algorithms

MDPI

Review

Insights into Image Understanding: Segmentation Methods for Object Recognition and Scene Classification

Sarfaraz Ahmed Mohammed and Anca L. Ralescu *

Department of Computer Science, College of Engineering and Applied Science, University of Cincinnati, Cincinnati, OH 45221-0030, USA; mohammsm@mail.uc.edu
* Correspondence: ralescal@ucmail.uc.edu

Abstract: Image understanding plays a pivotal role in various computer vision tasks, such as extraction of essential features from images, object detection, and segmentation. At a higher level of granularity, both semantic and instance segmentation are necessary for fully grasping a scene. In recent times, the concept of panoptic segmentation has emerged as a field of study that unifies semantic and instance segmentation. This article sheds light on the pivotal role of panoptic segmentation as a visualization tool for understanding scene components, including object detection, categorization, and precise localization of scene elements. Advancements in achieving panoptic segmentation and suggested improvements to the predicted outputs through a top-down approach are discussed. Furthermore, datasets relevant to both scene recognition and panoptic segmentation are explored to facilitate a comparative analysis. Finally, the article outlines certain promising directions in image recognition and analysis by underlining the ongoing evolution in image understanding methodologies.

Keywords: convolutional neural networks; image segmentation; computer vision; instance segmentation; semantic segmentation; panoptic segmentation; scene recognition; artificial intelligence; machine learning; deep learning; pre-trained networks

Citation: Mohammed, S.A.; Ralescu, A.L. Insights into Image Understanding: Segmentation Methods for Object Recognition and Scene Classification. *Algorithms* **2024**, *17*, 189. https://doi.org/10.3390/a17050189

Academic Editors: Laura Antonelli and Lucia Maddalena

Received: 25 March 2024
Revised: 21 April 2024
Accepted: 24 April 2024
Published: 30 April 2024

1. Introduction

Perception is an important aspect of computer vision, focused on understanding the complexities of images, including object recognition, scene interpretation from video surveillance devices, analyzing pedestrian movements on roads, and medical image analysis, among others. These devices generate a very high volume of data. Consequently, the fields of machine learning and big data analytics have assumed a pivotal role in extracting and comprehending patterns within these data, with the aim of developing intelligent algorithms capable of executing tasks such as object detection, segmentation, and classification.

These tasks are central to computer vision, for example, the study and implementation of techniques that include feature selection/extraction, object detection and segmentation, object labeling, to performing segmentation of image scenes from videos in real time [1]. The primary objective of Image Understanding (IU) is to accomplish three key tasks: (i) object identification (referred to as instance segmentation), (ii) object labeling, and (iii) providing a segmentation mask (bounding box) around each detected object. Image labeling represents another intriguing facet [2], facilitating the automatic assignment of labels to objects and thereby imparting significance to each object within a frame.

Semantic segmentation refers to pixel-to-pixel segmentation and relies on Fully Convolutional Networks (FCNs). In this technique, the final conventional fully connected CNN layer is substituted with a deconvolutional layer to categorize each individual pixel. For example, given a coarse image structure, FCNs excel in fine tuning the segmentation aspect by leveraging deconvolutional and pooling layers. The segmentation performance can be enhanced even further by adapting the FCNs to suit specific requirements. For instance, in

the case of U-Net [3], the approach involves augmenting the deconvolutional layers and efficiently mapping the data to higher resolutions. Conversely, the SegNet network [4] employs an encoder–decoder model to refine deconvolutional layers, by extracting indices from the max-pool layers.

However, it is worth noting that many networks share architectures that are comparable on the encoder side but mainly differ in their decoder configurations. Most recently, a probabilistic graphical model known as the fully connected random field (CRF) has been adopted by Deep Lab [5] as a substitution for the deconvolutional layers. To address the issue of information loss, the literature has explored numerous semantic segmentation approaches. These methods concentrate on feature extraction, employing either multi-scale feature aggregation [6] or end-to-end structured prediction [7].

Instance segmentation on the other hand, segments objects and identifies the object boundaries irrespective of whether the objects belong to the same category. In situations where multiple objects fall in a similar category, instance segmentation promises to differentiate each object from the other by drawing object boundaries whereas semantic segmentation relies on segmenting the collective information from these objects. Two important strategies, namely, a segment-first and instance-first strategy, have been proposed [8]. The segment-first strategy segments each object by classification results whereas the instance-first strategy identifies the ROIs for every instance and performs classification and segmentation task in parallel to each ROI. By incorporating this idea, it is inferred that Mask-RCNN outperforms other models on the COCO based instance segmentation [9].

Authors in [10] hold the view that, although semantic and instance segmentation may seem similar at first glance, the metrics and datasets associated with these two visual recognition tasks exhibit substantial differences. These differences are akin to the difference between "stuff" and "things". To gain a better understanding between semantic and instance segmentation models, Kirillov et al. [10] emphasized the need of distinguishing between "stuff" and "things", a crucial aspect in various visual recognition tasks, "stuff" referring to countable objects within an image (instance segmentation), including entities like people, animals, trucks, sky, road, and grass, while "things" (semantic segmentation), account for identification of regions with similar textures. We see a strong dichotomy between two concepts and the question of attaining a unified vision system that can perform segmentation that is coherent to meet the needs of real-world applications is still an important concern.

Panoptic segmentation (PS) is the first framework proposed in 2019 by [10] that unifies both semantic and instance segmentation. This unified segmentation has immense potential to open doors for researchers to come up with novel algorithmic solutions. When performing image segmentation, each pixel is provided with a semantic label along with an instance id. This unified approach, to combine scene-level and subject-level understanding, is driving the design and development of panoptic segmentation models. Industry leaders such as Apple, Facebook, Tesla, and Uber have been in the forefront to come up with pioneering vision systems that could provide a comprehensive view of the broader panoptic segmentation landscape.

In a similar vein, this article aims to present an overview of various PS methods for scene classification and object detection and the potential challenges that revolve around it till date. The evaluation metrics applied in this context and some of the prominent PS models such as the Panoptic Feature Pyramid Network (FPN), Attention-guided unified network for PS, Seamless scene segmentation, Panoptic Deep lab, Unified PS network, and Efficient PS are presented. Finally, the paper outlines certain key future directions in deep learning-based segmentation research.

Main Contributions of the Paper

Considering the evolution of the deep learning-based image segmentation and its anticipated future development, this article is organized as follows:

i. An overview of the research advancements in deep learning-based image segmentation with focus on panoptic segmentation (PS) is presented.
ii. The article draws attention to several interesting works towards PS that include Panoptic Feature Pyramid Network, Attention-guided network for PS, Seamless Scene Segmentation, Panoptic Deep lab, Unified panoptic segmentation network, and Efficient panoptic segmentation. A top-down approach to PS is discussed and suggested improvements in predicted outputs are highlighted.
iii. Research efforts by leading companies supporting the bigger picture of developing computer vision models for PS are presented.
iv. Performance metrics of both scene recognition and panoptic segmentation models are discussed. Several comparisons have been performed to measure the performance using different datasets under different metrics and highlight the potential benefits and challenges.

From this point on, the paper is organized as follows: after Section 1 (the current section), Section 2 discusses the literature review. Section 3 presents the concept of PS along with the metrics and discusses some interesting models used to date. It also presents a top-down view of PS and discusses certain challenges and further improvements. Section 4 presents the research efforts to date by companies such as Apple, Facebook, and Tesla in developing computer vision models using PS. For example, the use of on-device PS to enhance the camera vision that use transformers, the subject lifting network architectures by Apple, and Detectron2 by Facebook AI research, and self-driving PS support vehicles by Tesla are presented. Section 5 introduces the publicly available datasets and benchmarks to both scene recognition and PS models. Finally, Section 6 concludes the paper with future directions.

2. Literature Review

Deep learning represents a unique category within the broader domain of artificial neural networks (ANNs) that has gained immense popularity in the areas of image processing, computer vison, speech processing to name a few. It involves hundreds and thousands of neurons with millions and billions of connections and requires lots of computational power. With this, the emergence of GPUs as computational devices came into forefront, and it turns out that GPUs are great for neural networks (NNs) as they are parallel systems. The parallel nature of NNs allows us to exploit GPUs to speed up the computations on NN models. Convolutional neural networks (CNNs) are examples of deep learning models and serve as powerful tools for image analysis, video analysis, and speech recognition. Deep convolutional neural networks (DCNNs) have gained immense popularity in the research community to understand and study the different approaches to scene recognition. Based on the features extracted from an image, the authors in [11] classified scene recognition algorithms into broadly six different categories: (i) global attribute descriptors, (ii) patch feature encoding, (iii) spatial layout pattern learning, (iv) discriminative region detection, (v) object correlation analysis, and (vi) hybrid deep models.

Visual scene understanding is another interesting aspect of computer vision [12] and many research efforts have contributed to the current state of the art in this area. In general, scene understanding refers to the understanding of the intrinsic details of a scene at a very detailed level of granularity (for example, single scene category) and hence providing a global description of the image [13,14]. Object detection includes the localization of objects using bounding boxes [15–17], and focuses on identifying the object instances and categories within a scene/image. There is a strong dichotomy between semantic and instance segmentation. The semantic segmentation emphasizes a much finer grained representation and prediction of the semantic category that each pixel belongs to [5,18], whereas with instance based semantic segmentation it is arduous to identify pixels that comprises of each object instance thereby combining the integration of semantic segmentation with fine grained object detection [19].

The ongoing revolution in computer vision is driven by the success of deep learning [20] and algorithms pertaining to the understanding of visual scenes. It is important to remember that training of these deep learning models necessitates a substantial amount of training data and computational resources especially in the case of medical images and self-driving vehicles. Consider, for example, the case of autonomous self-driving cars, where decision making is critical to visual analysis and requires a reliable, real-time scene understanding [21]. Datasets such as Microsoft COCO [9], ImageNet [22], YouTube-8M [23], CamVid [24], KITTI vision benchmark suite [25], CityScapes [26], Leuven [27], are available on large scale suitable for the training of deep models for a specific task. These datasets support the use of both semantic and instance segmentation and each of these segmentation methods perform their own task and contains the information needed to perform panoptic segmentation.

3. Panoptic Segmentation

Panoptic segmentation (PS) is a computer vision task that combines semantic and instance segmentation to provide a distinction between "stuff" and "things". In general, it is a task for labeling each pixel in an image with a class category and a unique instance. It performs a good visualization of the scene components and provides a technique of performing object detection, categorization, and localization of scene components. This in turn may offer support to address the challenges of today's computer vision task such as understanding medical images, autonomous self-driving cars, video surveillance, and several others.

Alexander Kirillov et al. [10] introduced the initial framework for PS that presents the task format for PS wherein every pixel in an image is mapped to a pair that consists of a semantic class and an instance ID. Semantic labeling between stuff and things is performed by first deriving the subsets between these and pixels that belong to a single instance will have the same semantic class of pixel and the instance ID. The relationship of PS to semantic segmentation (SS) is the strict generalization as both PS and SS necessitates assigning each pixel to a semantic label. In situations where the ground truth fails to define instances or when all the classes are considered as stuff, the task formats become similar, although their metrics may vary. However, the things–classes may encompass multiple instances per image, thereby distinguishing PS from SS. While PS allows the assignment of a semantic label and an instance ID to each pixel without allowing overlapping segments, the IS task segments every instance in an image and permits the segmentation overlap.

3.1. Metrics for Panoptic Segmentation

PS is a joint segmentation task between SS and IS ("stuff" and "things") and earlier works focused on evaluating their performance metrics independently [6,28–30] (seen in Figure 1). The pursuit of a cohesive metric for unifying these distinct tasks introduces several algorithmic complexities. In their work [10], the authors introduced a trio of metrics: panoptic quality (PQ), segmentation quality (SQ) and recognition quality (RQ).

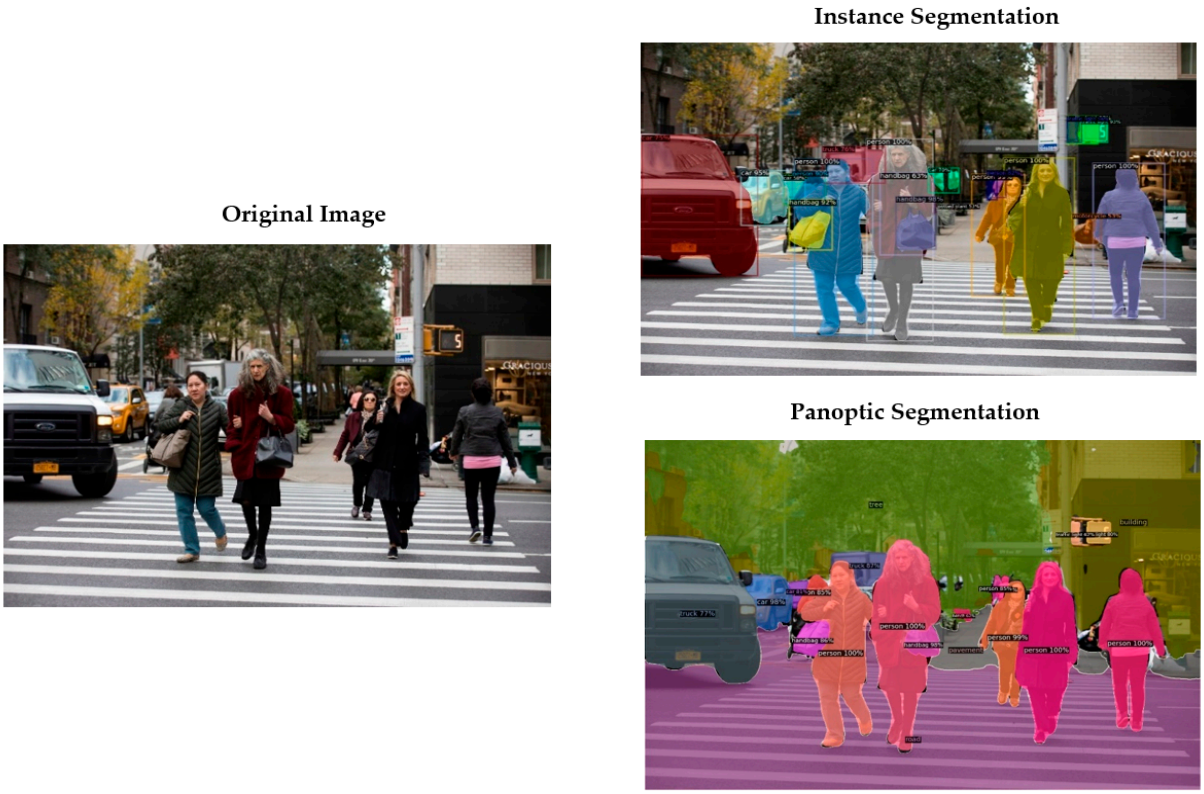

Figure 1. PS demonstrating the combined instance and semantic segmentation results. (Input image to the left is from Google search and the output to the right is extracted using Detectron2 [31]).

PQ measures the predicted PS quality relative to some ground truth. This involves tasks such as segment matching between the predicted segments and the ground truth. SQ is computed as the intersection over union (IoU) score (shown in equation below). RQ represents quality estimation in identification settings [32] and is the familiar F1 score [33]. It is assumed that both predicted and ground truth segments match only if the IoU exceeds a value of 0.5.

$$IoU = \frac{|Target \cap Predicted|}{|Target \cup Predicted|}$$

where, for a set A, $|A|$ denotes its size.

Panoptic quality (PQ): PQ is calculated independently for each class and an average over all the classes is obtained. The unique matching of each class divides the predicted and the ground truth segments into three categories namely: true-positives (TPs), false-positives (FPs), and false-negatives (FNs). These categories represent matched and unmatched predicted segments, and unmatched ground truth segments, respectively. PQ is calculated as

$$\mathbf{PQ} = \frac{\sum_{(p,g)\in TP} \mathbf{IoU}(p,g)}{|\ TP\ | + \frac{1}{2}|\ FP\ | + \frac{1}{2}\ |\ FN\ |}$$

where $\frac{1}{|TP|}\sum_{(p,g)\in TP} \mathbf{IoU}(p,g)$, is the average IoU of matched segments, and to penalize the segments that have no matches, $\left(\frac{1}{2}|FP| + \frac{1}{2}|FN|\right)$ is included in the denominator. It is important to know that regardless of their area, all the segments receive equal consideration. Also, on multiplying and dividing PQ proportional to the TP set size, PQ can be expressed as a product of SQ and RQ (SQ $\times$ RQ) as follows:

$$\mathbf{PQ} = \frac{\sum_{(p,g)\in TP} \mathbf{IoU}(p,g)}{|\ TP\ |} \times \frac{|\ TP\ |}{|\ TP\ | + \frac{1}{2}|\ FP\ | + \frac{1}{2}\ |\ FN\ |}$$

The first factor of the product is referred to as the segmented quality (SQ) whereas the second is referred to as the recognition quality (RQ), i.e., PQ = SQ × RQ. However, SQ and RQ are not independent as SQ is measured only over the segments matched. It is also interesting to consider the patches with void regions and group of instances [9].

3.2. Significant Advances in Panoptic Segmentation (PS) Achieved to Date

Research in PS is focused on evaluating the segmentation performance between "stuff" and "things" separately. To date, this field has witnessed several advancements, mainly by defining a unified metric that integrates both semantic and instance segmentation methods. Such unified metrics may help in achieving a PQ, improve the segmentation quality and recognition accuracy, thereby accurately differentiating the object instances within complex scenes, and semantically extracting the semantic context and relationships between the different objects and regions, therefore, enhancing the precision of segmented outputs. Several of the noteworthy achievements, include Panoptic Feature Pyramid Network (FPN) [34], Attention-guided unified network for PS [35], Seamless scene segmentation [36], Panoptic Deep lab [5], Unified PS network [7], and Efficient PS [37] are discussed in the following sub-sections.

3.2.1. Panoptic Feature Pyramid Network

The Feature Pyramid Network (FPN) aims to have a single network perform a unified prediction at an architectural level that combines the instance ("things") and semantic ("stuff") segmentation task by performing a shared computation. One of the most used instance segmentation methods named Mask R-CNN is combined with a branch of semantic segmentation using this shared FPN for improving the object detection and segmentation task. This further addresses the challenges of extracting relevant features at different scales by creating a feature pyramid that aids in extracting the features from distinct layers of a deep CNN.

As seen in Figure 2a, multi-scale features are extracted using the backbone model of FPN, used for object detection. In Figure 2b, to carry out instance segmentation, a region-based branch is used like the concept used in Mask R-CNN [38]. Subsequently, a lightweight dense pixel prediction branch is added in parallel, utilizing similar FPN features to perform semantic segmentation. Panoptic FPN is seen as a version of unifying the Mask R-CNN and FPN to ensure robust and precise segmentation and detection tasks.

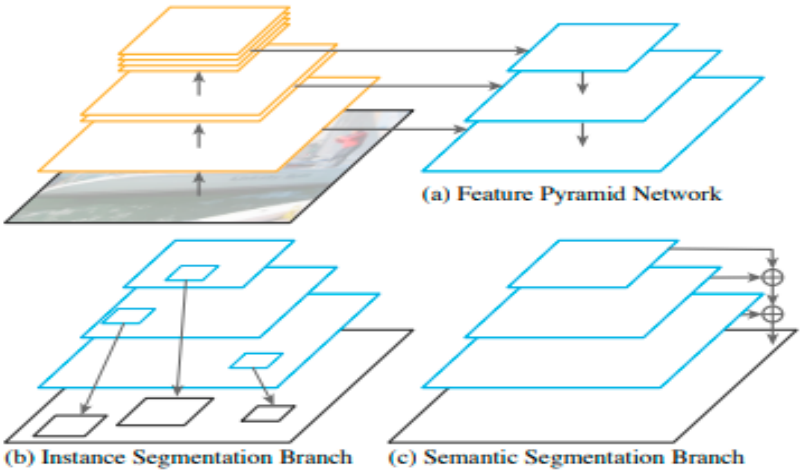

Figure 2. Panoptic Feature Pyramid Network (P-FPN) [34].

3.2.2. Attention-Guided Unified Network for Panoptic Segmentation

An attention-guided unified network is used for segmenting the foreground objects (instance level) and background objects (semantic level). Some works [35] emphasize extracting the cues (complimentary) from the foreground objects to achieve background

understanding. This network serves as an integrated framework with two branches, to segment the foreground and background simultaneously. A region proposal network (RPN) is added along with the foreground segmentation mask to the background branch to provide object-level and pixel level attentions. It is inferred that the network is evaluated on datasets such as MS-COCO, with PQ of 46.5%, and Cityscapes, with PQ of 59.0%, and achieves a uniform accuracy gain for both foreground and background segmentation, respectively.

The architecture seen in Figure 3 adopts panoptic FPN as a backbone and shares the relevant features in parallel with the three branches, namely, the foreground, background, and an RPN. It can be further viewed into two stages, namely, the training stage (where the network is fine-tuned in an end-to-end fashion) and inference stage (where panoptic results extracted by "things" and "stuff") stage. The PAM (proposal attention module) and the MAM (mask attention module) are the two models used for defining the complementary relation among the two branches.

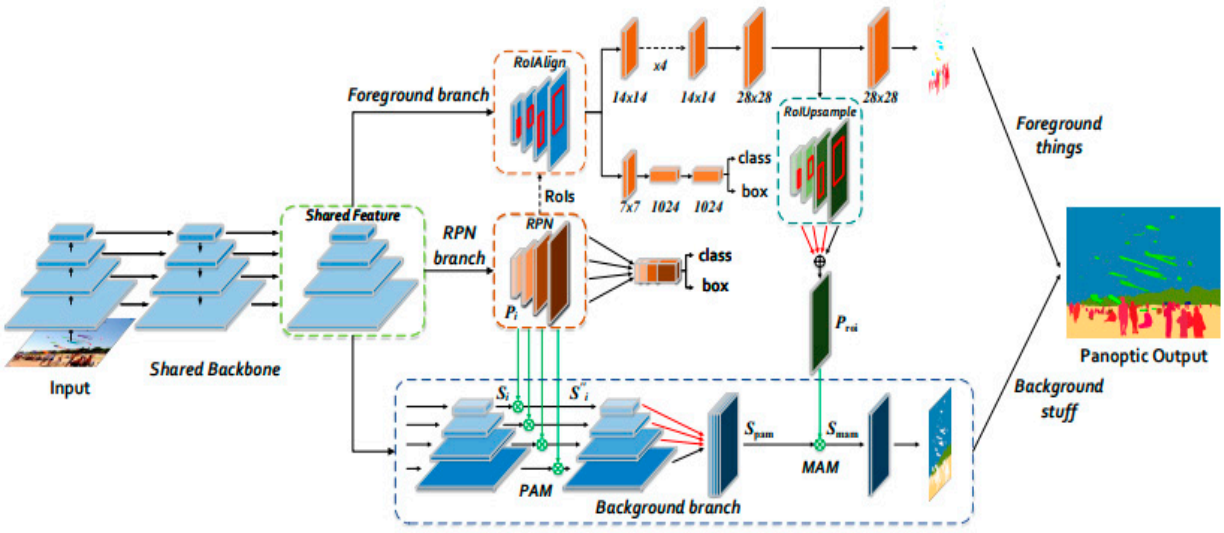

Figure 3. Attention-guided unified network architecture [35].

3.2.3. Panoptic DeepLab

Panoptic DeepLab [5] aims to address the semantic segmentation task based on deep learning by focusing on three important areas: (i) dense prediction tasks (using atrous convolution), (ii) atrous spatial pyramid pooling, and (iii) the localization of object boundaries. Atrous convolution is used as a convolution with up-sampled filters to perform dense prediction tasks and allows to explicitly control the resolution of different features computed with deep NNs. Atrous spatial pyramid pooling on the other hand, is used to segment the objects at multiple scales. An incoming convolutional feature layer with filters at different scales and varied sampling rates, are then used to generate a more effective field of view thereby capturing objects and the image context at multiple scales. To attain accurate localization of object boundaries, deep convolutional NN methods and probabilistic graphical models are then combined. An invariance is achieved with the combination of down sampling and max pooling but has an adverse effect on the localization accuracy. To overcome this, the responses at the last deep convolutional NN layer are combined with a fully connected Conditional Random Field (CRF). This way, the localization performance is improved both qualitatively and quantitatively.

Figure 4 illustrates the panoptic Deep Lab model that uses a deep convolutional NN (VGG-16 or a ResNet) for the semantic segmentation task by replacing all the fully connected layers by the convolutional layers and increasing the feature resolution using the atrous convolution and reducing the degree of signal down sampling from 32 pixels to 8 pixels. A bilinear interpolation is then used to up-sample the feature maps to the original

resolution of image. And to refine the segmentation result, a fully connected CRF is applied to capture the object boundaries.

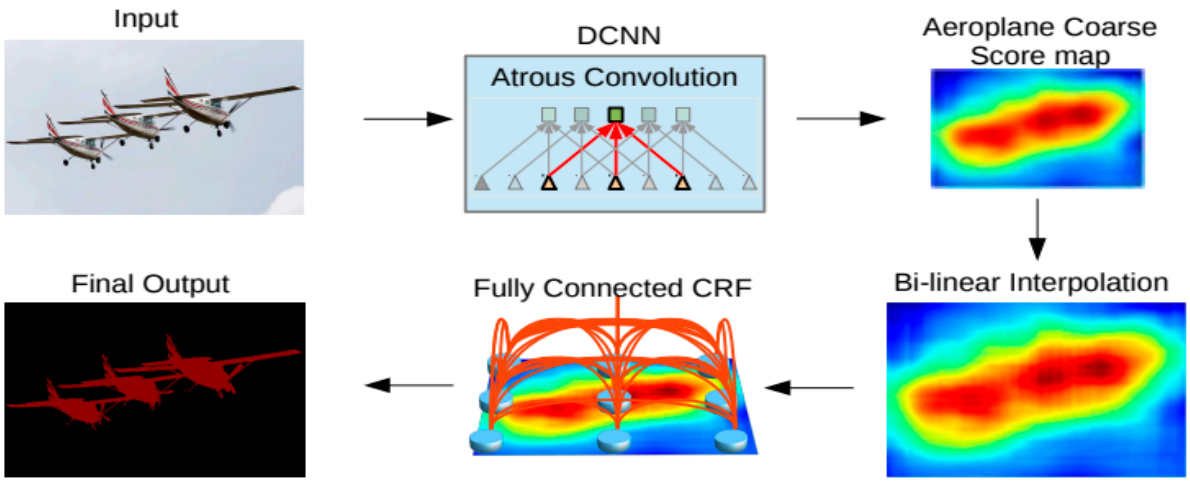

Figure 4. Illustration of the panoptic deep lab model [5].

3.2.4. Seamless Scene Segmentation

To cater for seamless scene segmentation, the architecture aims to use a convolutional NN that seamlessly integrates the multi-scale features generated by FPN with contextual information like a lightweight Deep Lab alike module. It overcomes the limitations of evaluating a non-instance category by learning a PQ metric only for the "things" classes and considers the "stuff" portion of an image as a unified instance. It is inferred that the predictions for "stuff" classes with the ground truth should not have an IoU > 0.5. With these parameters into consideration, the network is evaluated, and it seems to generate SOTA results on Cityscapes, Indian driving, and Mapillary Vistas datasets. It can be inferred that the novel CNN architecture proposed in [36], can generate a seamless scene segmentation output by jointly operating both semantic and instance segmented tasks on the top of a single network backbone.

3.2.5. Unified Panoptic Segmentation Network

The unified panoptic segmentation network is used to address panoptic segmentation tasks by designing a deformable convolution-based head that includes a semantic segmentation head with Mask R-CNN style instance segmentation head placed on the top of a sole backbone residual network. To resolve the semantic and instance segmentation, it expands the logic from the two heads and tries to come up with a representation to address the unknown class. The network aims to address the challenges posed by the variations in instances and allows backpropagation to the bottom modules in an end-to-end fashion. The experiment was carried out on a COCO and Cityscapes dataset and its performance is seen to draw faster inferences.

Figure 5 illustrates the architecture of a Unified Panoptic Segmentation Network that contains a backbone network with shared convolutional feature extraction with multiple heads stacked atop the network. Each head is seen as a sub-network that is designed for a specific task and leverages the features from the backbone. The Mask R-CNN serves as the backbone for feature extraction, employing a deep residual network (Res Net) with FPN.

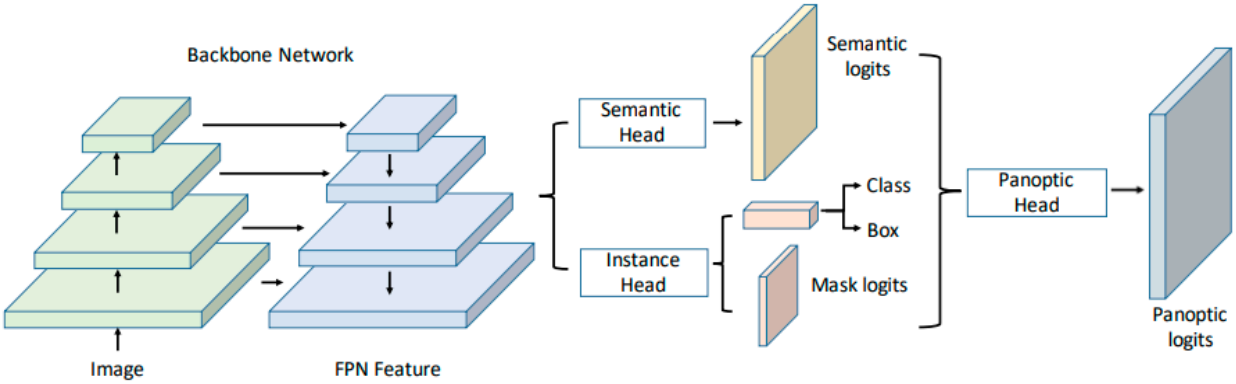

Figure 5. Architecture of a unified panoptic segmentation network [7].

3.2.6. Efficient Panoptic Segmentation

Comprehending a scene where an agent performs autonomous operation and whose performance is essential for ensuring its effective operation. It is important to understand and recognize the instances with general scene semantics to address the task of panoptic segmentation. An efficient panoptic segmentation architecture was designed to comprise a shared backbone and encode rich features at multiple scales. The paper [37] introduces the KITTI panoptic segmentation dataset [25] containing annotations for the KITTI challenge benchmark along with three other datasets, Mapillary Vistas, Cityscapes, and Indian Driving. A semantic head is used to aggregate fine, contextual features and an instance head that uses Mask R-CNN to achieve a seamless panoptic segmentation output.

Figure 6 illustrates the architecture of Efficient panoptic segmentation for PS. The architecture makes use of semantic prediction that contains a class, bounding-box, masks predictions. All these are then combined as input to the panoptic fusion step to get the resultant PS output.

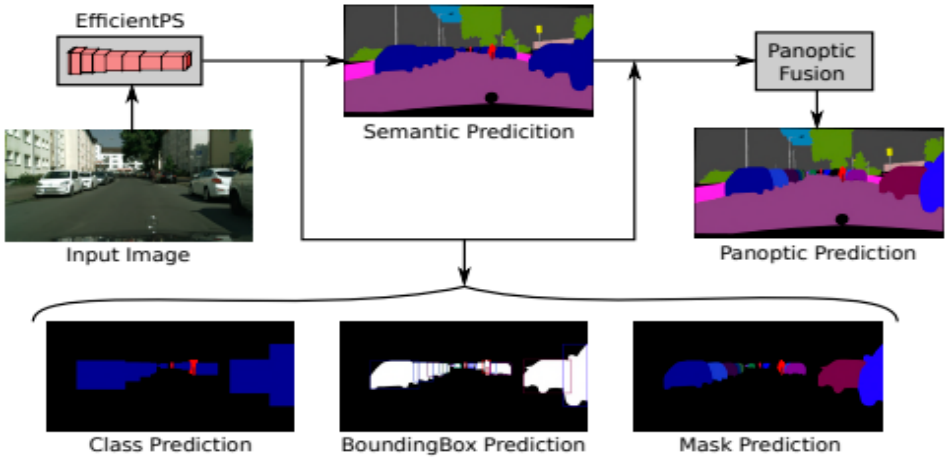

Figure 6. Efficient panoptic Segmentation Network architecture [37].

3.3. Top-Down Approach to Panoptic Segmentation

The PS is classified into top-down, bottom-up, single-path approaches, and other approaches [39]. But many deep learning methods follow the top-down approach, and this section elaborates the understanding of this approach. It is a simple approach of object detection and segmentation and is categorized into two stages namely one-stage and two-stage. In a one-stage approach, a one-stage detector is used to remove the proposals generation and make use of an anchor-free approach to perform object detection. The two-stage approach on the other hand, performs the proposals generation as its first step and then post-processing is done in the next step to achieve segmentation. Instance segmentation uses a Mask R-CNN [38] in this two-stage approach as seen in Figure 7.

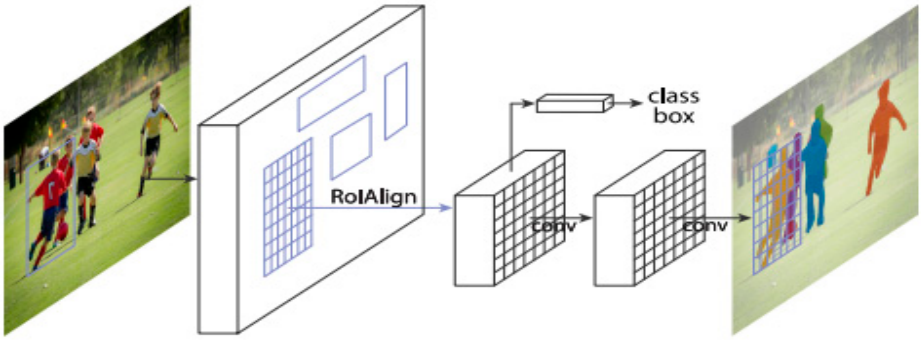

Figure 7. Mask R-CNN approach to perform Instance Segmentation [38].

The instance segmentation is built on a faster R-CNN, and it can be seen from Figure 7 that the ROI Pooling is substituted by ROI Align and minimizes the quantization errors. The technique is simple and straightforward, and based on the results of object detection by fast R-CNN, the Mask R-CNN utilizes a mask using smaller fully connected CNN to obtain the segmentation mask on a pixel-by-pixel basis along with the bounding boxes, output class, and mask logits. The issue lies not only in addressing conflicts between different branches such as the inconsistent class label predictions in both semantic and instance segmentation, but also within branches, addressing issues of overlapping and occlusion.

A single network is used by PS to combine the predictions of semantic and instance segmentation branches that are trained together using heuristics called the JSIS-Net (Joint semantic and instance segmentation) network [40]. This network architecture (seen in Figure 8) uses a feature extractor (ResNet-50 in this case) shared between the semantic and instance segmentation branch. Mask R-CNN is used by the instance segmentation to generate pixel clusters that are combined to obtain a normalized mask. However, a pyramid pooling module is used by the semantic segmentation branch to extract feature maps and to reshape the size prediction of the image input. The semantic branch's predicted 'thing' class is now substituted with the 'thing' class predicted from the instance branch. Subsequently, the network merges the predicted 'stuff' class from the semantic branch and the predicted 'thing' class from the instance branch to perform PS. This network architecture successfully resolves the differences between the semantic and instance segmentation branches via a post-processing module and overlooks the conflicts of intra-branch which is overlapping and occlusion.

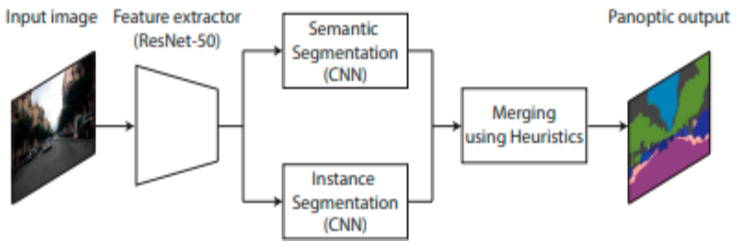

Figure 8. The Network architecture of Joint Semantic and Instance Segmentation [40].

3.3.1. Improvement in PS on Predicted Output Using Top-Down Approach

The conflicts that arise among the outputs of both semantic and instance segmentation branches should be studied comprehensively. It is seen that the top-down approach needs a refinement step to harmonize the results from the two branches seamlessly. But the conflicts are mainly noticed during the post-processing step as discussed in the previous section. The internal disagreements within the instance branches stem from overlapping and occlusion.

i. Overlapping and Occlusion

As already stated, in PS each pixel is labeled by the category to which it is assigned, and then segmentation is performed on each object instance. This is a challenging task as the current approaches make use of two independent models which do not share features and therefore the pipeline implementation is labor intensive. Since a heuristic approach is applied to merge the results, it is difficult to determine the extent of overlapping between object instances in the absence of sufficient contextual information at the time of merging. As instance segmentation is basically object detection, an overlap can be seen between the different instance masks being predicted. However, in the case of PS, a class label and an instance id are assigned to prevent this overlap. To avoid overlapping, non-maximum suppression (NMS) is utilized that primarily sorts the predicted segments with respect to its confidence scores. The bounding box with the highest scores is chosen and added to the list of outputs and is subsequently extracted from the candidate list. The IoU among all the candidate boxes is computed by the network and boxes whose IoU is greater than the threshold is discarded.

While resolving the overlapping issue, it is also important to view the instances sorted in descending fashion using detection scores and then placing the objects on some material canvas such that higher-score objects are placed higher. This approach can fail because of occlusion. To address this, a spatial sorting module named Occlusion Aware Network (OANet) [41] for PS is proposed at the post-processing step that can predict the 'stuff' and instance segmentation in a single network. This network maps the output of instance segmentation to a tensor; a large convolution is then applied to get a ranking score map. The ranking score map is optimized by using a cross-entropy loss and then OANet obtains the ranking score of each object instance and then executes spatial sorting. This way, the occlusion problem between the predicted instances is resolved.

Some improvements on PS include incorporating the information exchange module, incorporating certain methods based on attention, and improvements on the loss function to name a few, as seen in the literature.

4. Companies Developing Computer Vision Models for Panoptic Segmentation

Companies such as Apple, Facebook, and Tesla, have been in the forefront to come up with such computer vision system models that provide a comprehensive view of a bigger picture of panoptic segmentation. The research efforts by these companies up to this date include the following.

4.1. Apple

Understanding a range of scenes from an input image at a pixel-level (image segmentation) is central to any vision task and requires understanding of different features and segmenting them. Images captured from iPhones or iPads provide features to power the photographic styles that allow area adjustments and are guided by segmentation masks. The built-in features contain different image sharpening algorithms to render images of better quality and employ scene-level prediction of elements such as the roads, sky, buildings, etc., along with subject level prediction, for example, each person's body. As it is known, semantic segmentation provides a categorical label for each pixel, but lacks the ability to differentiate between various subject-level elements [42]. To this end, panoptic segmentation (PS) attempts to unify the scene and subject-level predictions, aiming to expand a range of predicted elements for a comprehensively parsed scene to larger number of categories.

i. On-Device Panoptic Segmentation model

In 2021, the machine learning research community at Apple was successful in designing an on-device PS model for enhancing their camera vision system on devices using transformers [42]. The idea was to divide each image into different segmentation masks, with a special focus on deriving the instance-aware segmentation masks (especially for

the persons category) presented as additional image channels. A neural architecture for PS was designed using transformers to be compatible to both the in-camera pipeline and offers the ability of achieving an efficient on-device execution with no impact on its battery life. To this end, a network was constructed that could run on the Apple Neural Engine (ANE). This ANE is an optimized co-processor designed for the energy-efficient execution of deep neural networks on Apple devices. For the execution of an intricate camera pipeline, containing multiple latency-sensitive workloads running concurrently to maximize the utilization of all available co-processors, they employed a single ANE segment. To achieve a high-resolution output image, a detection transformer (DETR) architecture was utilized that does not require post-processing and non-maximum suppression (NMS) for eliminating the anchor-based coordinate decoding and duplicate predictions. By doing so, DETR demonstrated high efficiency in assessing the regions of interest (ROIs) and utilized a two-stage approach. In the first stage, thousands of anchor-based ROIs are evaluated using Mask-RCNN. Subsequently, the top anchor-based proposals amounting to hundreds are forwarded to the next stage. The ROIs are constrained by an order of magnitude, typically set at 100 in the original DETR model and yet achieve a minimal degradation in detection performance for the target distribution of images of less than five people in the scene. The extension of DETR to the PS model introduces an additional convolutional decoder module batched along a sequence of dimension N, as seen in Figure 9.

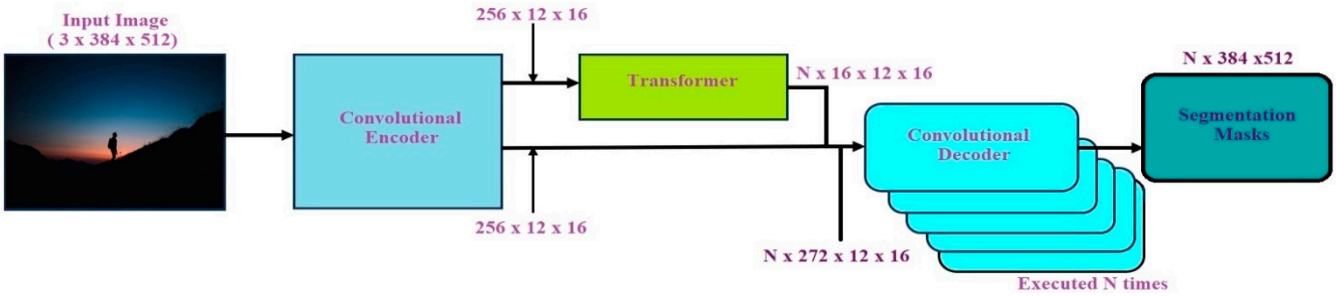

Figure 9. The DETR Network architecture (reproduced from [42]).

In the forward pass, each ROI generates a distinct segmentation mask. Its input comprises a unique set of feature maps generated from the transformer module, together with a shared set of feature maps generated from the convolutional encoder module. One of the performance bottlenecks in DETR is when many ROIs are processed together and when the output resolution is set to a relatively low value, causing the batched convolutional decoder module to become a performance bottleneck. The output resolution is set to as high as 384 $\times$ 512 to achieve a higher quality segmentation mask. To overcome this bottleneck of DETR at higher resolutions, and to scale the large number of object queries, the Hyper-DETR architecture shown in Figure 10 was proposed.

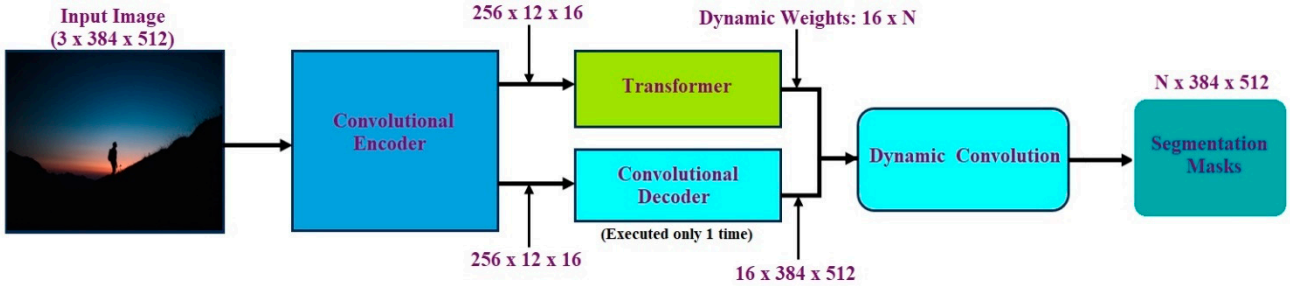

Figure 10. The Hyper-DETR Network architecture (reproduced from [42]).

The Hyper-DETR model integrates PS into DETR framework in an efficient way. Hyper-networks are considered as one of the meta-learning approaches and the tech-

nique works by decoupling the convolutional decoder–compute path from the transformer compute path. The outputs from the transformer module are then decoded as weight parameters of shape 16 × N and fed to the dynamic convolution layer that has a kernel size of 1 × 1. A higher-resolution feature map is generated by the convolutional 16 × 384 × 512 decoder, and finally the dynamic convolution layer linearly combines the two tensors into N unique 384 × 512 masks as the output of Hyper-DETR. The advantages of using Hyper-DETR are twofold. First, batching is not required to run a convolutional decoder, and this separates the intricacy of higher resolution mask synthesis from the length of ROI sequence. Second, categories of scene-level, for example, sky, are managed using the convolutional path, by skipping the transformer module execution in situations where subject-level attributes are not being sought. The primary limitation is the multitude of output channels that are not located at statically determined indexes. This way the Hyper-DETR PS architecture achieves a magnitude of higher output resolutions with a higher number of region proposals, thereby enabling the understanding of camera at pixel level and supporting a range of features without impacting the battery life.

ii. Fast Class-Agnostic salient-object segmentation

In 2023, Apple announced the launching of live stickers that support devices with subject lifting features [43]. For example, in Photo Apps, the subject lifting model executes only with user interactions such as touch and hold on a photo subject. To facilitate faster on-device segmentation and seamless integration of services, the model should exhibit an extremely low latency. Here, the image source is resized to 512 × 512 and subsequently fed as input to a convolutional encoder based on Efficient Net v2. The features extracted at different scales are subsequently fused and up sampled through a convolutional decoder. Two additional branches emerge from the terminal feature of the encoder: one branch that predicts an affine channel-wise reweighting for the decoded channels like the dynamic convolutional branch in [42]; and the other branch that predicts a scalar confidence score, estimating the likelihood of a salient foreground in the scene and is used for gating the segmentation output. The final output prediction is a single-channel alpha matte with the input resolution of 5 × 1512 as seen in Figure 11. The run-time on an iPhone 14 device is less than 10 milliseconds and on the older devices, the network runs on the GPU.

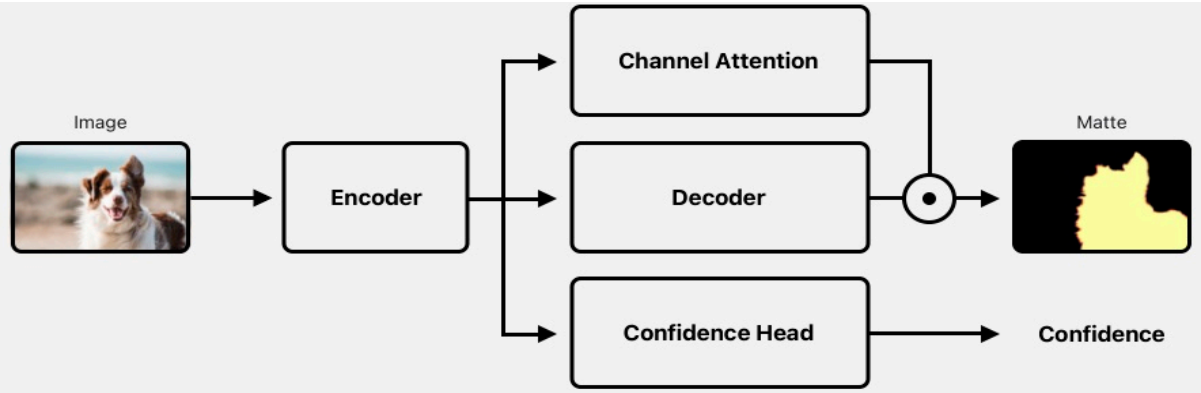

Figure 11. Subject lifting network architecture [43].

4.2. Facebook

Facebook AI Research (FAIR) released Detectron2 as part of their next-generation library that offers state-of-the-art models for computer vision tasks [31]. It is an open-source framework for object detection and segmentation framework built on top of PyTorch and provides a unified API for performing a variety of tasks such as object detection, instance segmentation, and panoptic segmentation. It provides high-quality implementations of the state-of-the-art algorithms like Mask-RCNN, Densepose estimation (see Figure 12, where the input image to the left is from Google search and the output result to the right

is extracted using Detectron2 [31]), RetinaNet, and rotated bounding boxes, to name a few, and is designed to be flexible, easy to use, and supports several research projects that have focus on enabling rapid research as it trains faster. It is considered as the successor to Detectron and the Mask-RCNN benchmark. Apart from this, it includes a model zoo with models for object detection, instance segmentation and many others.

Figure 12. Dense pose estimation of objects detected with prediction scores. (Input image to the left is from Google search and the output to the right is extracted using Detectron2 [31]).

Detectron2 uses a two-stage approach to object detection: the first makes use of the region proposal network (RPN) to generate a set of candidate regions. In the second stage, it uses a Mask R-CNN model to classify and segment the candidate regions. The RPN is a CNN that takes an input image and generates a set of candidate regions and is trained to predict a score for each identified object/region, indicating the likelihood of that object belonging to a certain category, and predicts a bounding box for each category (as seen in Figure 1 where the input image to the left is from Google search and the output to the right is extracted using Detectron2 [31]). The repository for Detectron2 has open-source model weights for algorithms such as instance segmentation, mask-RCNN, panoptic segmentation (PS), and dense pose estimation.

For PS, Detectron2 has been trained on COCO (common object in context) dataset that has been widely used particularly for visual detection tasks that constitute scene understanding. For training PS, Detectron2 uses 118 K images for training and 5 K images for testing, as the annotations in PS focus on "stuff" and "things" as discussed in earlier sections. Using this COCO dataset, 80 classes for "things" such as a person, an umbrella, a bicycle, etc., and 91 classes for "stuff" such as sky, road, pavements, etc., can be detected [44].

It is seen that PS can be performed only on a handful of datasets, like COCO, Cityscapes, Indian driving, Mapillary vistas, and ADE20k, that have required annotations. Moving forward, the next step in the process is to have an ecosystem where many different datasets are annotated, and PS could then become capable of addressing a wide spectrum of use cases. Training a PS on local machine without GPU support would hinder the performance as there are large images to train. To address this, it is important that the computer vision industry come up with some ways that are easier to implement, support custom datasets, and provide ways to expand the different use cases across multiple domains.

4.3. Tesla

In a 30 November 2021 tweet on the PS project, Andrej Karpathy, Senior Director of AI at Tesla, stated that "for autonomous driving vehicles, it is important be aware of the objects around the vehicle and on what surface it is driving on to navigate on the streets safely". This means that Tesla is very close to developing such a vision system where self-driving vehicles can identify both the roads and the relevant objects around it (see Figure 13).

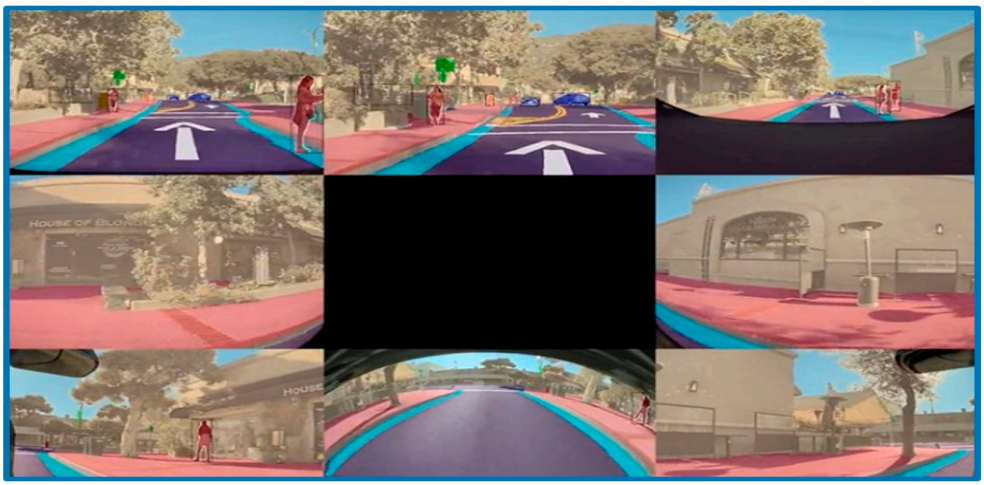

Figure 13. Autonomous driving using panoptic segmentation (Andrej Karpathy) [45].

Tesla has been on the verge of developing full scale, completely autonomous self-driving cars. To achieve this, it is necessary that the driverless vehicle consists of sensors for collecting real-world data and by using this dataset collected, it trains the neural networks that can support the auto-pilot features that enhance the capability of self-driving cars. It is important to have the labeled data: images collected are tagged with information such as people, vehicles, lanes, street signs, etc. If the images are labeled properly, then these images can be fed to the neural net vision system to perform recognition. The auto-pilot team at Tesla focuses on labeling the data. The latest project of Tesla's panoptic segmentation (PS) use would enhance the self-driving capabilities to next level and may attain level 4 and level 5 automation capabilities [45].

The PS results generated from the perception system of self-driving cars can be used by the planning and control modules to take informed driving decisions much better. For example, the detailed object shape and silhouette information may help improve object tracking, thereby resulting in a more accurate input for, steering, acceleration and many more tasks.

5. Publicly Available Datasets and Benchmarks

With the availability of large publicly available datasets that contain thousands of images with ground truth labels, it becomes easier for the models to learn from this huge collection of data. Some of the publicly available datasets that support PS and scene recognition are presented.

5.1. Datasets for Panoptic Segmentation

i. COCO 2020 Panoptic Segmentation

To push the SOTA in achieving a coherent scene segmentation, the COCO PS aims at unifying the semantic and instance segmentation tasks to address the needs of the current real-world computer vision systems (augmented reality, self-driving vehicles, etc.). As discussed earlier, "things" refers to as countable objects (animals, people, tools, etc.) and "stuff" refers to regions that belong to the same texture or material (sky, road, grass, pavements, etc.). Earlier Microsoft COCO models evaluated these two tasks separately. COCO PS works by assigning a semantic label and an instance id to each pixel in an image, thereby achieving a dense coherent scene segmentation. This PS task has been a part of the Joint COCO and LVIS recognition challenge workshop at ECCV 2020 [46]. PS utilizes all the annotated 123 k COCO images that are divided into 172 classes, of which 80 belong to the "thing" category and 91 to the "stuff" category. The panoptic quality (PQ) is used to evaluate the performance of the model.

ii. Cityscapes Panoptic–Semantic Labeling

Dataset of urban street scenes are captured by a vehicle in 50 German cities [47]. It is widely used to evaluate semantic, instance, and panoptic segmentation tasks. It comprises of around 5000 diverse frames with a high-quality pixel-level annotation. Together it constitutes 30 classes of which 10 classes belong to "thing" category (person, car, animals, etc.) and 20 classes belong to "stuff" category (sky, grass, and ground), respectively.

iii. BDD100K Panoptic Segmentation

BDD is a largest driving video dataset captured from various cities in the United States and contains 100 K videos that have pixel wise annotations and around 10 tasks for evaluating the exciting progress of image recognition algorithms in context to autonomous driving [48]. Together it constitutes 40 classes of which 10 classes belong to "thing" category (particularly non-stationary objects) and 30 classes belong to "stuff" category, respectively. This dataset takes into account environmental factors such as the weather, geographic diversity, and is useful for training models such these are less likely to be surprised by new conditions.

iv. Mapillary Vistas v2.0

Mapillary Vistas v2.0 supports semantic, instance, and panoptic segmentation and provides street-level images captured from six different continents. This dataset contains around 25,000 high resolution images with 124 semantic object categories, among which 70 belong to "thing" category and 54 belong to the "stuff" category [49]. Additionally, it captures various factors that include seasons, weather, time of the day, camera, and viewpoint. Moreover, the annotations consist of polygons and not bitmaps.

It is seen that this dataset earlier contained 65 classes in totality, among which 28 were from "stuff" classes and 37 from "thing" classes, respectively [50]. The dataset was further divided into three sets namely, the training, validation, and testing sets, and moreover, 18,000, 2000, and 5000 images were referred to as the size belonging to each category.

v. Semantic KITTI Panoptic Segmentation

Semantic KITTI is a 3D point cloud PS dataset that contain street scenes captured with LiDAR sensor or a stereo camera from Karlsruhe, Germany [51], with a 360 degrees field view. It consists of 11 driving sequences with PS labels and these labels use six "thing" classes and 16 "stuff" class categories.

vi. ScanNet

ScanNet is an RGB-D video dataset that contains 2.5 million views in 1513 scans, annotated with 3D camera poses, surface reconstructions, and instance-level semantic segmentations [52]. It uses 38 "thing" classes (items and furniture in rooms) and two "stuff" classes (such as walls and floor). Though it is not yet completely developed it covers around 90% of all surfaces.

vii. nuScenes

The nuScenes dataset is a large-scale dataset inspired by KITTI dataset for autonomous driving developed by Motional which was formerly known as nuTonomy [53]. It provides 1000 driving scenes from cities with dense traffic particularly from Singapore and Boston. To show an interesting set of traffic situations, driving maneuvers, and unexpected incidents, they extract scenes of 20 s length. For object detection and tracking, Motional provides 23 annotations for "thing" class and nine for the "stuff" classes and provides accurate 3D bounding boxes at 2 Hz. In 2019, this dataset was first released with 1000 scenes extracted from the sensor suite of an autonomous vehicle with one LiDAR, six cameras, five Radar, GPS. Comparatively, nuScenes has seven times more object annotations than the KIITI dataset. In the year 2020, Motional released nuScenes-lidarseg, where each lidar point from a keyframe present in nuScenes is annotated with one of 32 possible semantic labels. This way nuScenes-lidarseg provides 1.4 billion annotated points across 40,000-point clouds and 1000 scenes, i.e., 850 scenes for training and validation, and 150 scenes for testing.

5.2. Comparison of Various Panoptic Segmentation Models

Table 1 illustrates the performance of various panoptic segmentation models using different backbone networks found in the literature. On the Cityscapes dataset, the panoptic FPN comparison is performed with DIN [54]. Panoptic FPN surpasses DIN with a PQ margin of 4.3 points. DIN (Detect-to-Instance Network) is a substitute for region-based instance segmentation, initiating with pixelwise SS and performs grouping to retrieve instances [34]. Likewise, AUNet is compared with leading bottom-up methods (such as DWT [55], SGN [8]) and Mask R-CNN achieving a persistent improvement in accuracy with MS-COCO, and hence, serve as a new approach. The design of Panoptic DeepLab is straightforward and requires only three loss functions while training as discussed in [5], and incorporates minimal parameters to a contemporary semantic segmentation model. Note: the results of each performance metric presented in Table 1 are the results extracted from the respective models presented in [5,34–36].

Table 1. Performance of various PS models on different datasets found in the literature.

Models	Dataset	Back-Bone	PQ	PQ^{ST}	PQ^{TH}	Comparison
Panoptic Feature Pyramid Network (FPN) [34]	COCO	R50-FPN×2	39.2	27.9	46.6	On Cityscapes, the panoptic FPN comparison is performed with DIN [54] and it is inferred that panoptic FPN surpasses DIN with a 4.3-point PQ margin. Note: DIN is a substitute to region-based instance segmentation. It commences with pixelwise semantic segmentation and subsequently performs grouping to retrieve instances.
		R50-FPN	39.0	28.7	45.9	
		R101-FPN	**40.3**	**29.5**	**47.5**	
	Cityscapes	R50-FPN×2	57.7	62.4	51.3	
		R50-FPN	57.7	62.2	51.6	
		R101-FPN	**58.1**	**62.5**	**52.0**	
Attention-guided unified network for Panoptic Segmentation (AUNet) [35]	COCO	ResNet-101-FPN	45.2	31.3	54.4	AUNet is compared with the leading bottom-up methods (such as DWT [55], SGN [8]) and Mask R-CNN. It is inferred that a consistent accuracy gain is achieved with MS-COCO, and thereby a new state-of-the-art can be further achieved.
		ResNet-152-FPN	45.5	31.6	54.7	
		ResNeXt-152-FPN	**46.5**	**32.5**	**55.8**	
	Cityscapes	ResNet-50-FPN	55.0	57.8	51.2	
		ResNet-50-FPN	56.4	59.0	52.7	
		ResNet-101-FPN	**59.0**	**62.1**	**54.8**	
Panoptic DeepLab [5]	Cityscapes	VGG-16 based LargeFOV	40.3	49.3	33.5	The design of Panoptic DeepLab is simple and requires only three loss functions while training and incorporates minimal parameters to a contemporary semantic segmentation model.
	Mapillary Vistas	ResNet-101	65.5	-	-	
Seamless Scene Segmentation [36]	Cityscapes	ResNet-50-FPN	59.8	64.5	53.4	An effort to attain seamless scene segmentation involves the integration of semantic and instance segmentation methods, jointly operating on a sole network backbone.
	Mapillary Vistas	ResNet-101-FPN	37.2	42.5	33.2	
Unified panoptic segmentation network UPSNet [7]	COCO	ResNet-101-FPN	46.6	36.7	53.2	Three large datasets are used whose empirical results demonstrate that UPSNet attains SOTA performance with faster inference in comparison to other models.
	Cityscapes	ResNet-101-FPN	61.8	64.8	57.6	
	UPSNet dataset: MR-CNN-PSP	ResNet-50-FPN	47.1	49.0	43.8	
Efficient panoptic segmentation EPSNet [56]	COCO	ResNet-101-FPN	38.9	31.0	44.1	A one stage EPSNet is presented and achieves a significant performance on COCO dataset and outperforms other one stage approaches. Hence, EPSNet is notably faster than other existing PS networks.

To achieve seamless scene segmentation [36], semantic and instance segmentation methods are combined and jointly operate on top of a single network backbone. UPSNet [7] makes use of three large datasets such as COCO, Cityscapes, and UPSNet, whose empirical results illustrate that UPSNet attains SOTA performance with quicker inference in comparison to other models [57]. On the other hand, a one-stage EPSNet [37] was presented that achieves significant performance on COCO dataset and outperforms other one-stage approaches. Therefore, EPSNet is significantly faster than other existing PS networks.

5.3. Datasets for Scene Recognition

This section presents datasets for scene recognition from [11], and their associated benchmarks, as illustrated in Table 2. These datasets are a combination of widely used datasets and some of them are new. A comparison of the recognition accuracies among some of the representative algorithms is studied, followed by a comprehensive analysis.

Table 2. Comparison of accuracies of some of the scene recognition approaches on four different datasets from [11].

Scene Recognition Types	Method	Feature Retrieval	Scene-15	SUN-397	Indoor-67	Sports-8
Global Attribute Descriptors	GIST [13]	GIST	73.28	-	-	82.60
	LDBP [58]	LDBP	84.10	-	-	88.10
	mCENTRIST [57]	mCENTRIST	-	-	44.60	86.50
	CENTRIST [59]	CENTRIST	83.88	-	-	86.22
Patch Feature Encoding	SPM [60]	SIFT	81.40	-	34.40	81.80
	DUCA [61]	AlexNet	94.50	-	71.80	98.70
	HIK [62]	CENTRIST	84.12	-	-	84.21
	MOP-CNN [63]	AlexNet	-	51.98	68.88	-
	LScSPM [64]	SIFT	89.75	-	-	85.31
	NNSD [65]	ResNet-152	94.70	64.78	85.40	**99.10**
	LR-Sc+ SPM [66]	SIFT	90.03	-	-	86.69
Spatial Layouts Pattern Learning	RSP [67]	SIFT	88.10	-	-	79.60
	S2ICA [68]	VGG-16	93.10	-	74.40	95.80
	RS-Pooling [69]	AlexNet	89.40	-	62.00	-
Discriminative Region Detection	Object Bank [70]	Object Filters	80.90	-	37.60	76.30
	DSFL [71]	AlexNet	92.81	-	76.23	96.78
	VS-CNN [72]	AlexNet	**97.65**	43.14	80.37	97.50
	ISPRs [73]	HOG	91.06	-	68.50	92.08
Object Correlation Analysis	SDO [50]	VGG-16	95.88	73.41	86.76	-
	MetaObject-CNN [74]	Hybrid CNN	-	58.11	78.90	-
Hybrid Deep Models	DAG-CNN [75]	VGG-19	92.90	56.20	77.50	-
	Dual CNN-DL [76]	Hybrid CNN	96.03	70.13	86.43	-
	FOSNet [77]	SE-ResNeXt-101	-	**77.28**	**90.37**	-
	Hybrid CNNs [78]	VGG-19	-	64.53	82.24	-

i. Scene-15

This dataset contains 4485 gray images from 15 different categories that include both indoor and natural scenes [60]. The image sizes are relatively small, with 200 to 400 images per category. It is crucial to highlight that the recognition accuracy of certain CNN algorithms may be diminished for images lacking color information. In the absence of distinct training and test sets, random images, typically around 100 per category, are chosen for training, while the remaining images are employed for testing. For better evaluation results, it is suggested to perform random splits several times.

ii. SUN-397

SUN is an acronym for scene understanding. This dataset comprises 397 distinct scene categories and 108,754 color images, with at least 100 images per category. These categories contain various indoor and outdoor scenes with larger objects and alignment variance and hence impose huge complexity for scene recognition tasks. One hundred images per category are chosen as the standard protocol [79], fifty of which are used for training with the remaining fifty images used for testing.

iii. MIT Indoor-67 scenes

This dataset contains 15,620 color images distributed across 67 categories [80]. This is particularly complex as the indoor scenes suffer from huge intraclass variation, with confusing indoor scenes having similar backgrounds and sharing repeated objects.

iv. UIUC Sports-8

The UIUC dataset contains 1572 color images distributed over eight distinct categories covering various scenes of sports events [81], with 130 to 250 images per category. These are high-resolution images, i.e., from 800 $\times$ 600 to thousands of pixels per dimension; 70 images are randomly sampled for training with the remaining images used for testing each category.

v. Places

This dataset comprises around 2.5 million scene images [14] and contains an evaluation criterion based on top-5 error. Considering the case with the SUN-397 dataset, that provides scene images in good numbers, but each category still suffers from sufficient data required for feeding deep learning models. The Places-205 dataset, on the other hand, contains 205 scene categories that are similar with each category containing at least 5000 images for training; 100 and 200 images belonging to each category are utilized for validation and testing. The Places2 dataset [82] is an extension of the Places dataset and combines around 10 million images with scene categories containing more than 400. To date, this dataset seems to be probably more challenging to carry out scene recognition as it is based on the current occurrences of scenes. It is inferred that more than 4000 images are selected for each class across 365 categories for coming up with datasets like Places365-Standard and Places365-Challenge.

Table 2, drawn from [11], illustrates the accuracies comparison of some of the scene recognition algorithms with respect to feature extraction. These algorithms are grouped into six categories, as discussed in Section 2. Global attribute descriptors lead to worst recognition accuracy as these descriptors are obtained without training and by some predefined numerical calculation and are not likely to be suggested for the present scene recognition applications. In recent years, there has been extensive exploration of patch feature encoding, driven by its recognition accuracy, which shows similarity with methods such as patch features and codebook learning. As seen in Table 2, the features retrieved from patches achieve a recognition accuracy that is higher than the handcrafted features. However, advanced codebook learning methods may lead to better outputs due to their inherent relationships. An advantage of patch feature encoding is that these algorithms are trained to deal with cluttered backgrounds and with deformed objects in images to a certain degree. The sports-8 dataset achieves the highest recognition accuracy of 99.10% when compared to the other datasets. Patch feature encoding can be applied in situations where computational resources are limited with limited scene categories, and where the response time is of paramount importance compared to the accuracy of scene recognition. Spatial layout pattern (SLP) learning improves the recognition accuracy of the scene, but an excessive number of spatial partitions can negatively impact scene recognition accuracy. Consequently, while SLP learning proves effective for stable indoor and outdoor scenes, it may cause confusion in recognizing indoor scenes (that are highly lookalike) with similar spatial layouts.

These algorithms achieve moderate recognition accuracy with some minor changes to the existing CNNs and require additional computations in comparison to the current frameworks and utilize minimal inference time. Discriminative region detection addresses the shortcomings of extreme spatial partitions in context to SLP learning and focuses on the detection of regions of interest (RoIs) in complex scene categories and achieves better recognition results as seen in Table 2. To detect the RoIs, some algorithms make use of pretrained object detectors such as Object Bank [70], where the training process is more arduous and consumes a lot of time and increases the computational overhead. Also, the detection of discriminative regions is more sensitive to scene categories, and it is obvious

that for larger datasets there are larger objects with many different categories. In general, these kinds of discriminative detection algorithms achieve recognition accuracies that are higher even on smaller and intermediate datasets in a reasonably shorter time frames. Among all scene approaches to scene recognition, object correlation is more complex than object detection (identification of discriminative patches) is the first step for the analysis of subsequent correlations and require some real-time region proposal techniques. To this end, various probabilistic models have been developed to examine the connections among various objects and categories of scenes. Another disadvantage is that errors in detected objects or patches will have an impact on the subsequent correlations and hence, the recognition accuracy does not just relate to the object correlation but relies either on regional proposal techniques or the object detection model. Object correlation analysis achieves moderate accuracy in context to scene recognition task with slowest inference speeds due to computational overhead and serve as an optional choice for a purpose at hand. Hybrid deep models can achieve higher accuracies on massive scale datasets and combines the expressive powers of feature convolutions and the various methods to feature encoding such as CNN-DL [76], Hybrid CNNs [78], and VSAD [83], to name a few, that results in arduous training procedures, longer inference times and huge computational cost. Some tailored networks, for example, DAG-CNN [75] and FOSNET [77], that are end-to-end, attempt to simplify the training process by unifying extra information into the architecture. These hybrid deep models consume minimal computations and inference times to obtain a satisfactory recognition accuracy as they are favored with adequate computational resources.

Limitations and Challenges of Scene Recognition Algorithms

This section discusses the limitations and challenges of the scene recognition algorithms, namely, (i) global attribute descriptors (GADs), (ii) patch feature encoding (PFE), (iii) a spatial layout pattern learning (SPL), (iv) discriminative region detection (DRD), (v) object correlation analysis (OCA), and (vi) hybrid deep models (HDM) [11].

i. In early 2000, scene extractions from images were mainly carried out using the GAD's. These descriptors utilized low-level image features such as the semantic typicality (this measure groups the natural real-world images in terms of their similarity into six different scene categories that include forests, coasts, rivers/lakes, plains, sky/clouds, and mountains, and categorizes a given image into one of those categories along side the nine local semantic concepts based on the frequency of image occurrence. Here, a archetypal categorial form of representation is learnt from each scene category and the "typicality measure" proposed is further evaluated (qualitatively and quantitatively) by incorporating a cross-validation on images containing 700 natural scenes. Furthermore, as typicality is a measure of uncertainty of predictions based on given annotations, and the nature of real-world images resembling an obscure nature, it is imperative to pay attention to the modeling of scene typicality after carrying out manual annotations [84]); a GIST (which is an abstract representation of a scene for activating the memory representations of different scene categories, such as sky, city, mountains, etc.); a census-transform histogram (CENTRIST) [59] (a visual descriptor for identifying the scene categories or the topological places by encoding the structural properties in an image and by suppressing the detailed textual information. It is inferred that the model proved to be successful for both datasets related to scene categories and the topological places and has been noticeably faster); etc. These GAD's saw limited performance in scale for understanding the visual scene representations that are complex in nature.
ii. To improve the performance, PFE gained prominence in the research community. It made use of the local features (aka. local visual descriptors), for example, histogram of oriented gradients, scale-invariant feature transform [85], bag-of-visual words, and local binary patterns, to name a few. Researchers utilized the bag-of-visual words framework before deep learning took the center stage and comprised of three

different modules such as (i) feature extraction, (ii) code book learning, and (iii) coding processing. For any given image, the local features are extracted and are propagated to the code book learning module for extracting the visual words. This module uses k-Means clustering and extracts the k clusters by dividing the visual descriptors resembling the local features in terms of their Euclidean distances. Each cluster obtained represents a group of visual descriptors that share similar features whose center point is considered as the distinct visual word. This way all clusters containing the visual-words forms a code book. Finally, by incorporating all the learned features, the coding processing module predicts the contents of the entire image.

In PFE, the codebook structure has crucial implications on the scene recognition performace. One possible limitation of codebook learning as discussed in [11] relies on the dictionary size meaning, the amount of visual words belonging to each category. The codebook learning becomes exorbitant considering a humungous amount of scene images that it is required to deal with. Additionally, it is inferred that the dimensionality of the derived codes increases manifold with the increase in the number of scene categories, resulting in more complexity and very slow inference processes. One such solution is to learn a codebook that is compact while simultaneously maintaining a higher recognition accuracy. This can be possible by getting rid of the correlated words using an indicator function. By using an automatic compact dictionary learning (ACDL) [86], the size of the codebook can be reduced. Besides this, constraints such as selectivity, sparsity, and discrimination may be incorporated to assure certain specific characteristics from the derived codes. The authors in [87] present a comprehensive analysis of the various codebook learning methods by emphasizing on the main characteristics of those methods. However, choosing an appropriate characteristic depends on several factors, which need further research and exploration.

iii. The SPL pattern learning aims at increasing the scene recognition accuracy as some scenes may have certain specific spatial layouts. One such spatial layout utilizes the randomized spatial partitions [67] by considering both classification and optimal spatial partition as one single problem. Here, an input image is partitioned into a pool of several partitions, each representing a different size and shape. This is further transformed into a histogram based representation of features consisting of an ordered pair $p(I_i, \theta_j)$ where I_i represents the level and θ_j represents the partitioned patterns. Another spatial layout presented in [88] make use of class-specific spatial layouts that are obtained from spatial partitions based on the convolutional-feature maps. There have been several customized modules found in the literature that support various spatial structures, for example, randomized spatial pooling [69] and spatial pyramid pooling [89], to name a few.

To examine the spatial structures that are more flexible, there have been several research efforts found in the literature. One such effort is the use of a randomized spatial pooling layer proposed in [90] that embodies appropriate spatial layout information to the CNN.

iv. DRD is another way of independently extracting important regions or objects from the scenes. This is performed by using models such as deformable part based [90], and Object bank [70] to obtain the discriminative regions. However, to reduce the noisy features, important spatial pooling regions ISPR's are used in identifying and locating the discriminative regions. It is noted that ISPRs make use of part filters to preserve the quality of image regions.

DRD attempts aims to address the problems caused by the pooling regions in SPL for example, loss of certain salient features caused during the division of regions into several fragments. DRD focuses on the regions of interest. Using the deformable part-based models [90] with some latent SVM training, aids in discovering the common visual structures that helps in capturing the continual visual elements alongside salient objects. The Object bank [70], on the other hand, sits atop the response maps and contains several object-sensing filters that are pretrained on generic-labeled objects and integrates the local

semantic meanings into a complete image representation. The increase in the number of detected objects from scene images, results in an increase in the dimensionality of the response vector, and therefore, a regularized logistic regression can be used for activating several instances belonging to each class.

v. OCA models the relationship among the diverse assignment of objects and scene categories and is considered the most challenging tasks among several scene recognition approaches. Here, the discriminative patch identification serve as the first step in carrying out the subsequent correlation analysis and this depends either on the pre-trained object detectors or on other practical region proposal methods. To analyze and understand the relationship between the diverse assignment of objects and scene categories, several probability models have been introduced in the literature. With the humungous information, OCA achieves a moderate recognition accuracy, and is considered the slowest in terms of the inference speed because of heavy computational load. It is to be noted that OCA can serve as an alternative in situations when object detection is needed for the task.

vi. HDMs are considered as effective approaches to scene recognition. The intermediate layers in the CNN, on the one hand, capture the local features whereas the top layers capture the holistic features. In the end-to-end networks, multi-stage convolutional features should be considered for example the DAG-CNNs (directed acyclic graph CNNs) [75].

In general, features from layers close by contains unwanted information that is correlated, and hinders the scene recognition performance. These hybrid models though exhibit considerable advantages, but suffer from similar issues as with codebook learning. Also, codebook learning and deep models together consume huge memory and computational resources. Therefore, by combining the codebook approaches (existing) with deep networks serve as a potential solution and is considered an open issue.

6. Conclusions and Research Directions

This article presents a review of the scene recognition approaches and the role of PS as a breakthrough approach towards computer vision. With respect to feature extraction, the performance of the six categories of scene recognition algorithms such as global attribute descriptors, patch feature encoding, spatial layout pattern learning, discriminative region detection, object correlation analysis, and hybrid deep models, and the potential issues concerning it, are discussed [11]. Although multi-scale ensembles have proven to be effective in improving the recognition performance, they require more computational resources and are still task dependent. Also, many hybrid-deep learning models have been seen to emerge successful in scene recognition.

Panoptic segmentation that combines both semantic and instance segmentation is presented. Some of the important advances to panoptic segmentation along with some noteworthy research such as the Panoptic FPN, Attention-guided unified network for PS, Seamless Scene Segmentation, Panoptic Deep lab, Unified PS network, and Efficient PS is discussed. Of these models, some evaluate semantic and instance segmentation as separate entities and combine the joint results to produce panoptic segmentation, while some evaluate this as one unified model for PS. The PQ on the COCO and Cityscapes datasets (around 48% and 60% in each case) were seen to perform significantly better than all other models. But still there is room for further performance improvement. PS can provide support to numerous real-time applications such as bio-medical image analysis, pedestrian monitoring, online surveillance, self-driving cars, to name a few.

In addition, the field of PS seeks significant improvement especially in areas relating to complex scene backgrounds, issues pertaining to cluttered scenes, the dataset quality utilized and its associated computational costs [91]. Overlapping and occlusion is another intriguing issue as discussed under Section 3.3.1, wherein the issue of determining the extent of overlap between the two object instances in the absence of sufficient contextual information becomes difficult when merging them by using a heuristic approach. This is

challenging as the current approaches make use of two independent models which do not share features and therefore the pipeline implementation becomes labor intensive. In PS, a class label and an instance id are assigned to prevent the overlapping phenomenon by using an NMS that primarily sorts the predicted segments with respect to its confidence scores. While resolving the overlapping issue, it is also important to view the instances sorted in descending fashion using detection scores and then placing the objects on some material canvas such that objects with higher scores are placed higher. This approach can fail because of occlusion, and it is inferred how OANet for PS can be utilized at the post-processing step to predict the 'stuff' and instance segmentation in a single network, thereby resolving the problem of occlusion between the predicted segments to a certain extent. Some improvements on PS include incorporating the information exchange module, certain attention-based methods [92], and improvements on the loss function, to name a few, as seen in the literature. Moreover, concentrating on the performance improvements of various PS models may aid several applications in healthcare, autonomous self-driving vehicles, and many others.

Author Contributions: Conceptualization, S.A.M. and A.L.R.; methodology, S.A.M.; software, S.A.M.; validation, S.A.M.; formal analysis, S.A.M.; investigation, S.A.M.; resources, S.A.M.; data curation, S.A.M.; writing—original draft preparation, S.A.M.; writing—review and editing, S.A.M. and A.L.R.; supervision, A.L.R.; project administration, A.L.R. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Data sharing is not applicable to this article.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Wang, J.; Wang, Z.; Qiu, S.; Xu, J.; Zhao, H.; Fortino, G.; Habib, M. A selection framework of sensor combination feature subset for human motion phase segmentation. *Inf. Fusion* **2021**, *70*, 1–11. [CrossRef]
2. Cabria, I.; Gondra, I. MRI segmentation fusion for brain tumor detection. *Inf. Fusion* **2017**, *36*, 1–9. [CrossRef]
3. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015*; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2015.
4. Badrinarayanan, V.; Kendall, A.; Cipolla, R. SegNet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 122017. [CrossRef] [PubMed]
5. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *40*, 834–848. [CrossRef] [PubMed]
6. Yao, J.; Fidler, S.; Urtasun, R. Describing the scene as a whole: Joint object detection, scene classification and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012.
7. Xiong, Y.; Liao, R.; Zhao, H.; Hu, R.; Bai, M.; Yumer, E.; Urtasun, R. Upsnet: A unified panoptic segmentation network. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–17 June 2019.
8. Liu, S.; Jia, J.; Fidler, S.; Urtasun, R. SGN: Sequential grouping networks for instance segmentation. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; IEEE Computer Society: Los Alamitos, CA, USA, 2017.
9. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollar, P.; Zitnick, C.L. Microsoft COCO: Common objects in context. In *Computer Vision e ECCV 2014*; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2014; Volume 8693.
10. Kirillov, A.; He, K.; Girshick, R.; Rother, C.; Dollár, P. Panoptic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–17 June 2019.
11. Xie, L.; Lee, F.; Liu, L.; Kotani, K.; Chen, Q. Scene recognition: A comprehensive survey. *Pattern Recognit.* **2020**, *102*, 107205. [CrossRef]
12. Hoiem, D.; Hays, J.; Xiao, J.; Khosla, A. Guest editorial: Scene understanding. *Int. J. Comput. Vis.* **2015**, *112*, 131–132. [CrossRef]
13. Oliva, A.; Torralba, A. Modeling the shape of the scene: A holistic representation of the spatial envelope. *Int. J. Comput. Vis.* **2001**, *42*, 145–175. [CrossRef]
14. Zhou, B.; Lapedriza, A.; Xiao, J.; Torralba, A.; Oliva, A. Learning deep features for scene recognition using places database. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 8–13 December 2014.
15. Girshick, R. Fast R-CNN. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015.
16. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014.

17. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2016, Las Vegas, NV, USA, 26 June–1 July 2016.
18. Schwing, A.G.; Urtasun, R. Fully connected deep structured networks. *arXiv* **2015**, arXiv:1503.02351.
19. Li, Y.; Qi, H.; Dai, J.; Ji, X.; Wei, Y. Fully convolutional instance-aware semantic segmentation. *arXiv* **2016**, arXiv:1611.07709.
20. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [CrossRef]
21. Neuhold, G.; Ollmann, T.; Bulo, S.R.; Kontschieder, P. The Mapillary Vistas Dataset for Semantic Understanding of Street Scenes. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; IEEE Xplore: Piscataway, NJ, USA, 2017.
22. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.* **2015**, *115*, 211–252. [CrossRef]
23. Abu-El-Haija, S.; Kothari, N.; Lee, J.; Natsev, P.; Toderici, G.; Varadarajan, B.; Vijayanarasimhan, S. Youtube-8m: A large-scale video classification benchmark. *arXiv* **2016**, arXiv:1609.08675.
24. Brostow, G.J.; Shotton, J.; Fauqueur, J.; Cipolla, R. Segmentation and recognition using structure from motion point clouds. In *Computer Vision–ECCV 2008: 10th European Conference on Computer Vision, Marseille, France, 12–18 October 2008, Proceedings, Part I 10*; Springer: Berlin/Heidelberg, Germany, 2008.
25. Geiger, A.; Lenz, P.; Stiller, C.; Urtasun, R. Vision meets robotics: The KITTI dataset. *Int. J. Robot. Res.* **2013**, *32*, 1231–1237. [CrossRef]
26. Cordts, M.; Omran, M.; Ramos, S.; Rehfeld, T.; Enzweiler, M.; Benenson, R.; Franke, U.; Roth, S.; Schiele, B. The Cityscapes dataset for semantic urban scene understanding. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016.
27. Leibe, B.; Cornelis, N.; Cornelis, K.; Van Gool, L. Dynamic 3D scene analysis from a moving vehicle. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Minneapolis, MN, USA, 18–23 June 2007.
28. Tighe, J.; Lazebnik, S. Finding things: Image parsing with regions and per-exemplar detectors. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013.
29. Tighe, J.; Niethammer, M.; Lazebnik, S. Scene parsing with object instances and occlusion ordering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014.
30. Sun, M.; Kim, B.S.; Kohli, P.; Savarese, S. Relating things and stuff via object property interactions. *IEEE Trans. Pattern Anal. Mach. Intell.* **2014**, *36*, 1370–1383. [CrossRef] [PubMed]
31. Wu, Y. 2019. Available online: https://github.com/facebookresearch/detectron2 (accessed on 25 March 2024).
32. Martin, D.R.; Fowlkes, C.C.; Malik, J. Learning to detect natural image boundaries using local brightness, color, and texture cues. *IEEE Trans. Pattern Anal. Mach. Intell.* **2004**, *26*, 530–549. [CrossRef] [PubMed]
33. Van Rijsbergen, C. *Information Retrieval*; Butterworths: London, UK, 1979.
34. Kirillov, A.; Girshick, R.; He, K.; Dollár, P. Panoptic feature pyramid networks. Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–17 June 2019.
35. Li, Y.; Chen, X.; Zhu, Z.; Xie, L.; Huang, G.; Du, D.; Wang, X. Attention-guided unified network for panoptic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–17 June 2019.
36. Porzi, L.; Bulo, S.R.; Colovic, A.; Kontschieder, P. Seamless scene segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–17 June 2019.
37. Mohan, R.; Valada, A. EfficientPS: Efficient Panoptic Segmentation. *arXiv* **2021**, arXiv:2004.02307v3. [CrossRef]
38. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask R-CNN. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017.
39. Li, X.; Chen, D. A survey on deep learning-based panoptic segmentation. *Digit. Signal Process.* **2022**, *120*, 103283. [CrossRef]
40. de Geus, D.; Meletis, P.; Dubbelman, G. Panoptic Segmentation with a Joint Semantic and Instance Segmentation Network. *arXiv* **2019**, arXiv:1809.02110v2.
41. Liu, H.; Peng, C.; Yu, C.; Wang, J.; Liu, X.; Yu, G.; Jiang, W. An End-To-End Network for Panoptic Segmentation. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–17 June 2019.
42. On-Device Panoptic Segmentation for Camera Using Transformers. October 2021. Available online: https://machinelearning.apple.com/research/panoptic-segmentation (accessed on 25 March 2024).
43. Fast Class-Agnostic Salient Object Segmentation. June 2023. Available online: https://machinelearning.apple.com/research/salient-object-segmentation (accessed on 25 March 2024).
44. Prakhar Bansal. Panoptic Segmentation Explained. Available online: https://medium.com/@prakhar.bansal/panoptic-segmentation-explained-5fa7313591a3 (accessed on 25 March 2024).
45. Using Panoptic Segmentation to Train Autonomous Vehicles. Mindy News Blog. December 2021. Available online: https://mindy-support.com/news-post/using-panoptic-segmentation-to-train-autonomous-vehicles/ (accessed on 25 March 2024).
46. COCO 2020 Panoptic Segmentation. Available online: https://cocodataset.org/#panoptic-2020 (accessed on 25 March 2024).
47. Cityscapes Dataset. Available online: https://www.cityscapes-dataset.com/dataset-overview/ (accessed on 25 March 2024).
48. Yu, F.; Chen, H.; Wang, X.; Xian, W.; Chen, Y.; Liu, F.; Madhavan, V.; Darrell, T. BDD100K: A Diverse Driving Dataset for Heterogeneous Multitask Learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; Available online: https://arxiv.org/abs/1805.04687.pdf (accessed on 25 March 2024).
49. Mapillary Vistas Dataset. Available online: https://www.mapillary.com/dataset/vistas (accessed on 25 March 2024).

50. Cheng, X.; Lu, J.; Feng, J.; Yuan, B.; Zhou, J. Scene recognition with objectness. *Pattern Recognit.* **2018**, *74*, 474–487. [CrossRef]
51. Semantic KITTI Dataset. Available online: http://www.semantic-kitti.org/dataset.html (accessed on 25 March 2024).
52. Dai, A.; Chang, A.X.; Savva, M.; Halber, M.; Funkhouser, T.; Nießner, M. ScanNet: Richly annotated 3D Reconstructions of Indoor Scenes. In Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; IEEE: New York, NY, USA, 2017.
53. Caesar, H.; Bankiti, V.; Lang, A.H.; Vora, S.; Liong, V.E.; Xu, Q.; Krishnan, A.; Pan, Y.; Baldan, G.; Beijbom, O. nuScenes: A multimodal dataset for autonomous driving. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2020, Seattle, WA, USA, 13–19 June 2020; Available online: https://arXiv:1903.11027.pdf (accessed on 25 March 2024).
54. Arnab, A.; Torr, P.H. Pixelwise instance segmentation with a dynamically instantiated network. In Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.
55. Bai, M.; Urtasun, R. Deep watershed transform for instance segmentation. In Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.
56. Chang, C.-Y.; Chang, S.-E.; Hsiao, P.-Y.; Fu, L.-C. Epsnet: Efficient panoptic segmentation network with cross-layer attention fusion. In Proceedings of the Asian Conference on Computer Vision, Asian Conference on Computer Vision, Kyoto, Japan, 30 November–4 December 2020.
57. Xiao, Y.; Wu, J.; Yuan, J. mCENTRIST: A multi-channel feature generation mechanism for scene categorization. *IEEE Trans. Image Process.* **2014**, *23*, 823–836. [CrossRef]
58. Meng, X.; Wang, Z.; Wu, L. Building global image features for scene recognition. *Pattern Recognit.* **2012**, *45*, 373–380. [CrossRef]
59. Wu, J.; Rehg, J.M. CENTRIST: A Visual Descriptor for Scene Categorization. *IEEE Trans. Pattern Anal. Mach. Intell.* **2011**, *33*, 1489–1501.
60. Lazebnik, S.; Schmid, C.; Ponce, J. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, New York, NY, USA, 17–22 June 2006.
61. Khan, S.H.; Hayat, M.; Bennamoun, M.; Togneri, R.; Sohel, F.A. A discriminative representation of convolutional features for indoor scene recognition. *IEEE Trans. Image Process.* **2016**, *25*, 3372–3383. [CrossRef] [PubMed]
62. Wu, J.; Rehg, J.M. Beyond the Euclidean distance: Creating effective visual codebooks using the histogram intersection kernel. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009.
63. Gong, Y.; Wang, L.; Guo, R. Multi-scale order less pooling of deep convolutional activation features. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014.
64. Gao, S.; Tsang, I.W.H.; Chia, L.T.; Zhao, P. Local features are not lonely—Laplacian sparse coding for image classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, 13–18 June 2010.
65. Xie, L.; Lee, F.; Liu, L.; Yin, Z.; Chen, Q. Hierarchical coding of convolutional features for scene recognition. *IEEE Trans. Multimed.* **2019**, *22*, 1182–1192. [CrossRef]
66. Zhang, C.; Liu, J.; Tian, Q.; Xu, C.; Lu, H.; Ma, S. Image classification by non-negative sparse coding, low-rank and sparse decomposition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Colorado Springs, CO, USA, 20–25 June 2011.
67. Jiang, Y.; Yuan, J.; Yu, G. Randomized spatial partition for scene recognition. In Proceedings of the European Conference on Computer Vision, Florence, Italy, 7–13 October 2012; pp. 730–743.
68. Hayat, M.; Khan, S.H.; Bennamoun, M.; An, S. A spatial layout and scale invariant feature representation for indoor scene classification. *IEEE Trans. Image Process.* **2016**, *25*, 4829–4841. [CrossRef]
69. Yang, M.; Li, B.; Fan, H.; Jiang, Y. Randomized spatial pooling in deep convolutional networks for scene recognition. In Proceedings of the IEEE Conference on Image Processing, Quebec City, QC, Canada, 27–30 September 2015.
70. Li, L.J.; Su, H.; Fei-Fei, L.; Xing, E. Object bank: A high-level image representation for scene classification and semantic feature sparsification. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 6–11 December 2010.
71. Zuo, Z.; Wang, G.; Shuai, B.; Zhao, L.; Yang, Q.; Jiang, X. Learning discriminative and shareable features for scene classification. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014.
72. Shi, J.; Zhu, H.; Yu, S.; Wu, W.; Shi, H. Scene Categorization Model using Deep Visually Sensitive features. *IEEE Access* **2019**, *7*, 45230–45239. [CrossRef]
73. Lin, D.; Lu, C.; Liao, R.; Jia, J. Learning important spatial pooling regions for scene regions for scene classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014.
74. Wu, R.; Wang, B.; Wang, W.; Yu, Y. Harvesting discriminative meta objects with deep CNN features for scene classification. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015.
75. Yang, S.; Ramanan, D. Multi-scale recognition with DAG-CNNs. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015.
76. Liu, Y.; Chen, Q.; Chen, W.; Wassell, I. Dictionary learning inspired deep network for scene recognition. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018.
77. Seong, H.; Hyun, J.; Kim, E. FOSNet: An End-to End Trainable Deep Neural Network for Scene Recognition. *IEEE Access* **2020**, *8*, 82066–82077. [CrossRef]
78. Xie, G.; Zhang, X.; Yan, S.; Liu, C. Hybrid CNN and dictionary-based models for scene recognition and domain adaptation. *IEEE Trans. Circuits Syst. Video Technol.* **2017**, *27*, 1263–1274. [CrossRef]

79. Xiao, J.; Hays, J.; Ehinger, K.A.; Oliva, A.; Torralba, A. SUN database: Large-scale scene recognition from abbey to zoo. In Proceedings of the IEEE International Conference on Computer Vision, San Francisco, CA, USA, 13–18 June 2010.
80. Quattoni, A.; Torralba, A. Recognizing Indoor Scenes. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Miami, FL, USA, 20–25 June 2009.
81. Li, L.; Fei-Fei, L. What, where and who? Classifying events by scene and object recognition. In Proceedings of the IEEE International Conference on Computer Vision, Rio de Janeiro, Brazil, 14–21 October 2007.
82. Zhou, B.; Lapedriza, A.; Khosla, A.; Oliva, A.; Torralba, A. Places: A 10 million image database for scene recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *40*, 1452–1464. [CrossRef]
83. Wang, Z.; Wang, L.; Wang, Y.; Zhang, B.; Qiao, Y. Weakly supervised PatchNets: Describing and aggregating local patches for scene recognition. *IEEE Trans. Image Process.* **2017**, *26*, 2028–2041. [CrossRef]
84. Vogel, J.; Schiele, B. A semantic typicality measure for natural scene categorization. In *Pattern Recognition, Proceedings of the 26th DAGM Symposium, 30 August–1 September 2004*; Springer: Berlin/Heidelberg, Germany, 2004; pp. 195–203.
85. Amrani, M.; Jiang, F. Deep feature extraction and combination for synthetic aperture radar target classification. *J. Appl. Remote Sens.* **2017**, *11*, 042616. [CrossRef]
86. Song, Y.; Zhang, Z.; Liu, L.; Rahimpour, A.; Qi, H. Dictionary reduction: Automatic compact dictionary learning for classification. In *Computer Vision—ACCV 2016. ACCV 2016, Proceedings of the Asian Conference on Computer Vision, Taipei, Taiwan, 20–24 November 2016*; Springer: Cham, Switzerland, 2017.
87. Huang, Y.; Wu, Z.; Wang, L.; Tan, T. Feature coding in image classification: A comprehensive study. *IEEE Trans. Pattern Anal. Mach. Intell.* **2014**, *36*, 493–506. [CrossRef]
88. Weng, C.; Wang, H.; Yuan, J.; Jiang, X. Discovering class-specific spatial layouts for scene recognition. *IEEE Signal Process. Lett.* **2017**, *24*, 1143–1147. [CrossRef]
89. He, K.; Zhang, X.; Ren, S.; Sun, J. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2014**, *33*, 346–361.
90. Pandey, M.; Lazebnik, S. Scene recognition and weakly supervised object localization with deformable part-based models. In Proceedings of the IEEE International Conference on Computer Vision, Barcelona, Spain, 6–13 November 2011.
91. Elharrouss, O.; Al-Maadeed, S.; Subramanian, N.; Ottakath, N.; Almaadeed, N.; Himeur, Y. Panoptic Segmentation: A Review. *arXiv* **2021**, arXiv:2111.10250. [CrossRef]
92. Minaee, S.; Boykov, Y.; Porikli, F.; Plaza, A.; Kehtarnavaz, N.; Terzopoulos, D. Image Segmentation Using Deep Learning: A Survey. *arXiv* **2020**, arXiv:2001.05566v5. [CrossRef] [PubMed]