Survey on Segmentation techniques

sgc

June 5, 2022

1 Abstract

As computer vision becoming increasingly widely used in all kinds of fields, segmentation which is a important subject in this field also is a main hotspot draws many scholars' attention. And lots of solution to this problem is popping out.

2 Introduction

Segmentation is a method to extract or separate different part of an image or divide an image into constituent parts. There are 3 main types of segmentation: general segmentation refers to just separate pixels of different objects. Semantic segmentation gives semantic class to each region. Instance segmentation labels each object. And recently a novel and important research topic named panoptic segmentation appeared. It's main task is to give each pixel a semantic label and instance ID. Panoptic segmentation integrates semantic segmentation and instance segmentation in a sense, or it adds process on backgrounds compare with instance segmentation.
Top-level conferences and journals in Semantic segmentation or CV CVPR, ICCV, ECCV, Neurips, ICML, ICLR, AAAI, IJCAI.
Mostly used data set: Pascal VOC 2012consists of 20 kinds of objects including humans, mobiles, and others, can be used in segmentation.
Cityscapescity scene pictures of 50 cities.
Pascal Context400 indoors or outdoors pictures.
Stanford Background DatasetA set of outdoor scenes with at least one foreground object.
The standard index used to evaluate the performance of semantic segmentation model is the average IOU (intersection over union). IOU is defined as follows:

$$IOUs = \frac{Areaofoverlap}{AreaofUnion} = \frac{A_{pred} \cap A_{true}}{A_{pred} \cup A_{true}}$$

It can judge the accuracy of the model.

3 Methods of segmentaion

A general semantic segmentation architecture can be widely considered as an encoder network, followed by a decoder network: the task of the decoder is to project the recognition feature semantics learned by the encoder onto the pixel space to obtain dense classification.

Three main methods:

1-region based semantic segmentation:

The region based method first extracts and describes the free-form region from the image, and then classifies it based on the region. During testing, region based prediction is converted to pixel prediction, usually by marking pixels according to the highest scoring region containing the prediction.

2-full convolution network:

The original complete convolution network (FCN[1]) learns the mapping from pixel to pixel without extracting region recommendations. FCN network pipeline is an extension of classic CNN. The main idea is to make the classic CNN take images of any size as input. CNN only accepts and produces labels with specific size input. The restriction comes from the fully connected layer. In contrast, FCN has only convolution layer and pooling layer, which can take the input of any size.

3-weakly supervised semantic segmentation:

Most related methods in semantic segmentation rely on a large number of images with pixel level segmentation masks. However, manually annotating these masks is quite time-consuming, frustrating and commercial cost. Therefore, some weakly supervised methods have been proposed recently, which are committed to semantic segmentation by using annotated bounding boxes. Notably, 2-stage training is more and more popular. Its general frame is that in the 1st stage the model generates a CAM and the CAM[2] is used in the 2nd stage as pseudo label to commit a fully supervised segmentation.

Recently, more and more method based on Transformer merged and have good performance when applied to semantic segmentation.

The most important method ViT [3] proposed to implement transformer to CV and got great success. And many method find self-supervision(DINO[4]) and weak-supervision(Examples:GroupViT[5]) help to improve performance and lessen dependence on abundant human annotation or pixel level labels. MCTformer[6] improved ViT by adding more Class tokens to produce class-specific object localization maps as a kind of pseudo label to help do semantic segmentation.

3.1 Recent Works

3.1.1 Unsupervised Semantic Segmentation

FreeSOLO[7] is based on the authors' previous work SOLO. It consists of 2 parts: 1, Free Mask to produce coarse masks through self-supervise; 2, Self-supervised SOLO learn with the coarse masks and produce outputs. Notably, this model can be used to pre-train a model.
MaskFormer[8] is the first model that unifies semantic and instance level segmentation with the exact same model, loss, and training pipeline. It applies an extra transformer branch to help generating masks. It reached SOTA in 2021 and works well for panoptic segmentation too.
MaskContrast[9] made a first attempt to build a pixel level prior, and then use a contrastive framework to learn embeddings. This model mainly consists of 2 parts: first determine a prior by identifying objects in the images that can be grouped together with the saliency estimation, and then employ the obtained prior in a contrastive loss. The contrastive loss is calculated based on a D-dimensional super sphere.
MS-CLIP[10] proposed to share the transformer in conventional CLIP. The most intriguing thing is that by doing so the number of parameter is decreased and the performance of the model even increased. This model share most of the transformer except the first layer. And since it is hard to share the parameters of conventional transformer and ViT, a parallel branch is introduced, and this branch is can supplement the main branch with multi-scale feature when an image is taken as the input.

3.1.2 Self-supervised models

Self-supervised works are one of the most important field in unsupervised method research. Its main idea is trying to finish the task by delving into the features obtained from itself. Previous containing MoCo, SimCLR, SWaV and so on.
BEIT[11] is inspired of the success of BERT in NLP. The main idea is that generate 2 views of a image: the image patch processed by the transformer and some of them are masked; and the visual token generated by a discrete VAE. The objective is maximize the similarity of the correct tokens corresponding masked image patches and the predicted tokens.

3.1.3 Semi-supervised Semantic Segmentation

ST[12] inject strong data augmentation to unlabeled images to alleviate the overfitting on noisy data and decoupled predictions of T and S. ST++[12] further developed performance by take reliability into consideration.

3.1.4 Weak supervised Semantic Segmentation

CAM[2] is the basic work in recent WSSS. The main idea is using GAP to find the most discriminative regions corresponding particular category in the image.
RCA[13] is designed to take use of Inter-image information with a memory bank to store object pattern appearing in training data.
Sparsely annotation semantic segmentation(SASS) aim to train a segmentation networks from partly labeled pixels. TEL[14] provide Semantic Affinity of low-level and high-level for labeling and it is effective and easy to be incorporated into existing frameworks by combining it with a traditional segmentation loss.
A recent work[15] proposed pixel-to-prototype contrastive learning method to WSSS. This method uses 2 networks and extract Intra-View and Cross-View loss from their prototypes and values.
The first method to implement Transformer to WSSS is AFA[16]. AFA using affinity information from MHSA to refine pseudo labels generation. AFA is complemented with PAR which can refine labels considering information in its neighbor field.
There are usually unreliable pixels remain unused, so U^2PL[17] is designed to keep a memory bank to store unreliable pixels as negative samples to generate contrastive loss.
CLIMS[18] introduce natural language supervision to activate more complete object regions and suppress closely-related open background.
W-OoD[19] uses out-of-distribution(OoD) data to help to distinguish foreground and background pixels. The main idea is trying to ensure that the distance between input and in-distribution clusters is small, but the distance between input and OoD clusters P is large.
ReCAM[20] uses 2 stages of CAM generation layer, and convert the BCE loss task into SEC loss in FC2 through CAM from FC1 and get better performance by overcoming the limitation in BCE.
CDA[21] uses instance from other images to add extra limitation to train the modal and successfully decoupled contextual information(i.e., co-occurring background). It just uses simple pasting method to help to produce more reliable pseudo masks.
AdvCAM[22] is designed to increase the covering area of object since conventional CAM concentrate on the discriminative area. This method use adversarial climbing to increase the class score of all regions and notably effect undiscriminative area more than discriminative one. Regularization is also implemented to suppress the value of unrelated class and the value of region with a high class score. Furthermore this method can easily be attached to CAM based model without any modification.
AMN[23] is also designed to enlarge the activated area of the CAMs. AMN have 2 main learning objectives, a per-pixel classification loss to reduce the activation balance inside the foreground and provides the large gap between the foreground and the background and a lable conditioning module to eliminate the activation from the non-target classes.
A recent work[24] proposed a method to enlarge the activation map and produce

a well-defined prediction contours. It uses importance sampling to add new loss to constrain. And feature similarity loss to penalize dissimilar prediction of nearby similar pixels and vice versa thus produce well-defined contour.
POM and NSRM[25] proposed another method to enlarge the activated region. This paper firstly introduced a Global reasoning before the classifier to exploit the distant and disjoint information. Potential Object Mining is used to introduce extra useless regions in training. Non-Salient Region Masking is introduced to exploit information form the segmentation itself.
PPM[26] actually consists of several new methods to improve the performance. This work introduces some coefficient of variation smoothing from the distribution of the statistics to smooth the CAM and expand the activated area of the object. Proportional Pseudo-mask Generation is used to generate the pseudo mask with the smoothed CAM. Pretended Under-fitting Strategy then suppress the noise during training witch is really different with popular methods. In the end of this paper authors also propose to generate cyclic pseudo masks from the last trained model prediction.
BBAM[27] concentrates the bounding box annotation. First this method introduced a masking procedure to identify the important regions in the image to perform the object detection. Then use the BBAM of the trained object detector to generate pseudo labels. Firstly proposals are generated through jittering the GT box. Some of these proposals are positive samples. Pixels of which value in BBAM larger than a threshold is labeled as foreground.
A work conducted recently[28] deal with the bottleneck problem(loss large scale of information at the last layer of DNN while training). The proposed method is to refine the initial model which is trained with BCE loss with bottleneck-free method using RIB loss. And to further deal with overfitting, extra random fixed number of samples other than the objective image are introduced at each iteration. This paper also introduced a global non-discriminative region pooling (GNDRP) that selectively aggregates the values of spatial locations whose CAM scores are below a threshold.
BAP & NAL[29] are newly method designed to boost the performance. Background aware pooling(BAP) is introduced to discriminate foreground and background inside the bounding boxes. Instead of GAP, this method implements the weighted average pooling, the weight is the probability of a pixel being foreground. Noise-aware loss(NAL) is introduced to train CNNs for semantic segmentation that makes the network less susceptible to incorrect labels.

3.2 Few-Shot Segmentation

Most recent learn few shot tasks using meta learning which is limited for its bias to seen categories. BAM[30] applied a new branch the base learner to produce extra information for confusing regions of the query image. This work outperformed existing models in a considerable margin.

3.3 Domain Adaptive Semantic Segmentation

Domain adaptive semantic segmentation mean to transfer known structure related information to unseen domains. Then the model can be used to work on unseen images. And there 2 main works on this: UDA(unsupervised domain adaptation) and DG(domain generalization), the main difference between them is whether unseen domain is available during training.
[31] including new paradigm and test strategy to alleviate the problem that model works worse in unseen domains. And furthermore a memory bank is developed to obtain more accurate statistics for normalization. This model consists of 3 main parts: Model-agnostic meta-learning in DG, Target-specific normalization and the image bank in the test stage.
SAN and SAW[32] are two plug-and-play method in semantic segmentation. SAN performs category-level center alignment. SAW distributed alignment to achieve both domain-invariant and discriminative features.

4 Related Works in This Field

There are lots of detail problems can be handled and recently some excellent works revealed some intrinsic shortage of some existing methods; For conventional knowledge distillation, DKD[33] divided KL Loss into 2 parts (TCKD and NCKD) and found that NCKD is depressed and solve this by decouple $(1 - p_t)$ and NCKD.
As Siamese Network being widely used in SSL, people find that using random crop has 2 major problems: 1, It may generate bad pairs containing useless images(background); 2, It may generate similar pairs. So Contrastive crop [34] is designed to deal with that by implementing Semantic-aware localization and Center-suppressing sampling. MAE[35] is a classic model applying autoencoder or transformers to CV. The key idea is masking most of the image to enforce the encoder to learn useful relations between masked areas and unmasked areas. ConvMAE[36] is based on MAE aforementioned, and further improved it by applying pyramid structure. The 3-stage-encoder consist of 2 masked convolution block and a transformer. Input of the decoder is the combination of features generated by 3 stages to take both fine and coarse information into consideration.
Local-window self-attention suffers from non-overlapped windows and shares weights on channel dimension. So MixFormer[37] is proposed combining local window self-attention and depth-wise convolution in a parallel design. And they also designed a bi-directional interaction across the two branches. This model is designed to classification tasks but the architecture can easily adapt to segmentation tasks.
Prevalent works on semantic segmentation tasks mostly base on a learnable prototype per class and due to that these methods are limited. But thinking on prototype aspect[38] proposed a method that uses multiple of unlearnable

prototypes for each class.
TCL[39] uses 3 Loss models take inter and intra package information and local information to consideration and refine the self-supervision: cross-modal alignment (CMA), intra modal contrastive (IMC), and local MI maximization (LMI)
Self-Distillation from Last Mini-Batch (DLB)[40] proposed to use half of the generated soft labels in the last epoch to maintain the consistency information with little computation cost.
DiRA[41] is a model in the field of medical image analysis, the main contribution is that it combines discriminative model, restorative model and adversarial model together to foster collaborative learning for representation.It is easy and maybe can be used in other fields.

References

[1] 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. 3

[2] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2921–2929. 3, 3.1.4

[3] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020. 3

[4] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 9650–9660. 3

[5] J. Xu, S. De Mello, S. Liu, W. Byeon, T. Breuel, J. Kautz, and X. Wang, "Groupvit: Semantic segmentation emerges from text supervision," *arXiv preprint arXiv:2202.11094*, 2022. 3

[6] L. Xu, W. Ouyang, M. Bennamoun, F. Boussaid, and D. Xu, "Multi-class token transformer for weakly supervised semantic segmentation," *arXiv preprint arXiv:2203.02891*, 2022. 3

[7] X. Wang, Z. Yu, S. De Mello, J. Kautz, A. Anandkumar, C. Shen, and J. M. Alvarez, "Freesolo: Learning to segment objects without annotations," *arXiv preprint arXiv:2202.12181*, 2022. 3.1.1

[8] B. Cheng, A. Schwing, and A. Kirillov, "Per-pixel classification is not all you need for semantic segmentation," *Advances in Neural Information Processing Systems*, vol. 34, 2021. 3.1.1

[9] W. Van Gansbeke, S. Vandenhende, S. Georgoulis, and L. Van Gool, "Unsupervised semantic segmentation by contrasting object mask proposals," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10 052–10 062. 3.1.1

[10] H. You, L. Zhou, B. Xiao, N. C. Codella, Y. Cheng, R. Xu, S.-F. Chang, and L. Yuan, "Ma-clip: Towards modality-agnostic contrastive language-image pre-training," 2021. 3.1.1

[11] H. Bao, L. Dong, and F. Wei, "Beit: Bert pre-training of image transformers," *arXiv preprint arXiv:2106.08254*, 2021. 3.1.2

[12] L. Yang, W. Zhuo, L. Qi, Y. Shi, and Y. Gao, "St++: Make self-training work better for semi-supervised semantic segmentation," *arXiv preprint arXiv:2106.05095*, 2021. 3.1.3

[13] T. Zhou, M. Zhang, F. Zhao, and J. Li, "Regional semantic contrast and aggregation for weakly supervised semantic segmentation," *arXiv preprint arXiv:2203.09653*, 2022. 3.1.4

[14] Z. Liang, T. Wang, X. Zhang, J. Sun, and J. Shen, "Tree energy loss: Towards sparsely annotated semantic segmentation," *arXiv preprint arXiv:2203.10739*, 2022. 3.1.4

[15] Y. Du, Z. Fu, Q. Liu, and Y. Wang, "Weakly supervised semantic segmentation by pixel-to-prototype contrast," *arXiv preprint arXiv:2110.07110*, 2021. 3.1.4

[16] L. Ru, Y. Zhan, B. Yu, and B. Du, "Learning affinity from attention: End-to-end weakly-supervised semantic segmentation with transformers," *arXiv preprint arXiv:2203.02664*, 2022. 3.1.4

[17] Y. Wang, H. Wang, Y. Shen, J. Fei, W. Li, G. Jin, L. Wu, R. Zhao, and X. Le, "Semi-supervised semantic segmentation using unreliable pseudo-labels," *arXiv preprint arXiv:2203.03884*, 2022. 3.1.4

[18] J. Xie, X. Hou, K. Ye, and L. Shen, "Cross language image matching for weakly supervised semantic segmentation," *arXiv preprint arXiv:2203.02668*, 2022. 3.1.4

[19] J. Lee, S. J. Oh, S. Yun, J. Choe, E. Kim, and S. Yoon, "Weakly supervised semantic segmentation using out-of-distribution data," *arXiv preprint arXiv:2203.03860*, 2022. 3.1.4

[20] Z. Chen, T. Wang, X. Wu, X.-S. Hua, H. Zhang, and Q. Sun, "Class re-activation maps for weakly-supervised semantic segmentation," *arXiv preprint arXiv:2203.00962*, 2022. 3.1.4

[21] Y. Su, R. Sun, G. Lin, and Q. Wu, "Context decoupling augmentation for weakly supervised semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 7004–7014. 3.1.4

[22] J. Lee, E. Kim, and S. Yoon, "Anti-adversarially manipulated attributions for weakly and semi-supervised semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 4071–4080. 3.1.4

[23] M. Lee, D. Kim, and H. Shim, "Threshold matters in wsss: Manipulating the activation for the robust and accurate segmentation model against thresholds," *arXiv preprint arXiv:2203.16045*, 2022. 3.1.4

[24] A. Jonnarth and M. Felsberg, "Importance sampling cams for weakly-supervised segmentation," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 2639–2643. 3.1.4

[25] Y. Yao, T. Chen, G.-S. Xie, C. Zhang, F. Shen, Q. Wu, Z. Tang, and J. Zhang, "Non-salient region object mining for weakly supervised semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 2623–2632. 3.1.4

[26] Y. Li, Z. Kuang, L. Liu, Y. Chen, and W. Zhang, "Pseudo-mask matters in weakly-supervised semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 6964–6973. 3.1.4

[27] J. Lee, J. Yi, C. Shin, and S. Yoon, "Bbam: Bounding box attribution map for weakly supervised semantic and instance segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 2643–2652. 3.1.4

[28] J. Lee, J. Choi, J. Mok, and S. Yoon, "Reducing information bottleneck for weakly supervised semantic segmentation," *Advances in Neural Information Processing Systems*, vol. 34, 2021. 3.1.4

[29] Y. Oh, B. Kim, and B. Ham, "Background-aware pooling and noise-aware loss for weakly-supervised semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 6913–6922. 3.1.4

[30] C. Lang, G. Cheng, B. Tu, and J. Han, "Learning what not to segment: A new perspective on few-shot segmentation," *arXiv preprint arXiv:2203.07615*, 2022. 3.2

[31] J. Zhang, L. Qi, Y. Shi, and Y. Gao, "Generalizable model-agnostic semantic segmentation via target-specific normalization," *arXiv preprint arXiv:2003.12296*, 2020. 3.3

[32] D. Peng, Y. Lei, M. Hayat, Y. Guo, and W. Li, "Semantic-aware domain generalized segmentation," *arXiv preprint arXiv:2204.00822*, 2022. 3.3

[33] B. Zhao, Q. Cui, R. Song, Y. Qiu, and J. Liang, "Decoupled knowledge distillation," *arXiv preprint arXiv:2203.08679*, 2022. 4

[34] X. Peng, K. Wang, Z. Zhu, and Y. You, "Crafting better contrastive views for siamese representation learning," *arXiv preprint arXiv:2202.03278*, 2022. 4

[35] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," *arXiv preprint arXiv:2111.06377*, 2021. 4

[36] P. Gao, T. Ma, H. Li, J. Dai, and Y. Qiao, "Convmae: Masked convolution meets masked autoencoders," *arXiv preprint arXiv:2205.03892*, 2022. 4

[37] Q. Chen, Q. Wu, J. Wang, Q. Hu, T. Hu, E. Ding, J. Cheng, and J. Wang, "Mixformer: Mixing features across windows and dimensions," *arXiv preprint arXiv:2204.02557*, 2022. 4

[38] T. Zhou, W. Wang, E. Konukoglu, and L. Van Gool, "Rethinking semantic segmentation: A prototype view," *arXiv preprint arXiv:2203.15102*, 2022. 4

[39] J. Yang, J. Duan, S. Tran, Y. Xu, S. Chanda, L. Chen, B. Zeng, T. Chilimbi, and J. Huang, "Vision-language pre-training with triple contrastive learning," *arXiv preprint arXiv:2202.10401*, 2022. 4

[40] Y. Shen, L. Xu, Y. Yang, Y. Li, and Y. Guo, "Self-distillation from the last mini-batch for consistency regularization," *arXiv preprint arXiv:2203.16172*, 2022. 4

[41] F. Haghighi, M. R. H. Taher, M. B. Gotway, and J. Liang, "Dira: Discriminative, restorative, and adversarial learning for self-supervised medical image analysis," *arXiv preprint arXiv:2204.10437*, 2022. 4