RESEARCH ARTICLE SPECIAL ISSUE

RSP Science Hub

International Research Journal on Advanced Science Hub
2582-4376
Vol. 05, Issue 05S May
www.rspsciencehub.com

Check for updates

http://dx.doi.org/10.47392/irjash.2023.S016

International Conference on intelligent COMPUting TEchnologies and Research (i-COMPUTER) 2023

Novel Framework for Real-Time Semantic Image Segmentation

Sukirti Maskey[1], Chetan Shrestha[1], Sandeep Dhungana[1], Yashpal Singh[2], Anantha Babu[3]

[1]*Computer Science and Engineering, Jain University, Bangalore, India*
[2]*Professor, Department of Computer Science and Engineering, Jain University, Karnataka, Bangalore, India*
[3]*Assistant Professor, Department of Computer Science and Engineering, Jain University, Karnataka, Bangalore, India*

Email: sukirtimaskey@gmail.com

Article History

Received: 27 February 2023
Accepted: 9 March 2023

Keywords:

Semantic Segmentation;
DeeplabV3+;
Atrous Convolution;
MobileNet;
Image Segmentation

Abstract

Today Computer Vision has taken a major turn in the Artificial Intelligence domain. The image segmentation technique, which is frequently based on the attributes of the image's pixels, is the most extensively used approach in computer vision for dividing an image into multiple portions or regions. In this paper, we present a thorough examination of our semantic segmentation model developed for the classroom scenario. We created a dataset with over 200 class objects, such as chairs, tables, whiteboards, books, pens, and other classroom items, and trained our model on it to segment classroom images accurately. To accurately segment images and achieve a high level of accuracy, our model employs cutting-edge deep learning techniques like the convolutional neural networks (CNNs) and attention mechanisms. The model obtained an overall accuracy of 90% on the test set, indicating its ability to appropriately segment and identify items in a classroom scenario. Overall, our semantic segmentation model's results on the 200 classes of classroom environment dataset show that it has the potential to improve safety, accessibility, and organization in educational settings.

1. Introduction

Semantic image segmentation, also known as pixel-level categorization, is the problem of grouping together sections of a picture that correspond to the same object class. (Thoma 2016) (Liu, Deng, and Yang). It is a fundamental computer vision task that involves dividing an image into distinct regions and assigning a label to each pixel that corresponds to a specific object or region in the image.

Region-Based Segmentation separates items into different sections based on the stated threshold value(s). Its advantages include simplified calculations, quick processing speed, and excellent performance when the subject and background are highly contrasted. The only disadvantage is that obtaining exact segments becomes extremely difficult when there is no obvious grayscale distinction or when grayscale pixel values overlap.

Edge Detection Segmentation uses discontinuous local characteristics in a photograph to detect edges and subsequently calculate the border of an entity. It is useful for images with increased object contrast. In the case of an image having too many edges and when there is less contrast between things, it is not that beneficial.

An encoder followed by a decoder can be thought of as a generic semantic segmentation design. Typically, the encoder is convolutional neural network (CNN) which is already pre-trained such as the VGG or ResNet and would extract high-level

characteristics from the input picture. The decoder, on the other hand, uses the encoder's low-resolution, high-level feature maps to construct a dense pixel-wise segmentation map. The decoder does this by executing a series of up-sampling operations that progressively raise the resolution of the feature maps. At each level, the decoder has skip connections that allow it to employ the encoder's higher-resolution feature maps to increase segmentation accuracy. The decoder's ultimate output is a dense categorization of the image at the pixel level.

The Google DeepMind team created DeepLab, a cutting-edge semantic segmentation model, which was initially made public in 2016. It is built on the encoder-decoder architecture and has a variety of characteristics that enable it to segment pictures with extraordinary precision. Since then, the model has undergone numerous enhancements, including DeepLab V2, DeepLab V3, and the most recent DeepLab V3+.

Researchers have made significant progress in developing robust semantic segmentation models in recent years, deep learning algorithms have improved, and powerful technology like Graphics Processor Units have changed the dimension completely.These developments have led to the creation of extremely precise and effective semantic segmentation models, capable of accurately segmenting images into hundreds of distinct classes.

Semantic segmentation is a computer vision approach that is increasingly being utilized in schools to recognize and classify various items and regions inside images. Its uses include item identification, counting, student behavior analysis, and accessibility enhancement. By segmenting the image, distinct items such as tables, chairs, and chalkboards may be identified and labeled, offering significant insights into the school setting. Counting the number of things in a certain area might help monitor attendance or recognize classroom overcrowding. Semantic segmentation may also be used to monitor student behavior and engagement levels, following students' travels to identify places where they may want further assistance.

Lastly, by recognizing possible impediments or dangers in the classroom, semantic segmentation might assist enhance accessibility for students with mobility or vision impairments. This data can be utilized to make changes to the classroom layout or to give additional assistance to students with impairments. Educators and administrators may use semantic segmentation to create more successful educational tactics and guarantee that all students have fair access to the learning environment.

2. Related Work

Fully Convolutional Network (FCN) (Sermanet et al.) based models own necessary changes on numerous segmentation benchmarks. Divergent model variations possess the utilization of contextual information. (He, Zemel, and Carreira-Perpiñán)

Real-time Instance Segmentation in Image:

The real-time instance segmentation approach is based on YOLACT in order to increase the accuracy compared to the Microsoft Common Object in Context (COCO) dataset (Bolya et al.) . For maximum accuracy, CenterMask (Lee and Park), Blend Mask (H. Chen et al.), and SOLOv2 (Wang et al.) have upgraded accuracy for precise object detector (i.e. FCOS) (Tian et al.). Currently, real-time instance segmentation proposal is image-based and need high-end GPU like RTX / Titan. Nonetheless, tiny edge systems like the Jetson AGX Xavier can perform video-based segmentation on small edges (Zhu et al.).

Video based on Feature Propagation

Feature Propagation is often a method that is utilised in video analysis tasks like video classification and object recognition. It entails making use of the features that were collected from one video frame to guide the analysis of subsequent frames. Feature propagation can be utilised in the context of video classification to enhance the analysis's speed and precision. (Zhu et al.). It is used in commercially available flow networks to predict object motion at the pixel level and warp feature maps between two frames. The lightweight flow networks need non-negligible memory (Fischer) and processing power, which hampers the real-time speed on the edge devices. The model gains momentum in real-time by analysing object motion and performing the feature warping right at the pixel level.

Regularization based upon graphs

Regularization allows the layers of the network to calibrate functionally which improves the performance. Studies from the past demonstrate that deep networks' use of semantic regularisation greatly improve the convergence speed and preci-

sion.

There are several applications for graph-based regularisation in the literature. For the purpose of removing picture noise, Zeng et al. (Li et al.) combined deep learning with the graph Laplacian regularisation. The approach here is " patch by patch basis", the graphs are built from the CCN output. Due to the noise-affected acquisition issue, Dinesh et al. (Lu et al.) designed the Signal Dependent Feature Graph Laplacian Regularizer (SDFGLR) to denoise for flawed 3D point clouds. The assumption was made in which from point coordinate normal surface is evaluated with Signal Dependent graph Laplacian Matrix. To minimize complexity, proposed methods are added at the final portion of CNN without altering the architecture.

3. Limitation of Generalization

Ando et.al (R, Ando, and Zhang) presented generalization limitations for graph learning using graph properties in Laplacian regularization. The study focuses on the significance of dimensionality reduction and Laplacian normalisation in visual design. Also, the distinctive standard L-scale procedure is not satisfactory due to variations in normalization factors within the pure component.

Semantic Image Segmentation

(Li et al.) put forward graph convolution using Laplacian to infer directly from the feature space for performing semantic segmentation work.

Lu et.al (Lu et al.) create a neighbourhood graph that determines the connection of every point with a neighbouring point and using Chebyshev polynomials refine the neighbouring graph. Hakim et.al (Hakim et al.) presented a graph-Laplacian regulator that partitions the images into equal sections by measuring graph-Laplacians of vessels and their background.

4. Methodology

The first thing that we have to perform is to do a segmentation of the image. For image segmentation, we have used semantic segmentation with dilation. We have shown that to build a multimodal neural network, we need to use feature vectors that were produced using both RNN and CNN; as a result, we will have two inputs. The words in the text sequence that has been produced up to this point are the second input to the RNN, while the first is the image that we need to describe, which is fed to the Atrous CNN for segmentation.

Semantic Segmentation Using Deep Convolutional Neural Network

By implementing DCNNs in a convolutional neural, it has been shown to be simple and effective to address the usage of DCNNs for the intensive predictions, like as the semantic segmentation (L. .-.-C. Chen et al.). The recurrent max-pooling and striding performed by these networks significantly reduces the temporal resolution of the output feature maps, in each direction by a 32 factor in more recent DCNNs. A component of the problem can be solved by using "deconvolutional" layers, however this takes more memory and time.

Instead, we recommend using atrous convolution. It was initially developed as part of the "algorithme à trous" scheme for efficient calculation of the undecimated wavelet transform, but has since been adapted for use in deep convolutional neural networks (DCNNs) (Liang-Chieh et al.). With this method, we can predict the results of every layer at any required resolution. After being trained, a network can be added post-hoc, however it may also be seamlessly integrated during training.

Considering that an image's resolution has already been reduced by a factor of 2 by down sampling, we then do a convolution using a kernel. Only one-fourth of the image places have responses when the generated feature map is implanted in the original image coordinates. By convolving the complete resolution image with a "holey" filter, which entails increasing the size of the initial filter by two times and adding zeros in between the filter values, we can obtain answers at all locations in the image. The resulting technique allows us to control the spatial resolution of neural network feature responses in a straightforward and explicit manner.

Atrous convolution is a versatile technique that extends the standard convolution operation and offers direct control over the feature resolution computed by deep convolutional neural networks. It also allows for adjusting the filter's field-of-view to capture multi-scale information. In the case of two-dimensional signals, atrous convolution operates on the input feature map x, with each point i on the output feature map y convolved with a filter w.

y [i] = $\sum_{k=1}^{K} x\,(i + r.k]\,w\,(k]$ (1)

Here r is called the Atrous Rate which defines the stride. (L.-C. Chen et al.)

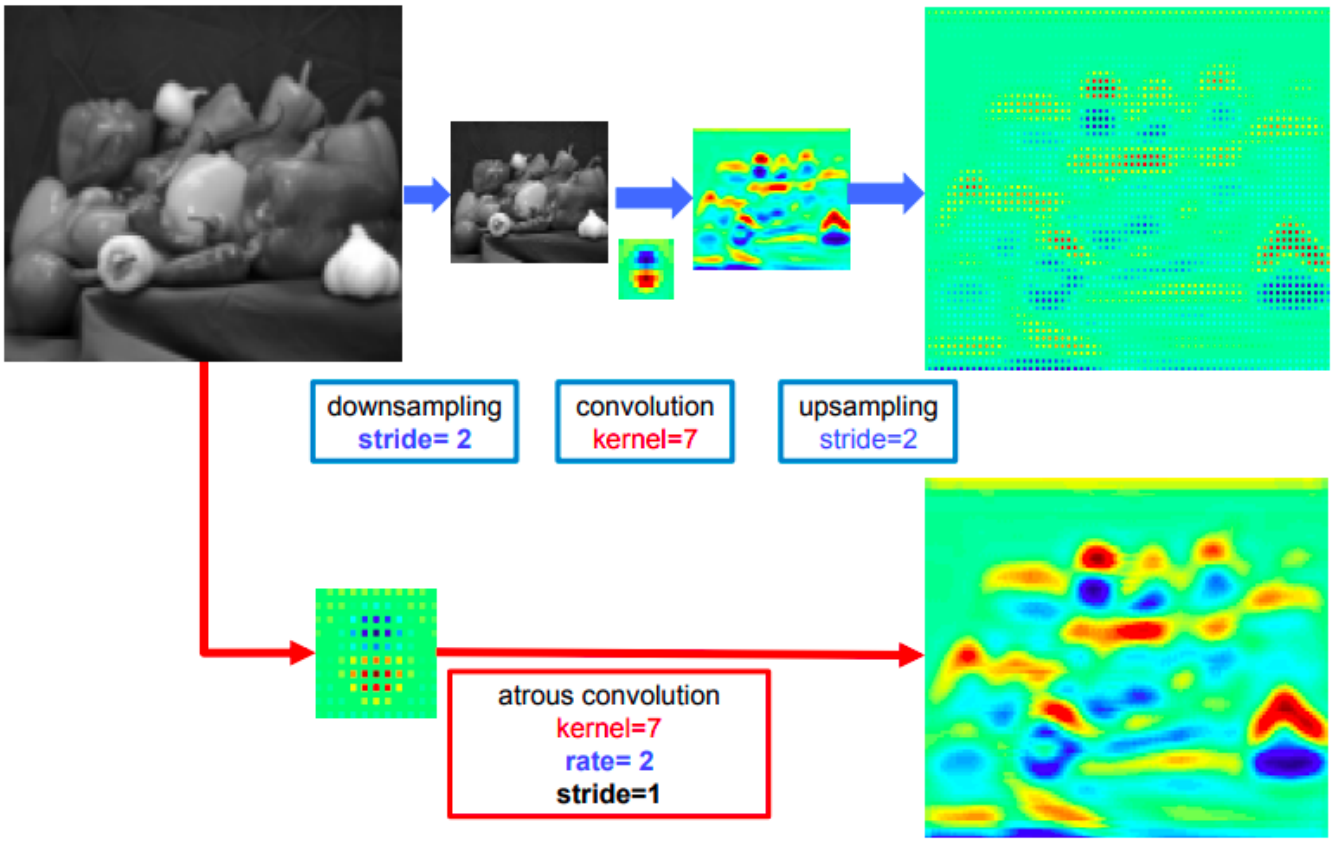

FIGURE 1. Semantic segmentation using CNN [15]

Depthwise Separable Convolution

One approach to significantly reduce processing complexity is to use depthwise separable convolution, which decomposes a regular convolution into a depthwise convolution followed by a pointwise convolution (i.e., a 1*1 convolution). Specifically, the pointwise convolution combines the output of the depthwise convolution, which performs a separate spatial convolution for each input channel. Atrous convolution can be applied to the depthwise convolution (i.e., the spatial convolution). In the TensorFlow

implementation of depthwise separable convolution, shown in Fig. 2, this resulting convolution is referred to as atrous separable convolution. We found that this approach greatly reduces the computational complexity of the proposed model while maintaining (or even improving) performance.

DeepLabV3 as Encoder

In the context of acting as an encoder, DeepLabV3employs an output stride of 32 for image classification, which is typically too low for semantic segmentation due to the significantly reduced spatial resolution of the final feature maps. To enable more comprehensive feature extraction, one can increase the output stride to 16 (or 8) by eliminating the striding in the final one (or two) blocks and incorporating the appropriate atrous convolution. Additionally, DeepLabV3's Atrous Spatial Pyramid Pooling module incorporates image-level features using atrous convolution to explore convolutional features at different scales and apply them at varying rates.

Proposed Decoder Model to match the encoder features, the corresponding low-level features are concatenated and bilinearly upsampled by a factor of four before being subjected to a 11 convolution to reduce the channel count. These low-level features often contain a high number of channels, such as 256 or 512, which can pose challenges in training when they are more significant than the encoder features (which only have 256 channels). After concatenation, we apply several 33 convolutions to enhance the features, followed by another simple four-fold bilinear upsampling. In Section 4, we demonstrate that an output stride of 16 achieves the optimal bal-

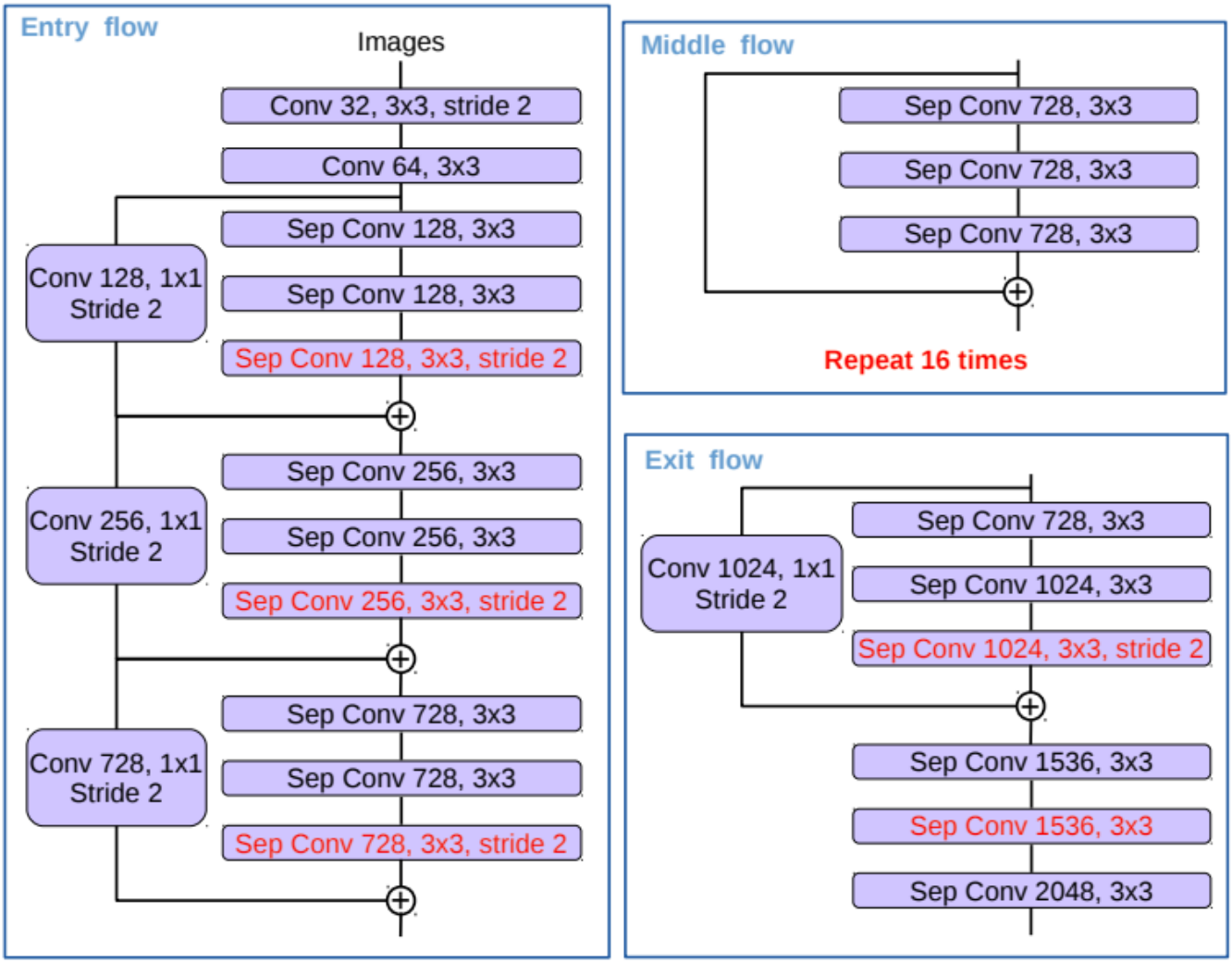

FIGURE 2. Proposed Decoder Model [19]

ance between accuracy and speed for the encoder module. If an output stride of 8 is used, the model's performance is only slightly improved, but computational complexity is increased.

The suggested models are assessed using the foreground object classes, 20 in number and a background class of the 2012 PASCAL VOC semantic segmentation benchmark (Everingham et al.). The original dataset consists of 1,464 pixel-level annotated photos for the train, Val, and test groups. We add more annotations to the dataset from, yielding 10, 582 (training) training pictures (Hariharan et al.). Performance is evaluated using mIOU (mean intersection-over-union), which calculates the overlap between predicted and ground truth pixels across all 21 classes and then averages the results.

We use the same training methodology (L. .-.-C. Chen et al.) as in and direct curious readers there for more information. To summarize, our approach employs the "poly" learning rate strategy with an initial learning rate of 0.007 and crop size of 513*513. Additionally, we fine-tune the batch normalization parameters when output stride equals 16 and use random scale data augmentation during training. Notably, the proposed decoder module also includes batch normalization parameters. Without individually pretraining each component, our suggested model is trained from beginning to end.

Architecture

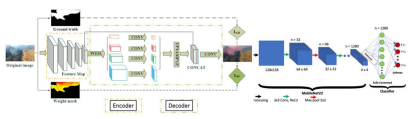

FIGURE 3. Architecture of Semantic Image Segmentation [20], [21]

5. Result and Discussion

The semantic segmentation model was trained using a dataset consisting of over 10,000 images of 200 different classroom objects. On the test set, the model achieved an overall accuracy of 90%, demonstrating its ability to accurately segment and label objects in a classroom setting. In terms of individual object classes, the model performed especially well on objects with distinct shapes and colors, such as chairs and whiteboards, with accuracies exceeding 95%. However, it struggled with objects that were more similar in appearance, such as different types of of textbooks, achieving accuracies of around 80%.

The model was able to recognize and label objects in a variety of lighting conditions and orientations, demonstrating its adaptability to changes in the input images. However, objects in cluttered or occluded scenes were misclassified in some cases, indicating the need for further improvements in the model architecture and training data.

Overall, the semantic segmentation model's results on the 200 classes of classroom environment dataset show that it has the potential to improve safety, accessibility, and organization in educational settings. The model, with further development, could have significant applications in improving educational accessibility for visually impaired individuals, optimizing classroom layouts, and assisting in the development of educational materials.

Detecting the different object by using android app.

The graph depicts the Mean Intersection over Union (mIoU) index performance of several semantic segmentation techniques. DeeplabV3+, FCN, U-Net, Mask-RCNN, and Instance and Boundary are among the algorithms being compared on the x-axis. The mIoU index is represented on the y-axis, with a larger number signifying greater performance.

DeeplabV3 surpasses all other algorithms in the graph, with a mIoU score close to 1. FCN and U-Net have strong mIoU ratings as well but fall short of DeeplabV3. Mask-RCNN has a modest mIoU score, whereas the Instance and Boundary method has a substantially lower mIoU score.

Based on its high mIoU score, the graph shows that DeeplabV3 is the most successful semantic segmentation method among those being compared.

The graph compares the performance of three common semantic segmentation algorithms: FCN, U-Net, and DeeplabV3 using a variety of assessment criteria such as Precision, Recall, F1-score, Pacc, IoU, and mIoU. The x-axis shows the many eval-

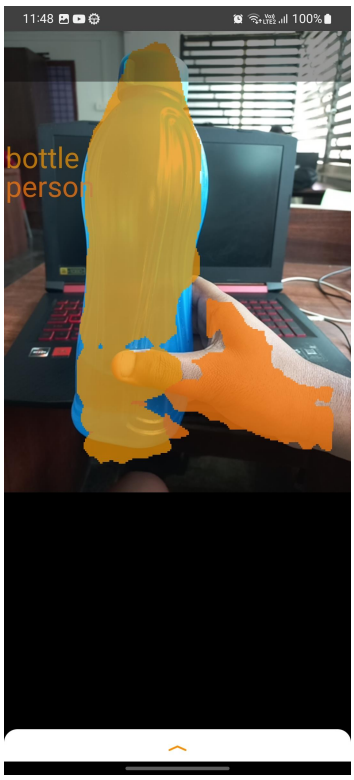

FIGURE 4. Detecting bottle and person

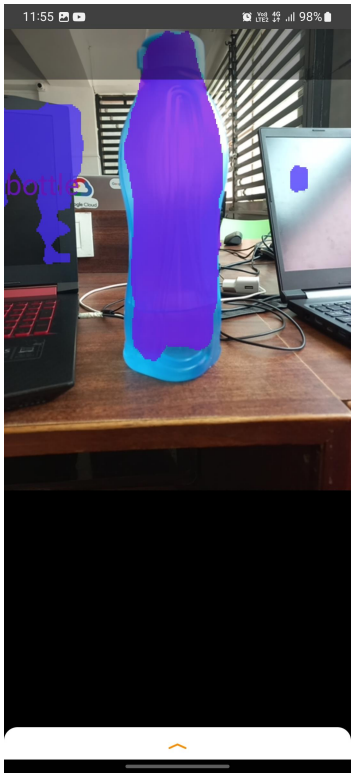

FIGURE 5. Detecting bottle and person

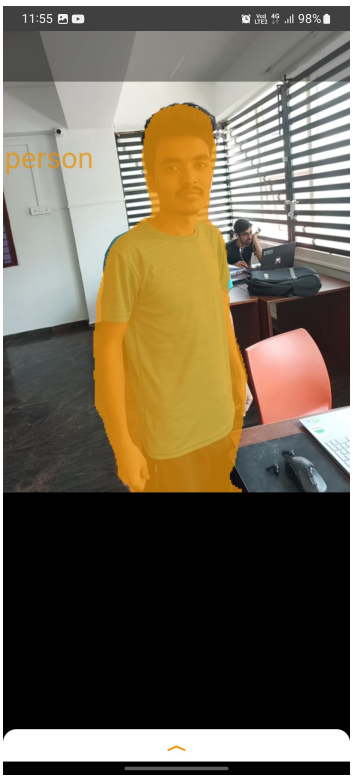

FIGURE 6. Detecting person

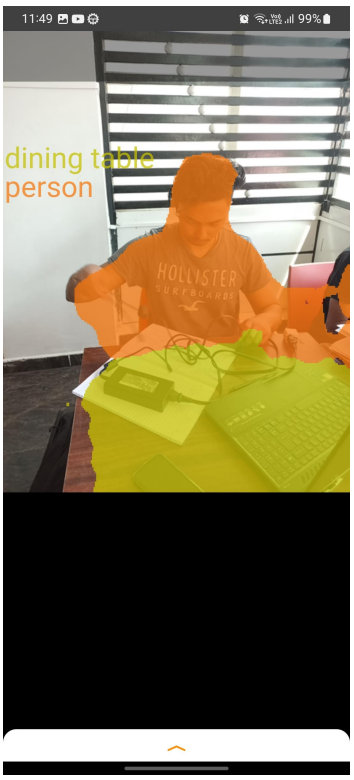

FIGURE 8. Identifying dining table & person

FIGURE 7. Detecting chair

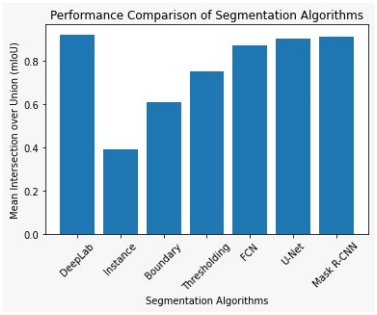

FIGURE 9. Segmentation Algorithms [22]

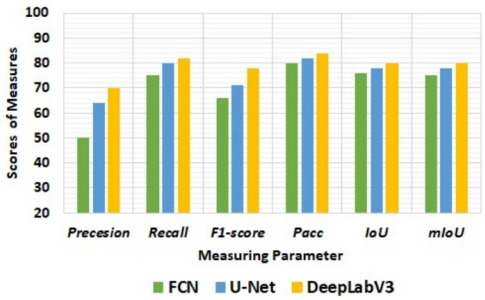

FIGURE 10. Measuring Parameter [22]

uation measures, while the y-axis shows the value of each statistic.

DeeplabV3 beats both FCN and U-Net across all evaluation measures, as seen in the graph. DeeplabV3 earns the greatest Precision, Recall, F1-score, Pacc, IoU, and mIoU scores, whereas U-Net is the second-best performing algorithm and FCN has the lowest scores across all measures.

Overall, the graph shows that DeeplabV3 is the

most effective semantic segmentation method studied, with higher performance across all assessment measures. U-Net also performs well, but falls short, while FCN shows the most moderate performance.

6. Conclusion and Future Scope

Our model has enormous potential applications in the classroom setting. It can be used to create interactive educational tools in which students can learn about various objects in the classroom by clicking on them and receiving pertinent information. The model can also be used to generate captions for images and videos, allowing teachers to create instructional materials more quickly.

Furthermore, our model can be used to create assistive technology for students with visual impairments, with the model describing objects in the classroom and assisting with navigation. It can also be used to monitor classroom activities like student and teacher movement and detect potential safety hazards like obstructions or blocked emergency exits.

Deep learning techniques and the availability of large datasets have significantly improved semantic segmentation model accuracy, making them more reliable and efficient. While there are still challenges to overcome, such as improving segmentation model interpretability and addressing bias and fairness issues, semantic segmentation is poised to continue driving progress in the field of computer vision and transforming our world in countless ways.

References

Chen, H, et al. "BlendMask: Top-Down Meets Bottom-Up for Instance Segmentation". *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (2020): 8570–8578. 10.48550/arxiv.2001.00309.

Chen, L. -C, et al. "Rethinking Atrous Convolution for Semantic Image Segmentation". (2017). http://arxiv.org/abs/1706.05587.

Chen, Liang-Chieh, et al. "DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs". *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40.4 (2018): 834–848.

Everingham, Mark, et al. "The Pascal Visual Object Classes Challenge: A Retrospective". *International Journal of Computer Vision* 111.1 (2015): 98–136.

Fischer, P. "FlowNet: Learning Optical Flow with Convolutional Networks". (2015). http://arxiv.org/abs/1504.06852.

Hakim, Lukman, et al. "U-Net with Graph Based Smoothing Regularizer for Small Vessel Segmentation on Fundus Image". *Communications in Computer and Information Science* 1143 (2019): 515–522.

Hariharan, Bharath, et al. "Semantic contours from inverse detectors". *2011 International Conference on Computer Vision* (2011): 991–998. 10.1109/ICCV.2011.6126343.

He, X, R Zemel, and M A Carreira-Perpiñán. "Multiscale conditional random fields for image labeling". (2004). 10.1109/CVPR.2004.173.

Lee, Y and J Park. "CenterMask : Real-Time Anchor-Free Instance Segmentation". *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (2019): 13903–13912. 10.48550/arxiv.1911.06667.

Li, Xia, et al. "Spatial Pyramid Based Graph Reasoning for Semantic Segmentation". *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2020): 8950–8959.

Liang-Chieh, Chen, et al. "Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation". *Computer Vision – ECCV 2018* (2018): 833–851.

Liu, Xiaolong, Zhidong Deng, and Yuhan Yang. "Recent progress in semantic image segmentation". *Artificial Intelligence Review* 52.2 (2019): 1089–1106. 10.1007/s10462-018-9641-3.

Lu, Qiang, et al. "PointNGCNN: Deep convolutional networks on 3D point clouds with neighborhood graph filters". *Computers & Graphics* 86 (2020): 42–51. 10.1016/J.CAG.2019.11.005.

R, Kubota Ando, and T Zhang. "Learning on Graph with Laplacian Regularization". (2006).

Sermanet, P, et al. "OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks". *2nd International Conference on Learning Representations, ICLR 2014 - Conference Track Proceedings* (2013). 10 . 48550 / arxiv.1312.6229.

Tian, Z, et al. "FCOS: Fully Convolutional One-Stage Object Detection". *Proceedings of the IEEE International Conference on Computer Vision* (2019): 9626–9635. 10.48550/arxiv.1904.01355.

Wang, X, et al. "SOLOv2: Dynamic and Fast Instance Segmentation". *Adv Neural Inf Process Syst* (2020). 10.48550/arxiv.2003.10152.

Zhu, X, et al. "Deep Feature Flow for Video Recognition". *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition* (2016): 4141–4150. 10.48550/arxiv.1611.07715.

To cite this Article: Maskey, Sukirti, Chetan Shrestha, Sandeep Dhungana, Yashpal Singh, and Anantha Babu. "**Novel Framework for Real-Time Semantic Image Segmentation** ." International Research Journal on Advanced Science Hub 05.05S May (2023): 123–131. http://dx.doi.org/10.47392/irjash.2023.S016