1. ***Abstract:***

Semantic segmentation is a fundamental task in computer vision, playing a crucial role in scene understanding and object recognition. Semantic segmentation's main goal is to give each pixel in an image a unique name, dividing the image into sections that have semantic significance. Through this approach, a greater understanding of the visual world is made possible, allowing machines to study and interpret images at a highly abstract level. This work offers an overview of deep learning techniques for semantic segmentation with applications in several fields. First, we go through the necessary underlying ideas and the terminology used in this discipline. The primary datasets and challenges are then made public to assist researchers in selecting the ones that best meet their objectives and needs. After that, a review of current approaches is conducted, emphasizing their contributions and field importance. Ultimately, a commentary on the findings is included after the quantitative results for the methodologies and datasets under description are provided. Lastly, we highlight many exciting upcoming projects and offer our assessments on the state of semantic segmentation with deep learning methodologies.

**Keywords**: Semantic Segmentation, Convolutional neural network, Weakly supervised method, Deep Learning, computer vision.

1. ***Introduction:***

Semantic segmentation is currently one of the main issues in computer vision, whether it be used to static 2D images, video, or even 3D or volumetric data. Semantic segmentation is one of the high-level tasks that leads to comprehensive scene knowledge, when seen in the broadest context [1]. Accurate scene interpretation is desperately needed, especially with the growing number of intelligent applications (such as mobile robots). Semantic segmentation has so attracted a great deal of attention in recent years as a necessary step towards this goal [2]. Applying deep learning-based Convolutional Neural Networks (CNN) approaches has led to notable progress in the field of semantic segmentation [3]. The fact that a growing number of applications rely on deriving knowledge from imagery highlights the significance of scene understanding as a fundamental computer vision problem. Among those uses are, to mention a few, human-machine interaction [5], autonomous driving [6] [7] [8], computational photography [9], picture search engines [10], and augmented reality.

Two typical concerns are: how to create effective feature representations to distinguish objects of different classes and how to use contextual information to guarantee pixel label consistency in order to achieve high-quality semantic segmentation [2]. Using hand-engineered features, like Scale Invariant Feature Transform (SIFT) [10] and Histograms of Oriented Gradient (HOG) [11], is advantageous for the majority of early approaches [12,13] when answering the first question. Utilizing learned features in computer vision tasks, including picture classification [14, 15], has been very successful in the last few years thanks to the emergence of deep learning [16, 17]. Consequently, the learnt features have received a lot of attention lately from the semantic segmentation field [18–21], where they are typically associated with Convolutional Neural Network (CNN or Convent) [22]. Using contextual models like Conditional Random Field (CRF) [23–25] and Markov Random Field (MRF) [26] is the most popular approach for the second problem, regardless of the feature employed.

This paper's primary goal is to present a thorough overview of semantic segmentation techniques, with an emphasis on examining the issues that are frequently raised and the associated solutions used. These days, semantic segmentation is a huge field with close ties to other computer vision tasks. The entire field cannot be covered by this review. There are already various evaluations on the state of the art in picture segmentation research, as well as semantic segmentation datasets and techniques [1,2].

The key contributions of our work are as follows:

* An extensive and well-structured analysis of the most important deep learning techniques for semantic segmentation, together with an overview of their history and contributions.
* Draw attention to the issues that need to be resolved by upcoming researchers.
* We offer an overview of available datasets that could be helpful for deep learning-based semantic segmentation projects.
* a comprehensive analysis of performance that collects numerical measurements for things like memory, execution time, and precision.

***This section will be changed:***

*The rest of this paper is structured as follows. First, the problem of semantic segmentation is introduced in Section 2, along with notation and standards that are frequently seen in the literature. Additional background ideas are also covered, such typical deep neural networks. Section 3 then goes on to discuss benchmarks, problems, and datasets that are currently available. Section 4 examines current approaches in a bottom-up complexity hierarchy according to their respective contributions. Rather of carrying out a quantitative evaluation, this section concentrates on outlining the theory and salient features of various approaches. In conclusion, Section 5 provides a succinct analysis of the techniques that have been provided, drawing on their quantitative outcomes using the previously mentioned datasets. Furthermore, directions for future research are also outlined.*

1. ***Background and Preliminaries:***
2. **Semantic Segmentation:**

Semantic segmentation is an essential computer vision approach that improves the efficiency with which machines evaluate and comprehend visual data. Comparing semantic segmentation to traditional image recognition techniques—which typically give an image a single label—reveals a significant improvement. Going one step further, semantic segmentation assigns a class or category to every pixel in an image according to what it symbolizes. Semantic segmentation achieves this by determining the semantic meaning of each pixel, resulting in a rich and comprehensive segmentation map that provides a finer and more accurate knowledge of the image. The foundation of many computer vision applications, including autonomous vehicles, medical imaging, and scene interpretation, is semantic segmentation.

1. **Labels or Classes:**

The terms "labels" or "classes" in the context of semantic segmentation refer to the predetermined categories or semantic identities that are given to every pixel in an image at the time of segmentation. Within the visual input, these labels denote the various regions, objects, or structures that the model has been trained to identify and distinguish. A unique label designating the semantic meaning or category to which each pixel in the segmented image belongs is assigned to it.

In a street scene, for instance, common labels or classes could be "car," "pedestrian," "road," "building," and so on. In order to provide a thorough and in-depth comprehension of the scene, semantic segmentation aims to precisely identify and outline each pixel in the image in accordance with these predetermined classes.

1. **Ground Truth:**

"Ground truth" in semantic segmentation refers to the manually annotated and labeled data that is the final source of reference for accurately segmenting pixels in an image. A semantic segmentation model's performance can be measured during both the training and testing stages using ground truth as a reference.

Human annotators carefully assign the appropriate semantic category or class to every pixel in an image in order to create ground truth. Annotators may designate pixels in a street scene, for instance, as belonging to the categories "car," "pedestrian," "road," or "building." The resultant annotated image serves as the ground truth for that particular image and is frequently referred to as a segmentation map.

In training, a dataset containing input images and the related ground truth annotations is used to teach a semantic segmentation model. The differences between the model's predictions and the labels that correspond to the ground truth are used to modify the model's internal parameters. The model is able to capture the complex features and patterns required for precise pixel-wise segmentation because of this iterative learning process.

During the testing or evaluation step, fresh, unobserved images are fed into the trained model, and its predictions are measured against the ground truth to gauge how well it performs. Common evaluation criteria that measure how well the model matches the real world include pixel accuracy and intersection over union (IoU).

1. ***Challenges:***

Some common challenges in semantic segmentation are:

* Dealing with varying scales and shapes of objects.
* Handling occlusions and overlapping objects.
* Distinguishing between similar classes or fine-grained categories.
* Achieving high accuracy and efficiency in segmentation models.

1. ***References:***
2. A. Garcia-Garcia, S. Orts-Escolano, S. Oprea, V. Villena-Martinez, and J. G. Rodr ́ıguez, “A review on deep learning techniques applied to semantic segmentation,” CoRR, vol. abs/1704.06857, 2017.
3. H. Yu, Z. Yang, L. Tan, Y. Wang, W. Sun, M. Sun, and Y. Tang, “Methods and datasets on semantic segmentation: A review,” Neurocomputing, vol. 304, pp. 82–103, 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231218304077
4. B. Emek Soylu, M. S. Guzel, G. E. Bostanci, F. Ekinci, T. Asuroglu, and K. Acici, “Deep-learning-based approaches for semantic segmentation of natural scene images: A review,” Electronics, vol. 12, no. 12, 2023. [Online]. Available: <https://www.mdpi.com/2079-9292/12/12/2730>
5. Y. Guo, Y. Liu, T. Georgiou, and M. S. Lew, “A review of semantic segmentation using deep neural networks,” International Journal of Multimedia Information Retrieval, vol. 7, no. 2, pp. 87–93, Jun 2018
6. M. Oberweger, P. Wohlhart, and V. Lepetit, “Hands deep in deep learning for hand pose estimation,” arXiv preprint arXiv:1502.06807, 2015.
7. M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, “The cityscapes dataset for semantic urban scene understanding,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 3213–3223.
8. A. Geiger, P. Lenz, and R. Urtasun, “Are we ready for autonomous driving? the kitti vision benchmark suite,” in 2012 IEEE Conference on Computer Vision and Pattern Recognition, June 2012, pp. 3354–3361
9. A. Ess, T. Muller, H. Grabner, and L. J. Van Gool, “Segmentation- ¨ based urban traffic scene understanding.” in BMVC, vol. 1, 2009, p. 2.
10. Y. Yoon, H.-G. Jeon, D. Yoo, J.-Y. Lee, and I. So Kweon, “Learning a deep convolutional network for light-field image superresolution,” in Proceedings of the IEEE International Conference on Computer Vision Workshops, 2015, pp. 24–32.
11. D.G. Lowe, Object recognition from local scale-invariant features, in: Proceedings of the IEEE International Conference on Computer Vision, 1999, pp. 1150–1157
12. N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2005, pp. 886–893.
13. N. Silberman, R. Fergus, Indoor scene segmentation using a structured light sensor, in: Proceedings of the IEEE International Conference on Computer Vision Workshops, 2011, pp. 601–608
14. S. Gupta, P. Arbelaez, J. Malik, Perceptual organization and recognition of indoor scenes from RGB-D images, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 564–571
15. A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, in: Proceedings of the Advances in neural information processing systems, 2012, pp. 1097–1105
16. K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv:1409.1556(2014).
17. G.E. Hinton, S. Osindero, Y.-W. Teh, A fast learning algorithm for deep belief nets, Neural Comput. 18 (7) (2006) 1527–1554.
18. Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (7553) (2015) 436–444
19. C. Farabet, C. Couprie, L. Najman, Y. LeCun, Learning hierarchical features for scene labeling, IEEE Trans. Pattern Anal. Mach. Intell. 35 (8) (2013) 1915–1929.
20. L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A.L. Yuille, Semantic image segmentation with deep convolutional nets and fully connected CRFs, arXiv:1412.7062(2014).
21. R. Girshick, J. Donahue, T. Darrell, J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 580–587.
22. D. Eigen, R. Fergus, Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 2650–2658.
23. Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, Proc. IEEE 86 (11) (1998) 2278–2324
24. J. Shotton, J. Winn, C. Rother, A. Criminisi, Textonboost: joint appearance, shape and context modeling for multi-class object recognition and segmentation, in: Proceedings of the European conference on computer vision, 2006, pp. 1–15.
25. P. Kohli, P.H. Torr, Robust higher order potentials for enforcing label consistency, Int. J. Comput. Vis. 82 (3) (2009) 302–324.
26. L. Ladick, C. Russell, P. Kohli, P.H. Torr, Associative hierarchical CRFs for object class image segmentation, in: Proceedings of the IEEE International Conference on Computer Vision, 2009, pp. 739–746.
27. J. Verbeek, B. Triggs, Region classification with Markov field aspect models, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2007, pp. 1–8.