

BFC-BL: Few-Shot Classification and Segmentation combining Bi-directional Feature Correlation and Boundary constraint

HaiBiao Yang
haibiao_yang@mail2.gdut.edu.cn

Guangdong University
of Technology, CN

Bi Zeng
zb9215@gdut.edu.cn

PengFei Wei*
wpf@gdut.edu.cn

JianQi Liu
liujianqi@ieee.org

Abstract

Few-shot classification and segmentation model realizes classification and segmentation by learning the feature correlation between a small number of samples. The lack of correlation learning between samples and the similarity of target foreground and background boundary pixels lead to segmentation errors, we propose Few-Shot Classification and Segmentation combining Bi-directional Feature Correlation and Boundary constraint(BFC-BL). Firstly, the correlation between query set and support set is calculated by cosine similarity to construct a 4D tensor. Then, a cross-scale bidirectional feature correlation fusion module (BFCF) is designed and embedded into the encoder structure to perform the interactive fusion of deep semantic correlation and shallow spatial correlation, while a bounding-constrained loss function is introduced to guide the model to learn the boundary information of the target foreground and background. Finally, a multi-level weight ratio loss function was constructed to make the network converge faster and generalize better. The experimental results show that compared with the ASNet method, the classification accuracy of the proposed method is increased by 1.7% and 1.8%, and the segmentation accuracy mean intersection over union ratio is increased by 1.3% and 1.4% on the *Pascal - 5ⁱ*. The code is publicly available at: <https://github.com/XIAO1HAI/BFC-BL>.

1 Introduction

In the era of big data and rapid technological advancements, artificial intelligence (AI) has found applications in various fields such as computer vision [13, 14], natural language processing [9], and medical imaging [26]. Deep learning, as a representative AI technique [21], has achieved remarkable success. However, the widespread adoption of deep learning comes with challenges. It requires complex model design, large-scale labeled training data, and

*Corresponding author.

high-performance computing infrastructure for accurate and efficient training. Moreover, the cost of data labeling and the limited availability of extensive datasets pose difficulties, leading to overfitting and reduced generalization in conventional deep learning models [6, 24]. Applying such models to scenarios with limited data becomes challenging. Nevertheless, the fundamental goal of AI is to mimic human intelligence, where humans can learn new knowledge and concepts from limited guidance and make accurate judgments. This ability aligns with the concept of few-shot learning, which aims to train models using a small number of samples [7, 8, 10, 12, 23]. Few-shot learning methods have gained attention and become a research hotspot in recent years.

In the field of computer vision [11, 9, 4, 14, 18], several methods have been proposed for Few-Shot Classification and Segmentation (FS-CS) [1, 16, 25]. These methods address the challenge of accurately classifying and segmenting multiple targets, even when the target category is missing in the query set. Most of these methods use metric meta-learning to extract and fuse features from the support set. In the fusion stage, the commonly used feature pyramid network [5, 10, 19] fuses shallow features and deep features. However, they often lack interaction between fine-grained semantic information and global spatial information. Additionally, the segmentation results often suffer from poor boundary segmentation due to the lack of learning boundary pixel semantic information. To overcome these challenges, we propose a few-shot classification and segmentation approach that combines bidirectional feature correlation and boundary constraint. Our method aims to improve the fusion of semantic and spatial information and enhance the segmentation accuracy at object boundaries. The main contributions of this paper are summarized as follows:

(1) This paper proposes bidirectional feature fusion for few-shot classification and segmentation. The cross-scale bidirectional semantic correlation fusion module (BFCF) integrates deep semantic and shallow spatial correlation measurements, facilitating rapid learning of the correlation between support and query samples.

(2) To enhance the segmentation performance of the model, we introduce a boundary constraint function, which is integrated with the region loss to construct a multi-level weight ratio loss function. This loss function guides the model to learn boundary information, thereby improving its ability to accurately segment target boundary pixels.

(3) By leveraging the two-way high-level semantic correlation between support and query samples, low-level spatial correlation, and the boundary constraints, we construct an end-to-end few-shot classification and segmentation network model. This model enables rapid and effective few-shot classification and segmentation, providing significant advancements in the field.

2 Problem Definition

The training samples of Few-shot Classification and Segmentation (FS-CS) usually contain N categories and each class has K samples, which is called N -way K -shot. First, assume that the target class set is C_s , and the data set is divided into a training set D_{train} and a test set D_{test} . Every time training will be randomly selected from a training set K sample I_s in each category and tag M_s as Support set $S = \{I_s^i, M_s^i \mid M_s^i \in C_s\}_i^{NK}$, The label M_s^i contains both classification labels (weak labels) and segmentation labels (strong labels). After selecting from a training set and support set is not the same sample I_q and label M_q as a Query set $Q = \{I_q^i, M_q^i \mid M_q^i \in C_s\}_i^{NK}$, including segmentation tags binary matrix that is the true seg-

mentation mask. For the query set image I_q , we aim to identify objects $\hat{y}_c \in \mathbb{R}^N$ of multiple categories and an accurate segmentation mask $Y_s \in \mathbb{R}^{H \times W \times (N+1)}$ of classes.

3 Model Architecture

We propose combining Bi-directional Feature Correlation and Boundary constraint Loss (BFC-BL) model is an encoder-decoder structure, and the overall architecture diagram is shown in Figure 1.

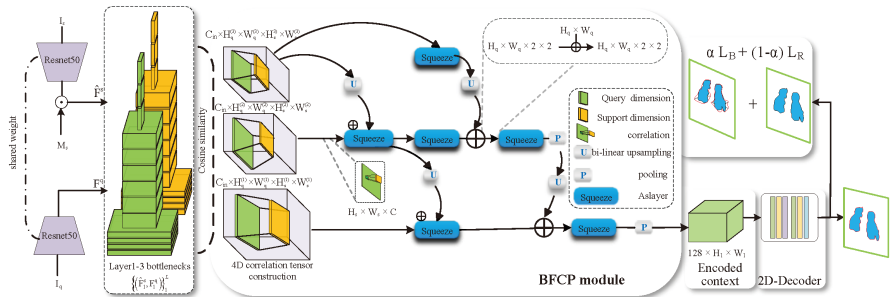


Figure 1: Overall architecture of the BFC-BL model. The model is mainly divided into three stages: (1) ResNet50 is used as the backbone network for feature extraction. This process involves computing the correlation between the query and support samples, resulting in a 4D tensor constructed using cosine similarity. (2) A cross-scale bidirectional feature correlation fusion module (BFCP) is designed and embedded into the encoder structure to perform an interactive fusion of deep semantic correlation and shallow spatial correlation features. (3) Utilizing the context information obtained from the encoder, a 2D decoder decodes the learned encoder knowledge and generates the final output.

3.1 Semantic correlation construction

Given the input support set image I_s and its corresponding mask information M_s , as well as the query set image I_q , represented as $I_s, I_q \in \mathbb{R}^{3 \times H \times W}$, we can obtain m layers of support feature maps denoted as F^s and query feature maps denoted as F^q . Here, m represents the number of layers in the bottleneck of the ResNet50 architecture. Furthermore, for each intermediate query feature map F^s in each layer, a Hadamard dot product operation is performed with the label mask M_s , as depicted in Equation 1 [□]. This operation helps calibrate the pixel positions of the true target class through the mask.

$$\hat{F}^s = F^s \odot BI(M_s), \quad (1)$$

Where $F^s \in \mathbb{R}^{C \times H \times W}$, $M_s \in \{0, 1\}^{H \times W}$, \odot is the Hadamard dot product, $BI(\cdot)$ is a bilinear interpolation function that interpolates the input mask to the size of the feature map, followed by an expansion along the channel dimensions such that $\mathbb{R}^{H \times W} \rightarrow \mathbb{R}^{C \times H \times W}$. Afterward, the cosine similarity calculation is performed on the feature map acquired at each layer to construct the correlation between the query and the support, and finally form the 4D

correlation tensor $Cl_l \in H_q \times W_q \times H_s \times W_s$, thus completing the construction of semantic correlation, the calculation Equation is:

$$Cl_l(I_q, I_s) = \text{ReLu} \left(\frac{F^q(I_q) \cdot \hat{F}^s(I_s)}{\|F^q(I_q)\| \|\hat{F}^s(I_s)\|} \right), \quad (2)$$

The correlation tensor has dimensions $H_q \times W_q$ to represent the query dimension, and $H_s \times W_s$ to represent the support dimension.

The m-layer support and query intermediate feature similarity calculations are formed L pairs of feature correlation tensors (L=m), L pairs of correlation tensors are divided into G groups according to the size of the spatial dimension, and are spliced into a correlation tensor along the new dimension, $\{Cl_l(g) \mid Cl_l(g) \in C_{in} \times H_q \times W_q \times H_s \times W_s\}_g^G$, the size of C_{in} depending on the number of correlation tensors of the g-th group, $H_q \times W_q \times H_s \times W_s$ is the spatial resolution of the g-th layer pyramid correlation.

3.2 BFCP module for bidirectional feature correlation fusion

Semantic features play a crucial role in guiding the learning process of network models, and their fusion is a key aspect. In this paper, we optimize and apply Bi-FPN[[19](#)] to few-shot classification and segmentation tasks for the first time. We propose a bidirectional feature correlation fusion module, called BFCP, for integrating the semantic correlation between the support and query images, as illustrated in Figure 2. The BFCP module effectively captures the bidirectional feature correlation within the semantic and spatial information interaction. To streamline the network and reduce computational overhead, we eliminate redundant inputs from the same compressed network layer and the input path from the low-level compressed network to the high-level. This ensures the effectiveness of the fusion process while reducing computational costs for the same correlation. Additionally, the BFCP module facilitates the integration of semantic feature correlation with the encoding stage. Initially, the correlation 4D tensor is input to the attention compression network, which performs feature compression and target positioning while preserving crucial correlation connections. The output of the compression network is then fused using the BFCP module and input into the encoder. Through the encoding stage, the final model effectively learns target-specific local and global context feature correlation information, leading to more accurate classification and segmentation of the target object during the decoding stage.

3.3 Multi-head Attention Compression Networks

To analyze the global context of each support image and preserve query dimension features, we reshape the correlation tensor into a matrix of size $H_q \times W_q$. This matrix retains the query dimension while extracting and representing support dimension features. Each element x_q in the correlation matrix corresponds to a support correlation, denoted as $Cl_l(x_q) \in \mathbb{R}^{H_s \times W_s \times C_{in}}$. This support correlation matrix is then embedded into the Query, Key, and Value components of the attention mechanism [[20](#)].

Firstly, the matrix multiplication of Query and Key is calculated to obtain the weight coefficient corresponding to the Key feature representation of each key, which is normalized by the softmax function. Then the foreground mask Y_s is used to mark the key pixels in the foreground with 0 and 1 so that more attention is paid to the foreground. The calculation

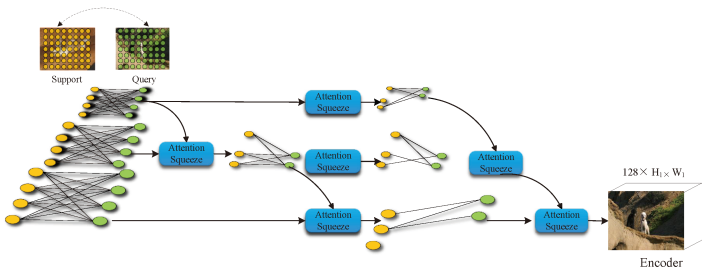


Figure 2: BFCP module for bidirectional feature correlation fusion. the green node represents the feature map of each layer of feature extraction, the blue node represents the feature processing layer (such as the attention layer), and the line of each node represents the feature fusion path.

Equation is as follows:

$$A = QK^T \in \mathbb{R}^{H'_s \times W'_s \times H'_s \times W'_s}. \quad (3)$$

$$\bar{A}(p_q, p_k) = \frac{\exp(A(p_q, p_k) Y_s(p_k))}{\sum_{p'_k} \exp(A(p_q, p'_k) Y_s(p'_k))}, \quad (4)$$

In Eq. 4,

$$Y_s(p_k) = \begin{cases} 1, & p_k \in F_g \\ -\infty, & p_k \notin F_g \end{cases}, \quad (5)$$

Where F_g is the foreground and $Y(p_k)$ is 1, if p_k is the foreground pixel; Otherwise, it is $-\infty$.

Next, the attention coefficient obtained from the foreground mask attention and the Value are utilized for the weighted summation of the aggregated embedding values. The equation representing this process is as follows:

$$O_A^s = \bar{A}V \in \mathbb{R}^{H'_s \times W'_s \times C_{hd}}, \quad (6)$$

To preserve the integrity of the network and prevent degradation, the query dimension feature correlation output undergoes a residual module using the original input. When the input and output dimensions do not match, the input is optionally fed to the convolutional layer W_I , and $\varphi(\cdot)$ is the activation function.

$$O_o^s = \varphi(W_o(O_A^s) + W_I(O^s)) \in \mathbb{R}^{H'_s \times W'_s \times C_{out}}. \quad (7)$$

Finally, the input is passed through the MLP layer to obtain the output, as follows:

$$O^{s'} = \varphi(W_{FF}(O_o^s) + O_o^s) \in \mathbb{R}^{H'_s \times W'_s \times C_{out}}. \quad (8)$$

3.4 Construction of Boundary Constraints

For objects with similar boundaries between the foreground and the background, it is easy to classify the foreground into the background, resulting in classification errors and subpar segmentation results. Hence, the learning of boundary information plays a crucial role in semantic segmentation. In our approach, we introduce a boundary constraint function that

measures distances in the contour space (specifically, the boundary of the target area) and integrates them along the region boundaries. This enables comprehensive learning of the target subject while placing additional emphasis on the learning of boundary features. Ultimately, we construct a loss function using multi-level weight ratios.

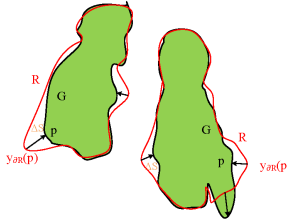


Figure 3: Diagram of the boundary loss function calculation. ∂G is the set of all points on the boundary of the ground truth region G , ∂R is the set of boundary points of the segmentation region R output by the network model, $D_G(q) = \|y_{\partial R}(p) - p\|^2$, D_G denotes the distance map concerning the boundary, $\|\cdot\|$ denotes the L2 norm, and p and $y_{\partial R}(p)$ denote the points on the true and predicted boundary.

Boundary loss function $\text{Dist}(\partial G, \partial R)$, partial R is in the region boundary space Ω distance metric, namely edge pixels for expression for predicting L2 distance between the boundary and real boundary, such as Eq. 9-12:

$$\text{Dist}(\partial G, \partial R) \approx 2 \int_{\Delta S} D_G(q) dq. \quad (9)$$

The calculation diagram for the boundary loss is depicted in Figure 3. To address the issue of non-differentiability in the differential calculation of contour points, the domain integral is employed to represent the boundary change. Eq. 10 demonstrates the validity of the domain integral.

$$\int_p^{y_{\partial R}(p)} 2D_G(q) dq = \int_0^{\|y_{\partial R}(p)-p\|} 2D_G dD_G = \|y_{\partial R}(p) - p\|^2. \quad (10)$$





Eq. 11 is derived from Eq. 9, where ΔS represents the area between the real and predicted boundary contours $\Omega \rightarrow \{0, 1\}$ is the binary indicator function on the area, $r(q)$ and $g(q)$ denote the indicator functions above R and G , respectively. If $q \in R$, $r(q)=1$, otherwise $r(q)=0$; $x_G \in G \rightarrow \mathbb{R}$ is the representation of the boundary ∂G , if $q \in G$, $x_G = -D_G(q)$, otherwise $x_G = D_G(q)$. The latter item in Eq. 11 does not contain model parameters and can be omitted during training. Use the softmax output $r_\theta(q)$ of the network to replace $r(q)$ in Eq. 11, Instant Eq. 12:

$$\frac{1}{2} \text{Dist}(\partial G, \partial R) = \int_R \Phi_G(q) dq - \int_G \Phi_G(q) dq = \int_\Omega \Phi_G(q) r(q) dq - \int_\Omega \Phi_G(q) g(q) dq, \quad (11)$$

$$L_B(\theta) = \int_\Omega \Phi_G(q) r_\theta(q) dq. \quad (12)$$

In addition to the learning of the object boundary contour, the learning of the central region is also more important. The central-region segmentation loss is calculated as the

Table 1: Experimental results of 1-way 1-shot and 2-way 1-shot for different network models.

Methods	1-way 1-shot										2-way 1-shot									
	classification 0/1 exact ratio (%)					Segmentation mIoU (%)					classification 0/1 exact ratio (%)					Segmentation mIoU (%)				
	5 ⁰	5 ¹	5 ²	5 ³	avg.	5 ⁰	5 ¹	5 ²	5 ³	avg.	5 ⁰	5 ¹	5 ²	5 ³	avg.	5 ⁰	5 ¹	5 ²	5 ³	avg.
PANet[	69.9	67.7	68.8	69.4	69.0	32.8	45.8	31.0	35.1	36.2	56.2	47.5	44.6	55.4	50.9	33.3	46.0	31.2	38.4	37.2
PFENet[	69.8	82.4	68.1	77.9	74.6	38.3	54.7	35.1	43.8	43.0	22.5	61.7	40.3	39.5	41.0	31.1	47.3	30.8	32.2	35.3
HSNet[	86.6	84.8	76.9	86.3	83.7	49.1	59.7	41.0	49.0	49.7	68.0	73.2	57.0	70.9	67.3	42.4	53.7	34.0	43.9	43.5
ASNet[	84.9	89.6	79.0	86.2	84.9	51.7	61.5	43.3	52.8	52.3	68.5	76.2	58.6	70.0	68.3	48.5	58.3	36.3	48.3	47.8
ours	87.4	89.4	81.1	88.5	86.6	52.9	62.7	44.6	54.2	53.6	70.6	77.0	60.4	72.1	70.1	50.2	59.2	37.8	49.7	49.2

average cross-entropy between the class distribution at each location and its corresponding true segmentation annotation:


$$L_R = -\frac{1}{N+1} \frac{1}{HW} \sum_{n=1}^{N+1} \sum_{p \in [H] \times [W]} Y_{gt}^{(n)}(p) \ln Y_S^{(n)}(p). \quad (13)$$

The overall loss function constraint can be expressed as follows:

$$L = (1 - \alpha)L_R(\theta) + \alpha L_B(\theta). \quad (14)$$


4 Experimental Results and Analysis

4.1 Experimental Setup and Metrics

The experiments in this paper were conducted on a server with an Intel i9-9900X 10-core processor and an NVIDIA RTX2080ti graphics card. The model was trained using the gradient descent method, and the Adam optimizer was utilized with a dynamically adjusted learning rate, initialized at 0.001. The dataset used for the experiments is the widely adopted *Pascal* – 5ⁱ dataset in the field of few-shot classification and segmentation. [].

To better evaluate the multi-label classification and segmentation effect of the model, this paper uses the 0/1 accuracy rate exact ratio $ER = 1 [y_{gt} = y_c]$ evaluation index to evaluate the multi-label classification and uses the average intersection and union ratio $mIoU = \frac{1}{C} \sum_c IoU_c$ evaluates the target segmentation (IoU_c represents the IoU value of the C-th category), where the evaluation examples cover All three cases of $C = \emptyset$, $C \subset C_s$, and $C = C_s$.

4.2 Experimental Result

This paper presents an analysis of recent few-shot classification and segmentation methods. The ResNet50 model is chosen as the backbone network for feature extraction, and its performance is compared with the few-shot segmentation model proposed in this paper on the widely used *Pascal* – 5ⁱ dataset. The models are evaluated within the IFSL framework of the ASNet[] network. The experimental results for the 1-way 1-shot and 2-way 1-shot scenarios are presented in Table 1.

The experimental results in Table 1 demonstrate the superior performance of our proposed model compared to other few-shot segmentation network models. Specifically, on the 1-way 1-shot task, our model outperforms the recent ASNet network model. The classification accuracy ratios and average intersection over union (mIoU) ratios on the 4-fold data exhibit improvements of +2.5%, -0.2%, +2.1%, +2.3%. The mIoU ratios also show improvements of +1.2%, +1.2%, +1.3%, +1.4%. However, there is a need for improvement in the

accuracy rate on the second dataset. Overall, the exact ratio and mIoU of our model show an increase of 1.7% and 1.3%, respectively. Furthermore, our model achieves better results on the 2-way 1-shot task with 4 folds, with an increase of 1.8% in the exact ratio and 1.4% in mIoU.

The performance improvement achieved by our proposed BFC-BL model can be attributed to the following factors: (1) The proposed bidirectional feature correlation fusion network enables the model to fuse the semantic and spatial correlation between different layers of learning samples; (2) The method of fusion of regions and boundaries is adopted to alleviate the problems of over-fitting and insufficient boundary information learning of the model for the central main area; (3) We combine the above two to construct an end-to-end model, which makes the convergence faster and the segmentation effect better.

4.3 Convergence and Visual analysis

The proposed BFC-BL network, which combines bidirectional feature correlation and boundary constraint, achieves superior experimental results and faster convergence compared to the baseline ASNet model. As shown in Figure 4(a) and 4(b), the convergence line graphs of the model training demonstrate that the accuracy rate and mIoU of our proposed model reach convergence between 200-250 epochs, outperforming the baseline. The experimental results in Figure 5 indicate that our model achieves satisfactory classification and segmentation performance, accurately identifying and segmenting objects. However, there is room for optimization in capturing finer details under challenging lighting conditions, such as overexposed or excessively dark backgrounds.

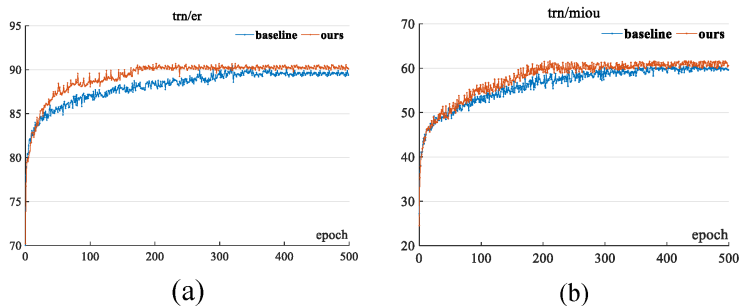


Figure 4: Comparison of model accuracy and convergence speed of mIoU training. Figure (a) with training accuracy on the left and Figure (b) with the convergence of sum and mIoU.

4.4 Ablation analysis

We propose a comprehensive loss function that incorporates both boundary semantics and region semantics with varying weight ratios. In our ablation experiments, we compare the hyperparameter α with four different weights: 0.001, 0.01, 0.05, and 0.1. The experimental results are presented in Table 2.

Table 2 reveals that when the weight α is set to 0.001 and 0.01, there is a slight improvement in learning boundary pixel information. However, the improvement is minimal, and in some folds, the results even decrease compared to the baseline. With the introduction

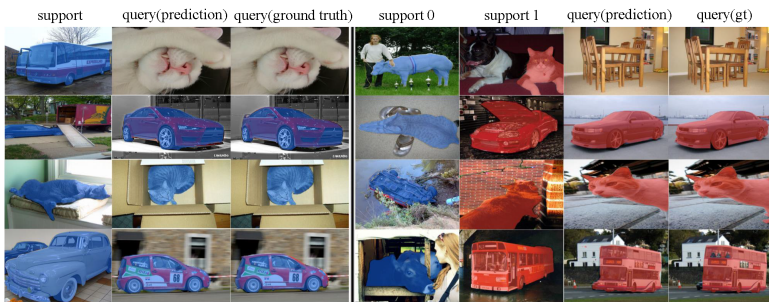


Figure 5: Figure of the experimental results of model 1-way 1-shot and 2-way 1-shot. The blue mask represents the first category of the support set, and the red mask represents the second category of the support set.

Table 2: Results of ablation experiments for hyperparameter α in boundary constraints.

Hyperparameters	classification 0/1 exact ratio (%)					Segmentation mIoU (%)				
	5^0	5^1	5^2	5^3	avg.	5^0	5^1	5^2	5^3	avg.
$\alpha=0.001$	85.9 $+1.0$	87.8 -1.8	80.1 $+1.1$	87.1 $+0.9$	85.2 $+0.3$	51.6 -0.1	61.9 $+0.4$	44.1 $+0.8$	53.1 $+0.3$	52.7 $+0.4$
$\alpha=0.01$	86.4 $+1.5$	88.1 -1.5	80.4 $+1.4$	87.6 $+1.4$	85.6 $+0.7$	52.1 $+0.4$	62.2 $+0.7$	44.1 $+0.8$	53.3 $+0.5$	52.9 $+0.6$
$\alpha=0.05$	87.4 $+2.5$	89.4 -0.2	81.1 $+2.1$	88.5 $+2.3$	86.6 $+1.7$	52.9 $+1.2$	62.7 $+1.2$	44.6 $+1.3$	53.9 $+1.4$	53.6 $+1.3$
$\alpha=0.1$	86.9 $+2.0$	89.5 -0.1	80.8 $+1.8$	87.8 $+1.6$	86.2 $+1.3$	52.5 $+0.8$	62.4 $+0.9$	44.2 $+0.9$	53.5 $+0.7$	53.1 $+0.8$

of $\alpha=0.05$, there is a significant enhancement in experimental results, with an average increase of 1.7% in accuracy rate (ER) and 1.3% in mIoU. The experiments demonstrate that an effective α allows the model to learn boundary information more effectively by increasing the weight of the boundary loss function, thereby focusing more on learning boundary pixels. It is worth noting that a larger weight for α does not necessarily lead to better results. Analysis of the proportion of target pixels indicates that the proportion of target pixels exceeds that of boundary pixels, suggesting that an excessively large boundary weight may not be an optimal learning strategy.

Table 3: Ablation experimental results of each module of the proposed model. "✓" means that the model adds this module, and "✗" means that the model removes this module

BFCP module	boundary constraints	classification 0/1 exact ratio (%)	Segmentation mIoU (%)
✓	✗	86.1 $+1.2$	52.5 $+0.2$
✗	✓	85.3 -0.8	53.2 $+0.9$
✓	✓	86.6 $+1.7$	53.6 $+1.3$

Based on Table 3, we comprehensively evaluate the impact of each module in the model on classification and segmentation performance. Adding only the BFCP module results in a 1.2% increase in classification accuracy, while the segmentation mIoU remains unchanged. Conversely, adding only boundary constraints leads to a decrease of 0.8% in classification accuracy, but an increase of 0.9% in segmentation mIoU. This indicates that the boundary constraints primarily enhance the mIoU of the segmentation boundary, while the classification accuracy is less affected. However, the best results of 86.6% and 53.6% are achieved when both modules are added simultaneously. The experiments demonstrate that the effectiveness of our model stems from the BFCP module's ability to capture the correlation between few-shot semantic features and sample correlations. Additionally, the inclusion of the boundary loss function enables the model to focus on learning target boundary pixels, resulting in overall improvements in both classification accuracy and segmentation mIoU.

5 Conclusion

Aiming at the problems of insufficient learning of correlation information between a small number of samples and the error of boundary pixel segmentation caused by the similarity of foreground and background, we propose Few-Shot Classification and Segmentation combining Bi-directional Feature Correlation and Boundary constraint (BFC-BL), which fuses deep semantic and shallow spatial feature correlation between a small number of samples, and constructs a multi-level weight ratio loss function. It makes the network convergence speed faster and generalization ability stronger.

Acknowledgements This work was supported in part by the Key Technology Project of Shunde District, Foshan City, under Grant 2130218003002.

References

- [1] Sergey Bartunov and Dmitry Vetrov. Few-shot generative modelling with generative matching networks. In *International Conference on Artificial Intelligence and Statistics*, pages 670–678. PMLR, 2018.
- [2] Henghui Ding, Hui Zhang, and Xudong Jiang. Self-regularized prototypical network for few-shot semantic segmentation. *Pattern Recognition*, 133:109018, 2023.
- [3] Linus Ericsson, Henry Gouk, and Timothy M Hospedales. Why do self-supervised models transfer on the impact of invariance on downstream tasks. In *The 33rd British Machine Vision Conference, 2022*, page 509. BMVA Press, 2022.
- [4] Honghao Gao, Junsheng Xiao, Yuyu Yin, Tong Liu, and Jianguang Shi. A mutually supervised graph attention network for few-shot segmentation: the perspective of fully utilizing limited samples. *IEEE Transactions on neural networks and learning systems*, 2022.
- [5] Golnaz Ghiasi, Tsung-Yi Lin, and Quoc V Le. Nas-fpn: Learning scalable feature pyramid architecture for object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7036–7045, 2019.
- [6] Olivier Henaff. Data-efficient image recognition with contrastive predictive coding. In *International conference on machine learning*, pages 4182–4192. PMLR, 2020.
- [7] Sunghwan Hong, Seokju Cho, Jisu Nam, Stephen Lin, and Seungryong Kim. Cost aggregation with 4d convolutional swin transformer for few-shot segmentation. In *Computer Vision—ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIX*, pages 108–126. Springer, 2022.
- [8] Dahyun Kang and Minsu Cho. Integrative few-shot learning for classification and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9979–9990, 2022.
- [9] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186, 2019.

- [10] Chunbo Lang, Gong Cheng, Binfei Tu, and Junwei Han. Learning what not to segment: A new perspective on few-shot segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8057–8067, 2022.
- [11] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.
- [12] Juhong Min, Dahyun Kang, and Minsu Cho. Hypercorrelation squeeze for few-shot segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6941–6952, 2021.
- [13] Seonghyeon Moon, Samuel S Sohn, Honglu Zhou, Sejong Yoon, Vladimir Pavlovic, Muhammad Haris Khan, and Mubbasis Kapadia. Hm: Hybrid masking for few-shot segmentation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XX*, pages 506–523. Springer, 2022.
- [14] Hang Qi, Matthew Brown, and David G Lowe. Low-shot learning with imprinted weights. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5822–5830, 2018.
- [15] Amirreza Shaban, Shray Bansal, Zhen Liu, Irfan Essa, and Byron Boots. One-shot learning for semantic segmentation. *The 28th British Machine Vision Conference, 2017*, pages 1–10, 2017.
- [16] Xiangwen Shi, Zhe Cui, Shaobing Zhang, Miao Cheng, Lian He, and Xianghong Tang. Multi-similarity based hyperrelation network for few-shot segmentation. *IET Image Processing*, 17(1):204–214, 2023.
- [17] Xinyu Shi, Dong Wei, Yu Zhang, Donghuan Lu, Munan Ning, Jiashun Chen, Kai Ma, and Yefeng Zheng. Dense cross-query-and-support attention weighted mask aggregation for few-shot segmentation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XX*, pages 151–168. Springer, 2022.
- [18] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 4080–4090, 2017.
- [19] Mingxing Tan, Ruoming Pang, and Quoc V Le. Efficientdet: Scalable and efficient object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10781–10790, 2020.
- [20] Zhuotao Tian, Hengshuang Zhao, Michelle Shu, Zhicheng Yang, Ruiyu Li, and Jiaya Jia. Prior guided feature enrichment network for few-shot segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 44(2):1050–1065, 2020.
- [21] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 6000–6010, 2017.

- [22] Kaixin Wang, Jun Hao Liew, Yingtian Zou, Daquan Zhou, and Jiashi Feng. Panet: Few-shot image semantic segmentation with prototype alignment. In *proceedings of the IEEE/CVF international conference on computer vision*, pages 9197–9206, 2019.
- [23] Zhitong Xiong, Haopeng Li, and Xiao Xiang Zhu. Doubly deformable aggregation of covariance matrices for few-shot segmentation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XX*, pages 133–150. Springer, 2022.
- [24] Xiaohua Zhai, Avital Oliver, Alexander Kolesnikov, and Lucas Beyer. S4l: Self-supervised semi-supervised learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1476–1485, 2019.
- [25] Qi Zhao, Binghao Liu, Shuchang Lyu, and Huojin Chen. A self-distillation embedded supervised affinity attention model for few-shot segmentation. *IEEE Transactions on Cognitive and Developmental Systems*, 2023.
- [26] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4*, pages 3–11. Springer, 2018.