Survey on Segmentation techniques

sgc

May 1, 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.

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

The most important method ViT [2] proposed to implement transformer to CV and got great success. And many method find self-supervision(DINO[3]) and weak-supervision(Examples:GroupViT[4]) help to improve performance and lessen dependence on abundant human annotation or pixel level labels. MCTformer[5] 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[6] 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[7] 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.

3.1.2 Semi-supervised Semantic Segmentation

ST[8] inject strong data augmentation to unlabeled images to alleviate the overfitting on noisy data and decoupled predictions of T and S. ST++[8] further developed performance by take reliability into consideration.

3.1.3 Weak supervised Semantic Segmentation

RCA[9] 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[10] 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[11] 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[12]. 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[13] is designed to keep a memory bank to store unreliable pixels as negative samples to generate contrastive loss.
CLIMS[14] introduce natural language supervision to activate more complete object regions and suppress closely-related open background.
W-OoD[15] 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[16] 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[17] 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.

3.2 Few-Shot Segmentation

Most recent learn few shot tasks using meta learning which is limited for its bias to seen categories. BAM[18] 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.

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[19] 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 [20] is designed to deal with that by implementing Semantic-aware localization and Center-suppressing sampling. MAE[21] 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.
Local-window self-attention suffers from non-overlapped windows and shares weights on channel dimension. So MixFormer [22] is proposed combining localwindow 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[23] proposed a method that uses multiple of unlearnable prototypes for each class.
TCL[24] uses 3 Loss models take inter and intra package information and local information to consideration and refine the self-supervision: cross-modal alignment (CMA), intramodal contrastive (IMC), and local MI maximization (LMI)
Self-Distillation from Last Mini-Batch (DLB)[25] proposed to use half of the generated soft labels in the last epoch to maintain the consistency information with little computation cost.

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] 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

[3] 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

[4] 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

[5] 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

[6] 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

[7] 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

[8] 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. 3.1.2

[9] 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. 3.1.3

[10] 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. 3.1.3

[11] Y. Du, Z. Fu, Q. Liu, and Y. Wang, "Weakly supervised semantic segmentation by pixel-to-prototype contrast," *arXiv preprint arXiv:2110.07110*, 2021. 3.1.3

[12] 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. 3.1.3

[13] 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. 3.1.3

[14] J. Xie, X. Hou, K. Ye, and L. Shen, "Cross language image matching for weakly supervised semantic segmentation," *arXiv preprint arXiv:2203.02668*, 2022. 3.1.3

[15] 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. 3.1.3

[16] 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. 3.1.3

[17] 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. 3.1.3

[18] 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*. 3.2

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

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

[21] 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

[22] 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

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

[24] 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

[25] 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