International Journal of Engineering and Management Research e-ISSN: 2250-0758 | p-ISSN: 2394-6962
Volume-13, Issue-1, (February 2023)
https://ijemr.vandanapublications.com https://doi.org/10.31033/ijemr.13.1.5

Dataset and Performance Metrics towards Semantic Segmentation

T.S. Rajalakshmi[1] and R. Senthilnathan[2]
[1]Assistant Professor, Department of Mechatronics Engineering, SRM Institute of Science and Technology, Kattankulathur, Chennai, Tamilnadu, INDIA
[2]Associate Professor, Department of Mechatronics Engineering, SRM Institute of Science and Technology, Kattankulathur, Chennai, Tamilnadu, INDIA

[1]Corresponding Author: rajalakt1@srmist.edu.in

ABSTRACT

Interest and ideas on semantic segmentation move on the increasing trend in the area of autonomous driving. This meets the rise in the deep learning approach. The first step in the training of a segmentation model is the dataset preparation. For this, RGB images and its corresponding segmentation images are required such that, the size of these remain the same. Each class in the image is assigned with a unique ID. The pixel value in the segmentation image denotes the class ID of the corresponding pixel. Moreover, as jpg format of the image is lossy, bmp or png formats are usually preferred. The success of the model is measured using metrics, which helps in grading the model. This paper deals with the examination of the widely used datasets in the field of semantic segmentation. The mIoU metric of the datasets on various models have been comparative studied at the end of the analysis.

Keywords-- 2D Dataset, 2.5D Dataset, 3D Dataset, Metrics, Semantic Segmentation

I. INTRODUCTION

Scene understanding is the process of perceiving, analysing and elaborating a scene interpretation. Understanding a scene aids in extracting semantic relationships as well as patterns. In fact, there is a huge increase in the number of applications which gain from the knowledge inferred from these images. Semantic segmentation paves a way to this scene understanding. Autonomous driving is one such application [1-3]. Growth in deep learning techniques has paved a way in resolving such computer vision based semantic segmentation problems, using convolutional neural networks [4-8], as they surpass other approaches in terms of accuracy and also in efficiency. This paper deals with a wide survey of datasets that are useful in semantic segmentation. It showcases the challenges and bench marks of the existing datasets along with their contribution.

II. BACKGROUND CONCEPT

Object recognition is the means to identify or understand the objects that are present in images and videos, similar to the way the humans do. Image classification is a means of outputting the classification label, with some probability for an input image. In case of object localization, the algorithm determines the object, labels the class. A bounding box is generated around the identified object. The bounding box location including the position, height and width will be the output corresponding to an input image. Object detection is a combination of image classification and object detection. For each input image, there might be multiple bounding boxes and class labels. The main issue with object detection is the shape of the bounding box. A rectangular bounding box may not be appropriate in the determination of those objects that have a curvy shape. Also, object detection cannot determine accurate measures of object area or perimeter of object.

Image segmentation, a further extension of object classification and detection, is the process wherein one image is being divided into multiple image segments. Here, each pixel in the image is related with an object type. This technique is more granular as it helps in determination of the shape of each and every object that is existing in the image. Image segmentation can be classified into two – semantic segmentation and instance segmentation. In semantic segmentation, all objects of the same type are indicated using one class label whereas in instance segmentation, each object of the same type gets its own separate label.

III. DATASET

Here, existing datasets are discussed with their evaluation based on quality, its popularity based on citation reports, usage as a benchmarking tool, its degree of significance and the level of impact in the corresponding field. Datasets are initially subdivided based on their data representation as 2D, 3D. They are also of any kind of representation say, gray scale or RGB.

3.1. 2D dataset

3.1.1 PASCAL-Visual Object Classes (VOC) –

This dataset [9-10] consists of annotated images, catering for 5 different tasks such as detection, classification, person-layout, action-recognition and segmentation. For segmentation purpose, this dataset contains 21 classes classified as vehicles, bicycle, car, bus, motorbike, aeroplane, boat, train, household, bottle, TV or monitor, chair, potted plant, dining table, sofa, person, animals, cow, cat, dog, sheep, horse and bird.

Any pixel not belonging to any of the mentioned classes is considered as background. The dataset is of 2sets – training set with 1464 images and validation set with 1449 images. For the actual challenge purpose, private dataset is available.

3.1.2 PASCAL-Context –

An extended version of the PASCAL VOC2010 is this dataset [11-12]. Here, the training images are pixel wise labels. There are more than 400 classes available, which includes almost 20 original classes. These classes can be categorized into 3 classes – objects, stuff, hybrids, along with background from PASCAL VOC. The down line with this dataset is that, many of them are too sparse.

3.1.3 PASCAL Part –

This dataset [13-14] is also an extension of PASCAL VOC for providing per pixel segmentation mask for each part of the object. It consists of labels for all PASCAL VOC training as well as validation images. Also, labels are available for testing images, which are 9637 in total.

3.1.4 Semantic Boundaries Dataset-

This dataset [15-16] is again an extension of the PASCAL VOC. It provides pixel labelling ground truth for the images that were not formally pixel labelled in the original VOC dataset. It consists of annotations for 11355 images. Category and instance level segmentations are provided by these annotations, along with boundary information. Moreover, SBD provides 8498 training images and 2857 validation images. Because of availability of large dataset for training purpose, this Semantic Boundaries Dataset is mostly used in place of the PASCAL VOC dataset.

3.1.5 Microsoft Common Objects in Context –

This dataset [17-18] is a segmentation dataset. This includes more than 80 classes catering for 82783 training images, 40504 validation images and 80000 test images. The test set is split into 4 sets each of 20000 images– test-dev for additional validation, test-standard for comparing state of art, test-challenge for evaluation challenge and test-reserve for protection from overfit issues.

3.1.6 SYNTHIA –

This dataset [19-20] also known as SYNTHetic Collection of Imagery and Annotations is semantically segmented providing pixel level annotation for 11 classes namely, building, pedestrian, sky, road, sidewalk, pole, fence, vegetation, car, sign, and cyclist. It consists of 13407 training set images, catering for different scenes – towns, highways, cities as well as seasons and weather conditions.

3.1.7 Cityscapes –

This dataset [21-22] is a large scale data base. Its main focus is on urban street scenes captured from 50 cities under good weather conditions. It provides pixel labels for 30 classes which are grouped into 8 categories – constructions, flat-surfaces, objects, humans, nature, vehicles, void and sky. This dataset consists of 5000, 20000 fine and coarse annotated frames respectively.

3.1.8 CamVid –

This dataset [23, 24, 26] is a dataset for understanding the driving scene understanding. These are manually annotated with 32 classes mainly, pedestrian, tree, cart, bicyclist, fence, driving and non-driving lane markings, sidewalk, parking, vehicle, animal, road, traffic light, sky, tunnel and other moving object. The partition by Sturgess et al. [25] divides the dataset into 367 train images, 100 val images and 233 test images.

3.1.9 KITTI –

This dataset [27] is for use in mobile robots as well as autonomous vehicle driving. It contains hours of recorded traffic scenes. Here, semantic segmented ground truth is not available for the original. But later, manual annotation has been carried out on these, by the. For 323 images, Álvarez et al. [28-29] generated ground truth from the road detection challenge with 3 classes: road, sky and vertical. Ros et al. [30] labelled 170 training and 46 testing images with a total of 11 classes namely, vehicle, road, sign, pedestrian, pole, fence, sidewalk, etc. Zhang et al. [31] annotated a total of 252 images, 140 for training and 112 for testing. These were annotated with 10 object categories.

3.1.10 Youtube-Objects –

It is a database [32] containing videos which were collected from you tube. They contains image from 10 PASCAL VOC classes. The original database were not pixel wise annotated. But a subset of 126 sequences were manually annotated by Jain et al [33]. Then a subset of frames were extracted so that, semantic labels will be generated. These totalled to 10167 annotated frames.

3.1.11 Adobe's Portrait Segmentation –

This dataset [34, 35] is obtained from Flickr. The images were those that were captured using front facing mobile cameras. The dataset consists of 1800 images - 1500 for training and 300 for testing, being binary annotated – person and background. Using face detector, the images were cropped to 600x800 and then were annotated manually using the photo-shop quick-selection.

3.1.12 Materials in Context –

This MINC dataset [36] is for classification of patch as well as entire scene segmentation. Here, the dataset annotates for 23 categories namely, food, material type, hair, painted, skin, sky, water, wood etc. It contains 7061/2500/5000 for training, validation and test images. These dataset are obtained from OpenSurfaces dataset [37], which was then augmented. Hence, the resolution for this dataset varies.

3.1.13 Densely-Annotated VIdeo Segmentation –

This DAVIS dataset [38-40] is for video object segmentation. It contains 4219 training frames and 2023 validation frames. Pixelwise annotations are provided frame wise, for 4 categories namely, animal, human, object and vehicle. This dataset is such that it does not have more different objects with significant motion.

3.1.14 Stanford background –

This dataset [41-42] contains outdoor scenery images from previous datasets like Geometric Context, PASCAL VOC etc. It consists of 715 images and there should be atleast one foreground object. The dataset is pixel annotated and it is used for the purpose of semantic supported scene understanding.

3.1.15 Siftflow –

This dataset [42, 44] consists of 2688 images, of 256x256 pixel. There are 33 semantic classes and they belong to one of these. The images are annotated as a subset from the database LabelMe[43]. The images are founded on 8 various outdoor settings, which includes beaches, buildings, fields and mountains.

3.1.16 ADE20K /MIT Scene Parsing –

This sceneParse150 dataset has its benchmark from the dataset ADE20K [45]. It provides training and evaluation for scene understanding and contains 22k images with 20k and 2k for training and validation. This dataset has almost 150 semantic categories.

3.1.17 Berkeley Segmentation Dataset –

This BSD dataset [46] is hand labelled from 1000 Corel dataset images consisting of 30 human subjects. Of this image segmentation, half were obtained from that of color image, while other half from that of grayscale image.

3.2. 2.5D dataset

These 2.5D datasets are RGB dataset along with the depth information. This is possible because of affordable range scanners.

3.2.1NYU-D-v2 –

This dataset [47, 48] captured by Microsoft Kinect device, contains 1449 dense labelled pair of depth and RGB images from more than around 450 scenes, which were taken from 3 different cities. These were later fused into 40 object classes for indoor, by Gupta et al. [49], 795 and 654 images for training and testing respectively. As this dataset consists of indoor objects, it is more useful towards robotic tasks at home. But compared to other datasets, this dataset is small.

3.2.2 Object Segmentation Database –

This OSD dataset is designed to segment unknown objects even under the case of partial occlusion. The dataset includes 111 entries, providing both depth as well as color images. Here, as the dataset is not able to differentiate category of different objects, the classes are derived to a give a binary set. This can be of objects and no-objects.

3.2.3 SUN3D –

This dataset [50-51] consists of large scale RGBD video dataset, containing 415 sequences shot in 41 different buildings, for about 254 different spaces. Each frame details the semantic segmentation of the object as well as the pose of the camera.

3.2.4 ScanNet –

This dataset [55] contains 2.5 million views, which are annotated with 3D camera poses. This dataset helped in achieving state of art performances on 3D object-classification and hence, scene-understanding. For data collection, a scalable RBGD capture system is designed, which includes reconstruction of surface, automated.

3.2.5 SUN-RGBD –

This dataset [52-53] captured with 4 RGBD sensors, consists of 10000 RGBD images at the same scale as that of the original PASCAL VOC. This dataset is annotated densely and it includes 2D polygon and 3D bounding box, – 146617 and 58657 respectively, suitable for tasks involving scene understanding.

3.2.6 UW-RGBD –

This RGBD Object dataset [54, 58] consists of 300 household objects, which were captured using Kinect style 3D camera. The dataset organised to 51 categories including 8 annotated video sequences. Here, the images captured are 640x480 pixel RGB and depth at 30Hz.

3.3. 3D Dataset

Volumetric representation such as point cloud and mesh provide the 3dimensional images. These are useful in robotic applications, medical image analysis or 3D scene understanding and in other construction oriented applications.

3.3.1 A Benchmark for 3D-Mesh Segmentation –

This dataset [59, 60] is designed by 380 meshes. These meshes are classified into 19 categories namely, airplane, animal, chair, cup, glass, hand, human, etc. Here, each mesh is manually segmented into functional parts.

3.3.2 Large-Scale Point-Cloud Classification Benchmark –

This dataset [61, 62] comprises 3D point cloud of diverse natural urban scenes such as church, castle, village, street etc. These are manually annotated and consists of statistically captured pointclouds with finer information and density. It consists of 15 large scale point-cloud for both training and testing. Each total to more than one billion labelled points.

3.3.3 ShapeNet Part –

This dataset [63, 64] is a subset of ShapeNet [65]. It concentrates on the Object segmentation of fine grains. It includes 31,693 meshes which are sampled from 16 categories and each shape class is labelled from 2 to 5 parts.

3.3.4 Stanford 2D-3D-S –

This dataset [66, 67] is captured in 6 indoor areas from 3 different educational and office buildings. With semantic annotation, it provides a wide variety from 2D to 2.5D to 3D. It is an extension of Stanford 3D Semantic Parsing work [68] and gives in a total of 271 rooms and 700 million points, which are annotated with labels from 13 categories - beam, board, bookcase, ceiling, chair, clutter, column, door, floor, sofa, table, wall and window

3.3.5 Sydney Urban Objects Dataset –

This dataset [69, 70] consists of various urban type road objects. They are scanned using a Velodyne HDK-64E LIDAR. It includes around 631 scan of

objects across a class of pedestrian, sign, tree and vehicles.

IV. SEGMENTATION MODEL METRICS

Metrics are used to evaluate the performance of the segmentation models to have a model that could contribute significantly to the field and also to enable fair comparison with the other existing methods. Also, the metric plays an important role in the validation of the model.

Pixel Accuracy – PA - It is the simplest metric which computes the ratio between the total number of pixels which are classified properly to the total number of pixels. If k is the number of foreground classes and k+1 implies an addition of background, then pixel accuracy is defined as

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

Mean Pixel Accuracy – MPA - This is an improved version of Pixel Accuracy wherein, the ratio of correctly classified pixels is calculated on the class basis. This is then averaged over the total number of classes found

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

TP, FP, FN - The IoU of a prediction target mask pair, if it exceeds a predefined threshold, is observed to have true positive. If the prediction mask has no associated ground truth, then a false positive is indicated. If the ground truth has no associated prediction mask, then a false negative is indicated.

Intersection over Union – IoU - Also known as Jaccard Index, this is defined as the ratio to the area of intersection between the prediction map and ground truth, to the area of union between the prediction map and ground truth. Suppose A is the ground truth and B is the predicted segmentation truth, then

$$IoU = J(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

Mean Intersection over Union – MIoU is a standard metric for segmentation that computes the IoU on class basis and then averages it. The MIoU can also be formulated as the ratio of number of true positive over the sum of true positive, false negative and false positive.

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

Precision - It indicates the purity of positive predictions with reference to the ground truth. It determines as to how many objects have matching ground truth annotations.

$$Precision = \frac{TP}{TP + FP}$$

Recall - It indicates the completeness of the positive prediction to that of the ground truth. It determines of all of ground truth annotations, how many are positive predictions.

$$Recall = \frac{TP}{TP + FN}$$

F1 score - This is the harmonic mean between precision and recall. It brings in a balance between precision and recall. A good F1 score implies less false positives and less false negatives

$$F1\ Score = \frac{2 * Precision * Recall}{Precisio + Recall}$$

Dice coefficient - This is defined as two times the overlap area between the prediction and ground truth map, divided by the sum of pixels in both prediction and ground truth map. Suppose A is the ground truth and B is the predicted segmentation truth, then

$$Dice = 2\ \frac{|A \cap B|}{|A| + |B|}$$

When applied to a Boolean data, Dice score is similar to that of the F1 score

$$Dice = F1 = \frac{2TP}{2TP + FP + FN}$$

Table 1: mIoU (%) metric details of various datasets on different models, collected from various papers as mentioned in the references

Model	ADE20k	PASCAL VOC	Cityscape	NYUD-v2	MS COCO
AC-Net [106]	45.9		82.3		40.1
APC-Net [82]		87.1			
BiSeNet [103]			78.9		
BoxSup [96]		75.1			
CascadeNet [109]	34.90				
CCN [81]					35.7
CCNet [91]			81.4		
CRF-RNN [71]		72			
DANet [89]			81.5		37.9
DeepLab-CRF [86]		79.7			
DeeplabV2 [85]			70.4		
DeeplabV3 [70]		85.7	81.3		
DeeplabV3+ [88]		87.8			
DenseASPP [87]			80.6		
DFN [93]		86.2	79.3		
DilatedNet [83]	32.31				
Dilation10 [84]			67.1		
DIS [100]		56.8			
DM-Net [80]	45.5	87.06			
DPN [74]		77	66.8		
DSSPN [107]	43.68				37.3
DUC-HDC [77]			77.6		
EfficientNet+NAS+FPN [113]		90.5			
EMANet [90]		57.7			39.9
EncNet [94]	44.64	S5.9			
Exfuse [99]		86.2			
FcaveaNet [101]			74.1		
FCN [72]	29.39	62.2	65.3	34	
GCN [97]		52.2	76.9		
GS-CNN [105]			82.8		
Hierarchical MSA [112]			85.1		
HPxNetV2+OCR (w/ASPP) [75]			83.7		
Ladder DenseNet [102]			73.7		
MSCI [114]		55			
OCR [75]			82.4		39.1
Piecewise [73]		78			
PSANet [92]	43.7		80.1		
PSPNet [79]	43.29	55.4	85.4		
RefineNet [95]	40.7	84.2	73.6		33.6
SAC [110]	44.3				
SDN [76]		86.6			
SGR [108]					39.1
UperNet [111]	42.6				
Wide ResNet [98]		84.9	78.4		

V. SUMMARY AND CONCLUSION

Datasets pertaining to semantic segmentation were described, indicating their need and characteristics. This could enable the decision making to choose the required dataset for a particular application. This study has led to the notion that there is no standard dataset, wherein all methods can relay their report. Many methods have infact, reported their results on non-standard datasets. This makes a comparative study difficult. Of the various metrics available, accuracy is of more importance for a real time application. Also, there is lack of information on other metrics like the execution time, memory foot print.

REFERENCES

[1] A.Ess, T. Müller, H. Grabner, & L.J. Van Gool. (2009). Segmentation-based urban traffic scene understanding, *BMVC, 1*.

[2] A.Geiger, P. Lenz & R. Urtasun (2012). Are we ready for autonomous driving? The KITTI vision benchmark suite. *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3354–3361. http://dx.doi.org/10.1109/CVPR.2012.6248074.

[3] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth & B. Schiele. (2016). The cityscapes dataset for semantic urban scene understanding. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3213–3223.

[4] F. Ning, D. Delhomme, Y. LeCun, F. Piano, L. Bottou & P.E. Barbano. (2005). Toward automatic phenotyping of developing embryos from videos. *IEEE Trans. Image Process. 14*(9), 1360–1371.

[5] D. Ciresan, A. Giusti, L.M. Gambardella & J. Schmidhuber. (2012). Deep neural networks segment neuronal membranes in electron microscopy images. In: *Advances in Neural Information Processing Systems*, pp. 2843–2851.

[6] C. Farabet, C. Couprie, L. Najman & Y. LeCun (2013). Learning hierarchical features for scene labelling. *IEEE Trans. Pattern Anal. Mach. Intell. 35*(8), 1915–1929.

[7] B. Hariharan, P. Arbeláez, R. Girshick & J. Malik (2014). Simultaneous detection and segmentation. In: *European Conference on Computer Vision, Springer*, pp. 297–312.

[8] S. Gupta, R. Girshick, P. Arbeláez, J. Malik, Learning rich features from RGB-D images for object detection and segmentation, in: European Conference on Computer Vision, Springer, 2014, pp. 345–360.

[9] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," International journal of computer vision, vol. 88, pp. 303–338, 2010.

[10] http://host.robots.ox.ac.uk/pascal/VOC/voc2012/.

[11] R. Mottaghi, X. Chen, X. Liu, N.-G. Cho, S.-W. Lee, S. Fidler, R. Urtasun & A. Yuille. (2014). The role of context for object detection and semantic segmentation in the wild. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[12] http://www.cs.stanford.edu/~roozbeh/pascal-context/.

[13] X. Chen, R. Mottaghi, X. Liu, S. Fidler, R. Urtasun & A. Yuille. (2014). Detect what you can: detecting and representing objects using holistic models and body parts. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[14] http://www.stat.ucla.edu/xianjie.chen/pascalpart dataset/pascal part.html.

[15] B. Hariharan, P. Arbeláez, L. Bourdev, S. Maji & J. Malik. (2011). Semantic contours from inverse detectors. In: *International Conference on Computer Vision, IEEE*, pp. 991–998.

[16] http://home.bharathh.info/home/sbd.

[17] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C.L. Zitnick, Microsoft coco: common objects in context, in: European Conference on Computer Vision, Springer, 2014, pp. 740–755.

[18] http://mscoco.org/.

[19] G. Ros, L. Sellart, J. Materzynska, D. Vazquez & A.M. Lopez. (2016). The synthia dataset: a large collection of synthetic images for semantic segmentation of urban scenes. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3234–3243.

[20] http://synthia-dataset.net/.

[21] M. Cordts, M. Omran, S. Ramos, T. Scharwächter, M. Enzweiler, R. Benenson, U. Franke, S. Roth & B. Schiele. (2015). The cityscapes dataset. *CVPR Workshop on The Future of Datasets in Vision*.

[22] https://www.cityscapes-dataset.com/.

[23] G.J. Brostow, J. Fauqueur, R. Cipolla, Semantic object classes in video: a high-definition ground truth database, Pattern Recognit. Lett. 30 (2) (2009) 88–97.

[24] G.J. Brostow, J. Shotton, J. Fauqueur, R. Cipolla, Segmentation and recognition using structure from motion point clouds, in: European Conference on Computer Vision, Springer, 2008, pp. 44–57.

[25] P. Sturgess, K. Alahari, L. Ladicky, P.H. Torr, Combining appearance and structure from

motion features for road scene understanding, BMVC 2012 – 23rd British Machine Vision Conference, BMVA (2009).
[26] http://mi.eng.cam.ac.uk/research/projects/VideoRec/CamVid/.
[27] A. Geiger, P. Lenz, C. Stiller, R. Urtasun, Vision meets robotics: the KITTI dataset, Int. J. Robot. Res. 32 (11) (2013) 1231–1237
[28] J.M. Alvarez, T. Gevers, Y. LeCun, A.M. Lopez, Road scene segmentation from a single image, in: European Conference on Computer Vision, Springer, 2012, pp. 376–389.
[29] G. Ros, J.M. Alvarez, Unsupervised image transformation for outdoor semantic labelling, in: 2015 IEEE Intelligent Vehicles Symposium (IV), IEEE, 2015, pp. 537–542.
[30] G. Ros, S. Ramos, M. Granados, A. Bakhtiary, D. Vazquez, A.M. Lopez, Vision-based offline-online perception paradigm for autonomous driving, in: 2015 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE, 2015, pp. 231–238.
[31] R. Zhang, S.A. Candra, K. Vetter, A. Zakhor, Sensor fusion for semantic segmentation of urban scenes, in: 2015 IEEE International Conference on Robotics and Automation (ICRA), IEEE, 2015, pp. 1850–1857.
[32] A. Prest, C. Leistner, J. Civera, C. Schmid, V. Ferrari, Learning object class detectors from weakly annotated video, in: 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2012, pp. 3282–3289.
[33] S.D. Jain, K. Grauman, Supervoxel-consistent foreground propagation in video, in: European Conference on Computer Vision, Springer, 2014, pp. 656–671.
[34] X. Shen, A. Hertzmann, J. Jia, S. Paris, B. Price, E. Shechtman, I. Sachs, Automatic portrait segmentation for image stylization Computer Graphics Forum, vol. 35, Wiley Online Library, 2016, pp. 93–102.
[35] http://xiaoyongshen.me/webpage portrait/index.html.
[36] S. Bell, P. Upchurch, N. Snavely & K. Bala. (2015). Material recognition in the wild with the materials in context database. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3479–3487.
[37] S. Bell, P. Upchurch, N. Snavely & K. Bala. (2013). OpenSurfaces: a richly annotated catalog of surface appearance. *ACM Trans. Graph. (SIGGRAPH), 32*(4).
[38] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, A. Sorkine-Hornung, A benchmark dataset and evaluation methodology for video object segmentation, Computer Vision and Pattern Recognition (2016).
[39] J. Pont-Tuset, F. Perazzi, S. Caelles, P. Arbeláez, A. Sorkine-Hornung, L. Van Gool, The 2017 Davis Challenge on Video Object Segmentation, 2017 arXiv:1704.00675.
[40] http://davischallenge.org/index.html.
[41] http://refhub.elsevier.com/S1568-4946(18)30281-3/sbref0215.
[42] C. Liu, J. Yuen & A. Torralba. (2009). Nonparametric scene parsing: label transfer via dense scene alignment. In: *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1972–1979.
[43] B.C. Russell, A. Torralba, K.P. Murphy, W.T. Freeman, Labelme: a database and web-based tool for image annotation, Int. J. Comput. Vis. 77 (1) (2008) 157–173.
[44] http://dags.stanford.edu/data/iccv09Data.tar.gz.
[45] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba, "Scene parsing through ade20k dataset," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017.
[46] D. Martin, C. Fowlkes, D. Tal, and J. Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in Proc. 8th Int'l Conf. Computer Vision, vol. 2, July 2001, pp. 416–423.
[47] N. Silberman, D. Hoiem, P. Kohli, R. Fergus, Indoor segmentation and support inference from RGBD images, in: European Conference on Computer Vision, Springer, 2012, pp. 746–760.
[48] http://cs.nyu.edu/silberman/projects/indoorscenesegsup.html.
[49] S. Gupta, P. Arbelaez, J. Malik, Perceptual organization and recognition of indoor scenes from RGB-D images, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2013) 564–571.
[50] J. Xiao, A. Owens, A. Torralba, SUN3D: a database of big spaces reconstructed using SfM and object labels, 2013 IEEE International Conference on Computer Vision (2013) 1625–1632. http://dx.doi.org/10.1109/ICCV.2013.458.
[51] http://sun3d.cs.princeton.edu/.
[52] S. Song, S.P. Lichtenberg, J. Xiao, SUN RGB-D: a RGB-D scene understanding benchmark suite, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2015) 567–576.
[53] http://rgbd.cs.princeton.edu/.
[54] K. Lai, L. Bo, X. Ren, and D. Fox, "A large-scale hierarchical multiview rgb-d object dataset," in 2011 IEEE international conference on robotics and automation. IEEE, 2011, pp. 1817–1824.

[55] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner, "Scannet: Richly-annotated 3d reconstructions of indoor scenes," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 5828–5839.
[56] Richtsfeld, The Object Segmentation Database (OSD), 2012.
[57] http://www.acin.tuwien.ac.at/?id=289.
[58] http://rgbd-dataset.cs.washington.edu/.
[59] X. Chen, A. Golovinskiy, T. Funkhouser, A benchmark for 3D mesh segmentation, ACM Trans. Graph. (Proc. SIGGRAPH) 28 (3) (2009).
[60] http://segeval.cs.princeton.edu/.
[61] T. Hackel, J.D. Wegner, K. Schindler, Contour detection in unstructured 3D point clouds, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2016) 1610–1618.
[62] http://www.semantic3d.net/.
[63] L. Yi, V.G. Kim, D. Ceylan, I.-C. Shen, M. Yan, H. Su, C. Lu, Q. Huang, A. Sheffer, L. Guibas, A scalable active framework for region annotation in 3D shape collections, SIGGRAPH Asia (2016).
[64] http://cs.stanford.edu/ericyi/projectpage/part annotation.
[65] A.X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su, et al., Shapenet: An Information-Rich 3D Model Repository, 2015 arXiv:1512.03012.
[66] Armeni, A. Sax, A.R. Zamir, S. Savarese, Joint 2D-3D-Semantic Data for Indoor Scene Understanding, 2017 arXiv:1702.01105
[67] http://buildingparser.stanford.edu.
[68] Armeni, O. Sener, A.R. Zamir, H. Jiang, I. Brilakis, M. Fischer, S. Savarese, 3D semantic parsing of large-scale indoor spaces, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2016) 1534–1543.
[69] A. Quadros, J. Underwood, B. Douillard, An occlusion-aware feature for range images, in: IEEE International Conference on Robotics and Automation, 2012, ICRA'12, IEEE, 2012.
[70] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, Rethinking atrous convolution for semantic image segmentation, arXiv preprint arXiv:1706.05587, 2017.
[71] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. H. Torr, Conditional random fields as recurrent neural networks, in Proceedings of the IEEE international conference on computer vision, 2015, pp. 1529–1537.
[72] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 3431– 3440.
[73] G. Lin, C. Shen, A. Van Den Hengel, and I. Reid, "Efficient piecewise training of deep structured models for semantic segmentation," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 3194–3203.
[74] Z. Liu, X. Li, P. Luo, C.-C. Loy, and X. Tang, "Semantic image segmentation via deep parsing network," in Proceedings of the IEEE international conference on computer vision, 2015, pp. 1377–1385.
[75] Y. Yuan, X. Chen, and J. Wang, "Object-contextual representations for semantic segmentation," arXiv preprint arXiv:1909.11065, 2019.
[76] A. Chaurasia and E. Culurciello, "Linknet: Exploiting encoder representations for efficient semantic segmentation," in 2017 IEEE Visual Communications and Image Processing (VCIP). IEEE, 2017, pp. 1–4
[77] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in 2016 Fourth International Conference on 3D Vision (3DV). IEEE, 2016, pp. 565–571.
[78] Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and ´ S. Belongie, "Feature pyramid networks for object detection," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 2117–2125
[79] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 2881–2890.
[80] J. He, Z. Deng, and Y. Qiao, "Dynamic multi-scale filters for semantic segmentation," in Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 3562–3572.
[81] H. Ding, X. Jiang, B. Shuai, A. Qun Liu, and G. Wang, "Context contrasted feature and gated multi-scale aggregation for scene segmentation," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 2393–2402.
[82] J. He, Z. Deng, L. Zhou, Y. Wang, and Y. Qiao, "Adaptive pyramid context network for semantic segmentation," in Conference on Computer Vision and Pattern Recognition, 2019, pp. 7519–7528
[83] P. O. Pinheiro, R. Collobert, and P. Dollar, "Learning to segment object candidates," in Advances in Neural Information Processing Systems, 2015, pp. 1990–1998.

[84] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," arXiv preprint arXiv:1511.07122, 2015.
[85] M. Bai and R. Urtasun, "Deep watershed transform for instance segmentation," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 5221–5229.
[86] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully onnected crfs," IEEE transactions on pattern analysis and machine intelligence, vol. 40, no. 4, pp. 834–848, 2017
[87] M. Yang, K. Yu, C. Zhang, Z. Li, and K. Yang, "Denseaspp for semantic segmentation in street scenes," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 3684–3692
[88] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in Proceedings of the European conference on computer vision (ECCV), 2018, pp. 801–818.
[89] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in Proceedings of the IEEE Conference on Computer Vision and Pattern ecognition, 2019, pp. 3146–3154.
[90] X. Li, Z. Zhong, J. Wu, Y. Yang, Z. Lin, and H. Liu, "Expectation maximization attention networks for semantic segmentation," in Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 9167–9176.
[91] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, "Ccnet: Criss-cross attention for semantic segmentation," in Proceedings of the IEEE International Conference on computer Vision, 2019, pp. 603–612.
[92] H. Zhao, Y. Zhang, S. Liu, J. Shi, C. Change Loy, D. Lin, and J. Jia, "Psanet: Point-wise spatial attention network for scene parsing," in Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 267–283
[93] Yu, J. Wang, C. Peng, C. Gao, G. Yu & N. Sang. (2018). Learning a discriminative feature network for semantic segmentation. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1857–1866.
[94] H. Zhang, K. Dana, J. Shi, Z. Zhang, X. Wang, A. Tyagi, and A. Agrawal, "Context encoding for semantic segmentation," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7151–7160.
[95] G. Lin, A. Milan, C. Shen, and I. Reid, "Refinenet: Multi-path refinement networks for high-resolution semantic segmentation," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 1925–1934
[96] J. Dai, K. He, and J. Sun, "Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation," in Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1635–1643.
[97] Peng, X. Zhang, G. Yu, G. Luo, and J. Sun, "Large kernel matters–improve semantic segmentation by global convolutional network," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 4353–4361
[98] Z. Wu, C. Shen, and A. Van Den Hengel, "Wider or deeper: Revisiting the resnet model for visual recognition," Pattern Recognition, vol. 90, pp. 119–133, 2019.
[99] Z. Zhang, X. Zhang, C. Peng, X. Xue, and J. Sun, "Exfuse: Enhancing feature fusion for emantic segmentation," in Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 269–284
[100] P. Luo, G. Wang, L. Lin, and X. Wang, "Deep dual learning for semantic image segmentation," in Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2718–2726.
[101] X. Li, Z. Jie, W. Wang, C. Liu, J. Yang, X. Shen, Z. Lin, Q. Chen, S. Yan, and J. Feng, "Foveanet: Perspective-aware urban scene parsing," in Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 784–792.
[102] Kreso, S. Segvic, and J. Krapac, "Ladder-style densenets for semantic segmentation of arge natural images," in IEEE International Conference on Computer Vision, 2017, pp. 238– 245.
[103] Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "Bisenet: Bilateral segmentation network for real-time semantic segmentation," in European Conference on Computer Vision, 2018, pp. 325–341.
[104] B. Cheng, L.-C. Chen, Y. Wei, Y. Zhu, Z. Huang, J. Xiong, T. S. Huang, W.-M. Hwu & H. Shi. (2019). Spgnet: Semantic prediction guidance for scene parsing. In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 5218–5228.
[105] T. Takikawa, D. Acuna, V. Jampani & S. Fidler. (2019). Gated-scnn: Gated shape cnns for semantic segmentation. In: *IEEE International Conference on Computer Vision*, pp. 5229–5238.

[106] J. Fu, J. Liu, Y. Wang, Y. Li, Y. Bao, J. Tang & H. Lu. (2019). Adaptive context network for scene parsing. In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 6748–6757.

[107] X. Liang, H. Zhou & E. Xing. (2018). Dynamic-structured semantic propagation network. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 752–761.

[108] X. Liang, Z. Hu, H. Zhang, L. Lin & E. P. Xing. (2018). Symbolic graph reasoning meets convolutions. In: *Advances in Neural Information Processing Systems*, pp. 1853–1863.

[109] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso & A. Torralba. (2017). Scene parsing through ade20k dataset. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

[110] R. Zhang, S. Tang, Y. Zhang, J. Li & S. Yan. (2017). Scale-adaptive convolutions for scene arsing. In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2031–2039.

[111] T. Xiao, Y. Liu, B. Zhou, Y. Jiang & J. Sun. (2018). Unified perceptual parsing for scene understanding. In: *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 418–434.

[112] A. Tao, K. Sapra & B. Catanzaro. (2020). *Hierarchical multi-scale attention for semantic segmentation*. arXiv preprint arXiv:2005.10821.

[113] B. Zoph, G. Ghiasi, T.-Y. Lin, Y. Cui, H. Liu, E. D. Cubuk, and Q. V. Le, "Rethinking pre-training and self-training," *arXiv preprint arXiv:2006.06882*, 2020

[114] D. Lin, Y. Ji, D. Lischinski, D. Cohen-Or & H. Huang.(2018). Multiscale context intertwining for semantic segmentation In: *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 603–619.

[115] http://www.acfr.usyd.edu.au/papers/SydneyUrbanObjectsDataset.shtml.