notable margins. Our results suggest that it would be beneficial to further investigate the proposed architecture on other real-world image parsing datasets and explore ways to scale it to handle larger and more complex datasets. Our work is a significant contribution to the field of image parsing, and we believe it has the potential to lead to new advancements in this area.

Acknowledgment

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

References

1. M. Everingham, S. Eslami, L. Van Gool, C. Williams, J. Winn and A. Zisserman, "The Pascal Visual Object Classes Challenge: A Retrospective", International Journal of Computer Vision, vol. 111, no. 1, pp. 98-136, 2014.
2. X. Ding, C. Shen, T. Zeng and Y. Peng, "SAB Net: A Semantic Attention Boosting Framework for Semantic Segmentation", IEEE Transactions on Neural Networks and Learning Systems, pp. 1-13, 2022.
3. B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso and A. Torralba, "Scene Parsing through ADE20K Dataset", 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.
4. H. Caesar, J. Uijlings and V. Ferrari, "COCO-Stuff: Thing and Stuff Classes in Context", 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018.
5. L. Wang, L. Zhang, X. Qi and Z. Yi, "Deep Attention-Based Imbalanced Image Classification", IEEE Transactions on Neural Networks and Learning Systems, vol. 33, no. 8, pp. 3320-3330, 2022.
6. D. Lin, Y. Wang, L. Liang, P. Li and C. Chen, "Deep LSAC for Fine-Grained Recognition", IEEE Transactions on Neural Networks and Learning Systems, vol. 33, no. 1, pp. 200-214, 2022.
7. Y. Liu, Y. Wu, P. Wen, Y. Shi, Y. Qiu and M. Cheng, "Leveraging Instance-, Image- and Dataset-Level Information for Weakly Supervised Instance Segmentation", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 44, no. 3, pp. 1415-1428, 2022.
8. Z. Cao, G. Hidalgo, T. Simon, S. Wei and Y. Sheikh, "OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 43, no. 1, pp. 172-186, 2021.
9. R. Girshick, J. Donahue, T. Darrell and J. Malik, "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation", 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) , 2014.
10. S. Minaee, Y. Boykov, F. Porikli, A. Plaza, N. Kehtarnavaz and D. Terzopoulos, "Image Segmentation Using Deep Learning: A Survey", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1-1, 2021.
11. M. Seyed Hosseini and T. Tasdizen, "Semantic Image Segmentation with Contextual Hierarchical Models", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 5, pp. 951-964, 2016.
12. I. Kokkinos, G. Evangelopoulos and P. Maragos, "Texture Analysis and Segmentation Using Modulation Features,

Generative Models, and Weighted Curve Evolution", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 1, pp. 142-157, 2009.
13. A. Chuchvara and A. Gotchev, "Content-Adaptive Superpixel Segmentation Via Image Transformation", *2019 IEEE International Conference on Image Processing (ICIP)*, 2019, pp. 1505-1509.
14. M. Zand, S. Doraisamy, A. Abdul Halin and M. Mustaffa, "Ontology-Based Semantic Image Segmentation Using Mixture Models and Multiple CRFs", IEEE Transactions on Image Processing, 2016, vol. 25, no. 7, pp. 3233-3248.
15. E. Shelhamer, J. Long and T. Darrell, "Fully Convolutional Networks for Semantic Segmentation", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 4, pp. 640-651, 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. W. Liu, A. Rabinovich, and A. C. Berg, "Parsenet: Looking wider to see better," *arXiv.org*, pp. 1-11, (2015)
18. J. Fu, J. Liu, J. Jiang, Y. Li, Y. Bao and H. Lu, "Scene Segmentation With Dual Relation-Aware Attention Network", IEEE Transactions on Neural Networks and Learning Systems, vol. 32, no. 6, pp. 2547-2560, 2021.
19. Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei and W. Liu, "CCNet: Criss-Cross Attention for Semantic Segmentation", *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 603-612.
20. Y. Yuan, L. Huang, J. Guo, C. Zhang, X. Chen and J. Wang, "OCNet: Object Context for Semantic Segmentation", *International Journal of Computer Vision*, vol. 129, no. 8, pp. 2375-2398, 2021.
21. J. He, Z. Deng, L. Zhou, Y. Wang and Y. Qiao, "Adaptive Pyramid Context Network for Semantic Segmentation", *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
22. P. Sturgess, K. Alahari, L. Ladický, and P. H. S. Torr, "Combining appearance and structure from motion features for road scene understanding," British Machine Vision Conference, pp. 1–11, (2009).
23. Paszke, Adam, et al. "Enet: A deep neural network architecture for real-time semantic segmentation." arXiv preprint arXiv:1606.02147 (2016).
24. Y. Wu, T. Yang, J. Zhao, L. Guan, and J. Li, "Fully combined convolutional network with soft cost function for traffic scene parsing," in Intelligent Computing Theories and Applications, vol. 10361, pp. 725–731, (2017).
25. Y. Yang, Z. Li, L. Zhang, C. Murphy, J. Ver Hoeve, and H. Jiang, "Local label descriptor for example based semantic image labeling," in Lecture Notes in Computer Science, vol. 7578, pp. 361–375, (2012).
26. Hua, Cam-Hao, Thien Huynh-The, and Sungyoung Lee. "Convolutional networks with bracket-style decoder for semantic scene segmentation." 2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC). IEEE, 2018.
27. L. Ladický, P. Sturgess, K. Alahari, C. Russell, and P. H. S. Torr, "What, where and how many? Combining object detectors and CRFs," in Lecture Notes in Computer Science, vol. 6314 LNCS, pp. 424–437, (2010).
28. 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," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 40, no. 4, pp. 834-848,
29. Yu, Changqian, et al. "Bisenet: Bilateral segmentation network for real-time semantic segmentation." Proceedings of the European conference on computer vision (ECCV). 2018.
30. Amirul Islam, Md, et al. "Gated feedback refinement network for dense image labeling." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017 , pp. 3751-3759
31. Fisher Yu, Vladlen Koltun , "Multi-scale context aggregation by dilated convolutions" in International Conference on Learning Representations (ICLR), 2016, pp. 1-13
32. Zhou, Quan, et al. "AGLNet: Towards real-time semantic segmentation of self-driving images via attention-guided lightweight network." in Applied Soft Computing, vol. 96 (2020).
33. Kundu, Abhijit, Vibhav Vineet, and Vladlen Koltun. "Feature space optimization for semantic video segmentation." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.
34. J. M. Alvarez, Y. LeCun, T. Gevers, and A. M. Lopez, "Semantic Road segmentation via multi-scale ensembles of learned features," Lecture Notes in Computer Science, vol. 7584, pp. 586–595, (2012).
35. Elhassan, Mohammed AM, et al. "DSANet: Dilated spatial attention for real-time semantic segmentation in urban street scenes." in Expert Systems with Applications vol. 183 (2021).
36. Jin Liu, Xiaoqing Xu, Yiqing Shi, Cheng Deng, Miaohua Shi, "RELAXNet: Residual efficient learning and attention expected fusion network for real-time semantic segmentation," in Neurocomputing vol. 474, pp. 115-127 (2022).
37. F. Visin et al., "ReSeg: A Recurrent Neural Network-based Model for Semantic Segmentation," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 426–433, (2016).
38. X. Ren, L. Bo, and D. Fox, "RGB-(D) scene labeling: Features and algorithms," in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 2759–2766, (2012).
39. Y. Wu, J. Jiang, Z. Huang, and Y. Tian, "FPANet: Feature pyramid aggregation network for real time semantic segmentation." Applied Intelligence, vol. 52, pp. 3319-3336, (2021)
40. Cam-Hao Hua, Thien Huynh-The, Sung-Ho Bae, and Sungyoung Lee, "Cross-Attentional Bracket-shaped Convolutional Network for semantic image segmentation" in Information Sciences, vol. 539, pp. 277-294 (2020)
41. Yang, Zhengeng, Hongshan Yu, Qiang Fu, Wei Sun, Wenyan Jia, Mingui Sun, and Zhi-Hong Mao. "NDNet: Narrow while deep network for real-time semantic segmentation." IEEE Transactions on Intelligent Transportation Systems 22, no. 9 pp. 5508-5519 (2020).
42. Hu, Xuegang, Liyuan Jing, and Uroosa Sehar. "Joint pyramid attention network for real-time semantic segmentation of urban scenes." Applied Intelligence vol. 52, pp. 580-594 (2022).
43. Hao, Xiaochen, Xingjun Hao, Yaru Zhang, Yuanyuan Li, and Chao Wu. "Real-time semantic segmentation with weighted factorized-depthwise convolution." Image and Vision Computing vol. 114, (2021).
44. Dong, Genshun, Yan Yan, Chunhua Shen, and Hanzi Wang. "Real-time high-performance semantic image segmentation of urban street scenes." IEEE Transactions on Intelligent Transportation Systems vol. 22, pp. 3258-3274 (2020).

45. D. Munoz, J. A. Bagnell, and M. Hebert, “Stacked hierarchical labeling,” in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2010, vol. 6316 LNCS, no. PART 6, pp. 57–70
46. S. Gould, J. Rodgers, D. Cohen, G. Elidan, and D. Koller, “Multi-class segmentation with relative location prior,” International Journal of Computer Vision, vol. 80, no. 3, pp. 300–316, (2008).
47. A. Sharma, O. Tuzel, and D. W. Jacobs, “Deep Hierarchical Parsing for Semantic Segmentation,” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 530-538, (2015).
48. L. 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, (2017).
49. H.J. Rad and A. Szabo, “Lookahead adversarial learning for near real-time semantic segmentation." Computer Vision and Image Understanding, vol. 212, (2021).
50. M. P. Kumar and D. Koller, “Efficiently selecting regions for scene understanding,” in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2010, pp. 3217–3224.
51. V. Lempitsky, A. Vedaldi, and A. Zisserman, “Pylon Model for Semantic Segmentation.” pp. 1485–1493, 2011.
52. C. Farabet, C. Couprie, L. Najman, and Y. LeCun, “Learning Hierarchical Features for Scene Labeling,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 35, no. 8, pp. 1915–1929, 2013
53. P. Luc, C. Couprie, S. Chintala, and J. Verbeek, “Semantic Segmentation using Adversarial Networks,” In NIPS Workshop on Adversarial Training, 2016.
54. X. Zhu, X. Zhang, X.-Y. Zhang, Z. Xue, and L. Wang, “A Novel Framework for Semantic Segmentation with Generative Adversarial Network,” J. Vis. Commun. Image Represent., vol. 58, pp. 532–543, 2019.