Journal of Computer and Communications, 2026, 14(7), 12-45
https://www.scirp.org/journal/jcc
ISSN Online: 2327-5227
ISSN Print: 2327-5219

Deep Learning Based Semantic Image Segmentation: A Problem Driven Analysis of Architectures, Datasets, and Open Challenges

Mehadi Hasan[1], Amina Khatun[2], Sadia Afrin Juie[1], Sumaita Binte Shorif[2], Mohammad Shorif Uddin[2*]

[1]Department of Computer Science & Engineering, Daffodil International University, Dhaka, Bangladesh
[2]Department of Computer Science & Engineering, Jahangirnagar University, Dhaka, Bangladesh
Email: mehadihasan.cse@diu.edu.bd, amina_basher@juniv.edu, juie2305101907@diu.edu.bd, sumaita.bs@gmail.com, *shorifuddin@juniv.edu

How to cite this paper: Hasan, M., Khatun, A., Juie, S.A., Shorif, S.B. and Uddin, M.S. (2026) Deep Learning Based Semantic Image Segmentation: A Problem Driven Analysis of Architectures, Datasets, and Open Challenges. *Journal of Computer and Communications*, **14**, 12-45.
https://doi.org/10.4236/jcc.2026.147002

Received: June 12, 2026
Accepted: July 6, 2026
Published: July 9, 2026

Abstract

Semantic segmentation is a cornerstone of computer vision, playing a crucial role in scene understanding and object recognition. The main goal of semantic segmentation is to give each pixel in an image its own label breaking it up into meaningful areas. This method helps machines understand visual data better allowing them to look at and make sense of images at a higher level. This paper presents a comprehensive, problem-driven analysis of deep learning based semantic image segmentation, focusing on architectures, datasets, evaluation metrics, and open challenges. It emphasizes key problem domains such as multi-scale context modeling, boundary accuracy, computational efficiency, and data scarcity. We analyze widely used deep learning methods, including Fully Convolutional Networks (FCNs), DeepLab, SegNet, and recurrent-based models, highlighting their strengths and limitations in addressing these challenges. We provide a structured overview of benchmark datasets and discuss evaluation metrics such as Intersection over Union (IoU) and pixel accuracy in practical contexts. A meta-level comparative analysis is conducted to examine trade-offs between accuracy, efficiency, and generalization across different approaches. Furthermore, we propose a conceptual segmentation framework integrating attention mechanisms, multi-scale feature extraction, and feature fusion. Finally, we identify current research gaps and outline future directions, including real-time segmentation, multi-modal learning, and data-efficient approaches.

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 [4]. Among those uses are, to mention a few, human-machine interaction [5], autonomous driving [6]-[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. Although several papers have explored semantic segmentation, many existing surveys are primarily architecture-focused and lack a problem-driven analytical perspective. In particular, prior works often emphasize model descriptions without systematically linking them to key challenges such as multi-scale variation, boundary precision, computational efficiency, and limited annotated data. Furthermore, limited attention has been given to integrating architectures, datasets, and deployment constraints within a unified analytical framework. To address these gaps, this paper adopts a problem-driven approach, where segmentation methods are analyzed based on how effectively they address core challenges. Additionally, this study provides a meta-level synthesis across architectures, datasets, and evaluation strategies, offering a more holistic and practical understanding of semantic seg-

mentation. This study focuses specifically on deep learning-based semantic segmentation methods. Classical non-deep learning approaches are only briefly mentioned for context and are not analyzed in detail. Furthermore, domain-specific segmentation applications (e.g., medical-only or satellite-only systems) are not deeply explored unless they contribute to general methodological insights.

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 study. 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 this paper are as follows:

• An extensive and well-structured analysis of the most important deep learning techniques for semantic segmentation, along with their historical development and contributions.

• A detailed overview of benchmark datasets and evaluation metrics for performance analyses in terms of memory usage, execution time, and precision.

• A proposed conceptual semantic segmentation architecture that integrates attention mechanisms, multi-scale context extraction, and feature fusion, serving as a blueprint for future research.

This survey follows a structured literature review approach to analyze deep learning-based semantic segmentation methods. Relevant studies were collected from major scientific databases including IEEE Xplore, ACM Digital Library, SpringerLink, ScienceDirect, and Google Scholar. The review primarily covers publications from 2014 to 2025, corresponding to the emergence and evolution of deep learning-based semantic segmentation. Papers were selected based on the following criteria:

• Introduction of a novel semantic segmentation architecture;

• Significant methodological contribution in contextual modeling, boundary refinement, efficiency, or data-efficient learning;

• Evaluation on established benchmark datasets such as PASCAL VOC, Cityscapes, ADE20K, and COCO-Stuff;

• High citation impact or influence on subsequent segmentation research.

The selected studies were grouped into five categories:

• Encoder-decoder architectures,

• Multi-scale context modeling approaches,

• Context-aware and recurrent architectures,

• Efficiency-oriented architectures,

• Uncertainty-aware segmentation models.

This categorization supports the problem-driven analysis adopted throughout the paper.

The subsequent sections of this article are structured as follows to clarify the

overall flow of this paper: Section 2 provides the basic concepts of semantic segmentation. Section 3 describes various deep learning based architectures for semantic segmentation. Section 4 provides the details of available datasets for semantic segmentation. Section 5 presents the evaluation metrics along with the performance analyses of the important semantic segmentation techniques. Section 6 presents a number of previous experimental results reported in the literature based on different types of deep learning architecture and available datasets and proposed a conceptual diagram for semantic segmentation. Section7 highlights the challenges and future scope for the researcher in image segmentation. Finally, Section 8 draws the conclusion of this study.

2. Background and Preliminaries

2.1. Semantic Segmentation

Semantic segmentation is a fundamental task in computer vision that enables machines to interpret visual scenes by assigning a semantic label to every pixel in an image. Unlike traditional image classification methods that assign a single label to an entire image, semantic segmentation provides a detailed pixel-level understanding of visual content. This allows the model to identify and differentiate multiple objects and regions within a scene simultaneously. By producing dense prediction maps, semantic segmentation enables more precise scene understanding, which is essential for many real-world applications such as autonomous driving, medical image analysis, robotic perception, and augmented reality. The ability to accurately locate object boundaries and contextual relationships between objects makes semantic segmentation a key component in modern computer vision systems.

2.2. Labels or Classes

In semantic segmentation, labels or classes represent the predefined semantic categories assigned to individual pixels within an image. Each pixel is classified according to the object or region it belongs to, allowing the model to construct a detailed representation of the scene. For example, in an urban street image, typical classes may include road, pedestrian, vehicle, building, vegetation, and sky. The segmentation model learns to associate pixel-level features with these categories during training, enabling it to generate a segmentation map where each pixel corresponds to a specific semantic class.

2.3. Ground Truth

Ground truth refers to manually annotated data used as a reference for training and evaluating semantic segmentation models. In this process, human annotators assign a semantic label to every pixel in an image, creating a labeled segmentation map that represents the correct interpretation of the scene. During the training stage, the model learns by comparing its predicted segmentation map with the ground truth annotations. The difference between these two maps is used to up-

date the model's parameters through optimization algorithms. During evaluation, predicted results are compared with ground truth data using performance metrics such as pixel accuracy and Intersection over Union (IoU), which measure the similarity between predicted and actual segmentation outputs.

2.4. Transfer Learning

Training deep neural networks from scratch often requires very large datasets and substantial computational resources. In many practical scenarios, such datasets may not be available. It is frequently beneficial to begin with pre-trained weights rather than randomly started ones, even in cases when a sufficiently big dataset is available and convergence happens quickly [27] [28]. Transfer learning addresses this limitation by utilizing pre-trained models that have already learned useful feature representations from large-scale datasets. In semantic segmentation, transfer learning typically involves fine-tuning a pre-trained network by adapting its weights to a specific segmentation task. Since lower layers of neural networks often capture general visual features such as edges and textures, these layers are usually retained while higher layers are adjusted for the new task. This approach reduces training time and improves model performance, especially when the available training data is limited.

2.5. Data Preprocessing and Augmentation

Data preprocessing and augmentation play an important role in improving the performance and generalization capability of semantic segmentation models. Preprocessing techniques prepare raw images for training by ensuring consistency in data representation and reducing noise or variability in input images. Common preprocessing steps include image normalization, resizing, and cropping. Normalization scales pixel values to improve numerical stability during training, while resizing ensures uniform input dimensions across the dataset. Data augmentation increases the diversity of training data by applying transformations to existing images. Techniques such as image rotation, horizontal and vertical flipping, scaling, and color jittering introduce variations in object orientation, size, and lighting conditions. It is a widely used method that has been shown to help with deep architectures in particular and machine learning models in general. It can either accelerate convergence or function as a regularizer to prevent overfitting and improve generalization capabilities [29].

2.6. Super-Pixels

Superpixels are groups of neighboring pixels that share similar visual characteristics such as color, texture, or intensity. Instead of processing each pixel individually, superpixel segmentation divides an image into meaningful regions that preserve important structural information while reducing computational complexity. Algorithms such as Simple Linear Iterative Clustering (SLIC), Felzenszwalb segmentation, and QuickShift are commonly used to generate superpixels. By oper-

ating on these grouped regions rather than individual pixels, segmentation models can perform more efficient analysis while maintaining important spatial relationships within the image.

2.7. General Semantic Segmentation Method

Semantic segmentation typically follows a structured pipeline consisting of several stages. Figure 1 shows the structured process of general semantic segmentation process. The process begins with capturing or obtaining an input image that represents the scene of interest. The segmentation model then extracts relevant features from the image using convolutional neural networks or other deep learning architectures. These extracted features are analyzed to identify potential object regions and contextual information within the scene. The model subsequently assigns semantic labels to each pixel, generating a segmentation map that highlights different objects and regions in the image. Post-processing steps may also be applied to refine segmentation boundaries and improve prediction accuracy.

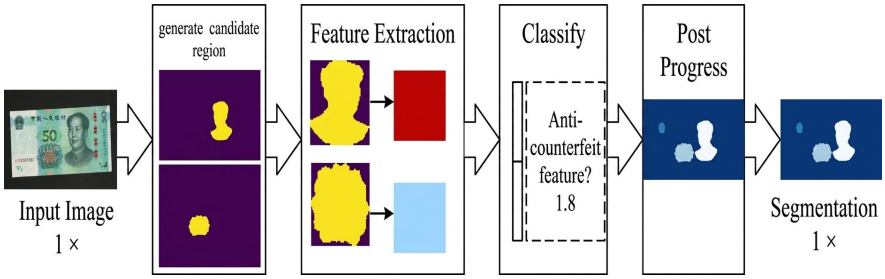

Figure 1. Schematic diagram for the general semantic segmentation approach [30].

2.8. Semantic Segmentation under the Complex Background

Segmenting objects in complex scenes remains a challenging task due to factors such as occlusions, varying lighting conditions, and overlapping objects. In such scenarios, segmentation models must effectively capture both local details and global contextual relationships to accurately distinguish objects from background regions. Advanced deep learning architectures address these challenges by incorporating mechanisms such as contextual modeling, multi-scale feature extraction, and encoder-decoder structures. These techniques allow models to analyze images at different spatial resolutions while maintaining detailed information about object boundaries. The segmentation workflow typically involves identifying image components, grouping related regions, assigning semantic labels, and verifying the consistency of the segmentation output. Figure 2 illustrates a typical segmentation pipeline used to process images with complex backgrounds.

3. Deep Network Architectures for Semantic Segmentation

We said before in the section that certain deep networks have become widely used benchmarks in the area due to their impressive performance. Among them are DeepLabv2, ResNet, VGG-16, MCG, AlexNet, and GoogLeNet. Because of their

immense power, these networks are frequently the foundation of numerous segmentation models. As such, this section will be devoted to their analysis.This section will also adopts a problem-driven perspective by analyzing how different models address key challenges in semantic segmentation, including multi-scale feature representation, contextual understanding, boundary accuracy, and computational efficiency. This approach enables a more analytical comparison across methods beyond simple architectural descriptions.

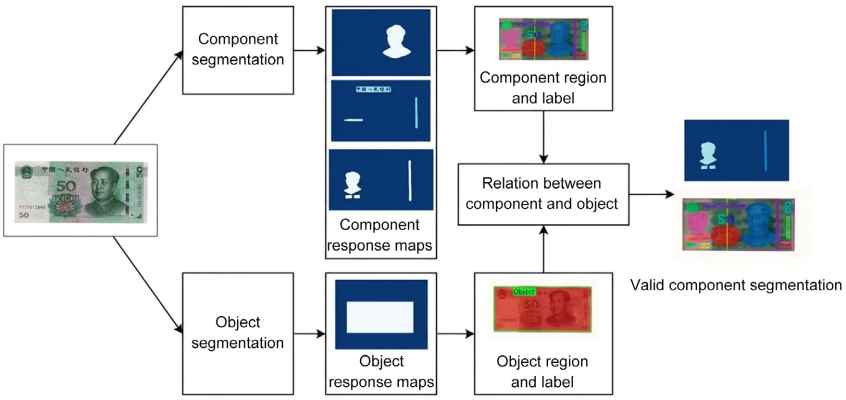

Figure 2. Segmentation pipeline used to process images with complex backgrounds [30].

3.1. DeepLab

A complex framework for semantic picture segmentation, the DeepLab architecture makes use of cutting-edge methods to boost feature extraction and increase segmentation accuracy. Atrous convolution [19] is a key component of its design because it provides fine control over feature response resolution, making it possible to capture multi-scale contextual information without adding more parameters or computing power. Using numerous parallel atrous convolutional layers with varied sampling rates, the design uses Atrous Spatial Pyramid Pooling (ASPP) to efficiently separate objects at different scales. Besides, DeepLab combines the advantages of probabilistic graphical models and deep convolutional neural networks (DCNNs) to refine segmentation findings and enhance border localization by integrating fully connected Conditional Random Fields (CRFs). DeepLab achieves state-of-the-art performance on multiple semantic segmentation benchmarks thanks to its improved feature representation, which is gained from its foundation in residual networks (ResNet).

3.2. Dilation

According to the study paper "Multi-Scale Context Aggregation by Dilated Convolutions," the idea of dilation [31] is essential for improving convolutional neural networks' performance, especially when it comes to semantic segmentation which is the process of giving each pixel in an image a name. Dilation is a method that increases convolutional layers' receptive fields without lowering the feature maps' spatial resolution. This is made possible by the use of dilated convolutions, which

let the network gather more contextual data by allowing filters to be applied at intervals rather than at each pixel. The mathematical formulation of the dilated convolution operator is expressed in Equation (1).

$$\left(F * l_k\right)(p) = \sum_{s+lt} F(s)k(t) \tag{1}$$

A discrete function is represented by F in this equation, a discrete filter by k, and the dilation factor by l. The dilation factor establishes the distance between filter applications, enabling a single filter to affect a greater portion of the input data. This capacity allows the model to collect data from a larger range of input without the requirement for downsampling, which can result in the loss of crucial spatial details. It is especially helpful for tasks that require comprehending the context around each pixel. The capacity of dilated convolutions to accommodate an exponential expansion of the receptive field is among its most important advantages. The model can successfully include multi-scale contextual information because as the dilation factor rises, the region of the input that contributes to the output grows rapidly. Understanding the interactions between pixels at different scales is important for semantic segmentation, as it can greatly improve prediction accuracy. The authors carry out a number of in-depth studies to confirm the efficacy of their strategy. They show that substantial accuracy gains can be achieved by integrating dilated convolutions into current semantic segmentation designs. By offering a more thorough comprehension of the input data, the context module significantly improves the performance of cutting-edge models by enabling the network to take use of contextual cues that are essential for producing precise predictions.

3.3. CRFasRNN

A sophisticated technique for semantic image segmentation called CRF-RNN (Conditional Random Fields as Recurrent Neural Networks) combines RNNs and CRFs into a single, end-to-end trainable neural network architecture. By improving pixel-by-pixel predictions and promoting smoothness in segmentation outputs where neighboring pixels with similar appearances are more likely to share the same label CRFs are probabilistic models that are used to enforce spatial dependencies. Traditionally, Convolutional Neural Networks (CNNs) produced the initial predictions, and then CRFs were applied as a post-processing step. On the other hand, CRF-RNN is novel in that it incorporates CRF inference as a recurrent process, enabling joint CNN and CRF training. The outputs of this integration are segmentation that is more precise and well-rounded. By using recurrent layers to update the segmentation mask iteratively, CRF-RNN is able to enhance performance on dense pixel labeling tasks and differentiate object borders more accurately. The method, which was first presented by Shuai Zheng et al. in their 2015 publication "CRF as RNN: Conditional Random Fields as Recurrent Neural Networks," [32] has been demonstrated to perform better than standalone CNNs, particularly in datasets where accurate segmentation is crucial, such as PASCAL

VOC and MS COCO.

3.4. ParseNet

Designed for semantic segmentation, ParseNet [33] is an end-to-end convolutional neural network that successfully resolves local ambiguities in pixel classification by incorporating global context. ParseNet incorporates global context directly into a fully convolutional network (FCN) without segmenting the image, in contrast to earlier techniques that depended on patch-based frameworks. This enables joint predictions of all pixel values. Using unpooling and selective application to various feature maps, the architecture creates a context vector by pooling feature maps over the entire image. This vector is then appended to the features sent to following layers. ParseNet uses L2 normalization and a scaling factor that is trained by backpropagation to handle the variations in feature scales, which greatly improves the FCN's performance. The authors show that adding global context results in accuracy that is on par with post-processing techniques that use detailed structure information, like Conditional Random Fields (CRFs). Segmentation results on the PASCAL VOC2012 test set fall within the DeepLab-LargeFOV-CRF standard deviation. With its resilience and simplicity of design, ParseNet is an easy and effective choice for semantic segmentation tasks, achieving accuracy on par with more complicated architectures with little additional training or inference time. ParseNet achieves near state-of-the-art performance on PASCAL VOC2012 and state-of-the-art results on SiftFlow and PASCAL-Context, according to extensive experimental validation. The authors are interested in combining ParseNet with structured training and inference methods to further improve performance.

3.5. FCN-8s

Fully Convolutional Networks (FCNs) [34] were first introduced for pixel-wise prediction tasks such as semantic segmentation in the paper "Fully Convolutional Networks for Semantic Segmentation" by Jonathan Long, Evan Shelhamer, and Trevor Darrell, which was presented at the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). The authors enabled dense output predictions by replacing fully connected layers with convolutional ones, converting conventional convolutional neural networks (CNNs) into fully convolutional architectures that allowed the network to generate pixel-level categorization maps. By upsampling feature maps using deconvolution (transposed convolution) layers to recover spatial resolution lost during pooling, FCNs effectively carry out end-to-end learning for segmentation. Skip connections allow the model to capture strong semantic knowledge while preserving fine features between deeper, low-resolution layers and early, high-resolution layers. The scientists created FCN variants of popular image classification networks like AlexNet, VGGNet, and GoogLeNet by utilizing pre-trained models for better segmentation results. On datasets such as PASCAL VOC, FCNs obtained state-of-the-art results, striking a signifi-

cant advantage over region-based techniques. This work continues to have an impact on applications in satellite imagery, autonomous driving, and medical imaging. It also lay the groundwork for contemporary segmentation systems such as U-Net, SegNet, and DeepLab. FCNs are an essential tool in computer vision, having revolutionized dense prediction problems with their introduction.

3.6. Multi-Scale-CNN-Eigen

The "Multi-scale-CNN-Eigen" [21] method represents a substantial development in computer vision, especially with regard to scene comprehension. This method allows for a thorough comprehension of the environment by integrating surface normal estimate, depth prediction, and semantic labeling into a single multi-scale convolutional network. This is important for applications like as augmented reality, robotics, and autonomous driving, where a thorough grasp of the surroundings is necessary for efficient navigation and interaction. The method examines images at various resolutions using a convolutional neural network (CNN) architecture to capture both fine-grained details and large contextual information. The "Multi-scale-CNN-Eigen" solution does away with the necessity for low-level segmentation approaches by simply regressing pixel maps from the input images. This streamlines the pipeline and improves the model's capacity to learn from unprocessed input. Using a series of scales, the architecture iteratively improves predictions, with coarser scales giving a broad picture of the world and finer scales concentrating on specific details. Tasks that demand precise and comprehensive outputs, such as semantic labeling, surface normal estimate, and depth prediction, benefit greatly from this multi-scale processing. For scholars and professionals in the field, the "Multi-scale-CNN-Eigen" approach presents a strong substitute since it provides a flexible and effective model that performs well on a variety of tasks. Future research and applications in computer vision will have a strong basis thanks to the field's continued evolution.

3.7. Bayesian SegNet

A more sophisticated iteration of the SegNet architecture called Bayesian SegNet [35] uses Bayesian inference to assess uncertainty in semantic segmentation tasks. This approach, which was introduced by Kendall et al., uses Monte Carlo dropout during inference, allowing for the approximation of epistemic uncertainty, or uncertainty resulting from the model's parameters. Bayesian SegNet provides confidence levels for pixel-wise classifications by aggregating predictions to estimate both mean and variance across repeated forward passes with dropout enabled. Understanding the model's confidence is essential for making sound decisions in safety-critical applications like autonomous driving and medical imaging, where this uncertainty estimation is very helpful. Bayesian SegNet improves prediction reliability by addressing both epistemic and aleatoric uncertainty (inherent noise in the data). This makes it useful for tasks that require accurate scene understanding and for directing active learning strategies in environments where the data is

noisy or unclear.

3.8. DAG-RNN

Directed Acyclic Graph Recurrent Neural Networks (DAG-RNNs) [36] are a novel technique to semantic segmentation, as presented in the paper "DAG-Recurrent Neural Networks for Scene Labeling" by B. Shuai, Z. Zuo, G. Wang, and B. Wang. By representing pixel associations as a directed acyclic graph (DAG), DAG-RNNs are engineered to capture long-range dependencies in images, in contrast to conventional Convolutional Neural Networks (CNNs) that mostly rely on local context. The fundamental idea is to multiply the ways in which contextual information spreads throughout the image, hence improving the network's understanding of global structures while preserving the local accuracy needed for pixel-level labeling. This system aggregates contextual information throughout the entire image by treating each pixel as a node in the DAG and passing information along edges between nodes. Through the use of a recurrent process, DAG-RNNs iteratively improve the accuracy of segmentation results by refining their predictions. The network performs especially well in complicated scenarios with unclear object borders or obscured objects. The architecture outperforms classic CNN-based techniques in scene labeling challenges by capturing contextual links between objects and their surroundings more effectively. Additionally, DAG-RNNs can be combined with current CNN designs to improve speed by incorporating the ability to represent long-range dependencies. Because of its great flexibility and scalability, the model may be applied to a wide range of scene understanding tasks that call for meticulous segmentation that takes into account both local and global image data.

3.9. rCNN

The publication "Recurrent Convolutional Neural Networks for Scene Parsing" presents rCNNs [37], which are a type of convolutional neural network that incorporates recurrent connections to capture spatial correlations in an input image. Recurrent layers are used by rCNNs, in contrast to regular CNNs, which process information locally. This allows the network to continually process features from earlier steps and accumulate contextual knowledge over a number of iterations. By capturing both local and global context for more precise segmentation, this method improves scene parsing by modeling long-range dependencies throughout the image. A convolutional layer and a recurrent layer, which are implemented using structures like Recurrent Neural Networks (RNNs) or Long Short-Term Memory (LSTM) units, make up the rCNN architecture. By feeding their output back into the network, these recurrent layers allow the network to gradually improve its predictions. The primary benefit of rCNN is in its capacity to manage intricate spatial relationships inside scenes, hence enhancing parsing accuracy for objects with diverse sizes and forms. Sequential spatial correlations may be accounted for by the model thanks to its recurrent structure, which is cru-

cial for tasks like semantic segmentation and object border identification.

3.10. 2D-LSTM

A unique method for semantic segmentation utilizing 2D-Long Short-Term Memory (LSTM) [38] networks is presented in the publication "Scene Labeling with LSTM Recurrent Neural Networks" by W. Byeon, T. M. Breuel, F. Raue, and M. Liwicki. Specifically designed for picture data, the 2D-LSTM architecture expands the capabilities of regular LSTMs, which are usually used to sequential data, to function over two-dimensional grids. To enable the network to comprehend pixel interactions in all directions, the 2D-LSTMs in this model process an image by capturing spatial dependencies both vertically and horizontally. This has significant importance in scene labeling tasks where correct object segmentation depends heavily on the context and spatial dependencies across various areas of the image. The way the 2D-LSTM works is that it processes each pixel in a way that resembles a grid, repeatedly going over the image in both horizontal and vertical directions while preserving hidden states that hold contextual data. The network can collect long-range dependencies throughout the image thanks to this recurrent mechanism, which helps it get over the drawbacks of conventional convolutional neural networks (CNNs), which usually rely on local receptive fields and struggle to capture more contextual information. Because the architecture uses 2D-LSTMs, it performs well in situations where precise scene understanding requires knowledge of both local details and global context, like occlusions or intricate item arrangements.

3.11. SegNet

A ground-breaking deep learning architecture designed for semantic picture segmentation is presented in the 2015 publication "SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation" by V. Badrinarayanan, A. Kendall, and R. Cipolla. With the help of its pre-trained weights, the encoder in SegNet's [39] encoder-decoder framework which is based on the VGG16 model is able to efficiently extract hierarchical features from input images. By using the pooling indices from the encoder, the architecture's decoder recreates high resolution segmentation maps while maintaining spatial information that is frequently lost in conventional upsampling methods. The decoder's structure is mirrored by the encoder. The authors highlight SegNet's effectiveness by showing how it can produce precise segmentation results without using as much processing power as deep learning models do. Because of this, SegNet is especially well-suited for real-time applications where accurate and timely picture segmentation is essential, such as robotic vision and autonomous driving. In comparison to cutting-edge segmentation techniques, extensive testing on benchmark datasets like the CamVid dataset shows that SegNet performs competitively and demonstrates its durability in a variety of contexts. The benefits of employing pooling indices, the effects of varying network depths, and the importance of training procedures to maximize

segmentation quality are just a few of the architecture's features that are covered in this research. In addition, SegNet laid the groundwork for other sophisticated segmentation structures, which has had a substantial impact on further research in the subject. SegNet is now a well-known point of reference for researchers and practitioners looking to enhance semantic segmentation in a variety of applications due to its effectiveness and clarity. This has established SegNet as a crucial component of the continuous advancement of deep learning techniques in computer vision.

3.12. ReSeg

F. Visin and colleagues introduced ReSeg [40], a semantic segmentation model based on an architecture based on recurrent neural networks (RNNs). It functions in an encoder-decoder architecture, where the encoder extracts high-level, coarse feature maps from input pictures using conventional convolutional neural networks (CNNs). ReSeg's decoder, which uses recurrent convolutional layers to iteratively improve the segmentation map, is its special power. Through the use of RNNs more especially, convolutional gated recurrent units, or ConvGRUs the model is able to repeatedly acquire and process spatial dependencies throughout the image. This is essential for increasing the accuracy of segmentation, particularly in complicated photos with occlusions or imprecise object borders. ReSeg can iteratively revisit its former states thanks to the recurrent structure, which progressively improves pixel-wise predictions and preserves fine details. More accurate segmentation results are produced by this iterative refining, which also helps to account for contextual information, improve object boundaries, and fix past errors. ReSeg shows how segmentation performance may be enhanced by combining the advantages of CNNs for feature extraction and RNNs for context-aware processing. This is especially useful in complex settings where spatial linkages or unclear object boundaries are significant.

3.13. ENet

ENet is a deep neural network architecture that was created especially for real-time semantic segmentation, with a heavy emphasis on speed and efficiency. It was first presented by A. Paszke, A. Chaurasia, S. Kim, and E. Culurciello [41]. In contrast to numerous other segmentation models that give precedence to accuracy over computational resources, ENet is tailored for settings with constrained processing capabilities, such real-time apps and mobile devices. In order to accomplish this, it uses an encoder-decoder structure while significantly lowering the computing burden on the encoder and decoder. The architecture minimizes the number of operations while retaining useful feature extraction by employing downsampling techniques early in the network to significantly reduce the spatial resolution of the input. ENet includes several asymmetric convolutions, bottleneck modules, and dilated convolutions that efficiently collect multi-scale information without adding to computational complexity in order to further boost ef-

ficiency. PReLU (Parametric Rectified Linear Units) activations are another technique that ENet employs to improve gradient flow during training and hasten convergence. While ENet loses some accuracy when compared to larger systems, it achieves the best possible speed and performance trade-off, which makes it perfect for real-time applications where quick processing is essential. The model demonstrates its feasibility in real-world scenarios where reasonable accuracy and speed are required, by achieving state-of-the-art segmentation performance on many datasets.

In summary, different deep learning architectures exhibit distinct strengths and limitations. For instance, models such as DeepLab effectively capture multi-scale contextual information using atrous convolution, while encoder-decoder architectures like SegNet and FCN excel at preserving spatial details. Lightweight models such as ENet prioritize computational efficiency, making them suitable for real-time applications, albeit with some trade-offs in accuracy. Recurrent-based approaches (e.g., CRF-RNN, DAG-RNN) enhance contextual modeling and boundary refinement but often increase computational complexity. These observations highlight an inherent trade-off between accuracy, efficiency, and model complexity, indicating that the choice of architecture should be guided by the specific application requirements.

Below Table 1 provides a comparative summary of major deep learning architectures for semantic segmentation, highlighting their strengths, limitations, and suitable application scenarios.

Table 1. Comparison among all major semantic segmentation models.

Model	Key Strength	Weakness	Best Use Case
FCN-8s	First pixel-wise segmentation model	Low boundary precision	Basic segmentation tasks
DeepLab	Strong multi-scale feature extraction	High computational cost	High-accuracy segmentation
SegNet	Efficient encoder-decoder structure	Slightly lower accuracy	Balanced performance tasks
Bayesian SegNet	Provides uncertainty estimation	Slower inference	Safety-critical applications
CRF-RNN	Improves boundary refinement	Complex training pipeline	Fine boundary segmentation
ParseNet	Incorporates global context effectively	Limited improvement	Context-aware tasks
Dilated CNN	Expands receptive field	Gridding artifacts	Multi-scale feature learning
ENet	Very fast and lightweight	Lower accuracy	Real-time applications
MSCNN-Eigen	Captures global/local features	Complex architecture	Scene understanding tasks
DAG-RNN	Captures long-range dependencies	High complexity	Complex scene labeling
rCNN	Models spatial relationships	Slower training	Context-heavy segmentation
2D-LSTM	Strong spatial dependency	Computationally expensive	Structured scene analysis
ReSeg	Combines CNN + RNN	Slower and complex	Detailed segmentation tasks

3.14. Mapping Architectures to Core Challenges

1) *Multi-Scale Context Modeling*: Multi-scale object representation remains a major challenge because objects appear at varying sizes. DeepLab addresses this challenge through atrous convolution and ASPP modules that enlarge receptive fields without reducing feature resolution. Multi-scale CNN Eigen similarly captures information at different scales through hierarchical processing. However, both approaches may still struggle with extremely small objects and complex scale variations.

2) *Boundary Precision*: Accurate boundary delineation is essential for pixel-level segmentation. CRF-RNN improves boundary quality by integrating Conditional Random Fields into an end-to-end trainable framework, while SegNet preserves spatial information through encoder pooling indices. Nevertheless, boundary errors remain common in cluttered scenes and under heavy occlusion.

3) *Computational Efficiency*: Real-time applications require efficient architectures. ENet addresses this challenge using lightweight bottleneck modules and aggressive downsampling, significantly reducing computational cost. However, efficiency gains often come at the expense of segmentation accuracy compared with larger architectures such as DeepLab.

4) *Data Scarcity*: Limited annotated data remains a major obstacle. Bayesian SegNet provides uncertainty estimation that can support active learning, while transfer learning and augmentation techniques discussed in Section III help mitigate annotation scarcity. Nonetheless, fully supervised segmentation continues to depend heavily on large labeled datasets.

4. Available Datasets

In deep learning-based semantic segmentation systems, the availability of large and well-annotated datasets is essential for training robust models. Compared with traditional machine learning techniques, deep neural networks require significantly larger volumes of data to effectively learn complex visual patterns. However, constructing such datasets is a challenging task that requires substantial time, domain expertise, and computational resources. The process involves collecting images, performing accurate pixel-level annotations, and organizing the data in a format suitable for machine learning algorithms.

Due to the difficulty of constructing large datasets, researchers often rely on publicly available benchmark datasets that have already been curated and annotated. These datasets allow fair comparison between different segmentation algorithms and are widely used in research communities. Many datasets are also associated with benchmarking challenges where evaluation is performed on hidden test sets, ensuring objective performance comparison across different methods.

Several publicly available datasets have become widely adopted benchmarks for semantic segmentation research.

The ADE20K dataset [42] is a large and diverse dataset designed for scene parsing and semantic segmentation tasks. It contains more than 20,000 images with

pixel-level annotations covering approximately 150 semantic categories. The dataset includes both indoor and outdoor environments and presents significant challenges due to the wide variation in object sizes and spatial relationships. ADE20K has become an important benchmark for evaluating segmentation algorithms and is widely used in scene understanding research.

Another important dataset is COCO-Stuff [43], which extends the original COCO dataset by including pixel-level annotations for "stuff" classes such as sky, road, and vegetation. While the original COCO dataset focuses mainly on object detection, COCO-Stuff provides annotations for both object categories and background regions, allowing a more complete understanding of visual scenes. The dataset includes annotations for 91 object categories and has become an important resource for training models that require comprehensive scene understanding.

The Pascal VOC dataset [44] is one of the earliest and most influential benchmarks in computer vision. It contains images annotated across 20 object categories including vehicles, animals, and everyday objects. Pixel-level segmentation annotations are provided, making it suitable for evaluating segmentation algorithms that require accurate object boundary detection. Due to its standardized training and testing splits, Pascal VOC has been widely used for benchmarking segmentation methods.

The Pascal Context dataset [45] extends the Pascal VOC dataset by providing dense pixel-level annotations for the entire image rather than focusing only on object regions. The dataset contains more than 400 semantic labels representing objects, background elements, and hybrid classes. With more than 10,000 annotated images, Pascal Context provides a richer representation of scene context and is widely used in research involving scene understanding and contextual reasoning.

The PASCAL Part dataset [46] further expands the Pascal VOC dataset by providing segmentation masks for object parts. It includes annotations for different components of objects such as head, torso, arms, and legs. The dataset contains 39 part classes across 20 object categories and includes more than 10,000 images used for training and evaluation. This dataset is particularly useful for tasks involving fine-grained object understanding and human body part segmentation.

Indoor scene understanding datasets are also important for semantic segmentation research. The NYU Depth V2 dataset [47] contains RGB and depth images captured using the Microsoft Kinect sensor. The dataset includes 1449 aligned RGB-D image pairs with dense pixel-level annotations. In addition, the dataset provides hundreds of thousands of unlabeled frames that can be used for further research. NYU Depth V2 is widely used in tasks such as semantic segmentation, depth estimation, and surface normal prediction.

Another large indoor dataset is SUN RGB-D [48], which contains more than 10,000 RGB-D images collected from multiple sensors across various indoor environments such as bedrooms, offices, classrooms, and kitchens. Each image is

accompanied by rich annotations including object polygons, 3D bounding boxes, object orientations, and scene categories. These annotations make the dataset useful for tasks such as semantic segmentation, scene recognition, and 3D object detection.

The SUN3D dataset [49] contains RGB-D video sequences capturing the full three-dimensional structure of indoor environments. It includes more than 400 sequences recorded in multiple buildings and rooms. Each frame contains semantic segmentation annotations along with camera pose information, making the dataset useful for research in tasks such as structure-from-motion, object recognition, and scene reconstruction.

The Semantic Boundaries Dataset (SBD) [50] focuses on predicting semantic object boundaries rather than full object segmentation. The dataset contains more than 11,000 images derived from the Pascal VOC dataset and includes detailed figure-ground masks and boundary annotations for 20 object categories. It is widely used in tasks involving edge detection and contour prediction.

Synthetic datasets have also been developed to address the difficulty of collecting large real-world datasets. The SYNTHIA dataset [51] is a large synthetic dataset designed for urban scene understanding in autonomous driving scenarios. It contains more than 200,000 images generated from a virtual city environment with different weather conditions, lighting settings, and seasons. The images include pixel-level annotations for 13 semantic classes such as road, building, pedestrian, and vehicle.

The Berkeley Deep Drive (BDD100K) dataset [52] is a large-scale driving dataset containing 100,000 videos recorded in diverse urban environments across the United States. The dataset supports multiple tasks including object detection, semantic segmentation, lane detection, and drivable area segmentation. Its large scale and diverse environmental conditions make it particularly useful for autonomous driving research.

Another dataset commonly used for road scene segmentation is the CamVid dataset [53]. It consists of video sequences captured from a dashboard-mounted camera and includes pixel-level annotations for 32 semantic classes. The dataset contains high-quality labeled frames extracted from driving videos and has been widely used for evaluating segmentation algorithms in urban environments.

The Cityscapes dataset [6] is one of the most widely used benchmarks for urban scene understanding. It contains images collected from 50 cities and provides annotations for 30 semantic classes. The dataset includes approximately 5000 finely annotated images and 20,000 coarsely annotated images. Cityscapes is widely used for evaluating segmentation models designed for autonomous driving applications.

Video-based datasets have also been developed to support segmentation tasks involving temporal information. The Youtube-Objects dataset [54] contains videos collected from YouTube covering ten object categories such as cars, dogs, cats, and airplanes. The dataset contains approximately 500,000 frames and supports

tasks such as object tracking and video segmentation.

The Materials in Context (MINC) dataset [55] focuses on material recognition and segmentation. It contains more than three million labeled samples across 23 material categories such as wood, glass, metal, and fabric. The dataset captures a wide variety of real-world materials under different lighting conditions and textures.

The KITTI dataset [7] is another widely used benchmark for autonomous driving and mobile robotics research. It includes multi-sensor data collected from real driving scenarios, including stereo cameras, LiDAR sensors, and GPS/IMU measurements. Although originally developed for tasks such as object detection and depth estimation, it has also been used in semantic segmentation research.

The Adobe Portrait Segmentation dataset [56] contains more than 20,000 images of individuals with fine-grained segmentation masks across nine classes such as skin, hair, eyes, mouth, and background. The dataset is widely used for portrait segmentation tasks such as background removal and facial editing applications.

The DAVIS dataset [57] is a high-quality dataset designed for video object segmentation. It contains 50 video sequences with pixel-level annotations across 3455 frames. The dataset captures challenging scenarios such as motion blur, occlusions, and appearance changes.

The SIFT Flow dataset [58] contains 2688 images with pixel-level annotations across 33 semantic classes and three geometric classes. The dataset is widely used for research involving dense scene correspondence and semantic scene understanding.

The Object Segmentation Dataset (OSD) [59] contains RGB-D images of indoor scenes with annotated objects. It includes color images, depth images, and ground truth segmentation masks, making it useful for evaluating algorithms that combine color and depth information.

The Stanford Background dataset [60] consists of 715 outdoor images annotated with both semantic and geometric labels such as sky, road, building, and vegetation. The dataset also includes horizon location information and supports research in scene understanding.

Finally, the RGB-D Object dataset [61] contains 300 household objects categorized into 51 classes. The dataset includes synchronized RGB and depth images captured using a turntable setup, enabling research in object recognition, pose estimation, and segmentation.

Below **Table 2** provides a comparative summary of major datasets for semantic segmentation, highlighting their release date, number of classes, and data type.

5. Evaluation Metrics

A segmentation system's performance needs to be rigorously assessed in order for it to be beneficial and truly make a substantial contribution to the area. Furthermore, that assessment needs to be carried out with recognized and conventional parameters that allow for equitable comparisons with current approaches. A sys-

tem's accuracy and execution time are just two of the many factors that need to be assessed in order to determine its validity and use. Certain metrics may be more important than others depending on the goal or context of the system; for example, accuracy may be sacrificed up to a degree in favor of speed of execution for real-time applications.

Table 2. Summary of available major segmentation datasets.

Name	Purpose	Year	Cls	Data	Res	Seq	Syn/Real	Train	Val	Test
PASCAL VOC 2012	Segmentation	2012	21	2D	Var	X	R	1464	1449	P
PASCAL-Context	Generic	2014	540	2D	Var	X	R	10,103	N/A	9637
PASCAL-Part	Generic-Part	2014	20	2D	Var	X	R	10,103	N/A	9637
SBD	Generic	2011	21	2D	Var	X	R	8498	2857	N/A
MS COCO	Generic	2014	80	2D	Var	X	R	82,783	40,504	81,434
SYNTHIA	Driving	2016	11	2D	960 × 720	X	S	13,407	N/A	N/A
Cityscapes	Urban	2015	30	2D	2048 × 1024	✓	R	2975	500	1525
CamVid	Driving	2009	32	2D	960 × 720	✓	R	701	N/A	N/A
KITTI	Driving	2012	3	2D	Var	X	R	323	N/A	N/A
KITTI-Ros	Driving	2015	11	2D	Var	X	R	170	N/A	46
Stanford BG	Outdoor	2009	8	2D	320 × 240	X	R	725	N/A	N/A
SiftFlow	Outdoor	2011	33	2D	256 × 256	X	R	2688	N/A	N/A
Youtube Obj	Objects	2014	10	2D	480 × 360	✓	R	10,167	N/A	N/A
Portrait Seg	Portrait	2016	2	2D	600 × 800	X	R	1500	300	N/A
MINC	Materials	2015	33	2D	Var	X	R	7061	2500	5000
DAVIS	Generic	2016	4	2D	480p	✓	R	4219	2023	2180
NYUDv2	Indoor	2012	40	2.5D	480 × 640	X	R	795	654	N/A
SUNRGBD	Indoor	2015	37	2.5D	Var	X	R	2666	2619	5050
RGB-D Obj	Objects	2011	51	2.5D	640 × 480	✓	R	207,920	N/A	N/A

5.1. Execution Time

In semantic segmentation, speed, or runtime, is an important parameter, particularly since many systems have stringent inference time requirements. Although the training time may be considerable for very sluggish operations, it is typically not as significant because training is typically done offline. But giving precise times for methods can be deceptive because direct comparisons are challenging because hardware and backend implementation play a major role on performance. Still, repeatability is important because it allows researchers to fairly compare and evaluate the effectiveness of their methods for specific applications by publishing comprehensive execution durations, hardware specs, and benchmark settings.

5.2. Accuracy

Many evaluation criteria have been proposed and are commonly used to evaluate the precision of semantic segmentation methods. These metrics are often Intersection over Union (IoU) and pixel accuracy variations. We highlight the measures that are most frequently used to assess the performance of per-pixel labeling techniques in this domain. Please take note of the following notation for clarity: A background or void class is one of the $k+1$ classes (L_0 to L_k), and the amount of pixels from class I that are anticipated to be class j is denoted by p_{ij} . In particular, true positives are indicated by p_{ii} , but false positives and false negatives are generally denoted by p_{ij} and p_{ji} , respectively.

5.3. Pixel Accuracy (PA)

Pixel Accuracy (PA) is a baseline metric in semantic segmentation that computes the ratio of correctly classified pixels to the total pixel count. By comparing predicted pixel labels directly with the ground truth, PA provides an intuitive measure of global segmentation accuracy, where higher values denote superior boundary and region alignment. The pixel accuracy for a system evaluating $k+1$ classes is mathematically defined in Equation (2):

$$\mathrm{PA} = \frac{\sum_{i=0}^{k} p_{ii}}{\sum_{i=0}^{k} \sum_{j=0}^{k} p_{ij}} \tag{2}$$

5.4. Mean Pixel Accuracy (MPA)

Mean Pixel Accuracy (MPA) extends the baseline PA metric to provide a class-balanced evaluation, mitigating the influence of dominant background classes in imbalanced datasets. Unlike global pixel accuracy, MPA computes the classification accuracy for each individual semantic class independently before calculating their arithmetic mean. For a dataset containing $k+1$ classes, the mathematical formulation of MPA is defined in Equation (3):

$$\mathrm{MPA} = \frac{1}{k+1} \sum_{i=0}^{k} \frac{p_{ii}}{\sum_{j=0}^{k} p_{ij}} \tag{3}$$

5.5. Intersection over Union (IoU)

IoU is a popular metric for assessing image segmentation models; it is often referred to as the Jaccard Index. The overlap between the ground truth and the anticipated segmentation is measured. By dividing the amount of overlap between the ground truth and predicted segments by the area of their union, the IoU value is computed. This statistic gives a clear picture of how closely the actual segmentation matches the projected segmentation. The IoU for a given class is mathematically expressed in Equation (4):

$$\mathrm{IoU} = \frac{\text{Area of Overlap}}{\text{Area of Union}} \tag{4}$$

5.6. Mean Intersection over Union (MIoU)

When evaluating image segmentation tasks, Mean Intersection over Union (MIoU) is a frequently used metric. Equation (5) refers to the mathematical formula of Mean Intersection over Union (MIoU). Calculating the gap between the predicted segmentation regions and the ground truth allows one to assess a segmentation model's accuracy. More specifically, it shows the proportion of the union (all pixels that are part of either the ground truth or predicted regions) to the intersection (pixels that are successfully predicted).

$$\mathrm{MIoU} = \frac{1}{k+1}\sum_{i=0}^{k}\frac{p_{ii}}{\sum_{j=0}^{k}p_{ij}+\sum_{j=0}^{k}p_{ji}-p_{ii}} \tag{5}$$

5.7. Frequency Weighted Intersection over Union (FWIoU)

An improved form of Mean Intersection over Union (MIoU) that takes into consideration the frequency of each class's occurrence in the dataset is called Frequency Weighted Intersection over Union (FWIoU). In situations when certain classes predominate the dataset, it is intended to provide a more balanced evaluation metric by assigning greater weight to classes that appear frequently and less weight to classes that appear infrequently. Ihe FWIoU for a given classes is mathematically expressed in Equation (6):

$$\mathrm{FWIoU} = \frac{1}{\sum_{i=0}^{k}\sum_{j=0}^{k}p_{ij}}\sum_{i=0}^{k}\frac{p_{ii}\sum_{j=0}^{k}p_{ij}}{\sum_{j=0}^{k}p_{ij}+\sum_{j=0}^{k}p_{ji}-p_{ii}} \tag{6}$$

In practical applications, the importance of each evaluation metric may vary depending on the segmentation task. For example, in medical image segmentation, recall is particularly important because failing to detect a critical region may lead to serious consequences. In contrast, tasks such as autonomous driving often emphasize Intersection over Union (IoU) and pixel accuracy to ensure reliable scene interpretation. Therefore, the choice of evaluation metrics should be aligned with the objectives and constraints of the specific application domain.

6. Comparative Analysis of Existing Studies

The results presented in this section are compiled from previously published studies and are used to provide a comparative overview of semantic segmentation performance across different architectures and datasets. This section does not introduce new experimental findings; rather, it synthesizes results reported in the literature to elucidate performance trends, examine trade-offs between accuracy and efficiency, and assess the consistency of segmentation performance across benchmark datasets, as summarized in Table 3.

6.1. Result and Performance Evaluation

Different methods achieve different accuracy levels on the same, and also other datasets. The table below illustrates this with some examples of how certain meth-

ods work better for particular subsets. This yields varying results for the each method, and evaluates their performance on a dataset by examining these accuracy metrics (the strength and weaknesses of applying techniques). By comparing these results, one can decide which methods are optimal with the choice of dataset leading to higher performance or more reliable outcomes in data analysis. This comparison highlights the significance of appropriate method selection methods should complement data characteristics.

Table 3. Performance of different applied models using different datasets.

Dataset	Method	IoU (%)	Dataset	Method	IoU (%)
PASCAL VOC-2012	DeepLab [19] [62]	79.70	CamVid	DAG-RNN [36]	91.60
	Dilation [31]	75.30		Bayesian SegNet [35]	63.10
	CRFasRNN [32]	74.70		SegNet [39]	60.10
	ParseNet [33]	69.80		ReSeg [40]	58.80
	FCN-8s [34]	67.20		ENet [41]	55.60
	Multi-scale CNN Eigen [21]	62.60	CityScapes	DeepLab [19] [62]	70.40
	Bayesian SegNet [35]	60.50		Dilation10 [31]	67.10
PASCAL-Context	DeepLab [62]	45.70		FCN-8s [34]	65.30
	CRFasRNN [32]	39.28		CRFasRNN [32]	62.50
	FCN-8s [34]	39.10		ENet [41]	58.30
SiftFlow	DAG-RNN [36]	85.30	PASCAL-Person-Part	DeepLab [19] [62]	64.94
	rCNN [37]	77.70	Stanford Background	rCNN [37]	80.20
	2D-LSTM [38]	70.11		2D-LSTM [38]	78.56

Table 4. Summary of several deep-learning-based architectures for semantic segmentation techniques [1].

Model & Ref.	Architecture	Accuracy	Efficiency	Training	Code	Contribution
SegNet [39]	VGG-16 + Decoder	★★★	★★	★	✓	Encoder-decoder design
Bayesian SegNet [35]	SegNet	★★★	★	★	✓	Uncertainty modeling
DeepLab [19] [62]	VGG-16/ResNet-101	★★★	★	★	✓	Atrous convolution + CRF integration
CRFasRNN [32]	FCN-8s	★	★★	★★★	✓	CRF formulated as RNN
Dilation [31]	VGG-16	★★★	★	★	✓	Dilated convolutions
ENet [41]	ENet bottleneck	★★	★★★	★	✓	Lightweight bottleneck design
Multi-scale CNN Eigen [21]	Custom	★★★	★	★	✓	Multi-scale refinement
ParseNet [33]	VGG-16	★★★	★	★	✓	Global context fusion
ReSeg [40]	VGG-16 + ReNet	★★	★	★	✓	CNN + RNN refinement
2D-LSTM [38]	MDRNN	★★	★★	★	×	Spatial context modeling
rCNN [37]	MDRNN	★★★	★★	★	✓	Multi-scale context learning
DAG-RNN [36]	Elman network	★★★	★	★	✓	Graph-based context modeling

A brief overview of the several deep learning architecture-based semantic segmentation techniques is given in Table 4. In this table, important goals like accuracy, efficiency and training performance of each architecture are shown. Each goal is rated on a three-star scale (★), where a higher number of stars indicates greater emphasis on that aspect. A check mark (✓) denotes that the source code is publicly available, while a cross mark (×) indicates that it is not available. The table also highlights the key contributions of each model, such as context modeling, uncertainty modeling, and encoder-decoder design. Notable architectures include ReSeg, which extends ReNet for semantic segmentation; DeepLab, which incorporates atrous convolutions and standalone Conditional Random Fields (CRFs); and SegNet along with its probabilistic variant, Bayesian SegNet. To enhance context modeling, methods such as CRFasRNN and DAG-RNN integrate Conditional Random Fields (CRFs) with recurrent and graph-based structures to better capture spatial dependencies and contextual information.

6.2. Proposed Conceptual Architecture

We propose a synthesized conceptual reference framework that integrates commonly adopted components from modern semantic segmentation architectures, including attention mechanisms, multi-scale feature extraction, feature fusion, and contextual modeling. The framework is intended to serve as an analytical reference model rather than a new algorithmic contribution. A pictorial representation of this conceptual diagram is shown in Figure 3.

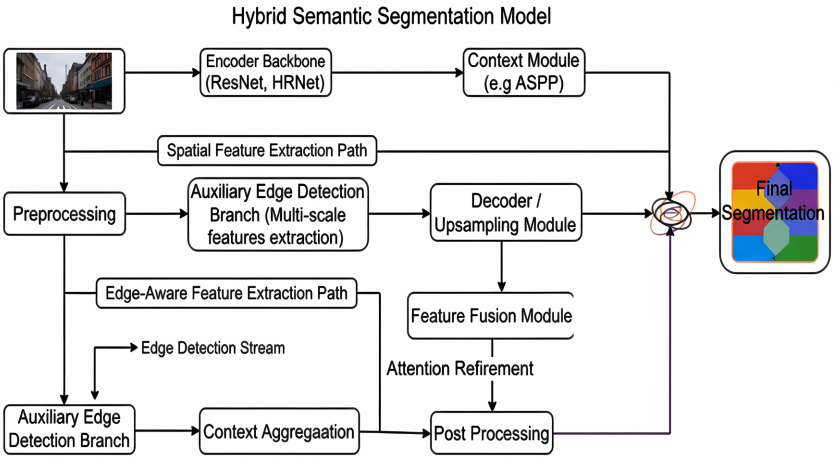

Figure 3. Proposed deep learning based semantic segmentation architecture.

An attention module is used to highlight semantically important areas after a neural encoder obtains deep feature representations from the input image. To maintain spatial accuracy, a feature fusion module blends shallow and deep features. We provide a block called Atrous Spatial Pyramid Pooling (ASPP) to manage different object scales and contextual dependencies. Through the use of upsampling and skip connections, the decoder progressively restores resolution, leading to a segmentation head that generates predictions at the pixel level. Lastly,

for crisper edges and lower noise, a post-processing block may optionally refine the output using methods like CRF. The suggested approach is appropriate for a variety of uses, such as autonomous driving, remote sensing, and medical imaging, thanks to its modular and expandable design.

• **Input Image:** An RGB image containing the scene to be examined serves as the segmentation model's input. Usually scaled and normalized during preprocessing to meet the network input requirements, this image forms the basis for all further processing. The clarity and quality of this input are critical to accurate segmentation.

• **Preprocessing:** The raw image is prepped for feature extraction by the encoder through preprocessing. It involves resizing to a set shape, normalizing pixel values (e.g., scaling between 0 and 1), and using data augmentation techniques including flipping, rotating, cropping, and adding noise. By preventing overfitting, these augmentations enhance generality.

• **(Backbone CNN) Encoder:** Hierarchical features are extracted from the input image by the encoder using a pretrained CNN backbone, such as ResNet, VGG, or EfficientNet. While deeper layers capture semantic characteristics and global context, first layers catch fine textures and edges. To increase the receptive field while maintaining spatial resolution, dilated convolutions might be employed.

• **Attention Module:** To improve the network's focus on important spatial areas or feature channels, this component makes use of attention processes. Transformer-like self-attention, spatial attention maps, and squeeze-and-excitation blocks (channel-wise attention) are a few examples. This aids the model in focusing on crucial areas and reducing unimportant background noise.

• **Fusion of Features:** Feature fusion captures both semantic richness and spatial precision by combining shallow information (from early encoder layers) with deeper ones. It is usual practice to employ skip connections from FPN-style or UNet-style architectures. The network's capacity to segment both small and complex objects is enhanced by this fusion.

• **ASPP (Atrous Spatial Pyramid Pooling):** To capture multi-scale characteristics, ASPP uses several concurrent convolutions with varying dilation rates. It is particularly helpful when objects differ significantly in size and aids the network in combining local and global context. Semantic consistency throughout the image depends on this module.

• **Decoder:** Using the compressed form that the encoder learnt, the decoder recreates the full-resolution feature map. It makes use of upsampling methods such as transposed convolutions and bilinear interpolation. In order to help create more accurate segmentation boundaries, decoder layers frequently incorporate fused information from the encoder.

• **Segmentation Head:** Using a 1×1 convolution, this last layer converts the upsampled feature map into a pixel-by-pixel class prediction. The final prediction mask is usually obtained by passing the output through a sigmoid (for binary) or softmax (for multi-class) activation.

• **Post-processing:** Post-processing methods improve the unprocessed segmentation results. Erosion and dilation are examples of morphological procedures that help clean up fragmented or disconnected areas, whereas CRFs (Conditional Random Fields) can rectify noisy borders. The segmentation map's visual and analytical quality is enhanced by this phase.

• **Output Segmentation Map:** Each pixel in the final segmentation map is given a class label, creating a thorough comprehension of the image content. Agricultural analysis, medical diagnostics, autonomous driving, and other downstream applications can all use this map as input or for direct visualization.

The proposed conceptual architecture integrates several commonly used components in modern segmentation pipelines, including multi-scale feature extraction, contextual modeling, and feature fusion. The goal of this conceptual framework is not to introduce a new architecture but to illustrate how these components are typically combined to address common segmentation challenges such as boundary precision and contextual understanding.

7. Challenges and Future Opportunities

7.1. Challenges

• The paper "Focal Loss for Dense Object Detection" [63] addresses the obstacle of class imbalance in image processing tasks, particularly in the context of dense object detection and semantic segmentation. The main challenges posed by class imbalance include training inefficiency, loss of discriminative information, model degradation, biased predictions, and evaluation difficulties. To overcome these obstacles, the paper introduces the Focal Loss, a novel loss function designed to address the extreme foreground-background class imbalance encountered during training of dense detectors. By reshaping the standard cross entropy loss to down-weight the loss assigned to well classified examples, the Focal Loss focuses training on hard examples and prevents easy negatives from overwhelming the detector during training.

• **Limited Annotated Data:** In order to segment images for biomedical applications, the research [64] presents the U-Net architecture, a deep convolutional network. Lack of labeled training data is one of the primary challenges in image processing, particularly in the biomedical domain. The authors offer a training approach that effectively utilizes the existing annotated examples by relying mostly on data augmentation in order to overcome this difficulty. Using training pictures with elastic deformations, the network may learn to be invariant to these transformations without requiring a large amount of annotated data.

• **Context Understanding:** Semantic segmentation accuracy depends on an understanding of context since contextual information is typically provided by surrounding pixels. But balancing computational efficiency with the appropriate integration of local and global context is still a problem.

• **Real-time Inference:** Semantic segmentation in real time is crucial for numerous applications, including augmented reality and driverless vehicles. A major

problem is to enable real-time performance on resource-constrained devices by striking a balance between segmentation accuracy and processing efficiency.

• **Challenges of Data Availability in Algorithm Training:** Large volumes of labeled data are needed for some of the better methods. This implies that under certain scenarios, the algorithms won't work because the labeled datasets aren't available. Though the training set size is more likely to be in the thousands for most applications, viable datasets for scene classification generally contain millions to hundreds of millions of training photos. Can deep learning algorithms be designed with fewer examples needed if the domain experts find it difficult or impossible to create very big training sets?

• **Challenges in Assessing Algorithm Generality for General Imagery:** On broad images, the efficacy of top algorithms is still unknown. Frequently, the most effective techniques are tailored to particular circumstances or environments, making their applicability vague. It is imperative that the research community tackle this dilemma.

• **Challenges in Achieving High Accuracy with Limited Computing Resources:** Several of the more advanced techniques involve a significant amount of training on computers that are not always available, such as near supercomputers. That is why a lot of scholars are thinking about the following question: What is the most accuracy possible given a given set of parameters?

• **Contextual Challenges in Accuracy and Segmentation:** While increasing accuracy is a good thing, it's also critical to know what happens when segmentations go wrong. It is not uncommon to run into segmentation issues that weren't covered by the training dataset in specific circumstances, like driving a car in a city. It would be very helpful to have a very accurate image segmentation. Nevertheless, it's unclear if we have reached that stage yet.

• **Dealing with Varying Scales and Shapes of Objects:** Semantic segmentation involves many obstacles, particularly when dealing with objects of different sizes and shapes. One model cannot fully segment all of the variables in natural settings due to the wide range of sizes and forms of items. It is crucial, but difficult, to capture contextual information at various resolutions since it calls for advanced multi scale feature extraction techniques. Complexity also arises from the need to dynamically modify receptive fields to different sizes and forms using methods such as deformable convolutions. Robust data augmentation procedures are necessary for training models to be resilient to scale fluctuations, but they can be challenging to put into practice. An additional layer of complexity is introduced by using Region Proposal Networks (RPNs) to generate precise item suggestions of various sizes.

• **Managing Overlapping Objects and Occlusions:** Because it might be difficult to discern and segment specific items that partially conceal each other, handling occlusions and overlapping objects in semantic segmentation is a substantial problem. These situations are often difficult for traditional methods to handle, which results in inaccurate segmentation or blending of objects. This problem is

addressed by sophisticated approaches that provide various labels for overlapping objects, such as instance segmentation, which distinguishes between instances of the same class. The model's capacity to distinguish obscured objects according to their spatial relationships is improved by methods that use depth information, such as RGB-D datasets. Furthermore, context-aware networks and attention mechanisms assist the model focus on pertinent areas of the image, which aids in distinguishing between overlapping and obscured areas.

7.2. Opportunities

• **Real-time Semantic Segmentation [65]:** For applications like augmented reality, robotics, and autonomous driving, it is critical to create lightweight and effective models that can process and segment pictures in real-time. For these models to function well on platforms with constrained hardware resources, such as mobile and edge devices, they must strike a compromise between computational efficiency and accuracy.

• **Handling Varying Scales and Complex Environments [66]:** Improving the resilience of the model is crucial to manage objects with different scales and intricate backgrounds. This entails creating methods for precisely separating tiny and large items inside a single scene as well as guaranteeing dependable operation in a variety of congested and varied environments—both of which are typical of real-world applications.

• **Integration with Depth and Multi-modal Data [67]:** The accuracy of segmentation can be greatly increased by combining typical RGB data with additional data modalities (such as LiDAR or infrared imaging) and depth information. Richer contextual information can be obtained by using this multi-modal method, which can aid in differentiating items that share similarities in appearance but differ in depth or thermal signatures.

• **Medical Imaging Applications [68]:** Semantic segmentation models can help with tumor detection and organ segmentation, as well as enhance diagnostic accuracy when tailored and optimized for different medical imaging modalities like MRIs, CT scans, and X-rays. In order to do this, models must be sensitive to the unique qualities and noise patterns found in medical images.

• **Self-supervised and Unsupervised Learning [69]:** Using self-supervised and unsupervised learning techniques, one can democratize access to high performing segmentation models by reducing dependence on big labeled datasets. These methods lower the expense and labor involved in manual annotation by allowing models to learn from the large amount of unlabeled data.

• **Edge Computing and IoT Applications [70]:** Optimizing segmentation models for constrained computational and memory resources is necessary for deploying them on edge devices for Internet of Things applications. This facilitates applications like smart surveillance and industrial automation by enabling real-time analysis and decision-making in smart cameras, drones, and other linked devices.

• **3D Semantic Segmentation [71]:** It is imperative to develop methods for semantic segmentation in three dimensions for applications such as robots, augmented reality, and autonomous driving. In order to separate objects in three dimensions and provide a more comprehensive picture of the environment, this entails analyzing point clouds or volumetric data.

• **Transfer Learning and Domain Adaptation [72]:** Time and resources can be saved by enhancing model performance by modifying pre-trained models for use in new tasks and domains. Domain adaption techniques guarantee that models generalize effectively across many contexts and situations, whereas transfer learning enables models trained on broad, generic datasets to be fine-tuned for specific applications.

• **Interactive and User-Guided Segmentation [30]:** Creating interactive tools that let users direct the process of segmentation can improve precision and usefulness. With the use of these technologies, users can offer suggestions or edits throughout the segmentation process, which improves efficiency and allows the process to be customized for particular needs, particularly in creative industries like graphic design and video editing.

8. Conclusions

Semantic segmentation has become a cornerstone of modern computer vision, enabling advancements in autonomous driving, medical imaging, robotics, and augmented reality. This paper highlighted the transition from traditional methods to deep learning-based approaches, emphasizing the impact of convolutional neural networks (CNNs) in revolutionizing pixel-level image understanding. Architectures such as Fully Convolutional Networks (FCNs), U-Net, and DeepLab leverage hierarchical feature extraction and contextual modeling to achieve state-of-the-art performance across diverse applications.

Despite these advancements, several challenges remain. Handling objects of varying scales and shapes continues to be difficult, often addressed through multi-scale and pyramid pooling strategies but still yielding inconsistent results. Occlusions and overlapping objects demand more sophisticated modeling of spatial relationships and context. Integrating multi-modal data, including depth and thermal information, presents promising opportunities to enhance segmentation accuracy, although standardization and further research are required.

Reducing dependence on large labeled datasets through weakly-supervised, semi-supervised, and unsupervised learning is another critical research direction. These strategies are essential for enabling robust segmentation models in domains where annotation is costly or limited. Moreover, real-time segmentation remains crucial for robotics and mobile applications, necessitating models that balance speed and accuracy. Future work should also focus on improving model robustness to domain shifts and adversarial conditions, enhancing interpretability, and integrating segmentation with complementary computer vision tasks to build more comprehensive and adaptable systems.

While deep learning based semantic segmentation has achieved remarkable progress, continued research is needed to overcome remaining challenges and unlock its full potential. Addressing these issues will facilitate the development of more precise, efficient, and flexible segmentation models, broadening the scope of applications and advancing automated visual understanding.

Authors' Contributions

All authors have contributed equally and have given their consent to publish the manuscript.

AI Disclosure Statement

During the preparation of this work, the authors used ChatGPT and QuillBot to improve readability, polish prose, and improve language fluency. After using these tools, the authors reviewed and edited the content as needed and assume full responsibility for the content of the publication.

Conflicts of Interest

The authors declare that they have no conflict of interest.

References

[1] Garcia-Garcia, A., Orts-Escolano, S., Oprea, S., Villena-Martinez, V. and Rodrıguez, J.G. (2017) A Review on Deep Learning Techniques Applied to Semantic Segmentation. arXiv: 1704.06857. http://arxiv.org/abs/1704.06857

[2] Yu, H., Yang, Z., Tan, L., Wang, Y., Sun, W., Sun, M., *et al.* (2018) Methods and Datasets on Semantic Segmentation: A Review. *Neurocomputing*, **304**, 82-103. https://doi.org/10.1016/j.neucom.2018.03.037

[3] Emek Soylu, B., Guzel, M.S., Bostanci, G.E., Ekinci, F., Asuroglu, T. and Acici, K. (2023) Deep-Learning-Based Approaches for Semantic Segmentation of Natural Scene Images: A Review. *Electronics*, **12**, Article 2730. https://doi.org/10.3390/electronics12122730

[4] Guo, Y., Liu, Y., Georgiou, T. and Lew, M.S. (2017) A Review of Semantic Segmentation Using Deep Neural Networks. *International Journal of Multimedia Information Retrieval*, **7**, 87-93. https://doi.org/10.1007/s13735-017-0141-z

[5] Oberweger, M., Wohlhart, P. and Lepetit, V. (2015) Hands Deep in Deep Learning for Hand Pose Estimation. arXiv: 1502.06807.

[6] Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S. and Schiele, B. (2016) The Cityscapes Dataset for Semantic Urban Scene Understanding. arXiv: 1604.01685. http://arxiv.org/abs/1604.01685

[7] Geiger, A., Lenz, P. and Urtasun, R. (2012) Are We Ready for Autonomous Driving? the KITTI Vision Benchmark Suite. 2012 *IEEE Conference on Computer Vision and Pattern Recognition*, Providence, 16-21 June 2012, 3354-3361. https://doi.org/10.1109/cvpr.2012.6248074

[8] Ess, A., Mueller, T., Grabner, H. and Gool, L.V. (2009) Segmentation-Based Urban Traffic Scene Understanding. *Proceedings of the British Machine Vision Conference* 2009, London, 7-10 September 2009, 84.1-84.11. https://doi.org/10.5244/c.23.84

[9] Yoon, Y., Jeon, H., Yoo, D., Lee, J. and Kweon, I.S. (2015) Learning a Deep Convolu-

tional Network for Light-Field Image Super-Resolution. 2015 *IEEE International Conference on Computer Vision Workshop* (*ICCVW*), Santiago, 7-13 December 2015, 57-65. https://doi.org/10.1109/iccvw.2015.17

[10] Lowe, D. (2001) Object Recognition from Local Scale-Invariant Features. *Proceedings of the IEEE International Conference on Computer Vision*, Kerkyra, 20-27 September 1999, 1150-1157.

[11] Dalal, N. and Triggs, B. (2005) Histograms of Oriented Gradients for Human Detection. 2005 *IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (*CVPR*'05), San Diego, 20-25 June 2005, 886-893. https://doi.org/10.1109/cvpr.2005.177

[12] Silberman, N. and Fergus, R. (2011) Indoor Scene Segmentation Using a Structured Light Sensor. 2011 *IEEE International Conference on Computer Vision Workshops* (*ICCV Workshops*), Barcelona, 6-13 November 2011, 601-608. https://doi.org/10.1109/iccvw.2011.6130298

[13] Gupta, S., Arbelaez, P. and Malik, J. (2013) Perceptual Organization and Recognition of Indoor Scenes from RGB-D Images. 2013 *IEEE Conference on Computer Vision and Pattern Recognition*, Portland, 23-28 June 2013, 564-571. https://doi.org/10.1109/cvpr.2013.79

[14] Hinton, G., Krizhevsky, A. and Sutskever, I. (2012) ImageNet Classification with Deep Convolutional Neural Networks. *Advances in Neural Information Processing Systems*, **25**, 1097-1105.

[15] Simonyan, K. and Zisserman, A. (2014) Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv: 1409.1556.

[16] Hinton, G.E., Osindero, S. and Teh, Y. (2006) A Fast Learning Algorithm for Deep Belief Nets. *Neural Computation*, **18**, 1527-1554. https://doi.org/10.1162/neco.2006.18.7.1527

[17] LeCun, Y., Bengio, Y. and Hinton, G. (2015) Deep Learning. *Nature*, **521**, 436-444. https://doi.org/10.1038/nature14539

[18] Farabet, C., Couprie, C., Najman, L. and LeCun, Y. (2013) Learning Hierarchical Features for Scene Labeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **35**, 1915-1929. https://doi.org/10.1109/tpami.2012.231

[19] Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K. and Yuille, A. (2014) Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs. arXiv: 1412.7062.

[20] Girshick, R.B., Donahue, J., Darrell, T. and Malik, J. (2013) Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. arXiv: 1311.2524. http://arxiv.org/abs/1311.2524

[21] Eigen, D. and Fergus, R. (2015) Predicting Depth, Surface Normals and Semantic Labels with a Common Multi-Scale Convolutional Architecture. 2015 *IEEE International Conference on Computer Vision* (*ICCV*), Santiago, 7-13 December 2015, 2650-2658. https://doi.org/10.1109/iccv.2015.304

[22] Lecun, Y., Bottou, L., Bengio, Y. and Haffner, P. (1998) Gradient-Based Learning Applied to Document Recognition. *Proceedings of the IEEE*, **86**, 2278-2324. https://doi.org/10.1109/5.726791

[23] Shotton, J., Winn, J., Rother, C. and Criminisi, A. (2006) TextonBoost: Joint Appearance, Shape and Context Modeling for Multi-Class Object Recognition and Segmentation. In: Leonardis, A., Bischof, H. and Pinz, A., Eds., *Computer Vision—ECCV* 2006, Springer, 1-15. https://doi.org/10.1007/11744023_1

[24] Kohli, P., Ladický, L. and Torr, P.H.S. (2009) Robust Higher Order Potentials for En-

forcing Label Consistency. *International Journal of Computer Vision*, **82**, 302-324. https://doi.org/10.1007/s11263-008-0202-0

[25] Ladicky, L., Russell, C., Kohli, P. and Torr, P.H.S. (2009) Associative Hierarchical CRFs for Object Class Image Segmentation. 2009 *IEEE* 12*th International Conference on Computer Vision*, Kyoto, 29 September-2 October 2009, 739-746. https://doi.org/10.1109/iccv.2009.5459248

[26] Verbeek, J. and Triggs, B. (2007) Region Classification with Markov Field Aspect Models. 2007 *IEEE Conference on Computer Vision and Pattern Recognition*, Minneapolis, 17-22 June 2007, 1-8. https://doi.org/10.1109/cvpr.2007.383098

[27] Oquab, M., Bottou, L., Laptev, I. and Sivic, J. (2014) Learning and Transferring Mid-Level Image Representations Using Convolutional Neural Networks. 2014 *IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, 23-28 June 2014, 1717-1724. https://doi.org/10.1109/cvpr.2014.222

[28] Yosinski, J., Clune, J., Bengio, Y. and Lipson, H. (2014) How Transferable Are Features in Deep Neural Networks? arXiv: 1411.1792. http://arxiv.org/abs/1411.1792

[29] Wong, S.C., Gatt, A., Stamatescu, V. and McDonnell, M.D. (2016) Understanding Data Augmentation for Classification: When to Warp? 2016 *International Conference on Digital Image Computing*: *Techniques and Applications* (*DICTA*), Gold Coast, 30 November-2 December 2016, 1-6. https://doi.org/10.1109/dicta.2016.7797091

[30] Huang, J., Guixiong, L. and He, B. (2021) Fast Semantic Segmentation Method for Machine Vision Inspection Based on a Fewer-Parameters Atrous Convolution Neural Network. *PLOS ONE*, **16**, e0246093. https://doi.org/10.1371/journal.pone.0246093

[31] Yu, F. and Koltun, V. (2016) Multi-Scale Context Aggregation by Dilated Convolutions. arXiv: 1511.07122.

[32] Zheng, S., Jayasumana, S., Romera-Paredes, B., Vineet, V., Su, Z., Du, D., Huang, C. and Torr, P.H.S. (2015) Conditional Random Fields as Recurrent Neural Networks. arXiv: 1502.03240. http://arxiv.org/abs/1502.03240

[33] Liu, W., Rabinovich, A. and Berg, A.C. (2015) Parsenet: Looking Wider to See Better. arXiv: 1506.04579. http://arxiv.org/abs/1506.04579

[34] Long, J., Shelhamer, E. and Darrell, T. (2014) Fully Convolutional Networks for Semantic Segmentation. arXiv: 1411.4038. http://arxiv.org/abs/1411.4038

[35] Kendall, A., Badrinarayanan, V. and Cipolla, R. (2015) Bayesian SegNet: Model Uncertainty in Deep Convolutional Encoder-Decoder Architectures for Scene Understanding. arXiv: 1511.02680. http://arxiv.org/abs/1511.02680

[36] Shuai, B., Zuo, Z., Wang, G. and Wang, B. (2015) Dag-Recurrent Neural Networks for Scene Labeling. arXiv: 1509.00552. http://arxiv.org/abs/1509.00552

[37] Pinheiro, P.H.O. and Collobert, R. (2013) Recurrent Convolutional Neural Networks for Scene Parsing. arXiv: 1306.2795. http://arxiv.org/abs/1306.2795

[38] Byeon, W., Breuel, T.M., Raue, F. and Liwicki, M. (2015) Scene Labeling with LSTM Recurrent Neural Networks. 2015 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Boston, 7-12 June 2015, 3547-3555. https://doi.org/10.1109/cvpr.2015.7298977

[39] Badrinarayanan, V., Kendall, A. and Cipolla, R. (2015) SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. arXiv: 1511.00561. http://arxiv.org/abs/1511.00561

[40] Visin, F., Kastner, K., Courville, A.C., Bengio, Y., Matteucci, M. and Cho, K. (2015) ReSeg: A Recurrent Neural Network for Object Segmentation. arXiv: 1511.07053.

http://arxiv.org/abs/1511.07053

[41] Paszke, A., Chaurasia, A., Kim, S. and Culurciello, E. (2016) ENet: A Deep Neural Network Architecture for Real-Time Semantic Segmentation. arXiv: 1606.02147. http://arxiv.org/abs/1606.02147

[42] Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A. and Torralba, A. (2017) Scene Parsing through ADE20K Dataset. 2017 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Honolulu, 21-26 July 2017, 5122-5130. https://doi.org/10.1109/cvpr.2017.544

[43] Caesar, H., Uijlings, J.R.R. and Ferrari, V. (2016) Coco-Stuff: Thing and Stuff Classes in Context. arXiv: 1612.03716. http://arxiv.org/abs/1612.03716

[44] Everingham, M., Eslami, S.M.A., Van Gool, L., Williams, C.K.I., Winn, J. and Zisserman, A. (2014) The Pascal Visual Object Classes Challenge: A Retrospective. *International Journal of Computer Vision*, **111**, 98-136. https://doi.org/10.1007/s11263-014-0733-5

[45] Mottaghi, R., Chen, X., Liu, X., Cho, N., Lee, S., Fidler, S., *et al.* (2014) The Role of Context for Object Detection and Semantic Segmentation in the Wild. 2014 *IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, 23-28 June 2014, 891-898. https://doi.org/10.1109/cvpr.2014.119

[46] Chen, X., Mottaghi, R., Liu, X., Fidler, S., Urtasun, R. and Yuille, A.L. (2014) Detect What You Can: Detecting and Representing Objects Using Holistic Models and Body Parts. arXiv: 1406.2031. http://arxiv.org/abs/1406.2031

[47] Silberman, N., Hoiem, D., Kohli, P., Fergus, R. (2012) Indoor Segmentation and Support Inference from RGBD Images. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y. and Schmid, C., Eds., *Computer Vision—ECCV* 2012, Springer, 746-760. https://doi.org/10.1007/978-3-642-33715-4_54

[48] Song, S., Lichtenberg, S.P. and Xiao, J. (2015) SUN RGB-D: A RGB-D Scene Understanding Benchmark Suite. 2015 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Boston, 7-12 June 2015, 567-576. https://doi.org/10.1109/cvpr.2015.7298655

[49] Xiao, J., Owens, A. and Torralba, A. (2013) SUN3D: A Database of Big Spaces Reconstructed Using SFM and Object Labels. 2013 *IEEE International Conference on Computer Vision*, Sydney, 1-8 December 2013, 1625-1632. https://doi.org/10.1109/iccv.2013.458

[50] Hariharan, B., Arbelaez, P., Bourdev, L., Maji, S. and Malik, J. (2011) Semantic Contours from Inverse Detectors. 2011 *International Conference on Computer Vision*, Barcelona, 6-13 November 2011, 991-998. https://doi.org/10.1109/iccv.2011.6126343

[51] Ros, G., Sellart, L., Materzynska, J., Vazquez, D. and Lopez, A.M. (2016) The SYNTHIA Dataset: A Large Collection of Synthetic Images for Semantic Segmentation of Urban Scenes. 2016 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Las Vegas, 27-30 June 2016, 3234-3243. https://doi.org/10.1109/cvpr.2016.352

[52] Yu, F., Xian, W., Chen, Y., Liu, F., Liao, M., Madhavan, V. and Darrell, T. (2018) BDD100K: A Diverse Driving Dataset for Heterogeneous Multitask Learning. arXiv: 1805.04687. http://arxiv.org/abs/1805.04687

[53] Brostow, G.J., Fauqueur, J. and Cipolla, R. (2009) Semantic Object Classes in Video: A High-Definition Ground Truth Database. *Pattern Recognition Letters*, **30**, 88-97. https://doi.org/10.1016/j.patrec.2008.04.005

[54] Prest, A., Leistner, C., Civera, J., Schmid, C. and Ferrari, V. (2012) Learning Object

Class Detectors from Weakly Annotated Video. 2012 *IEEE Conference on Computer Vision and Pattern Recognition*, Providence, 16-21 June 2012, 3282-3289. https://doi.org/10.1109/cvpr.2012.6248065

[55] Bell, S., Upchurch, P., Snavely, N. and Bala, K. (2014) Material Recognition in the Wild with the Materials in Context Database. arXiv: 1412.0623. http://arxiv.org/abs/1412.0623

[56] Kapitanov, A., Kvanchiani, K. and Kirillova, S. (2023) EasyPortrait—Face Parsing and Portrait Segmentation Dataset. arXiv: 2304.13509. https://europepmc.org/article/PPR/PPR652963

[57] Perazzi, F., Pont-Tuset, J., McWilliams, B., Van Gool, L., Gross, M. and Sorkine-Hornung, A. (2016) A Benchmark Dataset and Evaluation Methodology for Video Object Segmentation. 2016 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Las Vegas, 27-30 June 2016, 724-732. https://doi.org/10.1109/cvpr.2016.85

[58] Liu, C., Yuen, J. and Torralba, A. (2011) SIFT Flow: Dense Correspondence across Scenes and Its Applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **33**, 978-994. https://doi.org/10.1109/tpami.2010.147

[59] Richtsfeld, A. (2012) The Object Segmentation Database(OSD). https://www.acin.tuwien.ac.at/en/vision-for-robotics/software-tools/osd/

[60] Gould, S., Fulton, R. and Koller, D. (2009) Decomposing a Scene into Geometric and Semantically Consistent Regions. 2009 *IEEE* 12*th International Conference on Computer Vision*, Kyoto, 29 September-2 October 2009, 1-8. https://doi.org/10.1109/iccv.2009.5459211

[61] Lai, K., Bo, L., Ren, X. and Fox, D. (2011) A Large-Scale Hierarchical Multi-View RGB-D Object Dataset. 2011 *IEEE International Conference on Robotics and Automation*, Shanghai, 9-13 May 2011, 1817-824. https://doi.org/10.1109/icra.2011.5980382

[62] Chen, L., Papandreou, G., Kokkinos, I., Murphy, K. and Yuille, A.L. (2016) DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. arXiv: 1606.00915. http://arxiv.org/abs/1606.00915

[63] Lin, T., Goyal, P., Girshick, R.B., He, K. and Doll'ar, P. (2017) Focal Loss for Dense Object Detection. arXiv: 1708.02002. http://arxiv.org/abs/1708.02002

[64] Ronneberger, O., Fischer, P. and Brox, T. (2015) U-Net: Convolutional Networks for Bio-Medical Image Segmentation. arXiv: 1505.04597. http://arxiv.org/abs/1505.04597

[65] Zhao, H., Qi, X., Shen, X., Shi, J. and Jia, J. (2017) ICNet for Real-Time Semantic Segmentation on High-Resolution Images. arXiv: 1704.08545. http://arxiv.org/abs/1704.08545

[66] Chen, W., Miao, Z., Qu, Y. and Shi, G. (2024) HRDLNet: A Semantic Segmentation Network with High Resolution Representation for Urban Street View Images. *Complex & Intelligent Systems*, **10**, 7825-7844. https://doi.org/10.1007/s40747-024-01582-1

[67] Hazirbas, C., Ma, L., Domokos, C. and Cremers, D. (2017) FuseNet: Incorporating Depth into Semantic Segmentation via Fusion-Based CNN Architecture. In: Lai, S.H., Lepetit, V., Nishino, K. and Sato, Y., Eds., *Computer Vision—ACCV* 2016, Springer, 213-228. https://doi.org/10.1007/978-3-319-54181-5_14

[68] Gao, Y., Jiang, Y., Peng, Y., Yuan, F., Zhang, X. and Wang, J. (2025) Medical Image Segmentation: A Comprehensive Review of Deep Learning-Based Methods. *Tomography*, **11**, Article 52. https://doi.org/10.3390/tomography11050052

[69] He, K., Fan, H., Wu, Y., Xie, S. and Girshick, R.B. (2019) Momentum Contrast for

Unsupervised Visual Representation Learning. arXiv: 1911.05722. http://arxiv.org/abs/1911.05722

[70] Lane, N.D., Bhattacharya, S., Mathur, A., Georgiev, P., Forlivesi, C. and Kawsar, F. (2017) Squeezing Deep Learning into Mobile and Embedded Devices. *IEEE Pervasive Computing*, **16**, 82-88. https://doi.org/10.1109/mprv.2017.2940968

[71] Qi, C.R., Su, H., Mo, K. and Guibas, L.J. (2016) PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation. arXiv: 1612.00593. http://arxiv.org/abs/1612.00593

[72] Chen, Y., Li, W., Sakaridis, C., Dai, D. and Gool, L.V. (2018) Domain Adaptive Faster R-CNN for Object Detection in the Wild. arXiv: 1803.03243. http://arxiv.org/abs/1803.03243