SSA-ECNet:Semantic Segmentation Architecture with Enhanced Cross-Attention Mechanism

Minghui Li
School of Mathematics and Computing Science
Guilin University of Electronic Technology
Guilin, China
calyli@mails.guet.edu.cn

Zengmin Xu[1,2,*]
[1]School of Mathematics and Computing Science, Guilin University of Electronic Technology
[2]Anview.ai
Guilin, China
[*]Corresponding author: xzm@guet.edu.cn

Ruxing Meng
Anview.ai
Guilin, China
mengruxing@163.com

Lingli Wei
School of Artificial Intelligence
Guilin University of Electronic Technology
Guilin, China
lingliwei@mails.guet.edu.cn

Weisen Luo
School of Mathematics and Computing Science
Guilin University of Electronic Technology
Guilin, China
18877075712@163.com

2024 7th International Conference on Advanced Algorithms and Control Engineering (ICAACE) | 979-8-3503-6144-5/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICAACE61206.2024.10548981

***Abstract*—This paper presents a novel Convolutional Neural Network (CNN) semantic segmentation architecture for detecting water leakage defects in house images. Recent semantic segmentation architectures have predominantly focused on RGB images, where water leakage traces are often vague and surface features insufficiently distinct. Traditional semantic segmentation architectures exhibit insufficient edge clarity.**

These challenges have spurred the proposal of an enhanced model for multispectral image segmentation. To benchmark our approach, we established an RGB thermal dataset and devised a new fused image attention module to better extract features. Our findings indicate a significant improvement in segmentation accuracy by incorporating thermal infrared information.

Keywords-Semantic Segmentation, Information Fusion, House defects, Convolutional Neural Network

I. INTRODUCTION

Long-term water leaks in houses can cause damage to the structural integrity, particularly in the case of wooden structures or certain types of building materials. Damp environments provide ideal conditions for the growth of mold and fungi. These organisms not only damage the interior of houses but can also pose health threats to the occupants, especially those with respiratory issues or allergies. Furthermore, severe cases can lead to significant safety hazards, including risks of electric shock and fire. To detect the damage in houses, thermal imaging cameras are used. In recent years, artificial intelligence has developed rapidly[1], and deep neural networks can analyze such thermal images. This non-destructive testing, based on infrared thermography, utilizes the physical property differences between materials. By examining the infrared wave field emitted by objects, the resulting thermal images can conveniently identify defects in building structures.

Deep neural networks are frequently utilized by researchers in areas such as defect detection, where these methods can bring industries highly efficient detection speeds and outstanding quality. The advent of CNN networks has accelerated the implementation of application projects [2] [3] [4]. However, these methods are unable to precisely locate defects, and the enhancement of accuracy is achieved by increasing the size of CNNs, implying that the time complexity of state-of-the-art CNN architectures is becoming increasingly large. These limitations have spurred the development of some more effective end-to-end networks.

In this article, we introduce a new network model for semantic segmentation, which integrates RGB images with thermal imaging to achieve robust and accurate semantic segmentation in architectural water seepage defect scenarios. Our primary work focuses on enhancing the Multi-spectral Fusion

Networks (MFNet) framework, employing AIC-AM that dynamically concentrates on pertinent features within both image modalities, highlighting areas with more pronounced defects. We have also maintained the encoder-decoder architecture [5]. Our main contributions are as follows:

1)We have created a semantic segmentation dataset for house water seepage defects and developed a novel deep neural network that integrates RGB images and infrared images for scenarios involving house water seepage defects.

2)We propose an improved loss function and AIC-AM, and our experiments demonstrate the effectiveness of our approach.

3)Comparative analyses with existing network models show that our network exhibits superior performance.

II. Related Work

In recent years, Zeng et al. [6] introduced an enhanced multiscale feature fusion method that improves the performance of small object detection. Yu et al. [7] proposed an efficient scale-aware network (ES-Net) to improve the effect of defect detection. These methods utilize object detection techniques to categorize defects. They are sensitive to data variations, and a single image cannot provide a deep analysis of defects. Researchers have proposed numerous network fusions to compensate for the lack of texture detail in RGB images. PIAFusion introduced a Comprehensive Mobility Discharge Assessment Framework (CMDAF) and a mid-way fusion strategy to integrate complementary information. CUFD [8] employed dual encoder-decoder networks to decompose the feature maps of infrared and visible light images into common and unique components.

Recent researchers have begun to use RGB images and infrared images as inputs for semantic segmentation networks. Pozzer et al. [9] utilized various deep neural network models to detect primary concrete anomalies in thermal and regular images, including defects such as spalling, cracks, and patches. MFNet [10] introduced a novel convolutional neural network architecture for semantic segmentation in multispectral scenes, considering the balance between performance and time consumption. RTFNet [11] demonstrates that utilizing thermal information can enhance semantic segmentation performance.

III. Network Framework

This section introduces our CNN network architecture for semantic segmentation, SSA-ECNet, which includes an enhanced cross-attention mechanism framework and a novel loss function we propose.

A. Network Structure

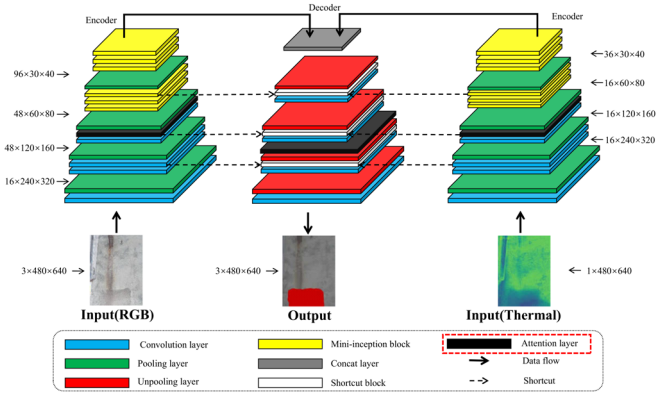

Figure 1. The diagram of the enhanced SSA-ECNet architecture, based on the MFNet [10], is presented. Multispectral images are partitioned into RGB and infrared images, serving as inputs, and segmentation images are produced as outputs. The mini-Inception blocks and shortcut connection blocks maintain the structural integrity of MFNet. We enhance the model's feature representation capability in both channels and spatial dimensions by incorporating an improved attention mechanism layer in Fig.3.

Our network framework is illustrated in Fig.1, where SSA-ECNet employs an encoder-decoder architecture. We have designed two encoders to extract features from RGB images and infrared images, respectively. The architectures of these encoders are identical, with the only difference being the number of input and output channels in the convolutional layers.

In the later stages of the encoder, we have incorporated the 'mini-inception' block proposed by MFNet, which utilizes dilated convolutions. We employ a 'shortcut block' in Fig.2 and then pass the upsampled feature maps through convolutional layers to generate dense feature maps. Batch normalization is applied after each convolutional layer. Before the third pooling layer in the encoding process, we integrate AIC-AM, as shown in Fig.3. Before the third upsampling in the decoder, we also introduce this mechanism.

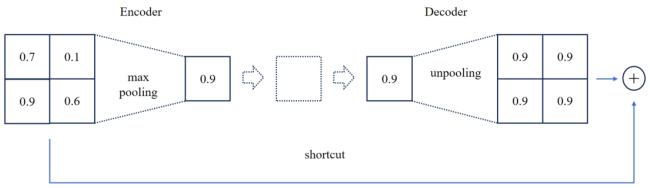

Figure 2. The diagram of the proposed SSA-ECNet architecture is shown. Multispectral images are split into RGB images and infrared images as inputs, and segmentation images are produced as outputs. The mini-inception and shortcut blocks still adhere to the structure of MFNet. In Fig.3, we illustrate the Attention layer introduced in our work.

B. A New Loss Function

In this section, we present a novel loss function that enhances the traditional Dice loss [12], by introducing category weights. The formula is depicted as follows (1):

$$\text{Dice Loss} = 1 - \frac{2 \cdot \sum_i (P_i \cdot T_i) + \varepsilon}{\sum_i P_i^2 + \sum_i T_i^2 + \varepsilon} \tag{1}$$

P_i and T_i represent the binary values of predicted and true labels, respectively. ε is a small constant introduced to prevent the denominator from being zero, thereby avoiding numerical stability issues. The conventional Dice loss function is suitable for situations with imbalanced pixel quantities and smaller target objects. It is more applicable to image segmentation tasks than the traditional cross-entropy loss function. However, due to its lack of consideration for distinctions between categories, it struggles to effectively differentiate between similar pixel values.

Additionally, we place a significant emphasis on the accuracy of segmentation edges, a critical factor for overall performance. The implementation of a dynamic smooth term allows for the adaptive adjustment of the smoothing coefficient, catering to targets of varying sizes. Lastly, the integration of $L_{2,1}$ -norm promotes row-level sparsity within the model parameters, mitigating overfitting and augmenting the model's generalization capability(2).

$$\text{Loss} = 1 - \frac{2 \cdot \sum_i (P_i \cdot T_i \cdot V_i) + S}{\sum_i (P_i \cdot V_i) + \sum_i (T_i \cdot V_i) + S} + \lambda_{2,1} \| \Theta \|_{2,1} \tag{2}$$

In this improved Dice loss function, $V_i = W_i \cdot E_i$, W_i denotes the class weight at position i, utilized to address class

imbalance issues. E_i is the edge weight, emphasizing the significance of image edges. We have opted to abandon ε as the smoothing coefficient. Instead, the smooth term S is defined as $S = smooth \cdot N$, where $smooth$ is a smoothing parameter, and N is the total number of elements in the target values. $\lambda_{2,1}$ is the coefficient for the $L_{2,1}$ regularization, controlling the strength of regularization. The $L_{2,1}$ regularization term, defined as $\| \Theta \|_{2,1} = \sum_i \sqrt{\sum_j \Theta_{ij}^2}$, encourages row-level sparsity in the model parameters, where Θ represents the parameters of the model.

C. A Improve Cross-Attention Mechanism

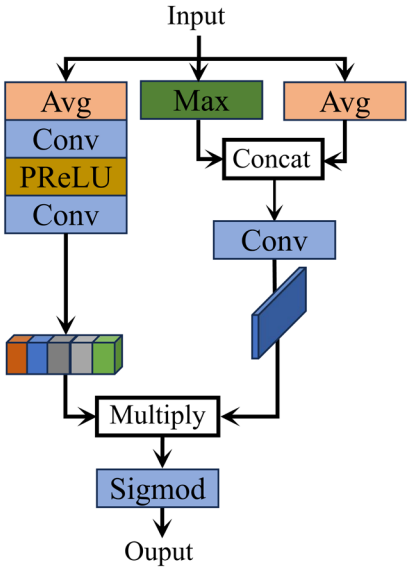

Figure 3. Incorporating A Improve Cross-Attention Mechanism (AIC-AM) enables the generation of the final attention weight maps. Within the encoder, RGB images are employed as input, while in the decoder, the input transitions to infrared images.

The framework for enhancing the Cross-Attention Mechanism, as illustrated in Fig.3, draws inspiration from the Convolutional Block Attention Module (CBAM) proposed by Woo et al. [13], which is utilized to integrate deep features of different modal images. Diverging from CBAM, we introduce the use of convolutional operations to replace the originally shared multi-layer perceptron (MLP), and connection operations replace the addition layer. This modification is motivated by the work of Wang et al. [14]. We present a novel parallel attention framework, which is also designed to be readily applicable as a plug-and-play module.

In the channel domain, initially, through the mean pooling layer, these vectors are then passed through two convolutional layers and a PReLU activation layer, connected, and fed into a convolutional layer to generate channel attention vectors. In the spatial domain, we obtain the initial spatial attention matrices through maximum and average pooling operations. These matrices are then concatenated into the convolutional layer to produce spatial attention matrices.

Subsequently, we element-wise multiply the channel attention vectors with the spatial attention matrices to obtain attention maps for the initial fused features. Next, we normalize these attention maps by applying the Sigmoid activation function to generate corresponding attention weights.

IV. Experimental Results and Discussions

A. The Dataset

In this section, we evaluate our proposed SSA-ECNet by training and predicting on our custom dataset. We compare it with networks proposed in recent years.

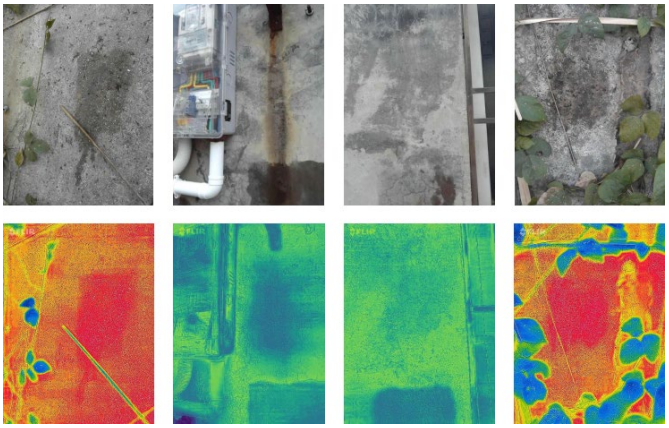

Figure 4. The top row contains four images in RGB format, whereas the bottom row has four images captured in the infrared spectrum. The infrared images with a reddish tint were taken in the summer, while those with a greenish tint were obtained in the autumn.

The data in Fig.4 are collected from rural houses built in the vicinity of cities for more than a decade. Due to their long construction history and lack of regular maintenance, such houses are prone to defects such as water leakage. The FLIR ONE PRO thermal imaging device was utilized for data collection, resulting in a total of 400 images. The image resolution on the dataset is 640 × 480.

B. Training Details

We use PyTorch 1.3.1, and our SSA-ECNet is trained on a machine equipped with an Intel 2.6GHz i5 CPU and an NVIDIA 4090 GPU running the Ubuntu operating system. The training involves the utilization of the simulated annealing algorithm with an initial temperature of 0.01 degrees, a final temperature of 0.0001 degrees, and a cooling rate of 0.9. We employ the Stochastic Gradient Descent (SGD) optimizer. Additionally, flip augmentation techniques are applied to enhance the training dataset. The network is trained until convergence, and no further reduction in loss is observed at the point of convergence.

C. Evaluation Metrics

We employ quantitative assessments for the semantic segmentation performance using two metrics. The second is the Intersection over Union (IoU) for each class (3). The first is the accuracy (Acc) for each class (4), also known as recall. The average values for all classes for both metrics are denoted as mIoU and mAcc, respectively. Their computation formulas are as follows:

$$\mathrm{mAcc} = \frac{1}{N}\sum_{i=1}^{N}\frac{TP_i}{TP_i + FN_i} \tag{3}$$

$$\mathrm{mIoU} = \frac{1}{N}\sum_{i=1}^{N}\frac{TP_i}{TP_i + FP_i + FN_i} \tag{4}$$

True Positive (TP) signifies the number of accurately predicted pixels or instances belonging to a specific class. False Negative (FN) denotes the count of pixels or instances genuinely belonging to a class but incorrectly predicted otherwise. False Positive (FP) represents the number of pixels or instances inaccurately predicted as belonging to a class when they do not.

D. Comparative Results

In this section, we compare SSA-ECNet with UNet[15], PSPNet[16], HRNet[17], SegNet, RTFNet, and MFNet. All networks have been trained until their loss converges to that of our SSA-ECNet. The result is in Table 1.

TABLE I. TEST SET PREDICTION RESULTS

Model	mAcc(%)	mIou(%)
SegNet	83.6	72.6
UNet	81.2	73.7
PSPNet	85.7	77.5
HRNet	86.9	74.3
RTFNet	88.9	79.7
MFNet	93.1	87.8
SSA-ECNet	**96.8**	**89.6**

V. CONCLUSIONS

In this paper, we present a novel CNN framework designed for the semantic segmentation of RGB thermal images to identify house leak defects. We introduce a new multispectral dataset with pixel-level annotations to facilitate the evaluation of segmentation performance. Our proposed method demonstrates higher precision when compared to state-of-the-art segmentation approaches. We devised a novel multimodal cross-attention mechanism to effectively extract deep features from both RGB images and infrared images. Additionally, we designed a more efficient loss function that incorporates edge pixel considerations. The improved smoothness term aids in achieving a better balance between small and large objects, enhancing overall segmentation performance. The utilization of the $L_{2,1}$ -norm further reinforces the robustness of the model.

ACKNOWLEDGMENT

This research was supported by the National Natural Science Foundation of China (61862015), the Guangxi Natural Science Foundation under grant (2024GXNSFAA010493), the National College Student Innovation Training Program (202310595053), and the Science and Technology Project of Guangxi (AD23023002).

REFERENCES

[1] Y. Jiang, X. Li, H. Luo, S. Yin, and O. Kaynak, "Quo vadis artificial intelligence?" Discover Artificial Intelligence, vol. 2, no. 1, p. 4, 2022.

[2] G. Kaur, R. Sinha, P. K. Tiwari, S. K. Yadav, P. Pandey, R. Raj, A. Vashisth, and M. Rakhra, "Face mask recognition system using cnn model," Neuroscience Informatics, vol. 2, no. 3, p. 100035, 2022.

[3] X. Xu, M. Zhao, P. Shi, R. Ren, X. He, X. Wei, and H. Yang, "Crack detection and comparison study based on faster r-cnn and mask r-cnn," Sensors, vol. 22, no. 3, p. 1215, 2022.

[4] W. Wang, Y. Chen, and P. Ghamisi, "Transferring cnn with adaptive learning for remote sensing scene classification," IEEE Transactions on Geoscience and Remote Sensing, vol. 60, pp. 1–18, 2022.

[5] S. Dalmia, D. Okhonko, M. Lewis, S. Edunov, S. Watanabe, F. Metze, L. Zettlemoyer, and A. Mohamed, "Legonn: Building modular encoder-decoder models," IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2023.

[6] N. Zeng, P. Wu, Z. Wang, H. Li, W. Liu, and X. Liu, "A small- sized object detection oriented multi-scale feature fusion approach with application to defect detection," IEEE Transactions on Instrumentation and Measurement, vol. 71, pp. 1–14, 2022.

[7] X. Yu, W. Lyu, D. Zhou, C. Wang, and W. Xu, "Es-net: Efficient scale-aware network for tiny defect detection," IEEE Transactions on Instrumentation and Measurement, vol. 71, pp. 1–14, 2022.

[8] H. Xu, M. Gong, X. Tian, J. Huang, and J. Ma, "Cufd: An encoder–decoder network for visible and infrared image fusion based on common and unique feature decomposition," *Computer Vision and Image Understanding*, vol. 218, p. 103407, 2022.

[9] S. Pozzer, E. Rezazadeh Azar, F. Dalla Rosa, and Z. M. Chamberlain Pravia, "Semantic segmentation of defects in infrared thermographic images of highly damaged concrete structures," *Journal of Performance of Constructed Facilities*, vol. 35, no. 1, p. 04020131, 2021.

[10] Q. Ha, K. Watanabe, T. Karasawa, Y. Ushiku, and T. Harada, "Mfnet: Towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2017, pp. 5108–5115.

[11] Y. Sun, W. Zuo, and M. Liu, "Rtfnet: Rgb-thermal fusion network for semantic segmentation of urban scenes," *IEEE Robotics and Automation Letters*, vol. 4, no. 3, pp. 2576–2583, 2019.

[12] X. Li, X. Sun, Y. Meng, J. Liang, F. Wu, and J. Li, "Dice loss for data-imbalanced nlp tasks," *arXiv preprint arXiv:1911.02855*, 2019.

[13] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.

[14] Z. Wang, W. Shao, Y. Chen, J. Xu, and L. Zhang, "A cross-scale iterative attentional adversarial fusion network for infrared and visible images," IEEE Transactions on Circuits and Systems for Video Technology, 2023.

[15] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. Springer, 2015, pp. 234–241.

[16] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2881–2890.

[17] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang *et al.*, "Deep high-resolution representation learning for visual recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 10, pp. 3349–3364, 2020.