Academic Journal of Science and Technology
ISSN: 2771-3032 | Vol. 9, No. 2, 2024

SegNet Network Architecture for Deep Learning Image Segmentation and Its Integrated Applications and Prospects

Chenwei Zhang[1, *], Wenran Lu[2], Jiang Wu[3], Chunhe Ni[4], Hongbo Wang[5]

[1]Electrical and Computer Engineering, University of Illinois Urbana-Champaign, Urbana, IL, USA
[2]Electrical Engineering, University of Texas at Austin, Austin, TX, USA
[3]Computer Science, University of Southern California, Los Angeles, CA, USA
[4]Computer Science, University of Texas at Dallas, Richardson, TX, USA
[5]Computer Science, University of Southern California, Los Angeles, CA, USA
[*] Corresponding author: zchenwei66@gmail.com

Abstract: Semantic image segmentation is a crucial task in computer vision, with applications ranging from autonomous driving to medical image analysis. In recent years, deep learning has revolutionized this field, leading to the development of various neural network models aimed at improving segmentation accuracy. One such architecture is SegNet, which we explore in this article.SegNet's architecture consists of an encoder network, a corresponding decoder network, and a pixel-wise classification layer. The encoder network, resembling VGG16 with 13 convolutional layers, extracts high-level features from input images. The innovation lies in the decoder network's approach to upsampling, utilizing pooled indices from the encoder's maximum pooling step to perform non-linear up sampling. This eliminates the need for additional learning during up sampling, making SegNet efficient in both storage and computation.SegNet represents an exciting advancement in deep learning image segmentation. Its efficient architecture, memory-conscious design, and potential for real-time applications make it a valuable tool in the field of computer vision with promising integrated applications and prospects.

Keywords: SegNet, Image segmentation, Deep learning, Computer vision.

1. Introduction

In recent years, with the widespread popularity of deep learning, numerous scholars have proposed various semantic segmentation algorithms based on different neural network models to enhance the accuracy of semantic segmentation. Semantic segmentation, a fundamental topic in computer vision, aims to assign semantic labels to each pixel in an image. Deep convolutional neural networks, particularly those based on fully convolutional neural networks, have demonstrated significant improvements over traditional systems that rely on manually crafted features.

The core of a trainable segmentation engine typically comprises an encoder network, a corresponding decoder network, and a pixel-by-pixel classification layer. The encoder network's architecture is topologically similar to that of the VGG16 network, consisting of 13 convolutional layers. The decoder network is responsible for mapping the low-resolution encoder feature map back to the full input resolution feature map for pixel-wise classification.

One of the noteworthy innovations introduced by SegNet is the approach it uses for upsampling the lower-resolution input feature map in the decoder. Specifically, the decoder performs nonlinear upsampling by using the pooled index calculated during the maximum pooling step of the corresponding encoder. This method eliminates the need for additional learning during the upsampling process. The upsampled map is sparse and is then convolved with a trainable filter to generate a dense feature map.

To evaluate the proposed architecture's performance, we conducted a comparison with widely used techniques such as FCN, as well as well-known architectures like DeepLab-LargeFOV and DeconvNet. This comparison reveals the trade-off between memory usage and segmentation accuracy, highlighting the importance of balancing these factors to achieve optimal segmentation performance.

SegNet was primarily inspired by applications related to scenario understanding. As a result, it was designed to be efficient in terms of both storage and computation time during inference. Additionally, it boasts a significantly smaller number of trainable parameters compared to competing architectures, and it can be trained end-to-end using stochastic gradient descent (SGD) optimization.Overall, SegNet represents a notable advancement in semantic segmentation, offering an efficient and effective approach for various computer vision tasks.

2. Related Work

"SegNet draws inspiration from scenario understanding applications, and as such, it is meticulously designed to excel in both storage and computational efficiency during inference. Furthermore, it boasts significantly fewer trainable parameters compared to competing architectures and can be trained end-to-end using stochastic gradient descent. We conducted controlled benchmark tests of SegNet and other architectures in road scene and SUN RGB-D indoor scene segmentation tasks. These quantitative evaluations demonstrate that, in comparison to other architectures, SegNet offers excellent performance while maintaining competitive inference times and highly efficient inference storage."

2.1. SegNet semantic segmentation

SegNet consists of a network of encoders and a network of

decoders, both of which are convolutional networks. The encoder network is typically extracted from a pre-trained VGG16 network with its fully connected layer removed. The decoder network is used to upsample the features obtained from the encoder network to produce an output of the same size as the input image. During the coding phase, when maximum pooling is performed, SegNet remembers the index of the maximum value of each pooled area. During the decoding phase, these indexes are used to upsample, providing a more accurate reconstruction of the structure. Since SegNet uses parameters only in the encoder and pooled indexes in the decoder for non-learning upsampling, it has fewer parameters than other models such as FCN.

(1) Using encoder-decoder method and multi-classification softmax classifier to calculate the classification confidence of each pixel, the image is segmented semantically;

(2) In the coding stage, only the maximum pooled index is recorded and stored, which greatly reduces the memory space occupied while depicting the boundary information;

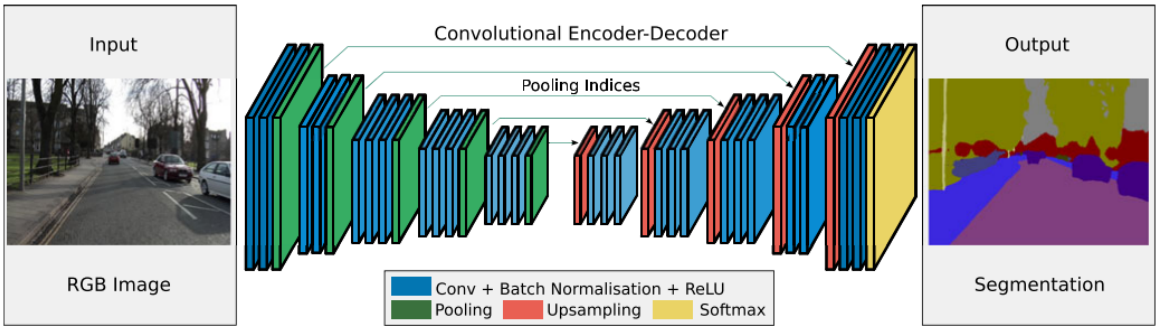

Figure 1. SegNet Network structure

The network structure is mainly composed of Encoder and Decoder. Its network structure is shown in the figure below. This network is adapted to VGG16, removing the last fully connected layer and retaining 13 convolutional layers, which exactly constitute the encoder network, that is, there are 5 coding blocks on the left (blue convolutional layer). Each block is downsampled through the maximum pooling layer (green pooling layer) and sent to the right side, then the feature map is up-sampled (red de-pooling layer), and then decoded through the decoding block (blue convolutional layer on the right side) after up-sampling, that is, decoded through the convolutional layer. Finally, the decoder output feature map is sent to Softmax classifier for pixelated classification. A completely symmetric coding-decoding structure is formed.

2.2. Pooling indices

"In the aforementioned description, the encoding layer employs the initial 13 layers of the VGG16 structure to extract features through convolution, batch normalization, and ReLU activation operations. Subsequently, downsampling is achieved using max-pooling layers with a kernel size of 2 and a stride of 2, ensuring translational invariance in the input images. However, such pooling and downsampling operations result in a loss of feature map resolution. As the network depth increases, the resolution of feature maps decreases, making it challenging to perform upsampling effectively and restore the fine details of the original image. Therefore, the authors have introduced optimizations within the encoder module.

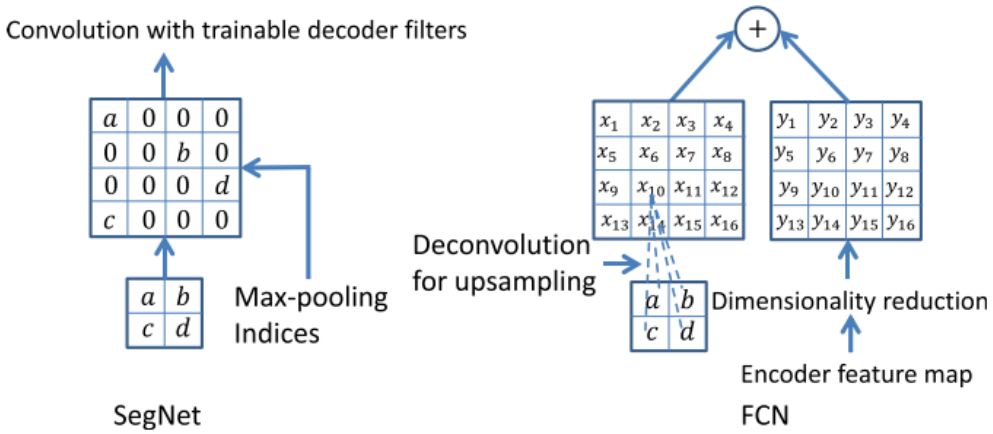

Figure 2. Deconvolution Deconvolution implementation based on FCN

To preserve crucial information during the downsampling process, the authors proposed a method for capturing and storing boundary information in the encoder's feature maps. This method involves preserving the pooling layer indices (referred to as pooling indices) and using positional information from the recorded pooling process instead of employing direct deconvolution operations. Specifically, as depicted in the figure below, SegNet employs max-pooling indices to map the features (a step that does not require learning) and then follows it with a trainable decoding filter (consisting of several convolutional layers). In contrast, FCN utilizes Deconvolution for

2.3. SegNet and deep learning applications

SegNet is a deep learning architecture for image segmentation. It improves FCN in some aspects. These

improvements not only improve the performance, but also reduce the computational complexity, reflecting the importance and application development of image segmentation technology based on deep learning. However, SegNet achieves upsampling by using the unpooling operation, rather than the deconvolution operation. These unpooling operations are performed based on the stored index without the need to learn parameters. This reduces the complexity and computation of the network, since deconvolution operations typically require learning a large number of parameters. Secondly, SegNet decoder uses the stored index when carrying out unpooling operation to maintain the integrity of high-frequency information.

In order to overcome the shortcomings of short-range CRF, fully connected CRF is adopted here. The basic form of its energy function (actually equivalent to the penalty function) is as follows:

$$E(\boldsymbol{x}) = \sum_i \theta_i(x_i) + \sum_{ij} \theta_{ij}(x_i, x_j) \quad (1)$$

$$\theta_{ij}(x_i, x_j) = \mu(x_i, x_j) \sum_{m=1}^{K} w_m \cdot k^m(\boldsymbol{f}_i, \boldsymbol{f}_j) \quad (2)$$

However, when unpooling a feature map with a lower resolution, the information between pixel neighbors may still be ignored. This shows that image segmentation remains a challenging task that requires a trade-off between computational efficiency and accuracy. The application of deep learning in the field of image segmentation has developed rapidly, and it has achieved great success in many fields. In medical image segmentation, for example, deep learning is already being used to identify tumors, organs, and lesions, helping doctors make more accurate diagnosis and treatment decisions. In the field of autonomous driving, image segmentation is used to detect and track roads, pedestrians and other vehicles, thereby improving vehicle safety and the performance of autonomous driving systems. In addition, image segmentation is also widely used in UAV, agriculture, geographic information system and other fields.

In conclusion, the application development of image segmentation technology based on deep learning is very important in different fields, which not only improves the accuracy of segmentation, but also makes the calculation more efficient. As the technology continues to evolve, we can expect more innovations and improvements to make image segmentation technology more successful in various fields.

3. Methodology

This paper mainly introduces a lightweight medical image segmentation network named SegNetr, and rethinks and improves the local - global interaction and long jump connection operation in traditional codec networks. In the field of medical image segmentation, U-Net type network has basically become the mainstream. However, U-Net still has some shortcomings in capturing long-distance context dependencies based on convolutional operations, which increases the parameter and computational complexity of the network or simply integrates the features of the encoder and decoder, ignoring the correlation between their spatial positions.

To solve the above problems, the paper introduces a novel SegNetr block, which can dynamically perform local-global interaction at any stage and has linear complexity

3.1. The parameters and model of SegNetr

SegNetr is a typical hierarchical U network, which includes SegNetr block and IRSC two important components. In order to make the network more lightweight, the authors base on MBConv as the underlying convolution building block. The SegNetr block enables dynamic local-global interactions at the encoder and decoder stages. Use patch merging to reduce resolution by up to twice without losing the original image information.

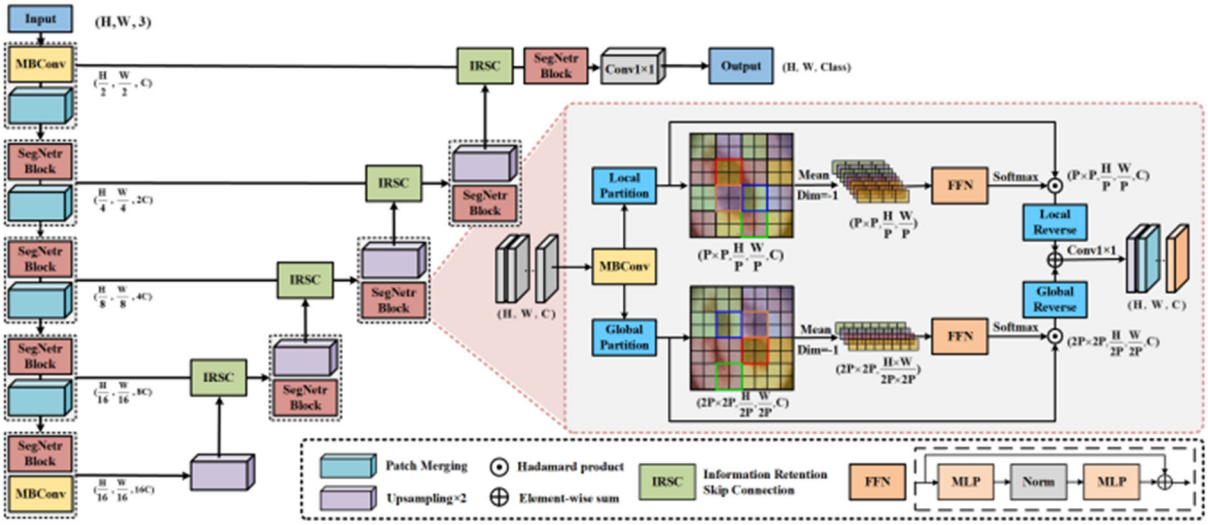

Figure 3. SegNetr model structure

In addition, IRSC is used in practical applications to fuse the characteristics of the encoder and decoder, reducing the loss of detail information as the network increases in depth.

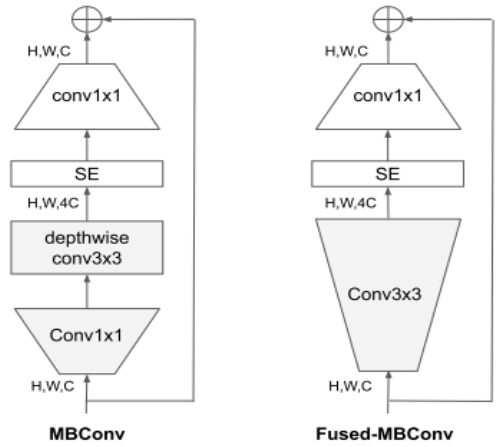

Figure 4. MBConv and Fused-MBConv in EfficientNetV2

3.2. Context capture

SegNetr is the core component of the entire network. It implements dynamic processing of features through local-global interaction. It uses MBConv as the base convolution module and introduces local and global branches to enable interaction. Therefore, the local interaction is realized by calculating the attention matrix of non-overlapping small patches when implementing the local branch of context capture. Global branches realize global interaction through aggregation and displacement of discontinuous patches in space. Local and global branches are finally fused by weighted summation. This design not only reduces the computational complexity, but also better captures the local and global information in the image.

The main process structure is as follows:

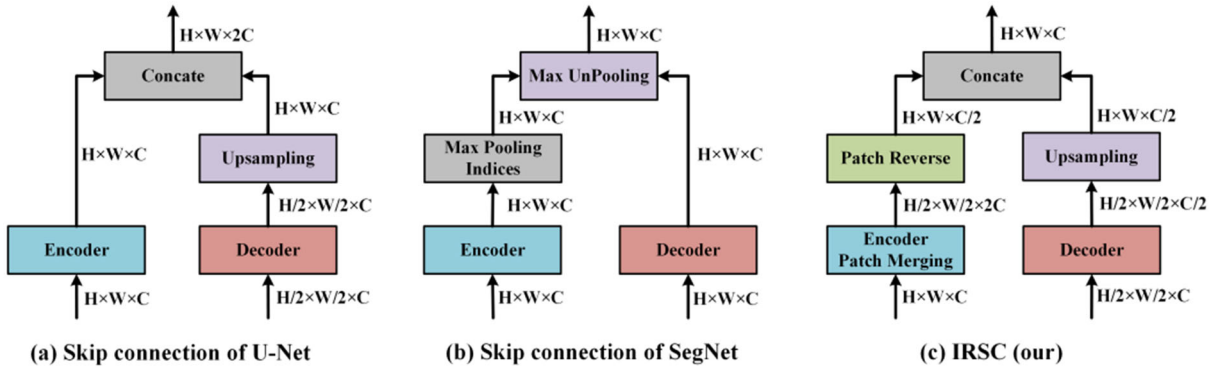

Figure 5. Context capture structure diagram

The information retention skip connection realizes the fusion of encoder and decoder features by Patch Merging and Patch Reverse. Among them, the specific operation of Patch Merging reduces the resolution of the input feature map, while expanding the channel dimension to retain more high-resolution detail information. Patch Reverse is used to restore the spatial resolution of the encoder and fuse it with the upsampled features of the decoder. In this way, the detail and position information of the feature map can be recovered better, and the accuracy of segmentation can be improved.

3.3. Data set

SegNetr and TransUNet achieved the highest IoU (0.775), 3.9% higher than the benchmark U-Net. Even SegNetr-S with fewer parameters can achieve similar segmentation performance to UNeXt-L. On the PH2 dataset, we observed that the Transformer-based method Swin-UNet has the worst segmentation performance, which is directly related to the amount of data in the target dataset. The proposed method achieves the best segmentation performance on the data set and keeps the computation cost low.

Table 1. Quantitative results on TNSCUI and ACDC datasets.

Network	TNSCUI	ACDC / IoU			Average	Params	GFLOPs
	IoU (Dice)	RV	Myo	LV	IoU (Dice)		
U-Net [1]	0.718 (0.806)	0.743	0.717	0.861	0.774 (0.834)	29.59 M	41.83
SegNet [20]	0.726 (0.819)	0.738	0.720	0.864	0.774 (0.836)	17.94 M	22.35
FAT-Net [18]	0.751 (0.842)	0.743	0.702	0.859	0.768 (0.834)	28.23 M	42.83
Swin-UNet [13]	0.744 (0.835)	0.754	0.722	0.865	0.780 (0.843)	25.86 M	5.86
TransUNet [12]	0.746 (0.837)	0.750	0.715	0.866	0.777 (0.838)	88.87 M	24.63
EANet [30]	0.751 (0.839)	0.742	0.732	0.864	0.779 (0.839)	47.07 M	98.63
UNeXt [15]	0.655 (0.749)	0.697	0.646	0.814	0.719 (0.796)	1.40 M	0.44
UNeXt-L [15]	0.693 (0.794)	0.719	0.675	0.840	0.744 (0.815)	3.80 M	1.08
SegNetr-S	0.707 (0.804)	0.723	0.692	0.845	0.753 (0.821)	3.60 M	2.71
SegNetr	0.767 (0.850)	0.761	0.738	0.872	0.791 (0.847)	12.26 M	10.18

Table results show that SegNetr's IoU and Dice are 1.6% and 0.8% higher than FATNet, respectively, while GFLOPs are 32.65 lower. In the ACDC dataset, the segmentation of the left ventricle is relatively easy, and U-Net has an IoU of 0.861, but it is 1.1% worse than SegNetr. The myocardium is located between the left and right ventricles in a circular pattern, and the IoU of the proposed method is 0.6% higher than the EANet focused on boundary segmentation.

3.4. Experiment Results

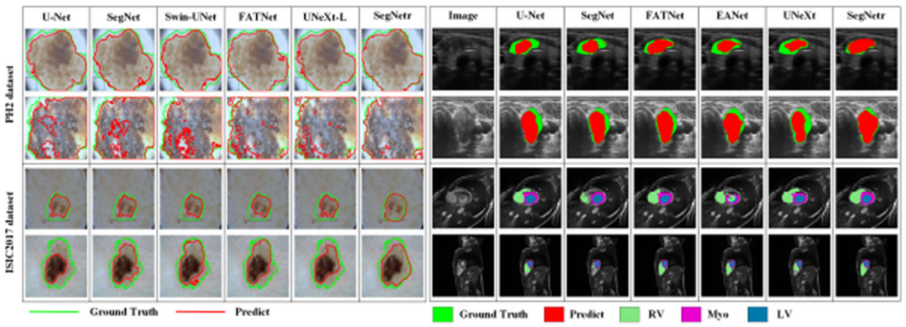

Figure 6. SegNetr Image segmentation results

As shown in the figure, it can be seen that SegNetr can accurately describe skin lesions with less data and achieve multi-class segmentation, minimizing the cases of undersegmentation and over-segmentation.

The experimental results show that SegNetr improves the segmentation performance of U-shaped networks by introducing SegNetr blocks and information retention skip connections. Among them, SegNetr blocks achieve better feature representation through local-global interaction, while information retention skip connections provide better feature fusion mechanism. These methods enable SegNetr to achieve segmentation performance comparable to or even better than traditional methods while reducing computational complexity.

4. Conclusion

This paper discusses one of the important tasks of deep learning in the field of computer vision, namely semantic image segmentation. Semantic image segmentation has a wide range of applications, including autonomous driving and medical image analysis. Deep learning has revolutionized the field in recent years, driving the development of various neural network models to improve segmentation accuracy. One such architecture is SegNet, which is discussed in this article. The architecture of SegNet consists of an encoder network, a corresponding decoder network, and a pixel-level classification layer. The encoder network is similar to VGG16, with 13 convolutional layers for extracting high-level features from input images. The innovation of decoder network lies in its upsampling method, which utilizes the pooling index in the encoder maximum pooling step to perform nonlinear upsampling. This eliminates the need for additional learning during the upsampling process, making SegNet more efficient in both storage and computational efficiency.

SegNet represents a significant advance in the field of deep learning image segmentation. Its efficient architecture, memory-optimized design, and potential for real-time applications make it a valuable tool in the field of computer vision with many potential integrated applications and prospects. Finally, SegNet, as an effective and efficient method, improves segmentation performance while reducing computational complexity, providing new possibilities for various tasks in the field of computer vision.

References

[1] "Implementation of Computer Vision Technology Based on Artificial Intelligence for Medical Image Analysis". International Journal of Computer Science and Information Technology, vol. 1, no. 1, Dec. 2023, pp. 69-76, https://doi.org/10.62051/ijcsit.v1n1.10.

[2] Dong, Xinqi, et al. "The Prediction Trend of Enterprise Financial Risk Based on Machine Learning ARIMA Model". Journal of Theory and Practice of Engineering Science, vol. 4, no. 01, Jan. 2024, pp. 65-71, doi:10.53469/jtpes.2024.04(01).09.

[3] "A Deep Learning-Based Algorithm for Crop Disease Identification Positioning Using Computer Vision". International Journal of Computer Science and Information Technology, vol. 1, no. 1, Dec. 2023, pp. 85-92, https://doi.org/10.62051/ijcsit.v1n1.12.

[4] Wang, Sihao, et al. "Diabetes Risk Analysis Based on Machine Learning LASSO Regression Model". Journal of Theory and Practice of Engineering Science, vol. 4, no. 01, Jan. 2024, pp. 58-64, doi:10.53469/jtpes.2024.04(01).08.

[5] "Based on Intelligent Advertising Recommendation and Abnormal Advertising Monitoring System in the Field of Machine Learning". International Journal of Computer Science and Information Technology, vol. 1, no. 1, Dec. 2023, pp. 17-23, https://doi.org/10.62051/ijcsit.v1n1.03.

[6] Yu, Liqiang, et al. "Research on Machine Learning With Algorithms and Development". Journal of Theory and Practice of Engineering Science, vol. 3, no. 12, Dec. 2023, pp. 7-14, doi:10.53469/jtpes.2023.03(12).02.

[7] Huang, J., Zhao, X., Che, C., Lin, Q., & Liu, B. (2024). Enhancing Essay Scoring with Adversarial Weights Perturbation and Metric-specific AttentionPooling. arXiv preprint arXiv:2401.05433.

[8] Tan, Kai, et al. "Integrating Advanced Computer Vision and AI Algorithms for Autonomous Driving Systems". Journal of Theory and Practice of Engineering Science, vol. 4, no. 01, Jan. 2024, pp. 41-48, doi:10.53469/jtpes.2024.04(01).06.

[9] Tianbo, Song, Hu Weijun, Cai Jiangfeng, Liu Weijia, Yuan Quan, and He Kun. "Bio-inspired Swarm Intelligence: a Flocking Project With Group Object Recognition." In 2023 3rd International Conference on Consumer Electronics and Computer Engineering (ICCECE), pp. 834-837. IEEE, 2023.DOI: 10.1109/mce.2022.3206678

[10] K. Jin, Z. Z. Zhong and E. Y. Zhao, "Sustainable Digital Marketing Under Big Data: An AI Random Forest Model Approach," in IEEE Transactions on Engineering Management, vol. 71, pp. 3566-3579, 2024, doi: 10.1109/TEM.2023.3348991.

[11] "The Application of Artificial Intelligence in Medical Diagnostics: A New Frontier". Academic Journal of Science and Technology, vol. 8, no. 2, Dec. 2023, pp. 57-61, https://doi.org/10.54097/ajst.v8i2.14945

[12] Pan, Yiming, et al. "Application of Three-Dimensional Coding Network in Screening and Diagnosis of Cervical Precancerous Lesions". Frontiers in Computing and Intelligent Systems, vol. 6, no. 3, Jan. 2024, pp. 61-64, https://doi.org/10.54097/mi3VM0yB.

[13] Liu, B., Zhao, X., Hu, H., Lin, Q., & Huang, J. (2023). Detection of Esophageal Cancer Lesions Based on CBAM Faster R-CNN. Journal of Theory and Practice of Engineering Science, 3(12), 36–42. https://doi.org/10.53469/jtpes.2023.03(12).06

[14] Liu, Bo, et al. "Integration and Performance Analysis of Artificial Intelligence and Computer Vision Based on Deep Learning Algorithms." arXiv preprint arXiv:2312.12872 (2023).

[15] Liu, Bo, et al. "Integration and Performance Analysis of Artificial Intelligence and Computer Vision Based on Deep Learning Algorithms." arXiv preprint arXiv:2312.12872 (2023).

[16] Yu, L., Liu, B., Lin, Q., Zhao, X., & Che, C. (2024). Semantic Similarity Matching for Patent Documents Using Ensemble BERT-related Model and Novel Text Processing Method. arXiv preprint arXiv:2401.06782.

[17] "The Application of Artificial Intelligence to The Bayesian Model Algorithm for Combining Genome Data". Academic Journal of Science and Technology, vol. 8, no. 3, Dec. 2023, pp. 132-5, https://doi.org/10.54097/ykhccb53.

[18] Jin, Keyan. "Impacts of Word of Mouth (WOM) on E-Business Online Pricing." JGIM vol.31, no.3 2023: pp.1-17. http://doi.org/10.4018/JGIM.324813

[19] Wei, Kuo, et al. "Strategic Application of AI Intelligent Algorithm in Network Threat Detection and Defense". Journal of Theory and Practice of Engineering Science, vol. 4, no. 01, Jan. 2024, pp. 49-57, doi:10.53469/jtpes.2024.04(01).07.

[20] Du, Shuqian, et al. "Application of HPV-16 in Liquid-Based Thin Layer Cytology of Host Genetic Lesions Based on AI Diagnostic Technology Presentation of Liquid". Journal of Theory and Practice of Engineering Science, vol. 3, no. 12, Dec. 2023, pp. 1-6, doi:10.53469/jtpes.2023.03(12).01.

[21] Xin, Q., He, Y., Pan, Y., Wang, Y., & Du, S. (2023). The implementation of an AI-driven advertising push system based on a NLP algorithm. International Journal of Computer Science and Information Technology, 1(1), 30-37.0

[22] Pan, Yiming, et al. "Application of Three-Dimensional Coding Network in Screening and Diagnosis of Cervical Precancerous Lesions". Frontiers in Computing and Intelligent Systems, vol. 6, no. 3, Jan. 2024, pp. 61-64, https://doi.org/10.54097/mi3VM0yB.

[23] "Enhancing Computer Digital Signal Processing through the Utilization of RNN Sequence Algorithms". International Journal of Computer Science and Information Technology, vol. 1, no. 1, Dec. 2023, pp. 60-68, https://doi.org/10.62051/ijcsit.v1n1.09.