Survey on Panoptic Segmentation

Tanmay Delhikar[1] and Peter Neigel[2]

[1] Technical University of Kaiserslautern
[2] DFKI GmbH

Abstract. Image segmentation is often viewed separately in terms of semantic and instance segmentation tasks. For modern applications like self-driving cars, robots with vision systems, interpreting rich scenes is extremely important. To address this issue, a new methodology called Panoptic Segmentation is gaining traction in the computer vision community that aims to view these two distinct tasks uniformly. This paper aims to compare the state of the art algorithms on popular data sets and measure performance in terms of a new metric called Panoptic Quality (PQ).

Keywords: image segmentation, panoptic segmentation, semantic segmentation, instance segmentation, object detection

1 Introduction

We, humans, have an amazing capacity to comprehend visual stimuli: images and videos. For years, we have tried to replicate this unique characteristic of us in machines. Computer vision is a scientific field that deals with making computers understand digital images or videos. Machines viewing and understanding the images as humans do is a critical task in many modern applications. In today's world, it is quite hard to imagine applications without vision systems. Hence, this field is key to advance many applications like robotics, autonomous vehicles. Segmentation of images is one of the tasks in computer vision that has high traction in the research community. Familiarity with the following concepts is helpful before understanding the task of Panoptic Segmentation [9].

Object Detection is the process of identifying the objects in an image and surrounding them with bounding boxes. The main challenges in this task are occlusion, viewpoint changes among others. The algorithm predicts the class and the bounding box location of each object in the given image relative to the ground truth as shown in Figure 1(a).

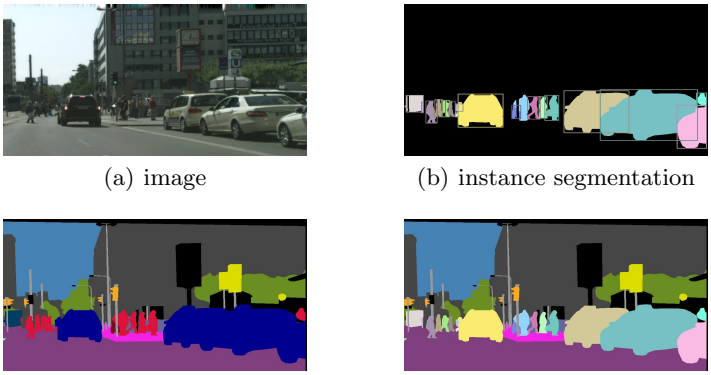
Instance Segmentation takes the previous task of object detection a step further. We are not only interested in the bounding boxes around each object instance, but also identifying masks and instances of each of the detected objects. Even if two objects in the image belong to the same class, Instance segmentation differentiates them with unique instance ids assigned to them. Occlusion, adhesion of objects in the scene are some of the challenges. As seen in Figure 1(b), cars are assigned different instances. Object masks are non-overlapping. It can be observed that generally this task only focuses on instances of objects that are countable.

Semantic Segmentation is the process of assigning a class to each pixel in the image. Pixels that share similar characteristics such as texture or material are assigned the same label/class. This method does not care about instances of the objects belonging to the same class. As seen in Figure 1(c), cars are assigned the same label and do not point out separate instances.

Panoptic Segmentation: Considering the previous methods of instance and semantic segmentation, the next evolutionary step is to identify class label and instance id for each pixel in the

image. In applications like self-driving cars, it is helpful to know the objects surrounding the vehicle as well as the surface the vehicle is driving on to make critical decisions. The word **Panoptic** in this context means a global view or a unified view of the above tasks. Countable objects are called 'Thing' classes and blobby regions of similar material or texture are called 'Stuff' classes. Typically, the Instance segmentation task is formulated as studying 'Thing' classes whereas Semantic segmentation is formulated as studying 'Stuff' classes. Also, note that 'Thing' classes are treated as 'Stuff' classes in Semantic segmentation task. Instance id for 'Stuff' classes are generally assigned as 'None'.

New algorithmic challenges arise when encompassing two different tasks in a global view. Semantic segmentation is achieved by widely used algorithms like Fully Convolutional Networks [15]. Differentiating object instances poses a challenge for these networks. The idea of Panoptic segmentation has not been well-known probably due to the lack of competitive challenges, and metrics to measure the quality of segmentation in a combined manner.



(a) image (b) instance segmentation

(c) semantic segmentation (d) panoptic segmentation

Fig. 1: An illustration of image segmentation. Panoptic segmentation 1(d) can be seen as a combination of instance segmentation 1(b) and semantic segmentation 1(c). Figure is taken from [9]

2 Datasets

Different datasets exist for both semantic and instance segmentation tasks. But, in recent times, additional datasets are available for Panoptic Segmentation of which many are built on existing semantic and instance segmentation datasets. Hence, Panoptic segmentation has gained traction due to the availability of these datasets relating to this specific task. Currently, the following datasets are popular for Panoptic Segmentation task:

- Cityscapes[6]: 5k images with 19 classes
- ADE20K[17]: Over 25k images with 150 classes
- Mapillary Vistas[14]: 25k images with 65 classes
- COCO[11]: Over 120k images with 80 classes

3 Metric

A widely used metric for semantic segmentation is IoU (Intersection over Union) [9] which is calculated based on pixel labels and ignores object-level labels completely. Average Precision (AP) [9]

is the standard metric for instance segmentation in which each object is associated with a confidence score to estimate precision/recall. Hence, AP is unsuitable to measure semantic segmentation quality.

Since Panoptic Segmentation views Semantic and Instance segmentation jointly, a new metric called Panoptic Quality (PQ) [9] is introduced to measure the segmentation quality with p and g as pixel prediction and ground truth respectively, and also considering True Positives (TP), False Positives (FP) and False Negatives (FN). The PQ term can be viewed as a combination of Segmentation Quality (SQ) and Recognition Quality (RQ) which is familiar to IoU and F1 score [9] respectively. PQ treats all 'stuff' and 'thing' classes in a combined way.

$$\mathrm{PQ} = \underbrace{\frac{\sum_{p,g \in TP} \mathrm{IoU}(p,g)}{|TP|}}_{\text{Segmentation Quality (SQ)}} \times \underbrace{\frac{|TP|}{|TP| + \frac{1}{2}|FP| + \frac{1}{2}|FN|}}_{\text{Recognition Quality (RQ)}} \tag{1}$$

4 Algorithms

4.1 Panoptic DeepLab

In Deep Convolutional Neural Networks, repeated combinations of max pooling and striding operations in sequential layers has a disadvantage of losing spatial resolution of the generated feature maps. To deal with this problem, generally, Deconvolution layers are used to up-sample the resulting feature maps. They cause additional overhead of memory in terms of parameters and the execution time. Instead of using Deconvolution layers, an alternate method is to use Atrous Convolution [2].

Atrous means holes. Up-sampled filters are used, meaning, $r-1$ number of zeroes are inserted between consecutive filter values (r is the atrous rate or dilation rate). This results in an enlarged field of view for the filters so that the larger context of the region considered is better understood by the network. This can be seen in Figure 2, controlling r value controls the field of view of the current filter. By keeping constant stride, and a larger field of view, the number of parameters are not increased. When $r = 1$, it is the same as standard convolution. This can be formulated as the equation below with input x, output y with a filter w of length K and atrous rate r.

$$y[i] = \sum_{k=1}^{K} x[i + r \cdot k] w[k]. \tag{2}$$

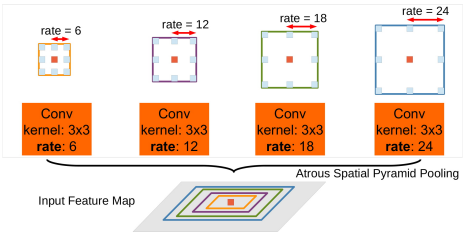


Fig. 2: In Atrous Spatial Pyramid Pooling(ASPP), atrous convolutions are applied at different atrous rate r in parallel which helps to account for objects at different scales. Different colors show effective fields of view. Figure is taken from [4]

In region proposal based methods (proposals of regions where an object may lie), generally, a two-stage pipeline is employed: localize each object with a bounding box and then predict an

object mask inside this bounding box. Therefore, performance is dependent on the quality of region proposals and also more computations are needed for having a two-stage pipeline.

Panoptic DeepLab [5] has two separate branches for semantic and instance segmentation tasks and finally merges them to generate Panoptic Segmentation as described below. Panoptic DeepLab essentially has four components:

a) Shared Encoder Backbone: Encoder-Decoder networks [1] have been generating excellent results in many computer vision tasks like semantic segmentation and object detection. For the two separate tasks of Semantic and Instance segmentation, a shared encoder is employed which is based on an ImageNet pre-trained network along with Atrous Convolutions [2] to extract dense feature maps as explained above. For segmentation tasks, objects in the scene and their locations are useful information.

b) Dual ASPP modules: ASPP [4] considers the multi-scale context of the objects in the scene as explained earlier. This essentially generates feature maps of the scene in different spatial resolutions. Intuitively, looking at the objects in the scene at different scales ensures that they can be differentiated from each other when generating semantic labels and their instances.

c) Dual Decoder modules: The Decoder is used to create output feature maps of the same size as an input image. Its goal is also to recover the boundaries of the objects in the scene. A semantic decoder and Instance decoder are employed next as shown in Figure 3. Applying 1x1 convolutions gradually recovers spatial information and also reduces the number of channels. This method is adopted from the previous version of DeepLab [3]. The idea is that encoded features from the previous module are up-sampled and corresponding low-level features in the encoder module are concatenated.

d) Semantic and Instance Prediction head: Each object instance is represented by 'center of mass' encoded by a 2D-Gaussian[5] which is inspired by the popular Hough Voting mechanism as described in [5]. With this, object centers are predicted and regressing every pixel to the corresponding center. The predicted semantic and instance segmentation is fused by majority vote [5].

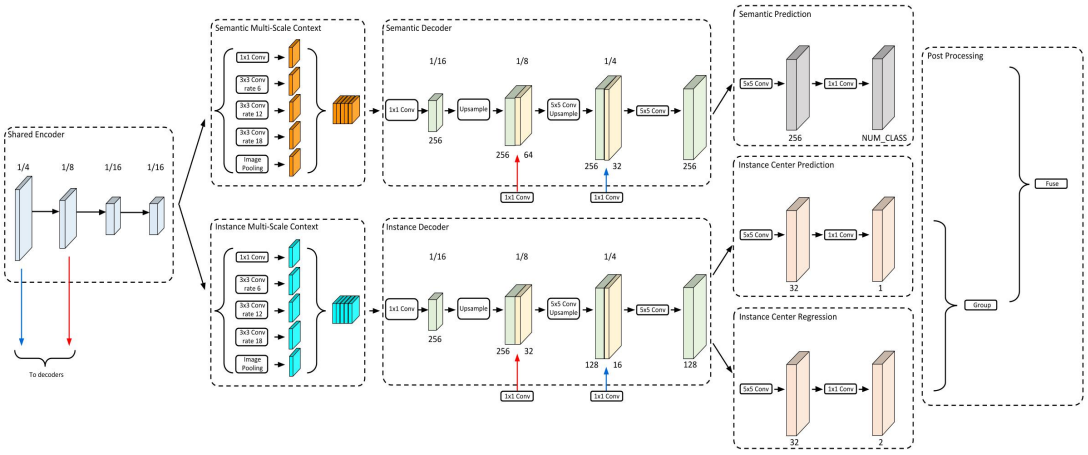


Fig. 3: In the architecture of Panoptic DeepLab, a shared common encoder is connected to two multi-scale contexts with ASPP for instance and segmentation tasks each to extract dense feature maps. The resulting semantic segmentation and instance segmentation maps are fused together to generate Panoptic Segmentation. Figure is taken from [5].

4.2 Seamless Scene Segmentation

Usage of different and separate models for semantic segmentation and detection tasks has a disadvantage of expensive computations due to a large number of parameters in separate models. Since

the Panoptic Segmentation requires to interpret 'Stuff' and 'Things' classes in the image as a global view, Seamless Scene Segmentation [12] method aims to achieve that by training seamless end to end network as shown in Figure 4(a) instead of treating semantic and instance segmentation tasks as different problems.

a) Shared backbone: The backbone used in this branch is a modified version of ResNet-50 [8] with Feature Pyramid Network(FPN) [10] on top of it. The main modification in ResNet-50 is the usage of LeakyRelu [16] instead of Relu. LeakyRelu has very small slopes for negative gradients, instead of absolute zero value in the Relu version. Hence this has contributed to a slight increase in accuracy of the network. The detection of very small objects is challenging in different scales of the image. The purpose of FPN in this network is to act as a feature extractor to generate multi-scale feature maps that have better quality information for object detection.

b) Instance Segmentation Branch: This branch follows the popular Mask R-CNN [7] architecture to solve the instance segmentation task which consists of two stages: It generates proposals about the regions where there might be an object in the input image. Next, it refines the bounding box of the object, predicts the class of the object in the box and also generates a pixel level mask of the object. This instance segmentation branch is again structured into Region Proposal Head (RPH) and Region Segmentation Head (RSH)

i The RPH uses an anchor (a region or a reference bounding box). These anchor boxes are used as a reference for box proposals that have pre-defined dimensions. These boxes are centered on spatial locations of the input image. The purpose of RPH is to do spatial transformations on the currently selected bounding boxes along with objectness score - useful for assessment of the validity of the region containing an object.

ii Region Segmentation Head: Each region proposal received from RPH is used as an input for RSH. It employs the ROIAlign method as mentioned in [7].

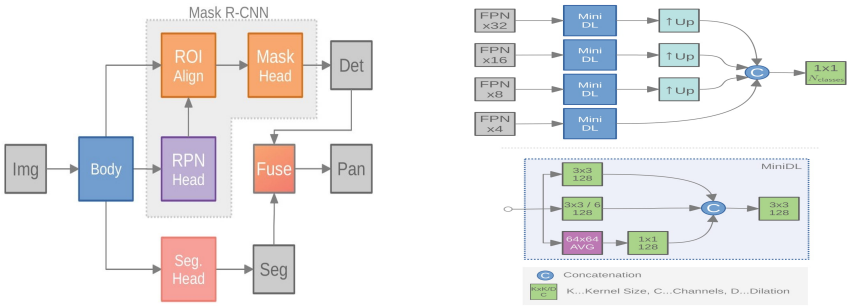
c) Semantic Segmentation Branch: The input of the backbone is fed to a variant of DeepLabV3 [4] head which is called MiniDL in this method. The difference from the original DeepLabV3 head and MiniDL here is that global pooling operation is replaced by Average pooling as shown in Figure 4(b), because it preserves translation equivariance, meaning translation of input features will result in equivalent translation of outputs. This helps the network to generalize shape, edge and texture detection at different locations.

d) Panoptic Fusion: The panoptic fusion is carried out by a Non Maximal Suppression(NMS) [12] like procedure where instances with low classification scores are removed. Then, iterating over the new sorted instances in descending order of confidence score, for each instance, pixels which have been assigned a segment before are removed and if at least 50% of the segment remains, this is accepted as a non-overlapping portion. The rest of the pixels are assigned a most likely class according to the predictions made by the semantic head, either 'stuff' class or 'thing' class if it belongs to 'void' class.

4.3 SSAP: Single-Shot Instance Segmentation With Affinity Pyramid

SSAP [13], a proposal free method, meaning there is no need for a separate region proposal method in the network to generate panoptic segmentation as explained below. This method consists of essentially two parts:

a) U-shaped Network For Semantic Segmentation And Affinity pyramid: A single backbone encoder-decoder modules are employed based on ResNet-50 [8] for semantic segmentation. Besides this branch, the Affinity branch is also incorporated to generate a pixel-pair affinity pyramid. Here affinity means similarity between two pixels similar to the cosine distance measure where the result is 1 if two pixels are the same and 0 otherwise. Affinities are used to distinguish different object instances based on the specification that two pixels belong to the same object. This is particularly calculated for each pixel in a small r x r size window, affinity matrix is generated. L2



(a) Architecute of Seamless Scene Segmentation

(b) MiniDL module

Fig. 4: In 4(a), with a common backbone, the results of Instance segmentation branch and Semantic segmentation branch are fused by NMS like procedure. In 4(b), a modification of DeepLabV3 head is used by replacing global pooling operation with average pooling. Figure is taken from [12].

loss [13] is employed to minimize errors with respect to the ground truth labeling (affinity in ground truth is equal to 1 if two pixels are from the same instance and 0 otherwise). This affinity pyramid is generated at different feature map resolutions as shown in the Figure 5. Hence the result of this module is that the semantic and affinity pyramid is jointly learned with mutual benefits as stated in [13].

b) Cascaded Graph Partition: With the jointly learned Semantic segmentation and affinity pyramid results from the previous module, the next step is to differentiate instances of objects in the scene. This is done by Cascaded graph partition as explained next. An undirected graph $G = (V, E)$ is generated with the pixel-pair affinity pyramid as learned in the previous module. V is the set of pixels and E is the set of pixel pair affinities with $a_{u,v}$ and $a_{v,u} \in (0, 1)$ are affinities for pixel pair u, v. The average affinity is calculated next and then transformed into a weighted score w as mentioned in the below equation. Average affinities are changed to positive if the average affinity is higher than 0.5 and 0 otherwise. This is transformed into a graph partition problem and can be deduced by following optimization problems as mentioned in [13]. A graph partition is generated from lower resolution(deep level in the network) and cascaded towards higher resolutions generating coarse-to-fine instance segmentation results hierarchically as shown in the Figure 5. The main disadvantage with this method is, if the number of nodes, pixels in this case is increased, the runtime is also significantly increased which is a problem for real-world applications.

$$\alpha_{u,v} = (a_{u,v} + a_{v,u})/2, \tag{3}$$

$$w_{u,v} = \log(\frac{\alpha_{u,v}}{1 - \alpha_{u,v}}). \tag{4}$$

5 Analysis

The Panoptic Quality metric for above three algorithms evaluated on Cityscapes dataset is shown in Table 1.

Panoptic DeepLab: This is the state of the art method with PQ score of 65.5. This method shows that it is very effective to adopt the Dual-decoder and Dual-ASPP module specific to the instance and semantic segmentation branches respectively because of being a single model without the need to fine-tune parameters on these two different tasks. Only single convolution is used in each upsampling stage. The usage of Mean Squared Error (MSE) over Cross Entropy Loss has resulted in 1.1% PQ improvement. Also, with the availablity of Mapillary-Vistas as additional data to the network, a slight improvement in PQ=65.5 is observed rather than just using Cityscapes

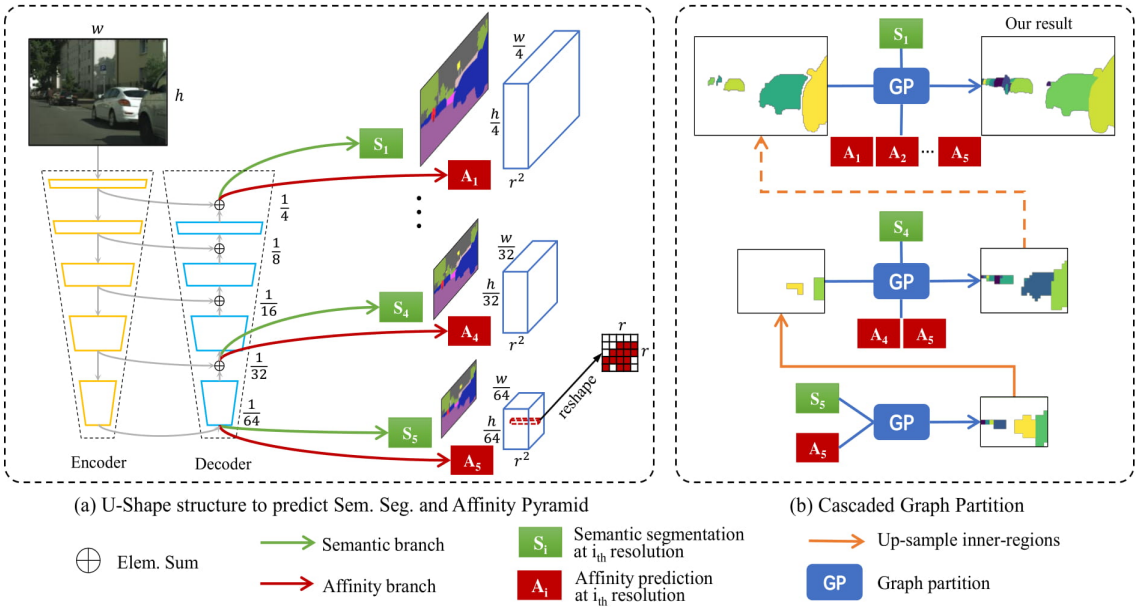


Fig. 5: Architecture of SSAP. Construction of affinity pyramid and semantic segmentation is shown in (a) and in (b) cascaded graph partition module starting from the deepest layer, refines instance prediction results. Figure is taken from [13].

data (PQ=62.3). 0.6% slight improvement is also observed when SGD momentum optimizer [9] is replaced with the Adam Optimizer [9].

Seamless Scene Segmentation: In this method, the DeepLabV3[4] like inspired module called as 'MiniDL' results in an improvement in the PQ. This is because replacing Global Pooling Operation with Average Pooling is better in two cases during training time:
a) If the input images are not cropped and without limiting the size, the context captured by global pooling operation is very wide and reduces the performance as experimented in [12]
b) If the input images are cropped, due to translation equivariance, features representations are changed significantly during training and test time.

Single-Shot Instance Segmentation With Affinity Pyramid (SSAP): The unique part of this method is it does not use any region proposal methods unlike Panoptic DeepLab and Seamless scene segmentation. The joint learning of affinity pyramids and semantic segmentation is mutually beneficial which further improves instance segmentation tasks. Comparing with non-cascaded algorithms to segment instances, this method resulted in a 5x speedup. Unlike previous methods, only single pass can generate instances. A disadvantage of this method is the inference time of segmenting instances increases with an increase in the number of nodes or pixels in the scene and is a problem for real-world applications.

6 Conclusion

After comparing the above algorithms, we conclude that Panoptic DeepLab is better with PQ as the metric. Due to the introduction of PQ and also the availability of datasets more than ever before, Panoptic Segmentation is getting more attention in the research community. Joint learning

Algorithm	PQ	Proposal-free	Backbone
Panoptic DeepLab	65.5	No	Xception-71
Seamless Scene Segmentation	62.6	No	ResNet-50
SSAP	58.9	Yes	ResNet-50

Table 1: Comparison of algorithms

architectures with the experimentations on proposal-free and proposal-based methods, the prospect is that it may drive significant innovations in terms of building deeply integrated models that view semantic and instance segmentation tasks in a combined way.

References

1. Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *CoRR*, abs/1511.00561, 2015.
2. Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *CoRR*, abs/1606.00915, 2016.
3. Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *CoRR*, abs/1606.00915, 2016.
4. Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *CoRR*, abs/1706.05587, 2017.
5. Maxwell D.; Zhu Yukun; Liu Ting; Huang Thomas S.; Adam Hartwig; Chen Liang-Chieh Cheng, Bowen; Collins. Panoptic-deeplab. 2019.
6. Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. *CoRR*, abs/1604.01685, 2016.
7. Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. Mask R-CNN. *CoRR*, abs/1703.06870, 2017.
8. Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.
9. Alexander Kirillov, Kaiming He, Ross B. Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. *CoRR*, abs/1801.00868, 2018.
10. Tsung-Yi Lin, Piotr Dollár, Ross B. Girshick, Kaiming He, Bharath Hariharan, and Serge J. Belongie. Feature pyramid networks for object detection. *CoRR*, abs/1612.03144, 2016.
11. Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312, 2014.
12. Aleksander Colovic Peter Kontschieder Lorenzo Porzi, Samuel Rota Bulò. Seamless scene segmentation. 2019.
13. Yupei Wang Xin Zhao Yinan Yu Ming Yang Kaiqi Huang Naiyu Gao, Yanhu Shan. Ssap: Single-shot instance segmentation with affinity pyramid. 2019.
14. Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulò, and Peter Kontschieder. The mapillary vistas dataset for semantic understanding of street scenes. In *International Conference on Computer Vision (ICCV)*, 2017.
15. Evan Shelhamer, Jonathan Long, and Trevor Darrell. Fully convolutional networks for semantic segmentation. *CoRR*, abs/1605.06211, 2016.
16. Bing Xu, Naiyan Wang, Tianqi Chen, and Mu Li. Empirical evaluation of rectified activations in convolutional network. *CoRR*, abs/1505.00853, 2015.
17. Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ADE20K dataset. *CoRR*, abs/1608.05442, 2016.