Inverse Pyramid Pooling Attention for Ultrasonic Image Signal Recognition

Zhiwen Jiang[1], Ziji Ma[1,2(✉)], Xianglong Dong[1], Qi Wang[1,2], and Xun Shao[3]

[1] College of Electrical and Information Engineering, Hunan University, Changsha 410082, China
zijima@hnu.edu.cn

[2] Greater Bay Area Insititute for Innovation, Hunan University, Guangzhou 511340, China

[3] Department of Electrical and Electronic Information Engineering, Toyohashi University of Technology, Toyohashi 4418580, Japan

Abstract. Ultrasound is commonly used for diagnosis and detection in a variety of fields, and the analysis of ultrasound echo signals presents a significant challenge in terms of the amount of time required by professionals to make subjective judgements. With the advances made in artificial intelligence technology on computers, more and more fields are being aided by it, not only increasing efficiency but also improving overall accuracy. In this paper, an inverse pyramid pooling of attention (IPPA) mechanism is proposed for images transformed from ultrasound echo signals. IPPA performs different pooling operations at multiple scale levels for each channel of the feature matrix, obtaining rich regional feature associations and thus improving the representation of the channels. In addition, different probability factors were assigned for the different pooling, and domain channel information was extracted by adaptive 1D convolution to enhance the adaptation range of the network model. Experimental results on a 10-class ultrasound hyperdata set (consisting of three sub-datasets) show that the sensitivity and robustness of the ResNet integrated with IPPA are improved over the original ResNet, with an accuracy of up to 99.68%.

Keywords: Convolutional neural network · Inverse pyramid attention · Combinatorial pooling · Ultrasound signal recognition

1 Introduction

Detection using ultrasonic signals is an important non-destructive testing technique that is now used in a variety of fields and has played a significant role in

This work is supported in part by the National Nature Science Foundation of China under Grant 61971182, in part by Nature Science Foundation of Hunan Province 2021JJ30145, in part by Hunan Province Enterprise Science and technology commissioner program 2021GK5021, in part by Guangxi key R&D plan project 2022AB41020.

C. Wu et al. (Eds.): MONAMI 2023, LNICST 559, pp. 94–107, 2024.
https://doi.org/10.1007/978-3-031-55471-1_8

several areas. In agriculture, ultrasound is commonly used to treat seeds, activate dormant seeds, and improve seed quality prior to planting [1,2]. Ultrasonic flaw detection, as a common non-destructive testing method in industrial inspection, is very frequently used. It utilizes ultrasonic echoes to reflect the internal properties of materials and can identify potential hidden problems [3]. In the medical field, ultrasound is widely used and considered a mature technology. Its development time is also longer. Ultrasound is used as a routine examination technique in medicine to scan a wide range of tissues and organs in the human body, providing a great deal of valid diagnostic value while causing less harm to the body [4–6].

Recently, artificial intelligence techniques, including deep learning, have been developed with the arithmetic power, algorithms and data base of computer technology [7–9]. The technology in the use of ultrasonic signal detection is also becoming digital and intelligent. Typically, the ultrasonic echo signal is converted into an image, and the interpretation of the target is performed by a trained professional relying on the ultrasound image. However, this approach presents a significant challenge in terms of diminishing accuracy and efficiency.

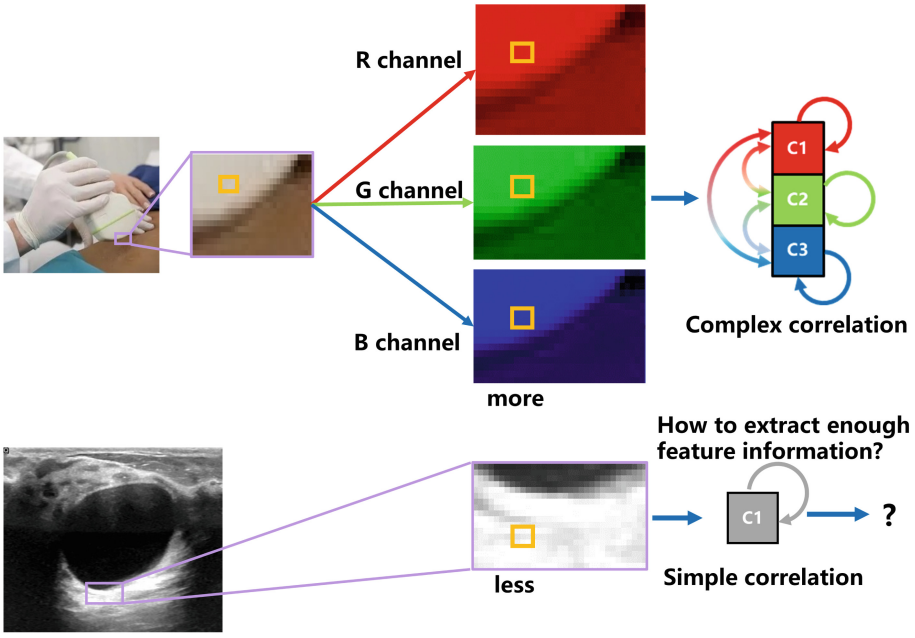

Fig. 1. Structural Disparities in Ultrasound and Optical Imaging Representations. In contrast to optical imaging, the ultrasound image (bottom left) has only a single channel and the pixel points can only contain neighbourhood relationships, lacking correlation information between channels and across channel neighbourhoods.

In addition, the image obtained after ultrasound imaging is a single-channel grayscale image, in contrast to the three-channel color image produced by conventional light imaging. Moreover, while natural light imaging captures multiple frequency bands of reflected light, ultrasound imaging operates at and around a single central frequency point during emission and reception. As a result, the ultrasound image contains significantly fewer informative features. Consequently, ordinary algorithms often struggle to perform effectively on both light images and ultrasound images. For a visual comparison, refer to Fig. 1. In this paper, we propose the Inverse Pyramid Pooling Attention (IPPA) algorithm to address this problem.

2 Related Works

Attention mechanisms have demonstrated remarkable efficacy across diverse domains within the realm of deep learning, captivating an ever-growing cohort of researchers. These attention mechanisms can be broadly categorized, based on their dimensions, into various types including channel attention, spatial attention, temporal attention, branching attention, and hybrid attention. Among these, one widely recognized network architecture is the Squeeze-and-Excitation Networks (SENet) [10]. In a notable application, Gao et al. [11] integrated SeNet within their implementation of YOLOv4 to detect microaneurysms, which serve as early symptomatic indicators of diabetes. This approach exhibited a substantial 13% enhancement in F-score and superior localization performance compared to the unmodified YOLOv4. Furthermore, a separate study conducted a comparative evaluation of SeNet, VGG6, ResNet50, ResNet-CBAM, and SKNet network models on liver pathology images across different stages of differentiation. The comprehensive analysis, encompassing metrics such as the confusion matrix, precision, recall, F1 score, and other relevant indicators, concluded that the SENet network model surpassed the others, establishing its superiority [12].

The spatial attention mechanism operates by treating the entire region of each feature channel as a unified entity and scrutinizing the weighting relationship between different regions. Representative networks embodying this approach include the Gather-Excite Net and Spatial Transformer Network (STN). In the context of cell imaging, some researchers have devised a deep learning auto-focus phase (DLFP) network to achieve rapid auto-focus. This method has demonstrated both efficiency and cost-effectiveness for cell imaging [13]. Nonetheless, the performance gains attained through the application of a single attention mechanism alone are often limited. Therefore, researchers often explore the amalgamation of two attention mechanisms. For instance, in [14], a neural network model that combines a BAM (Bottleneck Attention Module) module, a rectified linear unit (C.ReLU), and an initial module is proposed. This model was employed for the detection of human eye position and exhibited exceptional accuracy when evaluated across multiple datasets.

Recently, there has been a proliferation of different attention methods and there is room for different degrees of improvement in the different attention

mechanisms respectively [15]. Different types of attentional mechanisms have different focuses and can be applied differently, and a combination of them can be used to obtain more attentional power in more dimensions and thus improve the results.

3 Inverse Pyramid Pooling Attention Module

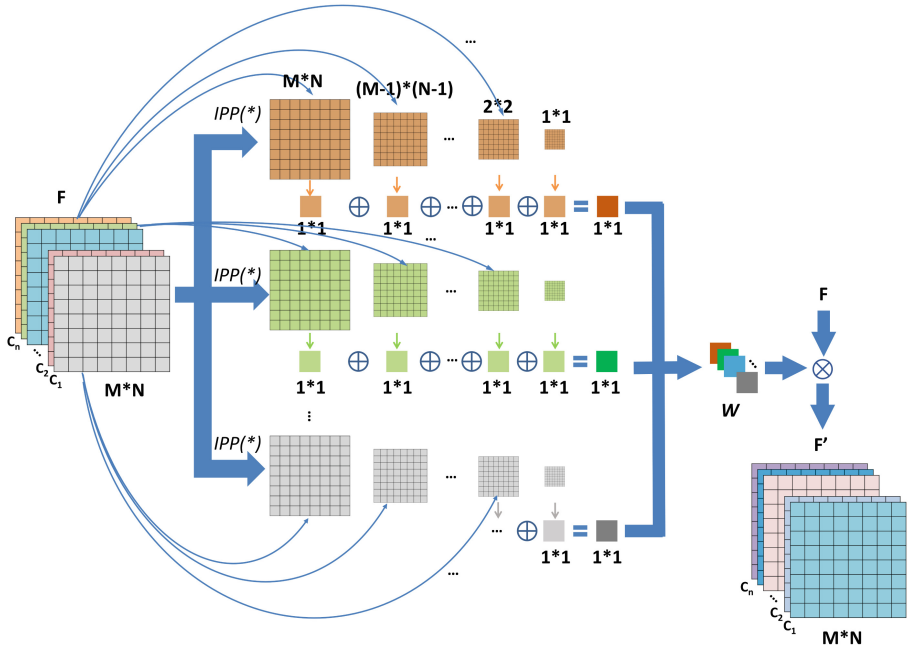

Fig. 2. The pipeline of IPPA. The input feature matrix F has C channels, each of which is subjected to an IPP operation. Multiple pooling operations are conducted on a single channel, where the size of the pooling decreases progressively from the outer regions to the center. Various pooling methods, corresponding to different sizes, are selected using a certain probability. The outcomes obtained from these multiple pooling operations are subsequently fused and summed to derive the original weight vector for the current channel.

The IPPA module offers the capability to integrate with arbitrary convolutional layers through adaptive transformations, thereby enhancing the representation of channel and spatial features. The module's comprehensive structure is depicted in Fig. 2. Within this structure, the initial adaptation vector $W \in R^{1*1*C}$ for each channel is constructed through pooling combinations with varying probabilities.

$$F^{'} = F * \sigma(\sum_{k=1}^{N}(IPP(F_c))) \tag{1}$$

where F_i is the C_i-th channel of the feature map F, $IPP(*)$ is the inverse pyramiding process and the sigma function is calculated as $\sigma(x) = \frac{1}{(1+e^{(-x)})}$.

3.1 Inverse Pyramid Pooling

A series of sets of images of progressively smaller sizes obtained by downsampling the same image several times in succession is an image pyramid. The process can be described as generating multiple images from a single graph, which carry different scale information. Pooling is widely used in convolutional neural networks and has important effects. However, pooling also poses some problems, such as the loss of a large amount of detailed pixel information and the retention of only a small number of important features. Given that ultrasound images contain less information than optical imaging images, there are limitations to the performance of conventional networks on ultrasound images. To solve the above problems and extract sufficient effective information from the feature maps, we propose a new set of inverse pyramid pooling (IPP) methods. Firstly, multiple pooling is performed in the feature matrix. Secondly, the window of each pooling operation is reduced by the maximum length and width of the feature matrix, in turn by a certain step in, and different pooling methods are chosen with a certain probability. Finally, multiple results can be obtained after multiple pooling of a single channel. In order to make the final result consistent with the number of channels, the summation and fusion of these pooling results are considered as the initial adjustment weights for that channel. The fused weights contain more hierarchical and scaling information compared to a single pooling. The overall pipeline process for IPP is to apply multiple pooling operations to the same feature map with non-equivalent size windows to form a corresponding single conditioning factor, called IPP. The overall pipeline process for IPP can be described as a fusion from multiple feature matrices of different sizes into a single representative weight, hence the term IPP.

The IPP process does not use any down-sampling operations when reducing the size. We reduce the selected area directly from the outer circle to the centre and pool accordingly, without using upsampling to revert to the original size, further reducing the computational effort. The IPP process for a single channel feature map is shown in Fig. 3. The resolution of the channel is $M * N$, after pooling (pooling window of $M * N$) to obtain W_1, reduce the pooling window to $M_1 * N_1$ (yellow box) to obtain W_2, until the smallest pooling window is $M_n * N_n$, the result is W_n. Each step in the reduction process produces a layer of feature maps of the corresponding size, from L1 to Ln layers, each of which can be individually selected for a different pooling operation. The results of multiple pooling are summed to obtain the final pooling result of the jth channel as W_j. The different colours indicate different pooling methods, such as maximum pooling, average pooling, etc. The pooling is done in different colours, such as maximum pooling, average pooling and minimum pooling. The calculations associated with the IPP to produce the initial conditioning weights for each channel are described below. First, the total number of layers F of the atlas that can be

generated from a feature map of a single channel is calculated by stepping down the pixel s by the size of the feature map to $M * N$.

$$F = \lceil (min(M, N) - 2)/(2 * S) + 1 \rceil \tag{2}$$

The size of the minimum layer map, M_n and N_n, is calculated as follows:

$$A = (min(M, N) - 2) \quad mod \quad 2 * S \tag{3}$$

$$M_n = \begin{cases} 2 * S, & \text{A=0} \\ M - 2 - |(M - 2)/2S| * 2S, & \text{otherwise.} \end{cases} \tag{4}$$

$$N_n = \begin{cases} 2 * S, & \text{A=0} \\ N - 2 - |(N - 2)/2S| * 2S, & \text{otherwise.} \end{cases} \tag{5}$$

To achieve adaptation for various scenes, we employ distinct step-reduction and feature map sizes. Furthermore, we calculate the total number of pixels, denoted as P_F, obtained through the IPP of the feature map for a single channel, using the aforementioned results F.

$$P_F = \sum_{i=1}^{F} (M - 2 * s * i)(N - 2 * s * i) \tag{6}$$

The detailed procedure steps are shown in Algorithm 1. The input feature map contains multiple channels (C), each of size M*N, which form a multilayer feature map after inverse pyramid. The inverse pyramid layers (F) discriminant factor of the channel is obtained after a 2-fold step-size residue calculation. The side with the smaller length and width of the feature map is used as the base for calculation to accommodate image features of different scales. The inverse pyramid layer (F) discriminant factor of the channel is obtained by 2-fold step residual calculation. When Min cannot divide T, the minimum feature map size is taken as the number of pixel points of the residual length and width, in which case the number of pyramid layers needs to be added one more layer. Finally traversal calculates the pooling result of each layer of the feature map, and the specific pooling method selection is determined by the random factor r. The summation result of the jth layer (F_j) of the ith channel is W_i, and the final channel weight matrix W is obtained from the weight contribution of each channel.

$$F = \lceil (min(M, N) - 2)/(2 * S) + 1 \rceil. \tag{7}$$

The size of the minimum layer map, M_n and N_n, is calculated as follows:

$$A = (min(M, N) - 2) \quad mod \quad 2 * S \tag{8}$$

$$M_n = \begin{cases} 2 * S, & \text{A=0} \\ M - 2 - |(M - 2)/2S| * 2S, & \text{otherwise.} \end{cases} \tag{9}$$

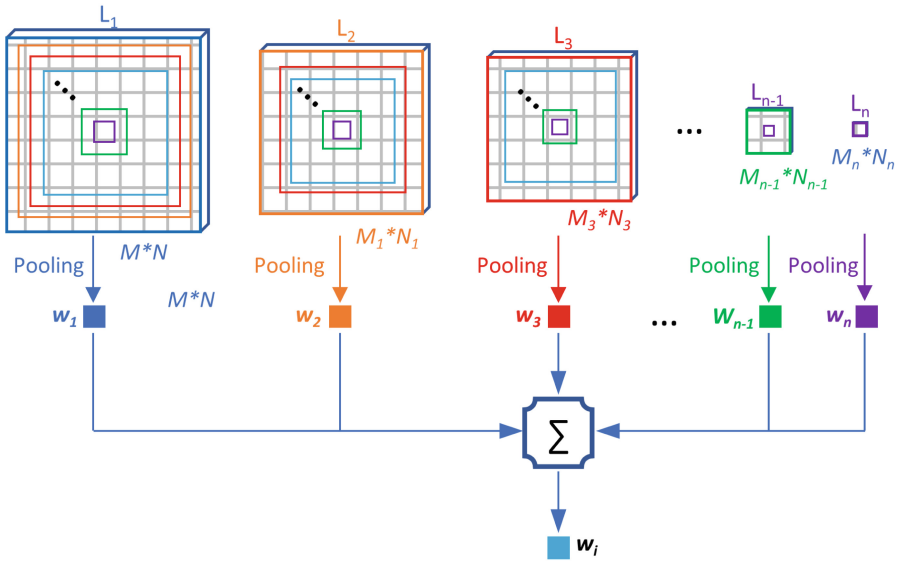

Fig. 3. Overview diagram of the single channel IPP operation.

Algorithm 1: Inverse Pyramid Pooling

Input: Standardised Channels Data C, step size T, random thresholdands $0<r_1<r_2<r_3<1$, and feature map resolution data $M * N$

Output: Weighting of standardised channels W

$Min = min(M, N)$;
$A = Min \ mod \ 2T$;
if $A == 0$ **then**
  $F = |Min/(2T) - 1/T|$;
else
  $F = |Min/(2T) - 1/T + 1|$;
for $i : C$ **do**
  $D = D + C_i/D_i$;
  for $j : F$ **do**
    $r = random(0, 1)$;
    if $0<r<r_1$ **then**
      $W_j = MaxPooling(F_j)$;
    else if $r_1<r<r_2$ **then**
      $W_j = AveragePooling(F_j)$;
    else
      $W_j = StochatiscPooling(F_j)$;
  $W_i = \sum_{j=1}^{F} W_j$;
$W = |W_1, W_2, ..., W_i, ..., W_C|$;

$$N_n = \begin{cases} 2 * S, & \text{A=0} \\ N - 2 - |(N-2)/2S| * 2S, & \text{otherwise.} \end{cases} \tag{10}$$

Taking a 128*128 size image with a vertical and horizontal indentation of 2 pixels, the inverse pyramid atlas is 32 layers with a minimum layer image size of 2*2, using a combined pixel count of nearly 45,000. Within the network, the feature maps exhibit varying numbers of channels, which differ across convolutional layers. Consequently, in the case of feature maps with multiple channels, the total number of pixels after pyramidalization can be denoted as P_C:

$$P_C = C * P_F \tag{11}$$

where F is the number of feature map layers, C is the number of channels, and P_C is the sum of the C channel pixels.

3.2 Inverse Pyramid Pooling ResNet

Since its inception, residual networks have been widely used in a variety of industries. The jump connections in residual network blocks [16] have guided the depth of the network body structure into a new level of hierarchy and also facilitated the development of deep learning. The inverse pyramid Pooling attention (IPPA) module given in this paper can be coupled to the convolutional layers of different convolutional neural networks as channel importance adjustment weights for the feature matrix to enhance the performance of the network. We take the "bottleneck" building block of a residual network with 50 layers or more as an example (as shown in Fig. 4), and couples the IPPA module to the last convolutional layer of the residual block, with the weight vector multiplied by Activation is multiplied by the corresponding channel. The input X_i is coupled by the IPPA to the ResNet output X_o, called IPPA-ResNet.

4 Experiments and Analysis

4.1 Experimental Preparation

Data: We employed three open datasets as the training sample pool for our model, comprising a cumulative sample size of approximately 2000 ultrasound images. These datasets include: the Dataset of breast ultrasound images (hereinafter referred to as DBUI) authored by Walid AI-Dhabyani et al. from Cairo University, Egypt [17]; the FAscicle Lower ultrasound image dataset for the prevention of calf muscle injury outlined in Michard et al.'s Leg Muscle Ultrasound Dataset (hereinafter referred to as FALLMUD) [18]; and a publicly accessible abdominal organ ultrasound image dataset (hereinafter referred to as AOUI) [19]. Furthermore, we conducted data augmentation techniques on the dataset, resulting in the final training library of ultrasound images.

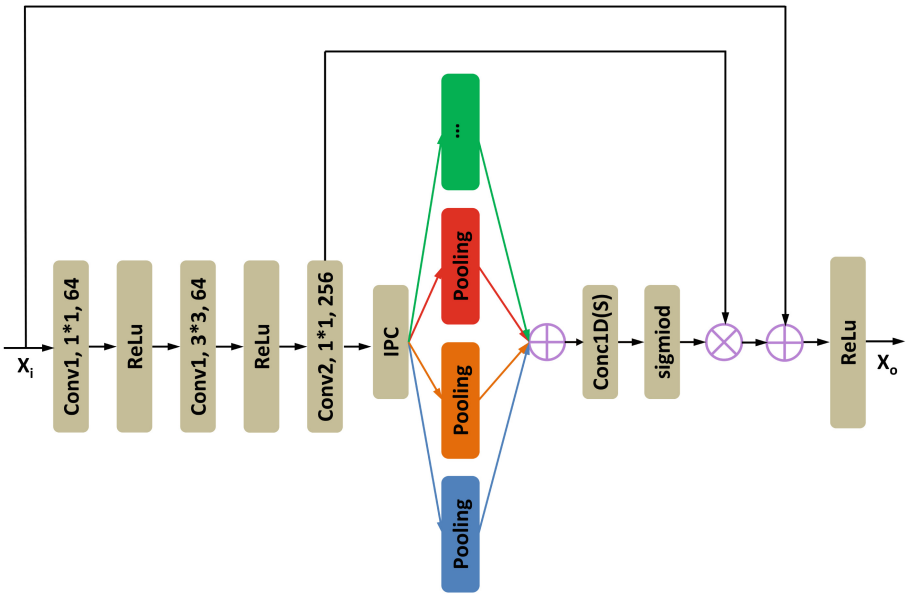

Fig. 4. The structure of IPPA-ResNet. The colored parts correspond to different pooling methods.

Devices and Parameters: We constructed an IPPA module and a 50-layer ResNet network based on the pyhton language and Pytorch's learning library. The IPP-ResNet was formed by embedding the IPPA module on the last convolutional layer of each residual block of the residual network. the training process was performed on a mobile configured with an Intel(R) Core(TM) i9-11900H @ 2.50 GHz CPU, 16G DDR4 RAM, and 8G RTX3070 GPU. The input to the network was a 64*64 sized ultrasound images with a bachsize of 48. The loss calculation was implemented by a cross-entropy loss function with 10 classifications, the optimizer was stochaastic gradient descent (SGD), and the learning rate used a thermal learning rate with an initial value of 0.005.

Assessment of Indicators: The performance of the network model is reflected by a number of metrics, namely accuracy (*Acc*), recall, precision, and F1 score, and the predicted sample types are classified as positive samples predicted by the model as positive class (TP), negative samples predicted by the model as negative class (TN), negative samples predicted by the model as positive class (FP), and positive samples predicted by the model as negative class (FN), which are calculated as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{12}$$

Table 1. Comparison of multiple indicators on the same or similar data sets with the latest methods

	Mathonds	Dataset	Precision (%)	Sensitivity (%)	F1-score (%)	Acc (%)
Xu et al. [17](2020)	VGG6	muscle ultrasound images (1498)	/	/	/	95.20
Reddy et al. [20](2021)	CNN+ Transfer learning	abdomenultrasound images (1096)	/	/	/	98.77
Gheflati et al. [21](2022)	Vision Transformers	DBUI	/	/	/	79.00
		DBUI+UDITA	/	/	/	86.70
Joshi et al. [22](2022)	VGG 19/ Yolo v3	DUBI	88.22	90.32	89.07	90.00
		UDITA	91.72	91.16	91.43	95.25
		DUBI+UDITA	96.36	96.71	92.99	96.31
Xu et al. [23](2022)	MTL-COSA	UDITA	89.36	82.17	90.41	87.08
		DUBI	95.05	92.20	93.59	91.48
Alireza et al. [24](2022)	Decision Tree Integration	DUBI	94.00	93.00	93.00	91.00
Ours	**IPPA-ResNet**	**DUBI**	**99.48**	**98.67**	**99.06**	**99.28**

$$Precision = \frac{TP}{TP + FP} \tag{13}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{14}$$

$$F1 - score = \frac{2 * Precision * Sensitivity}{Precision + Sensitivity} \tag{15}$$

4.2 Comparison with the Most Advanced Methods

We compare network models equipped with IPPA mechanisms with recent new results on the same dataset.

As can be seen from Table 1, our proposed IPPA-ResNet has better performance in terms of *Accuracy*, *Sensitivity*, $F1-score$ and *Prision* on the dataset compared to other recent methods, and is more adaptable to ultrasound images, which has better potential for development and contributes to the performance of the network on ultrasound images.

4.3 Comparison of Model Improvement Effects

The results of the experiments conducted on a dataset consisting of a combination of three datasets, AOUI, FALLMUD and DBUI, are represented in Fig. 5. The dataset contains 10 categories and we tested networks equipped with the IPPA mechanism and other attention mechanisms as well as some classical networks and their variants. SE-ResNet and ECA-ResNet are both ResNet with the attention mechanism inserted, ResNeXt is a variant of ResNet, and Inception-ResNetv2 also contains residual structures internally.

The diminishing fluctuation of the network's training results is observed as it approaches convergence. In order to provide a comprehensive evaluation of the network's performance, the mean value of *E-Acc* was calculated by considering the *Acc* of the last 10 epochs across all networks. The Inception-ResNet-v2 network does not exhibit a notable advantage when applied to the ultrasound dataset, primarily due to its lack of channel focus. Moreover, the network possesses a significant number of parameters, rendering it less practical in real-world applications. On the other hand, the ResNet, as presented in this paper, lacks the incorporation of IPPA, which results in a less smooth convergence process. A more visually appealing representation of this outcome is illustrated in Fig. 5.

The analysis depicted in Fig. 5 reveals a noticeable disparity in the effectiveness of various attention methods applied to the ultrasound image dataset. When compared to the attention mechanisms employed in ResNet-based models such as ECA-ResNet and SE-ResNet, the approach presented in this paper (IPPA-ResNet) demonstrates superior performance and quicker convergence. ECA-ResNet and SE-ResNet exhibit a smoother progression at approximately the 120th and 90th epochs, respectively. However, ECA-ResNet tends to encounter challenges related to the accurate determination of channel weights during the extraction of ultrasound image channel interactions. Conversely, SE-ResNet lacks

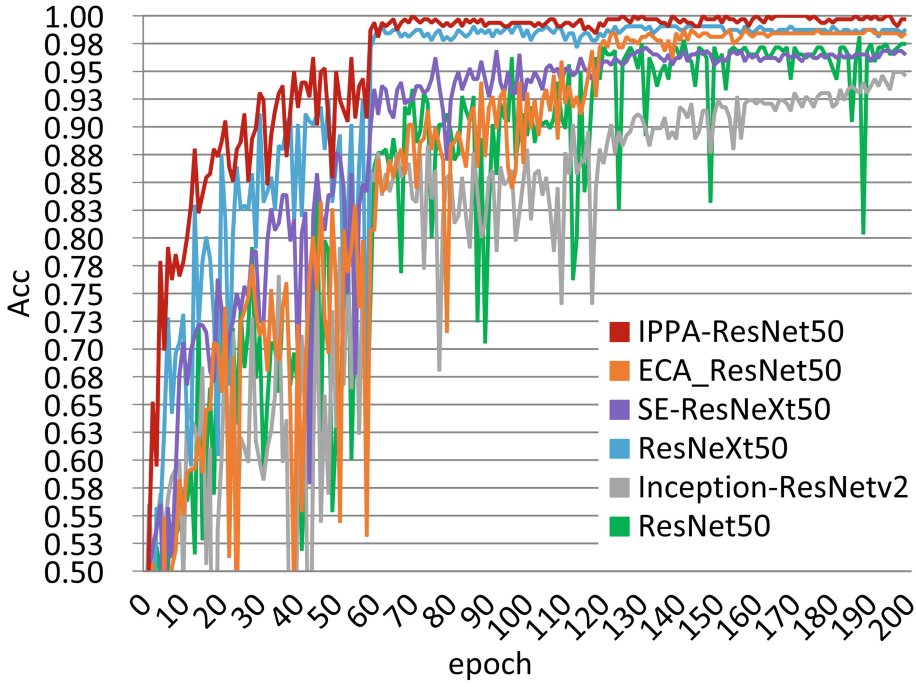

Fig. 5. Test set accuracy for six different networks.

the ability to extract multi-level weights for ultrasound image channels, resulting in an inadequate representation of channel differences.

In summary, the convolutional neural network incorporating the attention mechanism showcases enhanced performance in terms of Precision and F1-score compared to the original network. It demonstrates faster and smoother convergence. Moreover, our proposed IPPA further improves the network's generalization performance.

5 Conclusion

In the study in this paper, we developed a new attentional approach (IPPA) with 50 layers of ResNet as the core network to construct IPP-ResNet50. Extensive experiments were conducted on an ultrasound dataset (consisting of AOUI, DBUI together with FALLMUD), which included six classes of abdominal organs (gallbladder, bladder, liver, bowel, kidney, spleen), three types of breast (benign, malignant, normal) and calf muscles. The experimental results show that our proposed IPCPA-ResNet50 is superior in all metrics on the dataset compared to similar as well as the latest state-of-the-art methods. The integration of our proposed IPCPA onto a convolutional neural network enhances the extraction of the information density of each feature channel by the convolutional layer of the network. The fusion of multiple dimensional features of the extracted single channel obtains a better generalisation capability compared to single features. Moreover, probability factors are introduced into the feature extraction process of each individual channel to accentuate the differences between channels. Our approach embeds IPPA into a 50-layer ResNet network and introduces probabilistic elements, resulting in improved accuracy and convergence speed. The highest average accuracy (*E-Acc*) achieved is 99.68%, with a maximum sensitivity of 99.60%.

This work unveils the potential and application value of integrating IPPA into convolutional neural networks for ultrasound signal imaging classification. However, there is still room for improvement and refinement in the IPPA module, particularly in the inverse pyramid process, where smaller strides can generate more scales of feature maps compared to larger strides. The presence of a large number of feature maps results in increased computational overhead. Therefore, it is worth considering the possibility of retaining only the essential feature maps or adjusting the trade-offs for different scenarios. This approach can help reduce the computational burden without compromising accuracy. Lastly, we will endeavor to validate the performance of the proposed method by integrating it into more networks and evaluating it on diverse datasets.

References

1. Huang, S., Rao, G., Ashraf, U., et al.: Ultrasonic seed treatment improved morpho-physiological and yield traits and reduced grain Cd concentrations in rice. Ecotox. Environ. Safe. **214**, 112119 (2021)
2. Mo, Z., Liu, Q., Xie, W., et al.: Ultrasonic seed treatment and Cu application modulate photosynthesis, grain quality, and Cu concentrations in aromatic rice. Photosynthetica **58**(3), 682–691 (2020)
3. Kou, X., Pei, C., Liu, T., et al.: Noncontact testing and imaging of internal defects with a new Laser-ultrasonic SAFT method. Appl. Acoust. **178**, 107956 (2021). https://doi.org/10.1016/j.apacoust.2021.107956
4. Chen, X., Li, T., Dou, X., et al.: Reverse osmosis membrane combined with ultrasonic cleaning for flue gas desulfurization wastewater treatment. Water **14**(6), 875 (2022)
5. Chammas, M.C., Bordini, A.L.: Contrast-enhanced ultrasonography for the evaluation of malignant focal liver lesions. Ultrasonography **41**(1), 4–24 (2022)
6. Jiang, Z., Ma, Z., Wang, Y., et al.: Aggregated decentralized down-sampling-based ResNet for smart healthcare systems. Neural Comput. Appl. **75**, 1–13 (2021)
7. Shao, X., Asaeda, H., Dong, M., et al.: Cooperative inter-domain cache sharing for information-centric networking via a bargaining game approach. IEEE Trans. Netw. Sci. Eng. **6**(4), 698–710 (2019)
8. Liu, B., Fang, Z., Wang, W., et al.: A region-based collaborative management scheme for dynamic clustering in green vanet. IEEE Trans. Green Commun. Netw. **6**(3), 1276–1287 (2022)
9. Liu, J., Liu, H., Chakraborty, C., et al.: Cascade learning embedded vision inspection of rail fastener by using a fault detection IoT vehicle. IEEE Internet Things J. 1–12 (2021)
10. Li, X., Zhao, H., Ren, T., et al.: Inverted papilloma and nasal polyp classification using a deep convolutional network integrated with an attention mechanism. Comput. Biol. Med. **149**, 105976 (2022)
11. Gao, W., Shan, M., Song, N., et al.,: Detection of microaneurysms in fundus images based on improved YOLOv4 with SENet embedded. 2022, Sheng Wu Yi Xue Gong Cheng Xue Za Zhi, vol. 39, no. 4, pp. 713–720
12. Chen, C., et al.: Classification of multi-differentiated liver cancer pathological images based on deep learning attention mechanism. BMC Med. Inform. Decis. Mak. **22**(1), 176 (2022)
13. Liu, Y., Huaying, W., Zhao, D., et al.: Application of auto-focusing technology based on improved U-Net in cell imaging. China J. Lasers **49**(15), 1507302 (2022)
14. Nguyen, D.L., Putro, M.D., Vo, X.T., et al.: Convolutional neural network design for eye detection under low-illumination. In: Sumi, K., Na, I.S., Kaneko, N. (eds.) IW-FCV 2022. LNCS, vol. 1578, pp. 143–154. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-06381-7_10
15. Zhu, X., Cheng, D., Zhang, Z., Lin, S., et al.: An Empirical study of spatial attention mechanisms in deep networks . In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 6687–6696 (2019)
16. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
17. Xu, J., Xu, D., Wei, Q., et al.: Automatic classification of male and female skeletal muscles using ultrasound imaging. Biomed. Sig. Process. Control **57**, 101731 (2020)

18. Michard, H., Luvison, B., Pham, Q.C., et al.: AW-Net: automatic muscle structure analysis on b-mode ultrasound images for injury prevention. In: 12th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics (ACM-BCB 2021), pp. 1–9 (2021)
19. Li, K., Xu, Y., Zhao, Z., et al.: Automatic recognition of abdominal organs in ultrasound images based on deep neural networks and k-nearest-neighbor classification. IEEE-Robio 2021, pp. 1980–1985 (2021)
20. Reddy, D.S., Rajalakshmi, P., Mateen, M.A.: A deep learning based approach for classification of abdominal organs using ultrasound images. Biocybern. Biomed. Eng. **41**(2), 779–791 (2021)
21. Gheflati, B., Rivaz, H.: Vision transformers for classification of breast ultrasound images. In: Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference, vol. 2022, pp. 480–483 (2022)
22. Joshi, R.C., Singh, D., Tiwari, V., et al.: An efficient deep neural network based abnormality detection and multi-class breast tumor classification. Multimedia Tools Appl. **81**(10), 13691–13711 (2022)
23. Xu, M., Huang, K., Qi, X.: Multi-task learning with context-oriented self-attention for breast ultrasound image classification and segmentation. In: IEEE ISBI 2022, pp. 1–5 (2022)
24. Rezazadeh, A., Jafarian, Y., Kord, A.: Explainable ensemble machine learning for breast cancer diagnosis based on ultrasound image texture features. Forecasting **4**(1), 262–274 (2022)