Mathematical Modeling and Algorithm Application
ISSN: 3006-0842 | Vol. 8, No. 1, 2026

Review on the Application and Optimization of Unsupervised Edge Generation in Polyp Segmentation

Lufei Zhu[1], Min Jiang[1], Lixia He[1, *], Botao Liu[1, 2, *]

[1]College of Computer Science, Yangtze University, Jingzhou 434023, China;

[2]Hubei Key Laboratory of Oil and Gas Drilling and Production Engineering, Yangtze University, Jingzhou 434023, China;

[*] **Corresponding author**: Lixia He (Email: 66362643@qq.com), Botao Liu (Email: liubotao920@163.com).

Abstract: Colorectal Cancer (CRC) is a malignant tumor arising from glandular epithelium, and it ranks among the malignancies with the highest incidence and mortality rates worldwide. Colonoscopy serves as a reliable method for the detection and resection of colorectal polyps, which are precancerous lesions. However, colorectal polyps are associated with high rates of missed diagnosis and misdiagnosis, posing a significant challenge to clinical diagnosis. Computer algorithms and models based on deep learning are expected to play a crucial role in polyp diagnosis and segmentation, as well as in assisting physicians to automatically and accurately diagnose lesions, thereby reducing the rates of missed diagnosis and misdiagnosis. Nevertheless, these models require hundreds or thousands of standardized images for training. Medical image annotation is not only costly but also subject to significant subjective variability. Particularly for segmentation tasks, the useful information at image edges often suffers from inaccuracies or ambiguity. In recent years, as an emerging medical auxiliary technology, unsupervised edge ground truth generation technology automatically extracts edge information from existing images via algorithms. This technology avoids cumbersome and inconsistent manual edge annotation, thereby providing a more accurate tool for medical treatment.

Keywords: Unsupervised Learning; Edge Ground Truth Generation; Deep Machine Learning; Computer-Aided Diagnosis and Treatment; Medical Image Analysis.

1. Introduction

Early screening and diagnosis of Colorectal Cancer (CRC) can reduce its mortality rate. During colonoscopy, the accurate identification and segmentation of polyps form the basis for subsequent pathological analysis and endoscopic resection (1). However, polyp identification and segmentation are not straightforward. Polyps exhibit significant variability in size, shape, color, and texture, with no consistent characteristics. Additionally, they have low contrast with the surrounding intestinal mucosa and unclear boundaries. During examinations, complex lighting conditions, along with interferences such as mucosal folds and residues, impose substantial pressure on endoscopists. Consequently, even experienced physicians may miss polyps during detection. Clinical reports indicate that the polyp miss rate can be as high as 20%-30% (2, 3).

Therefore, to address this issue, deep learning-based Computer-Aided Diagnosis (CAD) methods, particularly semantic segmentation models, have been widely applied in polyp detection and segmentation tasks. Examples include U-Net (4), U-Net++ (5), ResUNet++ (6), as well as PraNet (7) and TGARNet (8) which are designed for detecting specific types of polyps. Their performance on various public datasets is comparable to, or even surpasses, that of human experts. Arguably, the most crucial reason for this lies in the availability of large volumes of high-quality pixel-level annotated data. However, such data is extremely difficult to acquire, with pixel-level boundary annotation being particularly challenging. This process requires experts to exercise meticulous care. Furthermore, ambiguous annotations are common, and even the same annotator may produce inconsistent labels at different times. The "variability" and "uncertainty" inherent in such annotations determine the performance limits of the models. Consequently, these models exhibit low accuracy in edge segmentation and weak generalization capabilities (9).

Against this backdrop, unsupervised edge ground truth generation technology has emerged. The core idea of this technology is that edge ground truth is not obtained through manual annotation; instead, edges are inferred from information between consecutive scene frames (10). Specifically, object contours are deduced from easily accessible, relatively simple annotated pixels and used as "ground truth" to train supervised segmentation models. Additionally, it ingeniously creates artificial noisy samples to overcome the challenges in recognition and segmentation. This paper primarily surveys unsupervised edge ground truth generation technology. Firstly, it introduces the background and research characteristics of unsupervised edge ground truth generation. Secondly, it presents and compares two types of unsupervised edge generation methods: traditional operators and deep learning-based approaches. Thirdly, it elaborates on the combination of generated ground truth and segmentation models, covering loss function design and multi-task learning, followed by a discussion of current existing problems and potential research directions. Finally, it concludes the content of this paper.

2. Characteristics of Unsupervised Edge Ground Truth Generation Technology

2.1. Motion Geometry Cue-based Methods Using Image Sequences or Videos

Methods in this category detect edges by utilizing information between consecutive frames in motion. The underlying theory is that motion or geometric discontinuities

always correspond to object boundaries.

Firstly, optical flow discontinuities are employed. In video sequences, pixels belonging to the same object tend to exhibit similar motion trajectories (i.e., optical flow). Therefore, large discontinuity points or boundaries in the optical flow field are typically edge points of objects. By calculating the optical flow field and detecting its edges, true edges at motion boundaries can be obtained. This method is effective for edge extraction of objects in dynamic scenes.

Secondly, multi-view geometric search methods are adopted. Specifically, depth information of the scene is acquired through techniques such as Structure from Motion (SfM) and stereo matching in multi-view images or videos. Depth discontinuity points—regions between object contours and the background—serve as reliable points for object edges (11). The edges computed by this method possess physical significance: they represent the boundaries of objects in three-dimensional space and do not respond to light and shadow variations.

2.2. Self-Supervised Methods Bsaed on Internal Statistical Characteristics Images

Methods in this category do not require acquiring data sequences from outside the image; instead, they directly extract data sequences from the image itself to generate edges. Their characteristics are as follows:

Firstly, utilization of patch similarity and patch dissimilarity. The underlying assumption is that if any two pixel patches in an image originate from the same region of the image (e.g., the surface of the same object), their features (color and texture) will be similar. In contrast, if the patches cross an edge, their feature differences will be significant (12). By systematically comparing a large number of pixel patch pairs in the image, an edge response map can be constructed.

Secondly, no requirement for pre-trained data (self-contained data).Once the pre-trained model is built, it can be directly applied to generate edges from new images without any edge adjustments based on annotated data. This is equivalent to possessing "zero-shot" capability, which makes the model highly convenient to use.

Thirdly, dependence on the quality of pre-training. The quality of generated edges depends on the architecture of the pre-trained model and the quality of pre-training data. The inherent bias of the model itself—such as high sensitivity to the classification of certain objects—directly affects the integrity of the generated edges.

2.3. Methods Based on Pre-trained Deep Models

This category leverages pre-trained Convolutional Neural Networks (CNNs) on large-scale datasets to learn generalized visual features.

First point, mining inherent edge information from pre-trained models. Although pre-trained models (e.g., image classification models) are not specifically designed for edge detection, the shallow and deep network layers of these models have adaptively learned edge filters that capture low-level features such as edge textures to fulfill recognition tasks (13). Obvious edge information can be observed from the gradients and activation rates of these feature maps.

Second point, “Zero-shot” detection capability. With a pre-trained model, this method can be directly applied to new images for edge detection without fine-tuning on any labeled data, thus possessing "zero-shot" detection capability. Therefore, it is well-suited for scenarios where annotated data is scarce.

3. Key Technical Pathways of Unsupervised Edge Ground Truth Generation

The core of unsupervised edge ground truth generation lies in extracting edge lines with single-pixel width and continuous closure from binarized polyp masks. Based on the technical means adopted, these methods can be mainly categorized into two types: those based on traditional image processing operators and those based on deep learning.

3.1. Methods Based on Traditional Image Processing Operators

This category is among the most commonly used and mature approaches. It does not rely on data-driven paradigms; instead, it directly employs mathematical morphology and differential operators for edge extraction.

Gradient and differential operators: Sobel, Prewitt, and Roberts are first-order differential operators. They approximate the rate of grayscale change in an image to represent gradients, and utilize two sets of operators to simultaneously estimate the direction and magnitude of gradients. This enables the calculation of gradient magnitudes for gradient operations or edge detection. These operators feature fast computation speeds, but they are relatively sensitive to noise and yield coarse results. As such, they require subsequent processing for edge refinement and are not directly used alone in generating edges for polyp segmentation (14).The Laplacian operator is also a type of differential operator, with the key distinction that it belongs to the second-order category. It exhibits higher sensitivity in distinguishing the second derivative of gradients at edges, which makes it more responsive overall and prone to generating double-pixel edges. Additionally, it reacts strongly to isolated noise points. Therefore, in the process of edge detection for medical images, it is typically combined with other methods rather than being used independently for ground truth generation (15).

Canny operator: The Canny edge detector is currently the most widely used tool for directly generating unsupervised edges in research, and it is also extensively applied in methods for direct edge ground truth generation in the field of polyp segmentation (16-17). The primary reason lies in its nature as a comprehensive, optimized multi-stage process consisting of several sequential steps. The specific process is as follows: First, first-order or second-order derivatives are employed to smooth the image and eliminate noise. Second, image edges are identified by calculating the gradient direction at each pixel. Third, gradient magnitudes are compared along the gradient direction, and only pixels with larger gradient magnitudes are retained. This ensures that edge lines are reduced to a single-pixel thickness. Fourth, two thresholds (a high threshold and a low threshold) are defined to classify pixels in the image into three categories: strong edges, weak edges, and non-edges. Only weak edges connected to strong edges are retained as final edges; all other weak edges are discarded. These four steps in the Canny edge computation process ensure that the Canny operator yields single-pixel edges with high continuity and noise resistance. Such edges are particularly well-suited for use as label data required by deep learning models (19). The precision of the

generated edge lines can be controlled by adjusting parameters such as the variance of the Gaussian filter or selecting appropriate double thresholds.

3.2. Deep Learning-Based Methods

Despite their effectiveness, traditional operators rely on fixed parameters and are content-agnostic. Specifically, they apply identical thresholds and filtering parameters to all images, without adaptively adjusting these parameters based on the content of individual images. This limitation prevents them from capturing the unique edge characteristics of polyps in different images.

In recent decades, numerous studies have focused on data-driven training to learn and extract authentic edges from natural or real-world images, encompassing unsupervised or self-supervised approaches. Several recently proposed unsupervised edge learning models, such as Holistically-Nested Edge Detection (HED) (20) and Richer Convolutional Features (RCF) (21), have been trained on large datasets of natural images. These edge detection networks are trained using ground truth in the form of manually annotated bounding boxes or sets of labeled pixels, enabling them to learn highly generalized features. When applied to medical images, these pre-trained networks (referred to as "edge extractors") can generate edges directly from within each polyp mask without requiring any input edge ground truth. This approach may learn edges that are more aligned with natural/real-world statistical properties and "semantic" meaning, differing from those produced by the Canny operator. Alternatively, Generative Adversarial Networks (GANs) (22) can be employed: a generator is trained to output a realistic edge map for either polyp masks or raw endoscopic images of polyp regions. However, such studies rarely focus on "unsupervised" "automatic generation" of "ground truth" for arbitrary image boundaries. Their concepts remain relatively simplistic and lack systematic development, but they can provide valuable insights for the current research. For instance, a cycle-consistent framework (23) can be established, allowing edge maps generated from masks to be converted back into the original masks via another network. This enables the learning of meaningful edge representations without the need for training image masks.

4. Collaborative Optimization of Generated Edge Ground Truth and Segmentation Models

After obtaining unsupervised generated edge ground truth, the key lies in how to effectively integrate it into the training process of segmentation models. This integration aims to enhance the performance of the models, particularly their segmentation accuracy in edge regions. There are two main integration strategies, as outlined below:

4.1. Multi-Task Framework

Under this framework, the model is designed to simultaneously perform two interrelated tasks: polyp region segmentation and polyp edge prediction. The network typically consists of a shared encoder (feature extraction backbone) and two separate decoder heads—one outputs the region segmentation map, and the other outputs the edge detection map. The two tasks share general features from the low and middle layers of the network, while task-specific features in the high layers are learned independently. The edge prediction task serves as an auxiliary task; its inclusion compels the shared encoder to learn edge-sensitive feature representations. These features, in turn, contribute to the performance improvement of the primary segmentation task. In the inference phase, either only the region segmentation head can be used, or the outputs of the two heads can be fused to refine the final result (24).

4.2. Structured Edge-Aware Modules

Edge information is treated as an explicit guidance signal (25-26), and dedicated network modules are designed to integrate edge features with contextual features. In this design, the Edge Feature Extraction (EFE) module either leverages unsupervised generated edge ground truth for supervised learning or directly extracts edge responses from intermediate features or inputs using traditional operators. Subsequently, in the Selective Edge Attention (SEA) module (27-28), these edge features are selectively fused with deep semantic features from the Cross-Layer Context Fusion (CCF) module (29). For instance, through an attention mechanism, the model autonomously learns which spatial locations require greater reliance on edge information (typically boundary-blurred regions) and which locations can depend more on regional contextual information (the interior of polyps).This "soft" and adaptive fusion method offers greater flexibility than simple loss function stacking, enabling more intelligent utilization of edge cues.

5. Challenges and Future Outlook

5.1. Key Challenges Currently Faced

Firstly, upper limit of generated groung truth quality. The quality of ground truth generated by unsupervised methods is entirely dependent on the algorithm itself. Traditional operators such as Canny lack the ability to interpret image semantics, which may lead to false edge generation in complex scenarios (e.g., misclassifying texture noise as edges) or the loss of weak true edges. The risk of "garbage in, garbage out" (30-32) remains, and low-quality edge ground truth may mislead model learning.

Secondly, bottleneck in the gain of final segmentation performance(33-34).Introducing edge supervision does not always yield significant performance improvements. When limitations such as insufficient model capacity or other issues with training data (e.g., lack of diversity) become the primary constraints, the benefits of edge optimization may hit a ceiling.

Thirdly, introduction of additional edge supervison (35-36). Whether through multi-task learning or complex perception modules, additional edge supervision increases the model's computational overhead and training complexity. While pursuing high accuracy, it is necessary to consider the feasibility of deploying such models in real-time clinical scenarios.

5.2. Outlook on Future Research Directions

Future development should move toward greater intelligence, so future research will tend to focus on developing more "intelligent" unsupervised edge generators. For instance, meta-learning or few-shot learning (37-38) can be utilized to enable models to automatically learn optimal edge extraction parameters based on a single image or a small number of images. Alternatively, more robust self-supervised edge learning models can be explored, allowing them to learn edge features directly from endoscopic images without

relying on manual masks. Furthermore, in-depth integration with emerging network architectures is a key direction—combining edge-aware mechanisms with advanced architectures such as Transformers (39) and diffusion models (40). For example, the global attention mechanism of Transformers can be used to model long-range dependencies between edges and inner-region areas; alternatively, the iterative denoising process of diffusion models can be leveraged to gradually refine edges.

6. Conclusion

Unsupervised edge ground truth generation technology automatically and batch-extracts objective, consistent edge labels from polyp masks via algorithms. This provides an effective solution to address the bottleneck problems in medical image segmentation, such as the difficulty of edge annotation and strong subjectivity. This paper systematically reviews the technological progress in this field, focusing on the analysis of two generation pathways—those based on traditional operators (with the Canny operator as the representative) and those based on deep learning. It also explores in depth various strategies for the collaborative optimization of generated ground truth with segmentation models, including loss function design, multi-task learning, and structured network modules.

Although existing technologies represented by the Canny operator have been proven in multiple studies to significantly improve the edge localization accuracy of polyp segmentation tasks, this field still faces inherent challenges, such as constraints on the upper limit of ground truth annotation quality and parameter sensitivity. Looking ahead to future development trends, key research directions to advance unsupervised edge ground truth generation technology will include developing intelligent edge generation methods with adaptive capabilities, introducing uncertainty estimation mechanisms, achieving in-depth integration with emerging architectures like Transformers, and striving to enhance cross-domain generalization ability. These advancements will not only promote the in-depth application of this technology in the field of polyp segmentation but also expand its use to a broader range of medical image analysis scenarios. Ultimately, with the gradual maturity of relevant technical routes, this technology will effectively drive the practical clinical application of high-precision, high-robustness computer-aided diagnosis systems, thereby providing important technical support for the early screening and diagnosis of colorectal cancer.

References

[1] Sung, H., et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA: A Cancer Journal for Clinicians, 2021.

[2] Leufkens, A. M., et al. Factors influencing the miss rate of polyps in a back-to-back colonoscopy study. Endoscopy, 2012.

[3] Wang, P., et al. Development and validation of a deep-learning system for polyp detection during colonoscopy. Nature Biomedical Engineering, 2020.

[4] Ronneberger, O., Fischer, P., & Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. MICCAI, 2015.

[5] Zhou, Z., et al. UNet++: A Nested U-Net Architecture for Medical Image Segmentation. IEEE TMI, 2018.

[6] Jha, D., et al. ResUNet++: An Advanced Architecture for Medical Image Segmentation. ISM, 2019.

[7] Fan, D. P., et al. PraNet: Parallel Reverse Attention Network for Polyp Segmentation. MICCAI, 2020.

[8] Xu, J., et al. TGA-Net: Token-Guided Attention Network for Polyp Segmentation. MICCAI, 2021.

[9] Tajbakhsh, N., et al. Embracing Imperfect Datasets: A Review of Deep Learning Solutions for Medical Image Segmentation. Medical Image Analysis, 2020.

[10] Bideau, P., & Eriksson, A. (2019). The Best of Both Worlds: Combining Data-Driven and Model-Based Edge Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

[11] Fortun, D., Bouthemy, P., & Kervrann, C. (2015). Optical Flow Modeling and Computation: A Survey. Computer Vision and Image Understanding, 134, 1-21.

[12] Xie, S., & Tu, Z. (2015). Holistically-Nested Edge Detection. In Proceedings of the IEEE International Conference on Computer Vision (ICCV).

[13] Luo, Z., Mishra, B., Achkar, A., Eichel, J., Li, S., & Jodoin, P. M. (2021). Non-Local Deep Features for Salient Object Detection. IEEE Transactions on Pattern Analysis and Machine Intelligence.

[14] Szeliski, R. Computer Vision: Algorithms and Applications. Springer, 2010

[15] Gonzalez,R. C.,& Woods, R. E. (2018).Digital Image Processing (4th ed.).Pearson. p. 182.

[16] Canny, J. A Computational Approach to Edge Detection. IEEE TPAMI, 1986.

[17] Wei, J., et al. Shallow Attention Network for Polyp Segmentation. MICCAI, 2021.

[18] Canny, J. (1986). A Computational Approach to Edge Detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, *PAMI-8*(6), 679–698.

[19] Szelski, R. (2010). Computer Vision: Algorithms and Applications. Springer Science & Business Media.

[20] Xie, S., & Tu, Z. Holistically-Nested Edge Detection. ICCV, 2015.

[21] Liu, Y., et al. Richer Convolutional Features for Edge Detection. IEEE TPAMI, 2019.

[22] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative Adversarial Nets. In Advances in Neural Information Processing Systems (Vol. 27).

[23] Zhu, J., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. In Proceedings of the IEEE International Conference on Computer Vision (pp. 2223-2232).

[24] Murugesan, B., et al. PGAN: A Progressive Generative Adversarial Network for Medical Image Segmentation. Medical Image Analysis, 2021.

[25] Takikawa, T., Acuna, D., Jampani, V., & Fidler, S. (2019). Gated-SCNN: Gated Shape CNNs for Semantic Segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 5229-5238).

[26] Yu, Z., Feng, C., Liu, M. Y., & Ramalingam, S. (2017). CASENet: Deep Category-Aware Semantic Edge Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 5964-5973).

[27] Woo, S., Park, J., Lee, J. Y., & Kweon, I. S. (2018). CBAM: Convolutional Block Attention Module. In Proceedings of the European Conference on Computer Vision (pp. 3-19).

[28] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is All You Need. In Advances in Neural Information Processing Systems (Vol. 30).

[29] Lin, T. Y., Dollár, P., Girshick, R., He, K., Hariharan, B., & Belongie, S. (2017). Feature Pyramid Networks for Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 2117-2125).

[30] Deng, R., Shen, C., Liu, S., Wang, H., & Liu, X. (2018). Learning to predict crisp boundaries. In Proceedings of the European Conference on Computer Vision (pp. 570-586).

[31] Bertasius, G., Shi, J., & Torresani, L. (2015). DeepEdge: A multi-scale bifurcated deep network for top-down contour detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 4380-4389).

[32] Northcutt, C. G., Jiang, L., & Chuang, I. L. (2021). Confident learning: Estimating uncertainty in dataset labels. Journal of Artificial Intelligence Research, 70, 1373-1411.

[33] He, T., Zhang, Z., Zhang, H., Zhang, Z., Xie, J., & Li, M. (2019). Bag of tricks for image classification with convolutional neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 558-567).

[34] Radosavovic, I., Kosaraju, R. P., Girshick, R., He, K., & Dollár, P. (2020). Designing network design spaces. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 10428-10436).

[35] Tan, M., & Le, Q. V. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. In International Conference on Machine Learning (pp. 6105-6114). PMLR.

[36] Howard, A., Sandler, M., Chu, G., Chen, L. C., Chen, B., Tan, M., ... & Adam, H. (2019). Searching for mobilenetv3. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 1314-1324).

[37] Wang, Y., Zhang, J., Kan, M., Shan, S., & Chen, X. (2020). Self-supervised scale-recurrent network for crowd counting. IEEE Transactions on Image Processing, 29, 8151-8163

[38] Hospedales, T., Antoniou, A., Micaelli, P., & Storkey, A. (2021). Meta-learning in neural networks: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(9), 5149-5169.

[39] Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., ... & Zhou, Y. (2021). TransUNet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306.

[40] Wolleb, J., Sandkühler, R., Bieder, F., Valmaggia, P., & Cattin, P. C. (2022). Diffusion Models for Implicit Image Segmentation Ensembles. In International Conference on Medical Imaging with Deep Learning (pp. 1336-1348).