Zero-Shot Semantic Segmentation with No Supervision Leakage

Subjects: Computer Science, Artificial Intelligence
Contributor: Yiqi Wang, Yingjie Tian

Zero-shot semantic segmentation (ZS3), the process of classifying unseen classes without explicit training samples, poses a significant challenge. Despite notable progress made by pre-trained vision-language models, they have a problem of “supervision leakage” in the unseen classes due to their large-scale pre-trained data.

Keywords: Semantic Segmentation ; ZS3

1. Introduction

Semantic segmentation is at the foundation of several high-level computer vision applications such as autonomous driving, medical imaging, and other areas involving identification and classification of objects within an image. Deep supervised learning has been instrumental in driving advancements in semantic segmentation [1][2][3][4]. However, fully supervised methods often require extensive labeled image databases with pixel-level annotations. They are typically designed to handle a pre-defined set of classes, restricting their application in diverse, real-world scenarios.

Some weakly supervised semantic segmentation (WSSS) approaches have been proposed for the above situation. These methods capitalize on easily accessible annotations like scribbles [5], bounding boxes [6], and image-level labels [7] and generate pseudo-ground-truths through visualization techniques [8][9]. However, this approach still relies on a certain degree of labeled data and needs to retrain the entire model if there are some new classes.

Humans possess an intuitive ability to recognize and classify new classes based solely on descriptive details, a powerful skill that current machine learning systems have yet to emulate fully. This observation has catalyzed the exploration of zero-shot semantic segmentation (ZS3) [10][11][12][13].

ZS3 aims to exploit the semantic relationships between image pixels and their corresponding text descriptions, predicting unseen classes through language-guided semantic information of the respective classes rather than the dense annotations. ZS3 techniques are broadly divided into generative and discriminative methods [14]. Generative ZS3 methods [15][16] usually train a semantic generator network which maps unseen class language embeddings into the visual feature space and fine-tunes the pre-trained classifier on these generated features. While these generative methods have demonstrated impressive performance, their effectiveness could be improved by a multi-stage training strategy. Discriminative methods directly learn the join embedding spaces for visual and language, like SPNet [17] and map the visual feature to the fixed semantic representations, bridging the gap between visual information and its corresponding semantic understanding. Similarly, JoEm [14] was proposed to optimize both the visual and semantic features within a joint embedding space.

2. Zero-Shot Semantic Segmentation with No Supervision Leakage

Semantic Segmentation: Semantic segmentation has made significant progress with the advent of deep learning technologies. Chen et al. [1], Long et al. [2], Ronneberger et al. [3], and Zhao et al. [4] have leveraged deep learning architectures to enhance the performance of semantic segmentation, making it more accurate and efficient. and fully supervised semantic segmentation, operate under the assumption of pixel-level annotations throughout all training data. The DeepLab model has notably augmented segmentation performance on renowned datasets like PASCAL VOC2012 [18] and MS-COCO [19], employing sophisticated techniques such as multiple scales [13][20] and dilated convolution [21][22]. Other algorithms, such as UNet [3] and SegNet [23], have also demonstrated commendable performance using a diverse set of strategies.

Furthermore, the transformative potential of the vision transformer (ViT) [24], as the pioneer in deploying transformer architecture for recognition tasks, cannot be overstated. Concurrently, the Swin Transformer took a leap forward, extrapolating the transformer's capabilities for dense prediction tasks and achieving top-tier performance in the process. However, it must be acknowledged that these cutting-edge methods are heavily reliant on costly pixel-level segmentation labels and presuppose the presence of training data for all categories beforehand.

In the quest to circumvent these obstacles, weakly supervised semantic segmentation (WSSS) methods have emerged, leveraging more readily accessible annotations such as bounding boxes [6], scribbles [5], and image-level labels [7]. A cornerstone in prevailing WSSS pipelines is the generation of pseudo-labels, chiefly facilitated by network visualization techniques such as class activation maps (CAMs). Some works employ expanding strategies to stretch the CAM ground-truth regions to encapsulate entire objects. Still, obtaining pseudo-labels that accurately delineate entire object regions with fine-grained boundaries continues to pose a significant challenge [25][26].

Zero-shot semantic segmentation: Zero-shot semantic segmentation (ZS3) models are primarily categorized into two main types: discriminative and generative. Discerning the nuances within these two categories provides a comprehensive understanding of the current strategies utilized in the field.

Discriminative methods encompass several noteworthy studies. For instance, Zhao et al. [10] pioneered a groundbreaking study that proposed a novel strategy for predicting unseen classes using a hierarchical approach. This strategy represents an effort to build upon the data's inherent structure, using hierarchies to draw insights into unseen classes. Another study, SPNet [17], adopted a different approach by leveraging a semantic embedding space. Here, visual features are mapped onto fixed semantic representations, bridging the gap between visual information and its corresponding semantic understanding. Similarly, JoEm [14] was proposed as a method that aligns visual and semantic features within a shared embedding space, thereby fostering a direct correlation between these two aspects.

On the other hand, some studies explore the generative landscape of ZS3. ZS3Net [11], for example, employed a Generative Moment Matching Network (GMMN) to synthesize visual features. However, this model's intricate three-stage training pipeline can potentially introduce bias into the system. To mitigate this issue, CSRL [13] employed a unique strategy that leverages relations of both seen and unseen classes to preserve these features during synthesis. Likewise, CaGNet [12] introduced a channel-wise attention mechanism in dilated convolutional layers, facilitating the extraction of visual features.

Recently, some works have explored the large-scale pre-trained model in zero-shot semantic segmentation [27][28][29]. Furthermore, the pre-trained data usually contain both seen and unseen labels (e.g., CLIP, WebImageText 400M) and have a supervision leakage problem. Supervision leakage is a crucial concern in machine learning, referring to the unintended incorporation of information about unseen classes during the training phase. Given that CLIP models are trained on a massive scale of approximately 400 million image–text pairs, it is conceivable that the text labels could encapsulate a diverse range of seen and unseen classes. Consequently, these models might unintentionally learn unseen classes during training, thus creating a form of supervision leakage. This situation could compromise the integrity of the zero-shot learning task, as the models are no longer genuinely learning from a "zero-shot" perspective.

In response to this significant challenge, the research introduces a distinct solution that effectively navigates the complexities of zero-shot learning while eliminating the risk of supervision leakage. By doing so, people enhance semantic segmentation models' reliability and adaptability. The approach offers a robust framework to accurately process and categorize visual data in a genuinely zero-shot learning context. People envisage this method, free of supervision leakage, becoming a cornerstone for future research and advancements in semantic segmentation. The work paves the way for more authentic and reliable zero-shot learning models, fostering a more resilient future for semantic segmentation in computer vision.

Visual-language learning: The domain of image–language pair learning has undergone a significant transformation marked by exponential growth. Several contributions have shaped the field, such as CLIP [30] and ALIGN [31]. Both models, pre-trained on hundreds of millions of image–language pairs, have marked substantial advancements in the field, pushing the boundaries of what is possible in image–language learning. Considering this, Yang et al. [32] put forth a unified contrastive learning method, successfully integrating both image–language techniques and image-label data. This method stands as an emblem of how these techniques can be effectively harnessed to push the frontier of the field further. In the ever-evolving domain of zero-shot learning, CLIP-based methods [27][30][33][34][35] have been recognized for their substantial contributions and potential to provide effective solutions. These models capitalize on the strength of large-

scale image–text pair datasets to deliver remarkable performance. However, a critical challenge that potentially undermines the legitimacy of their zero-shot learning capabilities is the risk of supervision leakage.

References

1. Chen, L.C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 801–818.
2. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
3. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015; Springer: Berlin/Heidelberg, Germany, 2015; pp. 234–241.
4. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid scene parsing network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2881–2890.
5. Sun, J.; Lin, D.; Dai, J.; Jia, J.; He, K.S. Scribble-supervised convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; Volume 26.
6. Dai, J.; He, K.; Sun, J. Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1635–1643.
7. Hou, Q.; Jiang, P.; Wei, Y.; Cheng, M.M. Self-erasing network for integral object attention. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS 2018), Montreal, QC, Canada, 3–8 December 2018; Volume 31.
8. Zeiler, M.D.; Fergus, R. Visualizing and understanding convolutional networks. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; Springer: Berlin/Heidelberg, Germany, 2014; pp. 818–833.
9. Zhang, D.; Zhang, H.; Tang, J.; Hua, X.S.; Sun, Q. Causal intervention for weakly-supervised semantic segmentation. Adv. Neural Inf. Process. Syst. 2020, 33, 655–666.
10. Zhao, H.; Puig, X.; Zhou, B.; Fidler, S.; Torralba, A. Open vocabulary scene parsing. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2002–2010.
11. Bucher, M.; Vu, T.H.; Cord, M.; Pérez, P. Zero-shot semantic segmentation. Adv. Neural Inf. Process. Syst. 2019, 32.
12. Gu, Z.; Zhou, S.; Niu, L.; Zhao, Z.; Zhang, L. Context-aware feature generation for zero-shot semantic segmentation. In Proceedings of the 28th ACM International Conference on Multimedia, Seattle, WA, USA, 12–16 October 2020; pp. 1921–1929.
13. Li, P.; Wei, Y.; Yang, Y. Consistent structural relation learning for zero-shot segmentation. Adv. Neural Inf. Process. Syst. 2020, 33, 10317–10327.
14. Baek, D.; Oh, Y.; Ham, B. Exploiting a joint embedding space for generalized zero-shot semantic segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 9536–9545.
15. Creswell, A.; White, T.; Dumoulin, V.; Arulkumaran, K.; Sengupta, B.; Bharath, A.A. Generative adversarial networks: An overview. IEEE Signal Process. Mag. 2018, 35, 53–65.
16. Kingma, D.P.; Welling, M. Auto-encoding variational bayes. arXiv 2013, arXiv:1312.6114.
17. Xian, Y.; Choudhury, S.; He, Y.; Schiele, B.; Akata, Z. Semantic projection network for zero-and few-label semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 8256–8265.
18. Everingham, M.; Van Gool, L.; Williams, C.K.; Winn, J.; Zisserman, A. The pascal visual object classes (voc) challenge. Int. J. Comput. Vis. 2010, 88, 303–338.
19. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; Springer: Berlin/Heidelberg, Germany, 2014; pp. 740–755.

20. Zhang, Z.; Zhang, X.; Peng, C.; Xue, X.; Sun, J. Exfuse: Enhancing feature fusion for semantic segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 269–284.
21. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Semantic image segmentation with deep convolutional nets and fully connected crfs. arXiv 2014, arXiv:1412.7062.
22. Chen, L.C.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking atrous convolution for semantic image segmentation. arXiv 2017, arXiv:1706.05587.
23. Badrinarayanan, V.; Kendall, A.; Cipolla, R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 2017, 39, 2481–2495.
24. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16 × 16 words: Transformers for image recognition at scale. arXiv 2020, arXiv:2010.11929.
25. Singh, K.K.; Lee, Y.J. Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; IEEE: Piscataway Township, NJ, USA, 2017; pp. 3544–3553.
26. Li, K.; Wu, Z.; Peng, K.C.; Ernst, J.; Fu, Y. Tell me where to look: Guided attention inference network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 9215–9223.
27. Ding, J.; Xue, N.; Xia, G.S.; Dai, D. Decoupling Zero-Shot Semantic Segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 11583–11592.
28. Xu, M.; Zhang, Z.; Wei, F.; Lin, Y.; Cao, Y.; Hu, H.; Bai, X. A simple baseline for open-vocabulary semantic segmentation with pre-trained vision-language model. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 24–28 October 2022; Springer: Berlin/Heidelberg, Germany, 2022; pp. 736–753.
29. Zhou, Z.; Lei, Y.; Zhang, B.; Liu, L.; Liu, Y. Zegclip: Towards adapting clip for zero-shot semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 11175–11185.
30. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning transferable visual models from natural language supervision. In Proceedings of the International Conference on Machine Learning. PMLR, Virtual, 18–24 July 2021; pp. 8748–8763.
31. Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.T.; Parekh, Z.; Pham, H.; Le, Q.; Sung, Y.H.; Li, Z.; Duerig, T. Scaling up visual and vision-language representation learning with noisy text supervision. In Proceedings of the International Conference on Machine Learning, PMLR, Virtual, 18–24 July 2021; pp. 4904–4916.
32. Yang, J.; Li, C.; Zhang, P.; Xiao, B.; Liu, C.; Yuan, L.; Gao, J. Unified Contrastive Learning in Image-Text-Label Space. arXiv 2022, arXiv:2204.03610.
33. Xu, J.; De Mello, S.; Liu, S.; Byeon, W.; Breuel, T.; Kautz, J.; Wang, X. GroupViT: Semantic Segmentation Emerges from Text Supervision. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 18134–18144.
34. Arnab, A.; Dehghani, M.; Heigold, G.; Sun, C.; Lučić, M.; Schmid, C. Vivit: A video vision transformer. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 6836–6846.
35. Xu, M.; Zhang, Z.; Wei, F.; Lin, Y.; Cao, Y.; Hu, H.; Bai, X. A simple baseline for zero-shot semantic segmentation with pre-trained vision-language model. arXiv 2021, arXiv:2112.14757.