Engineering and Technology Journal
Journal homepage: https://etj.uotechnology.edu.iq

Street Scene understanding via Semantic Segmentation Using Deep Learning

Amani Y. Noori [a*], Shaimaa H. Shaker [b], Raghad A. Azeez [c]

[a, b] computer science Engineering Dept., University of Technology-Iraq, Alsina'a street, 10066 Baghdad, Iraq.
[c] University of Baghdad, College of Education for Human Science-ibn rushed, Baghdad, Iraq.
*Corresponding author Email: amani.yousif@uomustansiriyah.edu.iq

HIGHLIGHTS

- Scene classification is an essential conception task used by robotics for understanding the environment.
- The deep learning technique has been proved as a great role in the challenging scene understanding application.
- Using data augmentation to increase dataset size
- Using K-means clustering as a preprocessor for the input dataset
- The proposed hydride model is generated by combined two of the deep, deep neural networks as an xception and U-net models.

ARTICLE INFO

Handling editor: Rana F. Ghani

Keywords: semantic segmentation; U-net; residual leaning; K-means clustering; RGB-D Xception.

ABSTRACT

Scene classification is an essential conception task used by robotics for understanding the environment. Like the street scene, the outdoor scene is composed of images with depth that has a greater variety than iconic object images. Image semantic segmentation is an important task for Autonomous driving and Mobile robotics applications because it introduces enormous information needed for safe navigation and complex reasoning. This paper provides a model for semantic segmentation of outdoor sense to classify each object in the scene. The proposed network model generates a hybrid model that combines U-NET with Xception networks to work on 2.5 dimensions cityscape dataset, which is used for 3D applications. This process contains two stages. The first is the pre-processing operation on the RGB-D dataset (data Augmentation and k- means cluster). The second stage designed the hybrid model, which achieves a pixel accuracy is 0.7874. The output module is generated using a computer with GPU memory NVIDIA GeForce RTX 2060 6G, programming with python 3.7.

1. Introduction

The deep learning technique has been proved as a great role in the challenging scene understanding application. It uplifts massive GPUs computing power for extracting features that combine with the input image semantics information. As a result, the interest in scene understanding community with the Deep Learning method increases. Many models have been proposed to classify outdoor scene tasks [1].

Semantic image segmentation is an important topic for understanding visual scenes. It is a classification for the multi-labeling problem. It aims to determine for every pixel in an image a class label. The semantic segmentation mixed the primary tasks of image segmentation and recognition of the object. This paper works on cityscape with depth dataset, which has many weakly labeled images. The challenge with working with 3D or 2.5D datasets is that they are not many accurate and high-resolution datasets in that field [2,3].

RGB-D image is equivalent to an RGB image and its matching depth image. A depth image is an image channel in which every pixel is associated with a distance in the midst of the object and matches the image plane in the RGB image. 3D Data represent in RGB-D representation using famous RGB-D sensors such as MICROSOFT KINECT. RGB-D describes three dimensions objects as 2,5 dimensions information, include two dimensions color information (RGB) with depth map (D) [4].

The Contribution to this paper is as follows:

http://doi.org/10.30684/etj.v40i4.2120
Received 13 April 2021; Accepted 23 July 2021; Available online 25 April 2022
2412-0758/©Publishing rights belongs to University of Technology's Press, Baghdad, Iraq

1) Using data augmentation to increase dataset size
2) Using K-means clustering as a preprocessor for the input dataset
3) Generate the hydride model by combining the U-Net and xception model.
4) Taking from xception network, the depthwise Separable convolution operation is followed by pointwise convolution.
5) Using residual learning rule as in xception network.

2. Related work

This section presents an overview of the networks used in semantic segmentation using deep learning.

2.1 Ikshananet-1

This network is based on the IKshana theory, which proposed a model for outdoor semantic segmentation work on the cityscape dataset. The IKshanaNet-1 architecture contains 3 blocks: glance module, projection module, and (1*1) convolution neural network layer. This network achieved MIOU:53.35. This model was trained on 2975 training images and 500 validation images[5].

2.2 SDC

This method (SDC) Semantic Divide and Conquer Network proposed to degrees monocular depth predication into that of a single semantic segment. A Deep Neural Network focuses on the divide and conquers semantic technique. This network subdivides the input scene into semantic segments for any instances object and background classes. Predicates shift and scale-invariant depth map for each semantic segment in the canonical space. Semantic segments for the similar class participate with the similar depth decoder for predication. The global depth is divided into a group of predications each category, which are easier to train and easy to generalize new scene kinds [6].

2.2.1 LightSeg

This module is faster and more efficient because it uses a new version of the atrous spatial pyramid pooling module, short and long residual connections, and depth-wise separable convolution. This network achieves an MIOU of 67.81% tested on the cityscape dataset [7].

2.2.2 ESPNet

This network depends on (ESP) efficient spatial pyramid modules, which use two types of convolutions: spatial pyramid of dilated convolutions and pointwise convolution. This network achieves MIOU: 60.3% tested on the cityscape dataset [8].

2.2.3 DSPNET

Propose an efficient approach for concurrent pixel semantic segmentation, depth estimation, and object detection by sharing the CNN structure. The introduced network model (DSPNet) named Driving Scene Perception Network uses multi-task learning and multi-level feature maps to improve single image input segmentation, object detection, and depth efficiency and accuracy. This method works on the cityscape dataset [9].

2.2.4 CMoDE

CMoDE (Convoluted Mixture of Deep Experts) is a fusion technique proposed for semantic segmentation tasks that used a multiple stream deep NN for features training from many complemented modalities. Deep Convolution Neural Network includes 3 components: experts who map spectra or modalities for the segment of the outputs, the CMoDE adjust weights class for one feature of expert networks by using the trained probability distributions, and the fusion segment that further learns complementary fused kernels [10].

2.2.5 ENet

ENet(Efficient neural network) used for outdoor semantic segmentation designed for mobile devices achieves MIOU:58.3tested on the cityscape dataset[11].

3. Methodology

Accurate semantic segmentation of outdoor scenes is important to increase the efficiency of scene understanding tasks. These are complicated and need planning and perception to classify objects used by robotics for many applications like self–driving – care and remote scene classification for military purposes. Semantic Segmentation provides information about the scene objects' components, classified into 8 categories: construction, flat, sky, human, object, void, vehicle, and nature [3]. The proposed algorithm for the proposed system contains many stages, as shown in the following diagram figure (1).

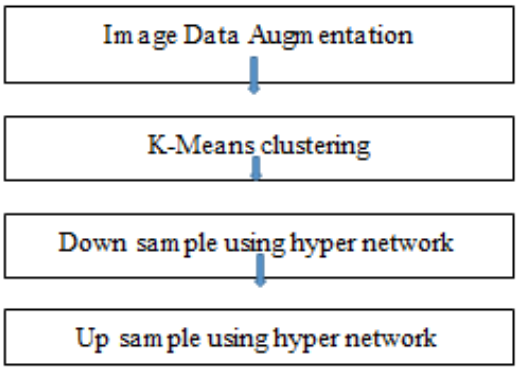

Figure 1: block diagram of the proposed system

The proposed system contains two main operations:

3.1 Image pre-processing

3.1.1 Image data augmentation

Image data augmentation is a technique used to enhance the performance of a deep learning model by increasing the quality and size of training data. It involves expanding the small dataset that generates different images to utilize the skill of big image data that enhances the model to generalize for learning new images, including many operations rotate, flip, zoom, clip [12].

3.1.1.1 Rotation

This operation rotates the image clockwise or counterclockwise by a number of degrees between 0 and 360. However, the effectiveness of rotation augmentation is hard to determine because it deepened on rotation parameters. For example, slight rotation done between -1 to -20 or 1 and 20 can be used for digit recognition [13].

3.1.1.2 Flipping

The flipping operation for an image means reversing the rows or columns of pixels in the case of a vertical or horizontal flip, respectively. Vertical axis flipping is less common than horizontal axis flipping; this operation is easy to implement [12].

3.1.2 K-Means clustering algorithm

it is an unsupervised clustering algorithm. This algorithm is used if having unlabeled data (data without any group or categories). This algorithm is used to search for any similarity in the data to form groups represented (K). It is also used to segment or partition the dataset into K-clusters or parts. The K-Means (K-centroids) can take any interesting area to segment the object from the background. Due to the compression of the JPEG standard, there are more than only a few discrete colors in the segmented image. K-Means is used to find the most important colors and identify similar colors [14].

3.2 The proposed hybrid networks

The proposed network provides 74 layers that are separated into two operations, the first step: encoder (contraction) downsampling operation and the second operation: decoder (expansion) upsampling operation, which contain the following operations: Convolution, separable Convolution, activation function, residual operation, and normalization. First, the image size (256*256) is input to the convolution operation to generate the (128*128) image size in the contraction path. Then Batch normalization, activation function RELU, separable convolution to downsample the image to (64*64) size, Relu activation function, separable convolution to generate (32*32) image size. Finally, the expansion path contains Relu activation function, convolution transpose operation to the reconstructed image by predicate pixels (64*64) image size, activation function RELU, separable convulsion to generate (128*128) image size, Relu activation function, convolution transpose operation to the reconstructed image by predicate pixels (256*256) image size.

3.2.1 Depth wise separable convolution [15]

Its name came from not dealing with just spatial dimensions but with depth dimensions (the numbers of channels). It is an efficient operation to decrease the computation cost, and adjusting parameter numbers compared to traditional CNN reduces overfitting. This operation can be done using two-step:

3.2.1.1 Depth-wise convolution

In this operation, the convolution operates on one channel at a time instead of multichannel. If you have an image with channel numbers(C) filters/kernels size (Ak *Ak *1), So the resulting output will be equal (Ap *Ap* C).

3.2.1.2 pointwise convolution

This operation uses (1*1) convolution operation on the C channels, the kernel/filter size (1*1*C). The resulting output is equal (Ap*Ap*K) where K = filter numbers, Ap numbers of slides horizontally and vertically. The equation (1) defines the complexity of depth-wise conv.

$$\text{Depth wise separable complexities} = C*Ap2 * (Ak2 + N) \quad (1)$$

3.2.1.3 Residual learning rule [16]

The deep neural network is very deep, which is training more difficult than another traditional neural network because the deeper network appears to be a degradation problem when it begins to converge. In this network, the depth will increase, leading to saturated accuracy and rapid degradation. However, this degradation does not happen to overfit because adding many layers increases training errors. The solution for this problem introduced the residual learning rule as expressed in equation (2) and appears in Figure (2). The x is the input vectors, y is the output vectors, and the function F (x, {wi}) acts as learned residual mapping.

$$y = F(x,\{ wi\}) + x \quad (2)$$

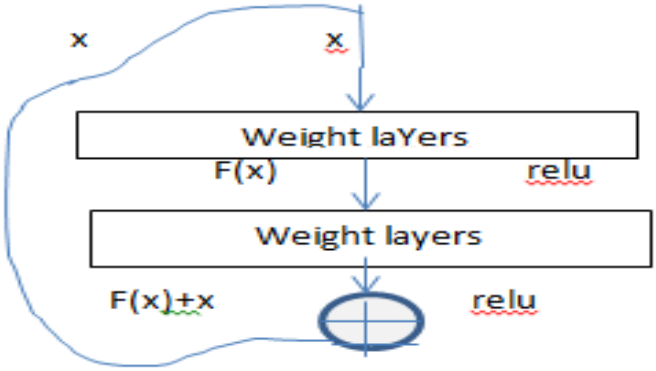

Figure 2: Block Diagram of Residual Learning Rule

3.2.2 Original U-Net

It is a fully connected network first introduced for segmented medical images. Its architecture contains contracting (downsample) used for low localization high features resolution. The second part upsampling operation expands the path used for propagating information of context to other high-resolution layers [17].

3.2.3 Xception network

Xception stands for “Extreme inception” the architecture of this convolution neural network is composed of a liner stacked of residual connections with depthwise separable convolution neural network layers [18]

4. Experimental result

In this paper, the cityscape dataset is used for training and testing the patterns to determine the best classification algorithm. Several operations will be conducted on this data, as illustrated in Figure (3). Firstly, the data will pre-process. Secondly, design a hybrid network for semantic segmentation of the image cityscape dataset. Finally, measure the selected data using the MIOU (Mean Intersection over Union).

4.1 Dataset

The cityscape dataset contains different urban street scenes collected from fifty cities at different times of the year, consisting of 5000 images, 2975 training images, and five hundred validation images. The size of every image is 256*512 pixels. The dataset contains the original image (input) at the left and the labeled image (segmentation mask with the dataset) at the right, as shown in Figure (3. a) and Figure (3. b) [19].

4.2 Pre-processing

Data augmentation's first operation includes random rotation and flipping, as shown in figure (3. c); the second operation means operation, as shown in figure (3.d).

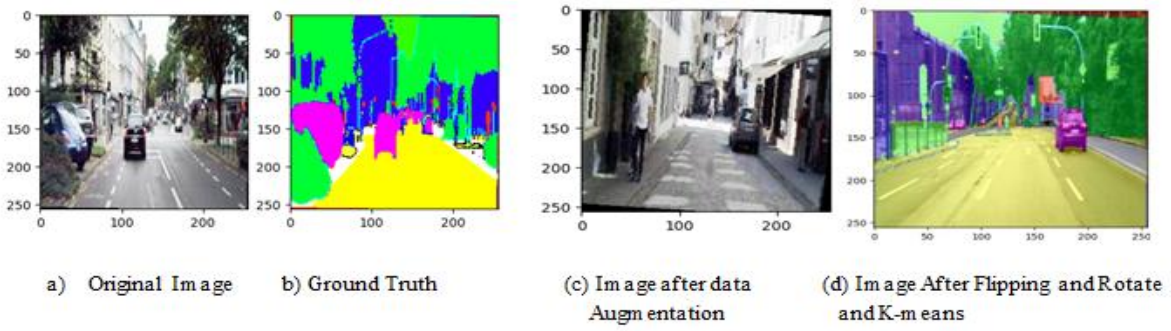

Figure 3: Cityscape Image Preprocessing

4.3 Comparison of U-net models

This work will compare with another 4 semantic segmentation networks works on the cityscape dataset as shown in table (1). The measure of comparison is MIOU (Mean Intersection Over Union). MIOU is the evaluation measure for the proposed system's predictive power. Its arithmetic mean is expressed in Equation 1. TP, FN, and FP are the false positive, false negative, and true positive numbers in the confusion matrix. The n = numbers of classes [20] [21] computed for the test dataset, the IOU can be defined in equation (3), the accuracy equation defined in Equation (5) [22] as follows:

$$IOU = (TP) / (TP + FP + FN) \quad (3)$$

$$MIOU = 1/n \sum^{n}_{I=1} IOU \quad (4)$$

$$\text{Accuracy } (\%) = (\text{correct outputs} / \text{ total numbers of inputs}) * 100\% \quad (5)$$

Table 1: comparison of proposed system results with 4 results semantic segmentation networks

methods	MIOU (%)
IKshanaNet-1 [5]	53.35%
Lightseg [7]	65.17%
ESPNet [8]	60.3%
ENet [11]	58.3%
Ours	69.0%

The proposed system achieves an accuracy of 0.7874 computed as defined in Equation (5). MIOU computed as defined in Equation (4) of all classes (no background, naïve mean):0.69, MIOU of all classes (with background, naive mean):0.064, MIOU of all non-absent classes (dropping background):0.065 and consuming time: 130878.43589234352 seconds, the output result is shown in figure (4) the left image is predicated image the output of the proposed model, the right side is the true labeled image (segmentation mask) (ground truth) within the cityscape dataset.

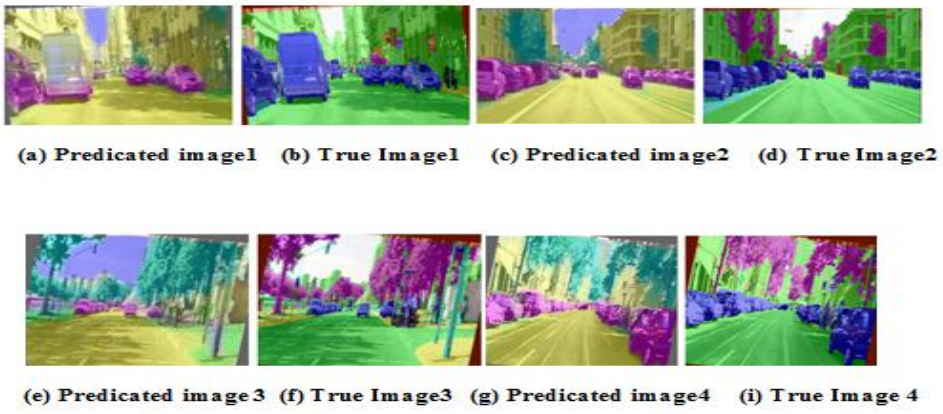

Figure 4: the output result images

5. Conclusion

This paper introduces semantic segmentation for outdoor city scenes; semantic image segmentation is an important task for Autonomous driving and Mobile robotics applications. This work uses two primary steps: pre-processing and semantic segmentation. The preprocessor operations include data augmentations for increments size of the dataset and k-means clustering to provide efficient color clustering. The semantic segmentation network is generated by a hybrid of two networks composed of U-NET and Xception. After comparing the semantic segmentation networks in the recent methods IKshanaNet-1, Lights, ESPNet, and ENet, our research achieves 69% MIOU, with an accuracy of 78% for the cityscape dataset. The difficulty with this module is fewer high-resolution 2.5D data sets that affect the accuracy of classification with deep learning.

Author contribution

All authors contributed equally to this work.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Data availability statement

The data that support the findings of this study are available on request from the corresponding author.

Conflicts of interest

The authors declare that there is no conflict of interest.

References

[1] H. K. A. H.N. Abdullah, Deep CNN Based Skin Lesion Image Denoising and Segmentation using Active Contour Method, *Eng. Technol. J.*, 37 (2019) 464–469.

[2] N. Zou, Z. Xiang, Y. Chen, S. Chen, and C. Qiao, Simultaneous semantic segmentation and depth completion with constraint of boundary, *Sensors (Switzerland)*, 20 (2020) 1–15, doi: 10.3390/s20030635.

[3] S. N. Hasan , Murat Gezer, Raghad Abdulaali Azeez, Sevinç Gülseçen , Skin Lesion Segmentation by using Deep Learning Techniques, Published in: 2019 Medical Technologies Congress (TIPTEKNO), IEEE *Xplore*: 11 November 2019, 10.1109/TIPTEKNO.2019.8895078.

[4] Y. Hu, Z. Chen, and W. Lin, RGB-D SEMANTIC SEGMENTATION: A REVIEW School of Remote Sensing and Information Engineering , Wuhan University , Wuhan , China Department of Electronic Engineering , Shanghai Jiao Tong University , Shanghai , China, *2018 IEEE Int. Conf. Multimed. Expo Work.*, pp. 1–6, 2018.

[5] V. S. S. A. Daliparthi, Ikshana: A Theory of Human Scene Understanding Mechanism, 2021, arxiv journal, [Online] Available: http://arxiv.org/abs/2101.10837.

[6] L. Wang, J. Zhang, O. Wang, Z. Lin, and H. Lu, SDC-Depth: Semantic Divide-And-Conquer Network for Monocular Depth Estimation, *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pp. 538–547, 2020, doi: 10.1109/CVPR42600.2020.00062.

[7] T. Emara, H. E. Abd, E. Munim, and H. M. Abbas, LiteSeg : A Novel Lightweight ConvNet for Semantic Segmentation, arxiv journal, arxiv journal, arXiv:1912.06683v1 [cs.CV] 13 Dec 2019.

[8] S. Mehta, M. Rastegari, and A. Caspi, "ESPNet : Efficient Spatial Pyramid of Dilated." arxiv journal, arXiv : 1803 . 06815v3 [cs . CV] 25 Jul 2018

[9] L. Chen, Z. Yang, J. Ma, and Z. Luo, Driving Scene Perception Network: Real-Time Joint Detection, Depth Estimation and Semantic Segmentation, *Proc. - 2018 IEEE Winter Conf. Appl. Comput. Vision, WACV 2018*, 2018 (2018) 1283–1291, doi: 10.1109/WACV.2018.00145.

[10] A. Valada, J. Vertens, A. Dhall, and W. Burgard, AdapNet: Adaptive semantic segmentation in adverse environmental conditions, *Proc. - IEEE Int. Conf. Robot. Autom.*, pp. 4644–4651, 2017, doi: 10.1109/ICRA.2017.7989540.

[11] A. Paszke, A. Chaurasia, S. Kim, and E. Culurciello, ENet : A Deep Neural Network Architecture for Real-Time Semantic Segmentation", arxiv journal, arXiv : 1606 . 02147v1 [cs . CV] 7 Jun 2016, pp. 1–10.

[12] J. Brownlee, How to Configure Image Data Augmentation in Keras, *Machine Learning Mastery*. 2019, [Online]. Available: https://machinelearningmastery.com/how-to-configure-image-data-augmentation-when-training-deep-learning-neural-networks.

[13] C. Shorten and T. M. Khoshgoftaar, A survey on Image Data Augmentation for Deep Learning, *Journal of Big Data*, 6 (2019), doi: 10.1186/s40537-019-0197-0.

[14] M. R. Khan, A. B. M. M. Rahman, G. M. A. Rahaman, and A. Hasnat, Unsupervised RGB-D Image Segmentation by Multi-layer Clustering, IEEE *Xplore*: 2016 5th International Conference on Informatics, Electronics and Vision (ICIEV) ,pp. 719–724, 2016.

[15] L. C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, Encoder-decoder with atrous separable convolution for semantic image segmentation, *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, 11211 (2018) 833–851, doi: 10.1007/978-3-030-01234-2_49.

[16] K. He, X. Zhang, S. Ren, and J. Sun, Deep residual learning for image recognition, *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2016 (2016) 770–778, doi: 10.1109/CVPR.2016.90.

[17] O. Ronneberger, P. Fischer, and T. Brox, U-net: Convolutional networks for biomedical image segmentation, *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, 9351 (2015) 234–241, doi: 10.1007/978-3-319-24574-4_28.

[18] F. Chollet, Xception: Deep learning with depthwise separable convolutions, *Proc. - 30th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR 2017*, 2017 (2017) 1800–1807, doi: 10.1109/CVPR.2017.195.

[19] M. Cordts *et al.*, The Cityscapes Dataset for Semantic Urban Scene Understanding, *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2016 (2016) 3213–3223, doi: 10.1109/CVPR.2016.350.

[20] S. Piao and J. Liu, Accuracy Improvement of UNet Based on Dilated Convolution, IOPScience: *J. Phys. Conf. Ser.*, 1345 (2019), doi: 10.1088/1742-6596/1345/5/052066.

[21] A. Y. Noori, S. H. Shaker, and R. A. Azeez, 3D scenes semantic segmentation using deep learning based Survey, *IOP Conf. Ser. Mater. Sci. Eng.*, 928 (2020), doi: 10.1088/1757-899X/928/3/032083.

[22] A. A. Abdulhussein and F.A. Raheem, Hand gesture recognition of static letters American sign language(ASL) using deep learning, Eng. Technol. J., 38 (2020) 926-937, DOI:https://doi.org/10.30684/etj.v38i6A.533.