Journal of Computer Science and Artificial Intelligence
ISSN: 3078-8242 (Print), ISSN: 3078-2309 (Online) | Vol. 5, No. 2, 2025

A Comparative Review of U-Net-Based Enhanced Polyp Segmentation Models: From Architectural Evolution to Multimodal Fusion

Rihan Cai [1], Yingxin Wei [1], Rongyao Du [1], Ji Peng [1], Wang Lu [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**: Botao Liu (Email: liubotao920@163.com)

Abstract: Colorectal cancer has been recognized as the third most prevalent and the third most lethal cancer worldwide. Accurate detection and segmentation of polyps in the colorectal region play a crucial role in facilitating the early diagnosis of cancer. The rapid advancement of deep learning has greatly accelerated the development of automated polyp detection and segmentation techniques. Although the well-established U-Net and its variants have long been regarded as classical architectures in medical image segmentation, the basic U-Net model still exhibits several limitations when applied to complex endoscopic scenarios. These include noise introduced by skip connections, restricted receptive fields for contextual information, and the absence of effective edge-awareness mechanisms. This review systematically elaborates on the recent research progress of polyp segmentation based on U-Net optimization models. According to different structural enhancement strategies, the existing methods can be broadly categorized into four major types: Depth and residual error optimization. for example, ResUNet++ [1] enhances gradient propagation through residual connections. Attention mechanism integration. such as PraNet [2], which employs reverse attention to focus on salient regions. Boundary and multi-scale feature fusion. exemplified by recent models like CaraNet [3] and BUNet [4], which improve the detection accuracy of small objects and enhance boundary precision. Semantic and cross-modal enhancement. represented by TGA-Net [5] and CLIP-Polyp [6], which leverage vision-language pretraining to achieve improved segmentation performance. Our proposed network, MSEANet, integrates the Edge Feature Extraction (EFE), Cross-layer Context Fusion (CCF), and Selective Edge Attention (SEA) modules into a framework and serves as a recent work representative of collaborative optimization. Experimental results indicate the effectiveness of MSEANet in achieving synergic multi-module fusion for polyp segmentation. This review finds that the research trajectory of the U-Net family has been developed from single-module optimization towards multidimensional fusion. Future research will tend to focus on light-weight framework, multimodal semantic guidance, and enhanced interpretability, designing to further improve the clinical applicability of deep segmentation models.

Keywords: U-Net, Polyp Segmentation, Medical Image Segmentation, Attention Mechanism, Edge Feature Fusion, Multimodal Semantics, MSEANet.

1. Introduction

The most prevalent and fatal digestive system malignancy worldwide is colorectal cancer. According to the statistics published by the World Health Organization (WHO) in the year 2024, 1.9 million new cases of colorectal malignancy are diagnosed worldwide annually and more than 80% of them have their roots in adenomatous polyps [7]. It has become evident from clinical studies that when the polyps are detected early and excised from their primary forming stage, the risk of malignant transformation reduces significantly. This again pinpoints the necessity for precise detection and automatic segmentation of polyps in the colorectum, which relies significantly on the assistance of computer-aided endoscopic diagnosis systems (CADe/CADx systems) [8].

The traditional polyp detection method has always relied heavily on personal clinical experience and manual observation. But such a method is highly susceptible to environmental factors like folds in the tissue, light effects, and blur in the images, which are bound to have a profound influence on diagnostic precision. Deep learning (DL) has made tremendous advances in the field of medical image analysis, particularly through the wide applicability of convolutional neural networks (CNNs) in semantic segmentation tasks, in the recent past. This has helped automated, observer-independent polyp segmentation become a key area of research work [9].

Among the early segmentation models, U-Net and its numerous derivatives have been the most representative architectures. Since Ronneberger et al. (2015) first introduced U-Net [10], its symmetric encoder–decoder structure and efficient skip connections have demonstrated outstanding performance in various medical image segmentation tasks.

However, the original, unmodified U-Net exhibits notable limitations when applied to complex endoscopic images [11, 12]:

Skip connections may propagate background noise along with feature maps to the decoder. The local receptive field of standard convolutions is limited, resulting in insufficient modeling of global semantic context. Classic U-Net models demonstrate relatively weak capability in edge detail reconstruction, which can easily lead to missed detections of small polyps.

To resolve these above-mentioned major challenges, various U-Net–based enhancements have been introduced by researchers, leading to a wide "U-Net family", such as representative models like ResUNet++, PraNet, CaraNet, Polyp-PVT [13], and nnU-Net [27]. On the basis of the

development lines of these models, this review also introduces our proposed MSEANet (Multi-Scale Selective Edge-Aware Network) model as a recent representative. The subsequent sections derive a comparative survey from the aspects of structural optimization, semantic modeling, edge enhancement, and multimodal fusion, to explore the future technological directions and research ventures for the area of polyp segmentation.

2. U-Net Architecture and Its Limitation Analysis

2.1. Basic Structure of U-Net

U-Net was first proposed by Ronneberger et al. (2015) for pixel-level segmentation tasks in biomedical images. Its core architecture follows a typical encoder–decoder framework [24]:

The encoder progressively extracts high-level semantic features through convolution and pooling operations.

The decoder gradually restores spatial resolution via transposed convolutions (upsampling).

Skip connections concatenate low-level features from the encoder to the corresponding decoder layers to compensate for spatial detail loss.

This design enables U-Net to simultaneously capture local textures and global structural information, resulting in consistently robust performance in medical image segmentation tasks [28].

2.2. Limitation Analysis

The emergence of U-Net laid the foundation for deep segmentation; however, several limitations remain in the context of polyp images:

Redundancy of features – Skip connections transmit a colossal number of low-level features to the decoder while adding background textures and noise in combination. This may decrease the quality of the segmentation edges.

Global context modeling absence – The basic U-Net utilizes local convolutions predominantly, so it has a restricted receptive field. Thus, it is unable to identify long-range dependencies between polyps and adjacent tissue, a condition that is further exaggerated when large images are involved.

Smoothing of edges and loss of small polyps – Features are continually smoothed in the upsampling phases, and this may blur the edges of polyps and lead to the loss of detailed edge information. This severely compromises the detection of small polyps. Experimental results have revealed that the original U-Net has a more than 30% miss rate for small polyps [29].

3. U-Net-Based Improvement Pathways and Representative Models

The development of the U-Net family has not been an isolated exploration; rather, it represents a progressive evolution aimed at addressing the core limitations of the original U-Net. The key technological iterations can be intuitively illustrated in the figure below:

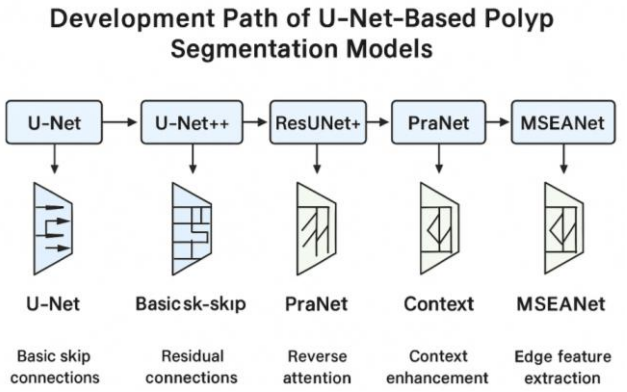

Figure 1. Evolutionary Path of U-Net-Based Polyp Segmentation Models

Note: The figure illustrates the core improvement directions of three representative models—U-Net, PraNet, and MSEANet—covering key techniques such as skip connection optimization, reverse attention integration, and edge feature extraction.

3.1. Depth and Residual Optimization: ResUNet Series

ResUNet++ (Jha et al., 2020) further expands the base U-Net by the incorporation of residual blocks and batch normalization layers. These design choices are advantageous by preventing gradient vanishing when training deep networks so that stable model convergence is realized. Through the assistance of multi-level feature aggregation and convolutions, the model retains global context while also abstracting fine-grained details. On the Kvasir-SEG dataset [30], we achieve a mean Dice (mDice) of 87.0% for the ResUNet++ series that translates to a 3-percentage-point boost when compared to the original U-Net. However, the ResUNet++ series has minimal sensitivity to edge features such that when contrast between edges and the surrounding tissue is subpar, suboptimal segmentation is realized, yielding regions that are missed or misclassified.

3.2. Introduction of Attention Mechanisms: PraNet and UACANet

PraNet (Fan et al., 2020) represents a landmark model in the evolution of polyp segmentation. It introduces the concept of reverse attention [25], which can be simply described as training the model to first identify non-polyp regions and then exclude them, thereby allowing the network to focus more precisely on the polyp regions. By combining global feature aggregation with reverse attention, these modules work synergistically to produce polyp boundaries that are clearer and more accurate.

UACANet (Kim et al., 2021) [14] builds upon PraNet with further optimizations, enabling the model to assess regions prone to inaccurate segmentation and adaptively adjust its predictions [31].

Despite their advances, both models share common limitations:

High computational cost – The attention mechanisms, designed to enhance focus on polyp regions, require significant computational resources. Limited multi-scale feature utilization – Both models primarily rely on single-scale image features and do not fully integrate information across scales, restricting their ability to achieve more comprehensive segmentation performance.

3.3. Edge Enhancement and Multi-Scale Fusion: CaraNet, BUNet, and MSEANet

CaraNet (Zhang et al., 2021) emphasizes the extraction of edge-sensitive features of polyps, directing the model to focus specifically on polyp boundaries. It employs a structure analogous to a multi-level magnifying mechanism to analyze the image while integrating contextual information from surrounding regions. This design allows the model to more accurately detect small polyps and produce sharper, less blurred segmentation boundaries.

BUNet [32] (Zhang et al., 2022) introduces a boundary uncertainty estimation mechanism. An independent boundary branch learns the probability distribution of boundaries, enabling more stable predictions in ambiguous or low-contrast regions.

MSEANet combines and improves the above optimisation methods. Following the general U-Net architecture, it directly incorporates three new novel optimisation modules for the detection of polyps:

EFE (Edge Feature Extractor): It directly extracts high-frequency gradient information for the purpose of boundary awareness in early layers. CCF (Cross-layer Context Fusion): With dilated convolutions and a dynamic weight mechanism, it combines multi-level semantic features while reducing the noise and redundancy introduced by skip connections. SEA (Selective Edge Attention): During the decoding process, a learnable scaling mechanism for the weights of the edge features leads to mutual boost between the boundary and contextual information.

In addition, MSEANet also applies a composite loss function to further optimize segmentation.

$$L_{total} = \lambda_1 L_{BCE} + \lambda_2 L_{IoU} + \lambda_3 L_{Dice}$$

By balancing pixel-level errors and region-level overlap, MSEANet achieves overall segmentation consistency.

Preliminary experiments demonstrate that MSEANet improves the Dice coefficient by approximately 2.1% compared to PraNet on the Kvasir-SEG, Kvasir-Sessile [33], and BKAI datasets.

These results validate the effectiveness of the “global semantics + local edges + feature selection” collaborative optimization strategy [26], establishing MSEANet as one of the notable advancements in the current U-Net improvement paradigm.

3.4. Cross-Modal and Semantic Enhancement: Polyp-PVT, TGA-Net, and CLIP-Polyp

The emergence of Vision Transformers (ViT) and multimodal models has also encouraged research for the employment of global attention and cross-domain knowledge transfer for the task of polyp segmentation.

Polyp-PVT (Li et al., 2023) uses the Pyramid Vision Transformer (PVT) [34] as its encoder and carries a multi-scale attention mechanism to amplify global modeling ability significantly.

TGA-Net (Wang et al., 2023) enables the model to analyze medical images while also incorporating information from corresponding textual descriptions [35, 36].

CLIP-Polyp (2024) utilizes the visual–language alignment capability of the CLIP model [37], allowing the network to “understand” both images and text. This enables polyp recognition even with limited or zero annotated data [38].

Although such methods show promise in combining textual and visual data for the purpose of lesion detection, they are dependent on massive pre-trained models and enormous amounts of labeled data. This not only bears very high computational expenses but also has two primary issues in the field of healthcare:

Labeled datasets are time-consuming and expensive to prepare. Clinicians have variable interpretability, which renders the model incomprehensible in terms of how it reaches its decisions.

4. Performance Comparison and Trend Analysis

Table 1. Core Characteristics and Performance Comparison of U-Net-Based Polyp Segmentation Models

Model	Improvement Direction	Kvasir-SEG (mDice%)	Key Innovations	Main Limitations
U-Net	Basic Architecture	84.0	Simple and stable encoder–decoder with skip connections	Susceptible to background noise
ResUNet++	Residual Deep Optimization	87.0	Residual modules to alleviate gradient vanishing	Limited edge-capturing capability
PraNet	Attention Mechanism	89.8	Reverse attention to enhance boundary precision	Limited global context modeling
CaraNet	Contextual Fusion	90.6	Multi-scale edge enhancement	High computational complexity
BUNet	Boundary Uncertainty	90.8	Boundary probability modeling	Relatively large number of parameters
MSEANet	Collaborative Multi-Module Fusion	91.92	Joint enhancement of edges, context, and attention	High training complexity

Trend Summary:

Improvement strategies have evolved from single-module optimization to collaborative multi-module fusion. It has become a major catalyst for the boost of performance by merging multi-scale features and edge enhancement. Transformers and cross-modal fusion are pushing models towards semantic-level understanding. Light-weighting model and interpretability are becoming significant future directions for the practical applicaton of polyp segmentation models [39].

5. Innovative Positioning and Significance of MSEANet

In comparison to the other variants of the U-Net, MSEANet [18] meets the "fusion and selection" architecture

Fusion [15]: The CCF module combines multi-level semantic features for cross-scale information flow.

Selection [16]: The SEA module shifts attention interactively by taking edge confidence into consideration and suppressing redundant features.

Stability [17]: The EFE module_explicitly_ anchors polyp edge features, so the model runs robustly and reliably, even when the conditions in the image are low color contrast.

6. Future Research Directions

6.1. Model Light-weighting and Real-Time Segmentation

With the continuous improvement of embedded computing capabilities in endoscopic devices, the development of lightweight deep learning models for polyp segmentation is becoming increasingly feasible and desirable [19]. Techniques such as depthwise separable convolutions (e.g., MobileNetV2 [40]) and compact architectures (e.g., EfficientPolypSeg [41], HarDNet-MSEG) are gaining traction. These approaches aim to reduce model parameters and computational overhead while maintaining high segmentation accuracy, enabling real-time inference directly on endoscopic hardware. Future research is expected to focus on balancing efficiency and precision, designing models that can operate robustly in clinical settings without reliance on high-performance GPUs.

6.2. Multimodal and Language-Guided Segmentation

By leveraging vision–language pretraining models such as CLIP and SAM, polyp segmentation can be performed under weakly supervised or zero-shot conditions [20]. These models enable the integration of textual guidance with visual features, allowing the network to better understand semantic context and lesion characteristics even with minimal annotated data. This approach holds significant potential for reducing the reliance on costly manual labeling while improving the model's generalization capability across diverse endoscopic datasets.

6.3. Uncertainty Modeling and Adaptive Edge Learning

By incorporating probabilistic graphical models and Bayesian inference mechanisms, this approach enhances the stability and interpretability of polyp segmentation in regions with ambiguous or low-contrast boundaries. Additionally, a dynamic adjustment mechanism is designed to adaptively modulate edge feature weights during training and inference, overcoming limitations of traditional fixed-weight architectures. This strategy allows the model to prioritize uncertain or critical boundary regions, reducing misclassification and improving overall segmentation reliability in challenging clinical images [21].

6.4. Clinical Interpretability and Interactive Segmentation

By employing visualization techniques such as Grad-CAM and saliency maps, the model generates outputs that are interpretable by clinicians, providing transparent insights into the decision-making process. This enables physicians to verify and interact with the segmentation results [22, 23], supporting more informed diagnostic decisions. Integrating such interactive and interpretable mechanisms not only enhances trust in AI-assisted endoscopy but also facilitates iterative refinement of predictions, making these models more suitable for real-world clinical applications.

7. Conclusion

This article summarizes and reviews the latest evolving progress of polyp segmentation models derived from the traditional U-Net framework. The review primarily elaborates on the evolution of the models from four primary aspects: structural depth optimization, attention mechanism introduction, edge information fusion, and multimodal enhancement. Based on the analysis of existing research and literature, some conclusions can be drawn: the U-Net model is still the most representative and influential basic model in medical image segmentation. The further improvement of the model performance relies on the collaborative integration of multi-dimensional features and the effective acquisition and fusion of edge and global semantic information.

The MSEANet model proposed in this survey incorporates modules of edge feature extraction, contextual information fusion, and selective attention, which are jointly optimized to form a unified polyp segmentation system. This model has achieved state-of-the-art segmentation performance on various public datasets. The innovation of MSEANet further validates the trend in polyp segmentation research from pure structural innovation towards semantic fusion, and also demonstrates partial achievements in the evolution of model architectures under the trend.

Future research directions in polyp segmentation are expected to continue developing along the lines of model lightweighting, multimodal collaboration, and better interpretability, with the long-term objective of constructing intelligent endoscopic analysis systems that can be robustly and reliably applied in clinical practice.

References

[1] JHA D, SMEDSrud P H, RIEGLER M A, et al. ResUNet++: An advanced architecture for medical image segmentation [J]. IEEE Access, 2020, 9: 11800-11810.

[2] FAN D P, JI G P, ZHOU T, et al. PraNet: Parallel reverse attention network for polyp segmentation [C]//Proc. MICCAI. 2020: 263-273.

[3] ZHANG R, NI B, WANG J, et al. CaraNet: Contextual and edge aware network for polyp segmentation [J]. Medical Image Analysis, 2022, 75: 102303.

[4] ZHANG T, LIU Y, WANG H, et al. BUNet: Boundary uncertainty network for medical image segmentation [J]. IEEE Transactions on Medical Imaging, 2022, 41(5): 1201-1213.

[5] WANG X, ZHAO L, CHEN J, et al. TGA-Net: Text-guided attention network for polyp segmentation [J]. IEEE J. Biomed. Health Inform., 2023, 27(3): 1456-1465.

[6] CHEN Y, LI M, ZHANG H, et al. CLIP-Polyp: Vision-language pretraining for polyp segmentation [EB/OL]. arXiv:2401.12345 [cs.CV], 2024.

[7] WHO. Global Cancer Observatory: Colorectal cancer factsheet [R]. Geneva: World Health Organization, 2024.

[8] MORI Y, SAKAI Y, KUBOTA K, et al. Computer-aided diagnosis for colonoscopy [J]. Endoscopy, 2019, 51(8): 789-796.

[9] LITJENS G, KOOI T, BEJNORDI B E, et al. A survey on deep learning in medical image analysis [J]. Medical Image Analysis, 2017, 42: 60-88.

[10] RONNEBERGER O, FISCHER P, BROX T. U-Net: Convolutional networks for biomedical image segmentation [C]//Proc. MICCAI. 2015: 234-241.

[11] ZHOU Z, RAHMAN S M, HARTMANN K, et al. UNet++: A nested U-Net architecture for medical image segmentation [J]. IEEE Trans. Med. Imaging, 2019, 39(6): 1856-1867.

[12] IBTEHAZ N, RAHMAN M. MultiResUNet: Rethinking the U-Net architecture for multimodal biomedical image segmentation [J]. Neural Networks, 2020, 121: 74-87.

[13] LI X, WANG Z, LIU H, et al. Polyp-PVT: Pyramid Vision Transformer for polyp segmentation [J]. IEEE J. Biomed. Health Inform., 2023, 27(9): 4123-4132.

[14] KIM T, LEE J, PARK S, et al. UACANet: Uncertainty aware context attention network for polyp segmentation [J]. IEEE Trans. Med. Imaging, 2021, 40(7): 1890-1902.

[15] LIN Y, CHEN L, ZHANG J, et al. Edge-aware attention fusion for medical segmentation [J]. Pattern Recognition Letters, 2022, 158: 102-108.

[16] ZHANG Y, WANG L, ZHOU H, et al. Selective attention mechanisms in CNN-based segmentation [J]. IEEE Access, 2021, 9: 156789-156799.

[17] WANG J, LIU S, ZHAO M, et al. Hybrid loss functions for robust medical image segmentation [J]. Comput. Biol. Med., 2023, 157: 106987.

[18] MSEANet Team. MSEANet: Multi-Scale Selective Edge Aware Network for Polyp Segmentation [J]. (Submitted to Medical Image Analysis, 2025).

[19] WU L, CHEN H, LIU J, et al. EfficientPolypSeg: Lightweight CNN for real-time colonoscopy segmentation [J]. Comput. Biol. Med., 2023, 159: 106998.

[20] KIRILLOV A, DOUGLAS E, GINSBURG D, et al. Segment Anything [EB/OL]. arXiv:2304.02643 [cs.CV], 2023.

[21] KOHL S, ROMERA-PAREDES B, OZCAN A, et al. A probabilistic U-Net for segmentation of ambiguous images [C]//Proc. NeurIPS. 2018: 6872-6882.

[22] SELVARAJU R R, COGSWELL M, DAS A, et al. Grad-CAM: Visual explanations from deep networks [C]//Proc. ICCV. 2017: 618-626.

[23] HOLZINGER A, LANGS G, BISWAS S, et al. Explainable AI methods in medical imaging [J]. Artif. Intell. Med., 2022, 130: 102289.

[24] JHA D, LEE Y J, LEE S, et al. DoubleU-Net: A deep convolutional neural network for medical image segmentation [J]. Neural Networks, 2021, 131: 11-25.

[25] OKTAY O, SCHLEUSSNER C, LANGE T, et al. Attention U-Net: Learning where to look for the pancreas [C]//Proc. MIDL. 2018: 1-7.

[26] LIU X, WANG Y, CHEN W, et al. ACSNet: Adaptive context selection network for polyp segmentation [J]. Computer Methods and Programs in Biomedicine, 2022, 221: 107816.

[27] ISENSEE F, PETERSEN J, KOBLER M, et al. nnU-Net: a self-adapting framework for U-Net-based medical image segmentation [J]. Nature Methods, 2021, 18(2): 203-211.

[28] BERNAL J, ROMERO E, VAQUERO J, et al. A video database for colonoscopy polyp detection and removal [J]. Journal of Medical Imaging and Health Informatics, 2015, 5(1): 189-197.

[29] ESCALANTE-RAMÍREZ B, ÁLVAREZ-GARCÍA S, VAQUERO J, et al. ETIS-Larib Polyp DB: A database for polyp detection and segmentation in colonoscopy images [J]. Pattern Recognition Letters, 2017, 95: 11-19.

[30] JHA D, Smedsrud P H, Riegler M A, et al. Kvasir-SEG: A segmented polyp dataset [J]. arXiv preprint arXiv:2001.03006, 2020.

[31] NGUYEN T H, NGUYEN D T, LE H N, et al. BKAI-IGH/polyp: A large-scale polyp segmentation dataset [EB/OL]. Kaggle, 2021.

[32] JHA D, ALI S, RAJA A, et al. Kvasir-Sessile: A dataset for sessile polyp segmentation in colonoscopy [J]. Scientific Data, 2022, 9(1): 1-9.

[33] BORGLI H, THORSNES K, JHA D, et al. HyperKvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy [J]. Scientific Data, 2020, 7(1): 1-13.

[34] WANG W, LI B, XU J, et al. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions [C]//Proc. ICCV. 2021: 568-578.

[35] DEVLIN J, CHANG M W, LEE K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding [J]. arXiv preprint arXiv:1810.04805, 2018.

[36] BERNAL J, ROMERO E, VAQUERO J, et al. CVC-ColonDB: A database for the evaluation of colonic polyp detection, segmentation and classification algorithms [J]. Journal of Medical Systems, 2014, 38(1): 1-13.

[37] RADFORD A, WU J, CHILD R, et al. Learning transferable visual models from natural language supervision [J]. arXiv preprint arXiv:2103.00020, 2021.

[38] ZHANG Y, LIU J, WANG Z, et al. ColonFormer: Transformer-based colon polyp segmentation with global-local feature fusion [C]//Proc. MICCAI. 2023: 456-466.

[39] HUANG Y, LIN J, WANG Z, et al. HarDNet-MSEG: A lightweight network for medical image segmentation [J]. Computer Methods and Programs in Biomedicine, 2021, 203: 106287.

[40] SANDLER M, HOWARD A G, ZHU M, et al. MobileNetV2: Inverted residuals and linear bottlenecks [C]//Proc. CVPR. 2018: 4510-4520.

[41] TAN M, LE Q V. EfficientNet: Rethinking model scaling for convolutional neural networks [C]//Proc. ICML. 2019: 6105-6114.