Few-Shot Semantic Segmentation of Batik Patterns via Attention-Weighted Hierarchical Decoding

Yuzhou Ma , Haolong Qian , Wei Li *

Posted Date: 14 February 2026

doi: 10.20944/preprints202602.1167.v1

Keywords: semantic segmentation; feature extraction; information fusion; transfer learning; batik images

 doi:10.20944/preprints202602.1167.v1

Article

Few-Shot Semantic Segmentation of Batik Patterns via Attention-Weighted Hierarchical Decoding

Yuzhou Ma [1], Haolong Qian [2] and Wei Li [2,*]

[1] School of Cyber Security and Defense, Anhui Police College, Hefei 238076, China

[2] College of Computer Science, Sichuan Normal University, Chengdu 610101, China

* Correspondence: liw@sicnu.edu.cn

Abstract

The digital preservation of batik, a world intangible cultural heritage, is hindered by the difficulty in performing accurate semantic segmentation on its complex patterns with limited annotated samples. To address this few-shot learning challenge, we constructed a few-shot batik pattern dataset, and proposed a novel network architecture centered on attention weighting and hierarchical decoding. Our method leverages a pre-trained ResNet101 backbone for transfer learning to establish a strong feature foundation. It incorporates a dual-attention module that combines spatial and channel attention to dynamically highlight semantically rich regions and intricate texture boundaries specific to batik. For multi-scale context aggregation, a lightweight module utilizing parallel dilated convolutions is introduced to efficiently capture features from varying receptive fields. Finally, a hierarchical decoder progressively integrates these enhanced, multi-scale features with high-resolution shallow features to reconstruct precise segmentation maps. Comprehensive evaluations on a dedicated batik dataset show that our model achieves state-of-the-art performance, with a mean Intersection over Union (mIoU) of 79.22% and a Pixel Accuracy (PA) of 92.47%. It notably improves over the strong DeepLabV3+ baseline by 3.3% in mIoU and 0.95% in PA, demonstrating its effectiveness for the task of batik pattern segmentation under data-scarce conditions.

Keywords: semantic segmentation; feature extraction; information fusion; transfer learning; batik images

1. Introduction

Semantic segmentation, as a core technology in the field of computer vision, has demonstrated significant value in scenarios such as autonomous driving [1,2], remote sensing monitoring [3,4], and medical diagnosis [5–7]. However, general-purpose segmentation models (e.g., U-Net [8], DeepLab [9]) perform poorly in the face of complex art images, and their limitations become more and more prominent, especially in the digitization of intangible cultural heritage such as batik patterns. Batik patterns are characterized by fine handmade textures, interlaced geometric patterns, and unique wax cracking effects [10,11], and traditional models are prone to losing details due to deep network downsampling, and the large number of parameters leads to high arithmetic demand, which makes it difficult to be adapted to low-power embedded process equipment [12]. How to achieve high-precision and lightweight semantic segmentation under limited labeled data has become a key challenge for the digital preservation and innovative application of traditional crafts.

The unique attributes of batik art require models with both texture sensitivity and cultural adaptability. For example, the hierarchical difference between traditional blue and white batik and multi-color gradient patterns, and the distinction between natural texture formed by handmade wax cracking and machine-printed patterns require the model to incorporate the understanding of artistic features in pixel-level prediction [10]. Existing research mostly focuses on natural images and lacks targeted design for the complexity of artistic patterns, resulting in fuzzy segmentation boundaries and low-detail reproduction. In addition, industrialized applications, textile printing equipment, or

 doi:10.20944/preprints202602.1167.v1

mobile design tools have urgent needs for model lightweight, while the high computational power dependence of general-purpose models makes it difficult to meet the requirements of real-time processing and low-cost deployment. These problems not only hinder the efficiency of the digital preservation of cultural heritage but also limit the innovative transformation of batik art in contemporary design. To address the above challenges, this paper contributes as follows:

- This paper presents the first study on semantic image segmentation of intangible cultural heritage batik patterns. To this end, we have constructed a few-shot batik pattern dataset (publicly available for download) and annotated the batik images. Our annotation methodology was jointly developed with batik experts to ensure the accuracy and scientific rigor of the dataset.
- Aiming at the problems of complicated texture details and blurred salient regions in batik images, this paper designs a multi-scale feature extraction network combining parallel dual-path structure and cavity convolution to realize feature capture and fusion from fine-grained to coarse-grained, which on the one hand effectively expands the sensory field of the network, and on the other hand makes use of auxiliary paths for feature compression and lightweighting, which reduces the amount of computation and guarantees the completeness of the key features at the same time.
- In the feature fusion stage, a dual-attention module combining channel attention and spatial attention is introduced, where channel attention can strengthen features with high information content, and spatial attention highlights the localization of key regions in the image, especially in the segmentation of complex patterns and abstract textures in batik, which shows higher accuracy, and effectively counteracts the defects of traditional methods that tend to ignore subtle features or key textures.
- Through the combination of hierarchical fusion decoding module and data enhancement strategy, not only the complementary relationship between high-level semantics and low-level details is fully explored during the decoding process, which improves the adaptability of the model to diversified interferences and deformations, but also the overfitting risk of few-shot samples is alleviated through various data enhancement operations such as rotating, flipping, cropping, etc., and the model's robustness and generalization ability are further strengthened.

Through comparative experiments conducted with mainstream semantic image segmentation methods such as Google's DeepLabV3+, the method in this paper has achieved significant improvement in both the segmentation accuracy of batik images and the attention to batik patterns, which provides a strong technical support for the automated processing and inheritance of batik images.

2. Related Works

Few-shot sample semantic segmentation refers to the mining of generalizable feature representations to achieve high-precision pixel-level segmentation of new categories of targets when the available annotation samples are extremely limited, to effectively address the segmentation needs in data-scarce scenarios. A series of achievements have already been made in the field of few-shot semantic segmentation. Guo et al [13] proposes a novel end-to-end model named GFormer to enhance generalization on unseen classes. Its key contribution is the introduction of an adversarial prototype generator that distinguishes between positive and negative examples. By using different but visually similar categories as negative samples against the target (positive) support classes, the model learns more discriminative features. Combined with a modified ViT encoder that reduces overfitting to image content and a prototype-guided decoder, this approach significantly improves segmentation accuracy for novel classes in few-shot settings. Lee et al [14] proposes a semi-supervised, few-shot learning framework for pointwise Vortex-Induced Vibration detection in structural health monitoring. Its core contribution is a practical methodology that requires only a small set of labeled real non-VIV samples and synthetic VIV data to bootstrap an initial model. This model then pseudo-labels large-scale unlabeled data and is refined through sequential transfer learning, effectively

 doi:10.20944/preprints202602.1167.v1

eliminating the need for extensive manual annotation. The integration of regularization techniques (high softmax threshold and label smoothing) ensures robustness against diverse, unseen environmental conditions, providing a scalable and cost-effective solution for real-world monitoring. He et al [15] proposed a new cross-domain few-shot sample semantic segmentation framework called APSeg, which realizes cross-domain adaptation without additional fine-tuning by using Segment Anything Model (SAM) with huge pretrained features and combining with Dual Prototype Anchor Transform (DPAT) and MetaPrompt Generation (MPG) modules, and robustly maps semantic features and automatically generates semantic features by fusing the support with the query samples' prototype information, and then automatically generates semantic features by fusing the support with the query samples' prototype information. By fusing support and query sample prototype information to robustly map semantic features and automatically generate cue vectors instead of manual input, it significantly improves segmentation accuracy and simplifies the reasoning process, providing an efficient and robust end-to-end solution for cross-domain semantic segmentation in low-labeling environments. Aiming at the association noise problem caused by cross-category local similarities and intra-class differences in few-shot sample semantic segmentation. Liu et al [16] proposed FECANet based on feature enhancement and context awareness: firstly, noise is suppressed and semantic consistency is enhanced by cross-attention and channel-attention, and then an association reconstruction module that integrates dense overall associations and multi-scale self-similar contextual information is constructed, which guides the network to captures more complete and fine-grained query-support matching patterns, and finally introduces a residual decoder to refine the output multiple times and extract the target boundary effectively. To address the challenge of few-shot sample semantic segmentation relying on large-scale annotations, Ding et al. [17] proposed the Self-Regularizing Prototype Network (SRPN): by comprehensively evaluating the prototyping effect in the support set, explicitly optimizing the prototype's representation of interclass distinguishability and intraclass completeness, and utilizing a distance metric named fidelity distance metric and iterative query inference mechanism to fully retain key semantic and detail information to effectively migrate to the query set under the premise of accurately reconstructing the semantics of the support set, to achieve more accurate and robust pixel-level segmentation of unseen classes. Cao et al. [18] proposed a support set prototype vector as the query of the Transformer decoder, which is combined with the Key-Value mechanism of the query image to generate an adaptive segmentation kernel (ProtoFormer), which significantly enhances the ability of spatial detail capture and target focusing, and thus achieves excellent segmentation performance under the condition of few-shot samples. Lu et al. [19] proposed a fractional normalization fusion and a cross-covariance Transformer-based prediction correction method (PCN) to deal with the problems of skewed prediction and over-confidence for the base and the new classes. Based prediction correction method (PCN), which effectively alleviates the bias while improving the generalized few-shot sample semantic segmentation accuracy.

The aforementioned few-shot segmentation methods generally face bottlenecks such as insufficient cross-domain adaptability, feature expression susceptible to noise interference, insufficiently fine-grained prototype generation and detail capture, etc. APSeg focuses on interactive cueing and out-of-domain generalization, while FECANet emphasizes feature augmentation and contextual reconfiguration, but relatively insufficiently balances the frequency-domain details and category imbalance; the auto-regularized prototype network (ARPN) inverts constraints on the quality of the prototypes but does not take full advantage of the complex texture and multi-scalar semantic segmentation. ProtoFormer and PCN focus on prototype matching or predictive correction transformers. Based on the above limitations and shortcomings of the few-shot sample semantic segmentation models for batik data.

Cross-domain transfer learning is a method of feature mapping and knowledge migration between source and target domains to mitigate data distribution differences and improve the learning performance of the target task. There are different schemes for cross-domain migration learning in different application scenarios, Gao et al [20], for the domain difference between ground

 doi:10.20944/preprints202602.1167.v1

and UAV images, based on deep convolutional network and differential preprocessing strategy to achieve accurate recognition of multi-source weed images by a single model, and successfully migrate the training results of the source domain to the target domain. Cuttano et al [21] proposed the CoRTe method, which can be applied to the target domain through relative confidence screening and a self-refined pseudo-labeling strategy. Confidence screening and self-refined pseudo-labeling strategy, which achieves reliable cross-domain migration learning from black-box source models to lightweight target models without direct access to source data and model details. Tan et al [22] proposed two new metrics, F-OTCE and JC-OTCE, which achieve accurate quantification of cross-domain cross-task migration learning through optimal transfer and conditional entropy without additional auxiliary tasks and apply the results of cross-domain cross-task migration learning to the target domain. Li et al. [23] proposed a single-model cross-domain migration framework for multi-target load monitoring by fixing the feature extraction layer and fine-tuning the output layer of the target domain to achieve efficient migration and adaptation of cross-domain power features, which can be used in different data sets. By fixing the feature extraction layer and fine-tuning the target domain output layer, efficient migration and adaptation of power features across domains can be realized, which significantly reduces the annotation requirement and maintains excellent decomposition accuracy on different datasets and equipment types. For the cross-modal domain problem caused by the difference between optical and SAR imaging, Gao et al [24] proposed the ADCG network, which uses lightweight dense connections and convolutional attention module to perform pixel-level feature conversion, effectively alleviating the insufficient labeling and category imbalance of SAR data and significantly improving the ship identification accuracy. Fang et al [25] proposed a new framework of deep cross-domain migration learning based on a particle swarm algorithm, which significantly improves the cross-domain anomaly detection accuracy by separating the conditional domain differences between normal and abnormal samples and integrating the edge distribution and conditional distribution. Wang et al [26] proposes a novel Cross-Dimensional Information Transfer Framework (CDITF) to bridge the fundamental gap between computer vision (CV) and bioelectrical signal domains. Its key contribution is a cross-task domain adaptation method (CDITF-CTDA) that jointly addresses the dimensional disparity (1D vs. 2D data) and task/label inconsistency between these domains. By using parallel encoding-decoding modules to map both data types into a shared feature space while preserving their discriminative characteristics, and by synergistically integrating feature and label information for adaptation, the framework enables effective knowledge transfer from label-rich CV data to label-scarce bioelectrical signal analysis. This achieves successful cross-dimensional, cross-domain learning with limited medical labels, consistently improving performance metrics like AUC.

Existing research focuses on cross-domain migration in general-purpose scenarios, including feature mapping between optical-radar images, cross-domain fusion of multi-source remote sensing images, non-intrusive load monitoring, multicategory target detection and identification, etc. Although it is effective in reducing annotation requirements and improving generalization ability, most of the methods rely on domains where the feature distribution is relatively stable or the size of the training samples is large. For the scene of batik pattern segmentation details and lack of annotation, these methods have limitations in feature extraction fineness and embedded deployment requirements.

To address the above issues, this paper proposes improved methods. In the field of few-shot semantic segmentation, this paper designs a two-way coding feature extraction architecture, a hierarchical fusion decoding architecture, and a two-phase fusion decoding architecture, to address the difficulties in batik images such as sparse samples, varied textures, and category imbalance. In cross-domain transfer learning, we employ a fine-tuning strategy leveraging pre-trained models—freezing low-level generic features, tuning high-level decision modules, and integrating targeted optimizations via data augmentation and lightweight architecture design.

3. Methods

 doi:10.20944/preprints202602.1167.v1

As a typical representative of intangible cultural heritage, the batik pattern dataset has few-shot samples, high complexity, and unique process characteristics. On the one hand, the handmade characteristics lead to limited data size, which is difficult to meet the comprehensive training requirements of deep learning models. On the other hand, the interweaving of the gradient texture formed by the halo-dye penetration of batik and the fine ice crack texture leads traditional segmentation models to easily fall into the dual predicament of overfitting and computational redundancy, making it difficult to balance segmentation accuracy and operational efficiency. Aiming at the above problems, this study proposes a lightweight semantic segmentation framework for few-shot sample scenarios. First, a cross-domain feature migration mechanism is constructed based on the idea of transfer learning, and deep semantic a priori features are extracted from a pre-trained model using ResNet101, which can effectively alleviate the problem of insufficient feature characterization under few-shot sample training and improve the robustness and generalization ability of the model at the same time. Secondly, to reduce the model parameter size while strengthening the ability to focus on the haloed boundaries and detail regions in the batik image, this paper designs a lightweight dual-attention module, which adopts spatial-channel separated weight calculation to strengthen the focus on the target region from multiple perspectives and suppress the background redundant information. In addition, to take into account the multi-scale feature capture of microstructure and macro-morphology, this study constructs a multi-branch inflationary convolutional structure, which reduces the computation amount while maintaining the comprehensive characterization capability of texture details and global semantics through parallel dilated convolutional layers. A progressive feature fusion strategy is adopted in the hierarchical decoder, which makes full use of the encoder's shallow texture features and deep semantic features for dynamic multiplexing, thus effectively improving the segmentation accuracy in the decoding stage.

3.1. Modeling Framework

The model framework of this paper is shown in Figure 1, and its core architecture adopts a hierarchical progressive feature processing mechanism to realize efficient extraction and reconstruction of semantic information through multi-module collaboration. The algorithmic process contains four modules: depth migration-based trunk feature extraction module, dual-attention feature enhancement module, lightweight multi-scale feature fusion module, and hierarchical semantic decoding module. This model framework reduces the parameter scale and computational complexity while guaranteeing the model accuracy.

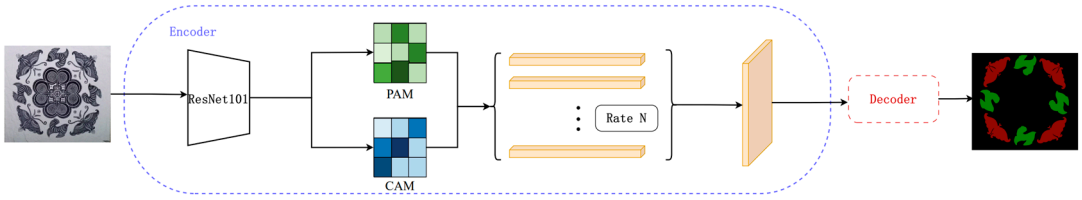

Figure 1. Overall Network Structure.

Aiming at the characteristic of the scarcity of batik pattern samples, this model adopts ResNet101 as the backbone network and effectively migrates the pre-trained knowledge to the target domain through the migration learning strategy. Dual-attention feature enhancement module, in the spatial dimension, PAM constructs the dynamic sensory field by deformable convolution and realizes pixel-level feature recalibration by using spatial location correlation matrix; in the channel dimension, CAM adopts a lightweight module to generate the channel weight vectors by global average pooling. The outputs of both are fused by adaptive weighting to form an enhanced feature map. This module introduces a parameter-sharing mechanism so that the dual-attention sub-network shares the underlying convolutional kernel, effectively controlling the model complexity. Adopt hollow space

 doi:10.20944/preprints202602.1167.v1

pyramid pooling for multi-scale context extraction of attention enhancement features, and use parallel convolution with differentiated expansion rate to capture different granularity features and reduce computational redundancy. The feature reorganization unit is designed to achieve channel compression by 1×1 convolution and cross-scale feature interaction using depth-separable convolution. To achieve a balance between detail preservation and semantic restoration, the upsampling phase of this model proposes a lightweight progressive feature decoding strategy.

3.2. Backbone Feature Extraction Module

ResNet101 is a deep convolutional neural network based on a residual learning framework, and its core design idea mitigates the gradient vanishing problem in deep networks through Skip Connection, thus supporting the stable training of ultra-hundred-layer networks. As shown in the model structure (Figure 2), the network adopts a staged feature extraction strategy, which compresses the feature map resolution by downsampling step by step and gradually increases the channel dimensions to capture multilevel semantic information.

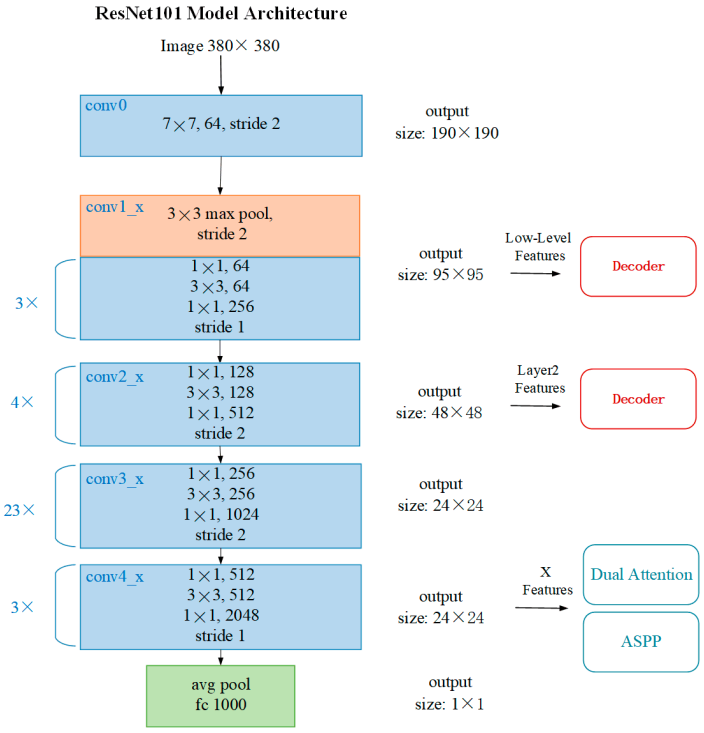

Figure 2. Backbone residual network diagram.

In the input stage, the conv0 layer uses a 7×7 large-size convolutional kernels (number of channels 64, stride size 2) to downsample the input image from the original resolution of 380×380 to 190×190. This design quickly captures the local texture features through a larger sensory field, and at the same time realizes the compression of spatial dimensions by utilizing stride size 2, which significantly reduces the amount of the subsequent computation. Subsequently, the conv1_x layer further reduces the feature map size to 95×95 by 3×3 maximum pooling (stride 2), an operation that effectively reduces redundant computation while preserving significant features. Low-level features (LLFs) then enter the decoder module, where shallow features may be multiplexed through cross-layer connections to enhance detail retention in few-shot sample scenarios.

The core residual module (conv2_x to conv4_x) adopts the classic Bottleneck structure, with the conv2_x stage downsampling the feature map from 95×95 to 48×48 with stride 2 while extending the number of channels from 64 to 512 through a 1×1/128→3×3/128→1×1/512 structure. The high-level feature extraction stage (conv3_x vs. conv4_x) strengthens the semantic characterization capability through progressive channel expansion (512→1024→2048). conv3_x uses stride 2 to downsample the

 doi:10.20944/preprints202602.1167.v1

resolution to 24×24, while conv4_x maintains the resolution constant through stride 1, at which point a Dilated Convolution is used to expand the receptive field without increasing the number of parameters. Number to extend the sensory field to improve the generalization ability in few-shot sample learning. Global Average Pooling (GAP) compresses the 24×24×2048 feature map into 2048-dimensional vectors, which are mapped to 1000 classes of classification results by a fully connected layer.

3.3. Dual Attention Feature Enhancement Module

In the dual attention mechanism shown in Figure 3, the green module PAM on the left (centered on the spatial dimensionality operations of Q, K, and V) is mainly responsible for capturing the long-range dependencies of features on the spatial domain. The input feature map $X \in R^{C\times H\times W}$ is rearranged into three tensors of queries (Q), keys (K) and values (V) for generating the spatial attention distributions: the queries and keys are transformed into $R^{HW\times\frac{C}{8}}$ versus $R^{\frac{C}{8}\times HW}$, respectively, to obtain a spatial attention distribution via the Matrix multiplication to obtain a correlation map of $R^{HW\times HW}$, followed by softmax normalization to obtain a spatial attention matrix A_p. This attention matrix can be viewed as a weighted mapping that establishes explicit associations between spatial locations to measure the degree of dependency between different pixels (or feature locations). The attention matrix is then multiplied with the numerical tensor $V \in R^{C\times HW}$ to obtain the feature representation $R^{C\times HW}$ that incorporates the global spatial context. After the corresponding reshape operation, it is multiplied element-by-element with the original feature map according to the channel dimension to obtain M_p, which enables the network to explicitly capture pixel correlations in the global context and feed such global information into the feature map. The spatial attention module effectively improves the sensitivity to the target or scene, enabling the network to have better performance in tasks such as parsing object edges, region consistency, and detail inference.

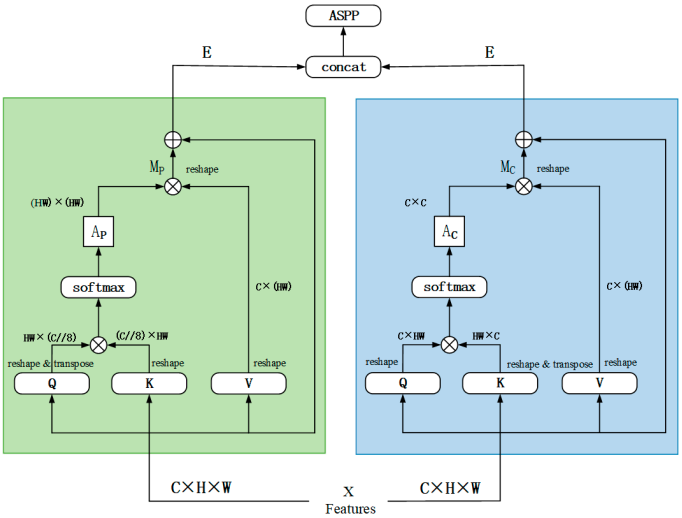

Figure 3. Dual Attention Feature Enhancement Module Diagram.

In contrast, the blue module CAM on the right side focuses on modeling global relationships on the channel domain, adopting a similar idea of attention computation as on the left side, but transforming the shapes of Q, K, and V into forms that are capable of interacting in the channel dimension: the query and key features are rearranged to $R^{C\times HW}$ and $R^{HW\times C}$, respectively, so that the resulting attention matrix A_c is computed as $R^{C\times C}$, is used to characterize the degree of inter-channel dependence. By softmax normalization, this channel attention matrix reveals which channels are more discriminative or more correlated with other channels. A_c is multiplied with the numerical tensor $V \in R^{C\times HW}$ to obtain the channel-enhanced contextual features $R^{C\times HW}$, which are then multiplied element-by-element with the original features to obtain M_c. After the two modules have captured the information in the spatial domain and the channel domain, respectively, the features

 doi:10.20944/preprints202602.1167.v1

are fused with the original features by concatenating them with subsequent operations, such as space pyramid pooling (ASPP). The features are then fused through concatenation and subsequent operations such as ASPP, thus giving the network the ability to model long-term dependencies at both the pixel and channel levels. The final output features can take into account the global context of spatial locations and channel relationships, which not only preserve the fine boundaries of the target but also strengthen the collaborative effect of the channels in determining the target category or scene semantics, which enables the model to achieve better representation in semantic segmentation visual tasks.

3.4. Lightweight Multi-Scale Feature Fusion Module

This paper adopts a lightweight multi-scale feature fusion module shown in Figure 4, whose core design is based on a parallel multi-branch structure with an efficient feature interaction mechanism, aiming to capture multi-level spatial context information with a low number of parameters and low computational complexity. The main body of the module consists of four heterogeneous branches: the main branch performs channel dimensionality reduction on the input features through 1×1 convolution to reduce the subsequent computation; the three parallel subbranches adopt the 3×3 depth-separable expansion convolution with expansion rates of 6, 12, and 18, respectively, and construct multi-level sensory fields by differentiating the expansion rates under the premise of avoiding explicit downsampling, to efficiently capture the local details, medium-scope semantics, and global context information. The last branch uses a global image pooling layer to extract spatially independent global statistical features, which are spliced with the outputs of the first four branches after adjusting the dimensionality by 1×1 convolution to form a cross-scale complementary feature representation.

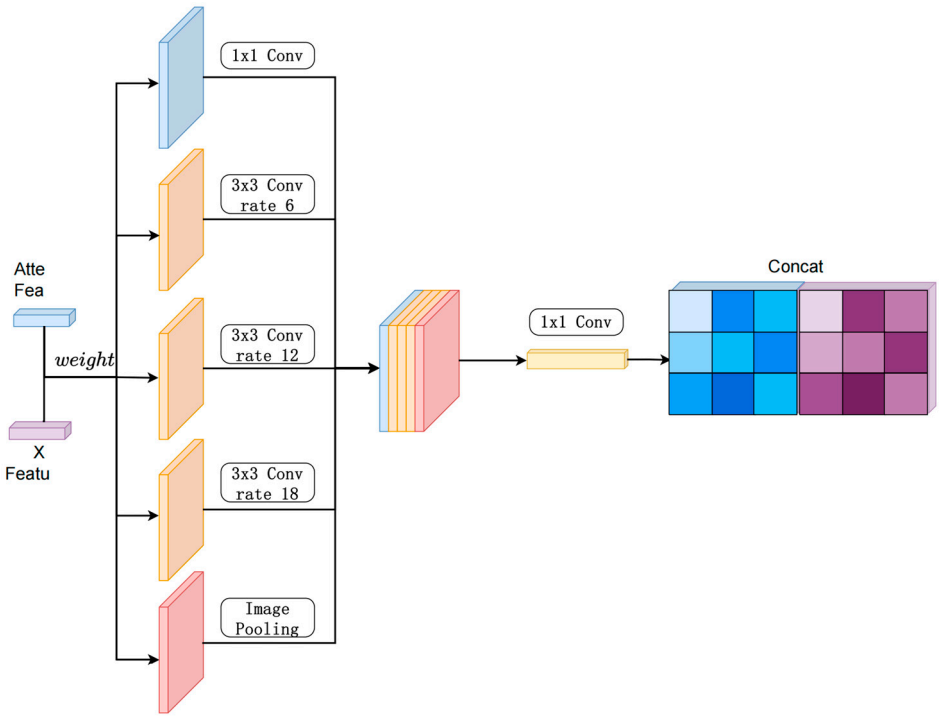

Figure 4. Lightweight Multiscale Feature Fusion Module Map.

In terms of lightweight design, the introduction of expansion convolution makes the network expand the effective sensory field without increasing the number of parameters. The feature fusion stage adopts channel splicing instead of element-by-element summation to avoid the problem of information suppression between features of different scales, while the output channels of each

 doi:10.20944/preprints202602.1167.v1

branch are compressed by the front 1×1 convolution to ensure that the total number of channels after splicing does not exceed the input dimensions, to control the memory occupation.

This module achieves multi-granularity feature decoupling at the microscopic level through heterogeneous expansion rate design and attention co-optimization. The nested structure of depthwise separable convolution and channel compression is used to reduce the computational density under the premise of guaranteeing the feature expressive power. The complementary design of global pooled branches and local convolutional branches enables the network to model long-range dependencies and local structural patterns at the same time.

3.5. Hierarchical Semantic Decoding Module

To better demonstrate the decoding process of the whole network, this model is multi-optimized in terms of feature fusion and up-sampling strategies. As shown in Figure 5, the decoder consists of four serial layers that combine the multilevel features from ResNet101 with the information provided by the dilated convolution (DC) and attention (DA) modules at different stages to achieve accurate recovery of the target region. The core idea of this decoding structure is: first reduce the computational cost by average pooling and upsampling while retaining the necessary spatial details; then extract the multiscale semantic representations by using the dilation convolution and enhance the key regions by combining with the attention module; and finally, fuse with the shallow features from ResNet101 to ensure the finesse and accuracy of the high-resolution output.

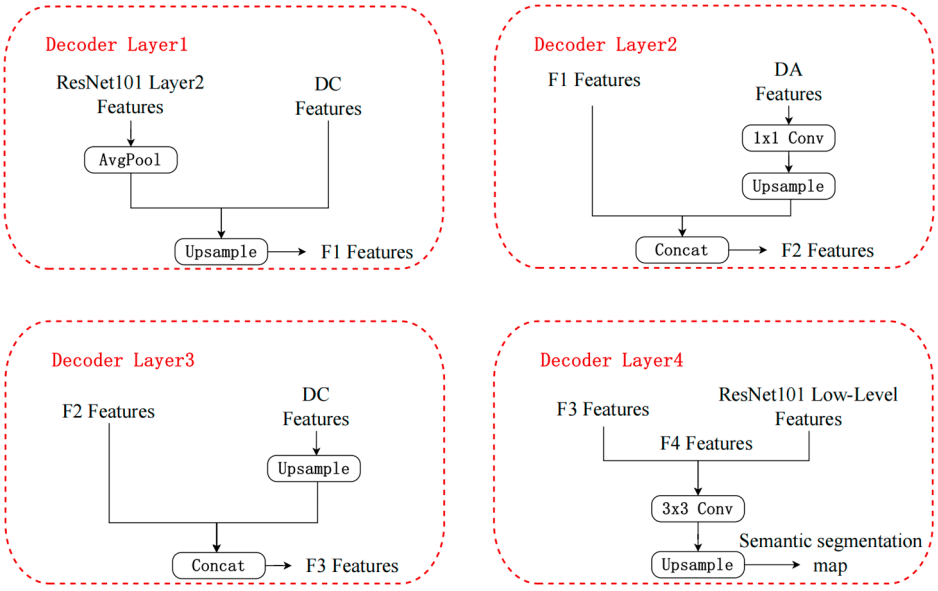

Figure 5. Hierarchical Semantic Decoding Module Diagram.

The first layer of the decoder extracts features mainly from the second layer of ResNet101 and performs dimensionality reduction through average pooling to reduce the subsequent computational complexity. After an upsampling operation, these features are scaled to a higher resolution.

Based on this, the features from the inflated convolution module are further fused and combined with the features from ResNet101 through the same up-sampling operation to form the F1 features. The main goal of this layer is to utilize the multi-scale features extracted by the expansion convolution and the low-level features extracted by ResNet101 to improve the spatial resolution through the up-sampling operation and provide richer feature information for the subsequent layers.

In the second layer, the decoder receives the F1 features from the previous layer and fuses them with the features from the attention module. DA features are processed by 1x1 convolution and then up-sampled. These processed DA features are subjected to concatenation operation with F1 features to generate F2 features. This process ensures that the decoder can fully utilize the useful information

 doi:10.20944/preprints202602.1167.v1

learned from the attentional feature module while combining the details of the F1 features to lay the foundation for subsequent feature fusion.

The main function of the third layer decoder is to fuse the F2 features with the output of the inflated convolution module. The DC features are up-sampled to restore spatial resolution. The F2 features are then concatenated with the up-sampled DC features to generate F3 features. The focus of this layer is to continue to utilize the multi-scale information to enhance the spatial details of the F2 features while keeping the semantic information of the original image undistorted.

The final layer of the decoder is responsible for combining the F3 features with inputs from ResNet101 low-level features. At this layer, the F3 features are first fused with the low-level features from ResNet101 to generate F4 features. The F4 features are processed by 3x3 convolution to further extract finer features. After the upsampling operation, the final semantic segmentation result is obtained. The core purpose of this layer is to recover more detailed information by fusing low-level features, which in turn improves the accuracy and quality of the semantic segmentation graph.

Through the above-layered design, the model can fully retain the key information from the deep and shallow layers in the decoding stage, which ensures the completeness of the segmentation result at the semantic level and also takes into account the accuracy of the spatial resolution. The synergistic effect of the expansion convolution and the attention mechanism enables the multi-scale context to be effectively utilized while suppressing the interference of background noise. Ultimately, by fusing with low-level features again at the last layer, the model can output a more detailed semantic segmentation map, providing robust feature representation and high-quality segmentation results for tasks such as scene understanding and target detection.

4. Experimental Evaluation

4.1. Datasets

A sample of the batik dataset is shown in Figure 6, where the same category of batik patterns contains a variety of styles that are richly textured and abstracted.

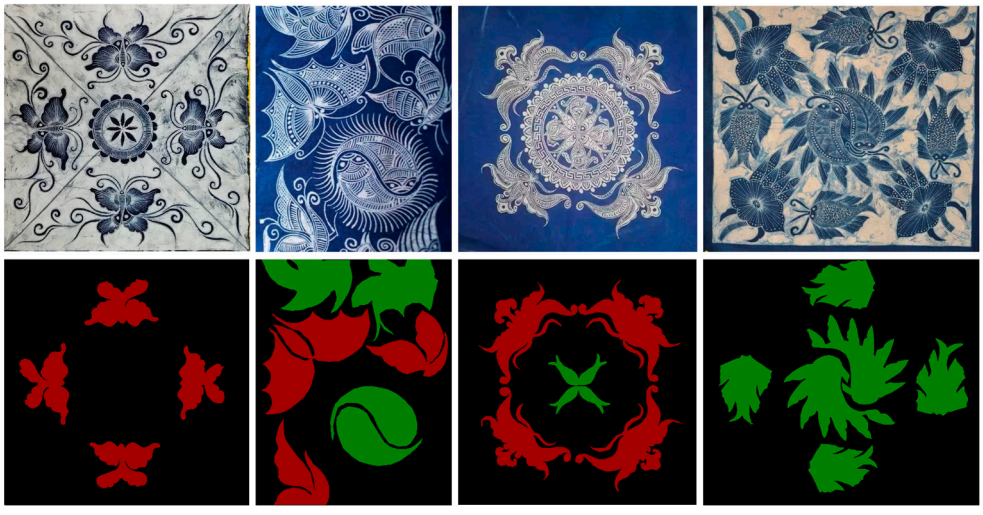

Figure 6. Examples of batik patterns and their labeled drawings.

This paper uses the LabelMe tool to label batik patterns and generate image labels in PASCAL VOC data format for subsequent training. In this paper, image data was collected and organized from various sources, which contain 106 images of butterfly patterns, 109 images of fish patterns, and 47 images of multi-category patterns. In the experiment, the data is divided into training images and validation images according to each category, with 80% of training images and 20% of validation images. The validation images include fish pattern images only, butterfly pattern images only, and composite images containing elemental patterns such as fish and butterflies, which cover every

 doi:10.20944/preprints202602.1167.v1

category and complexity, to better reflect the model's semantic segmentation ability and generalization. In this experiment, each image is cropped into 380×380 pixel image blocks as input images.

4.2. Experimental Environment and Parameter Settings

In this study, Linux Ubuntu 20.04 is used as the experimental environment, based on the hardware configuration of Intel(R) Core(TM) i7 13700F, 32GB of RAM, and a GeForce RTX 4090 GPU with 24GB of graphics memory. The software tools used include PyTorch 2.0.1 and CUDA.

To improve the diversity of the data and the generalization ability of the model, data augmentation operations such as scaling, rotating, and flipping were performed on the original images in this paper. Feature backbone extraction was performed using four convolutional layers and one average pooling layer of ResNet101 and pretrained on PASCAL VOC. To ensure fairness, the input image size of all comparison models is 380x380, the learning rate is 1e-2, the momentum is 0.9, the batch size is 8, and the training termination condition is set to 20 consecutive epochs of model performance without enhancement, and the optimal model parameters are saved.

4.3. Evaluation Indicators

To evaluate the semantic segmentation performance of the model, this paper analyzes and compares the experimental results in terms of both quantitative assessment and subjective evaluation. The quantitative evaluation mainly quantifies the segmentation accuracy and robustness of the model by calculating two evaluation metrics, Pixel Accuracy (PA) and Mean Intersection over Union (mIoU). The subjective evaluation mainly focuses on analyzing the model's segmentation effect on batik patterns of different categories and complexity by observing the resultant images of semantic segmentation.

Pixel Accuracy (PA), also known as Global Accuracy, is a metric for measuring the number of correctly classified pixels as a percentage of the total number of pixels. As shown in Equation (1), PA reflects the classification model's ability to predict all samples and a higher accuracy indicates a more accurate model. Where n_{jj} denotes the number of correctly categorized samples in category j, t_j denotes the total number of samples in category j, and the accuracy rate is defined as the ratio of the number of correctly categorized samples to the total number of samples in all categories.

$$PA = \frac{\sum_{j=1}^{k} n_{jj}}{\sum_{j=1}^{k} t_j} \tag{1}$$

The Intersection over Union (IoU) for a specific category j is given in Equation (2), where k denotes the number of categories, and n_{ij} represents the number of pixels with true label i that are predicted as label j. The IoU is calculated as the ratio of the intersection (i.e., correctly classified pixels n_{jj}) to the union (i.e., total pixels in either the ground truth or prediction for category j). This metric reflects the segmentation accuracy for each individual category.

$$\mathrm{IoU}_j = \frac{\mathrm{n_{jj}}}{\sum_{\mathrm{i=1}}^{\mathrm{k}} \mathrm{n_{ij}} + \sum_{\mathrm{i=1}}^{\mathrm{k}} \mathrm{n_{ji}} - \mathrm{n_{jj}}} \tag{2}$$

Mean Intersection over Union (mIoU), shown in Equation 3, is a metric used to evaluate the performance of a semantic segmentation model, which calculates the degree of overlap between the true and predicted labels. mIoU is higher, which indicates that the model is more capable of predicting all the categories. mIoU is the number of pixels that calculate the IoU for each category, and then find the average value of all the average of the category IoUs.

$$\mathrm{mIoU} = \frac{1}{\mathrm{k}} \sum_{\mathrm{j=1}}^{\mathrm{k}} \frac{\mathrm{n_{jj}}}{\sum_{\mathrm{i=1}}^{\mathrm{k}} \mathrm{n_{ij}} + \sum_{\mathrm{i=1}}^{\mathrm{k}} \mathrm{n_{ji}} - \mathrm{n_{jj}}} \tag{3}$$

4.4. Ablation Study

 doi:10.20944/preprints202602.1167.v1

4.4.1. Encoder Ablation Experiment

The results of the ablation experiments in Table 1 systematically reveal the mechanisms by which different encoder modules contribute to the semantic segmentation performance.

Table 1. Encoder ablation experiment results.

Backbone	Description	mIoU	PA
ResNet101	None	54.51%	82.37%
ResNet101	Dual Attention Feature Enhancement Module	76.02%	91.34%
ResNet101	Dual Attention + Multiscale Features	79.22%	92.47%

The baseline model based on ResNet101 achieves 54.51% and 82.37% in mIoU and PA metrics, respectively, a result that validates the effectiveness of the baseline architecture at the level of feature extraction, but also exposes its limitations in terms of insufficient feature characterization capability in complex scenarios. When the dualattention feature enhancement module is introduced, the mood is significantly improved to 76.02% (relative improvement of 21.51%), and the PA reaches 91.34% (relative improvement of 8.97%), which confirms that the module can effectively enhance the response strength of the key features through the dual-attention mechanism of channel and space. After further adopting the lightweight multi-scale feature fusion module, the model performance continues to improve to mIoU 79.22% and PA 92.47% and its 3.2 percentage point mIoU gain verifies the importance of cross-scale feature fusion for spatial detail preservation, and the module achieves a balance between sensory field expansion and local feature preservation through the feature fusion mechanism.

4.4.2. Decoder Ablation Experiment

In this section, to explore the effect of different decoders on the performance of semantic segmentation models, ablation experiments are done on the decoders. As shown in Table 2, this section systematically verifies the optimization effect of layered decoding on the performance of semantic segmentation by gradually introducing the layered decoder module.

Table 2. Decoder ablation experiment results

Backbone	Description	mIoU	PA
ResNet101	Decoder (Layer1+Layer4)	73.43%	90.65%
ResNet101	Decoder (Layer1+Layer2+Layer4)	75.65%	91.48%
ResNet101	Decoder (Layer1+Layer2+Layer3+Layer4)	79.22%	92.47%

When Layer1 (containing AvgPool and DC features of ResNet101 Layer2) and Layer4 (low-level feature fusion module) are used to form the base decoder, the model obtains 73.43% of mIoU and 90.65% of PA on the test set, which indicates that the combination of void convolution-based context modeling and shallow features has already possessed basic segmentation capability. After the introduction of Layer2 (with DA dual-attention mechanism and extended convolution), the mood is significantly improved by 2.22 percentage points to 75.65% and the PA is increased by 0.83 percentage points to 91.48%, which is attributed to the fact that the dual-attention module effectively enhances feature discriminative through channel and spatial co-modelling, while the extended convolution maintains the feature resolution while enlarging the sensory field, and the two of them synergistically optimize the semantic representation of middle and high-level features.

When the complete integration of Layer3 (with DC feature up-sampling and cross-layer connectivity) is used to build a four-level decoding architecture, the model performance achieves the most substantial improvement, with mIoU reaching 79.22% and PA rising to 92.47%, which is 5.79% and 1.82% higher than the baseline model, respectively. This result shows that Layer3 achieves dynamic alignment of deep semantic information with mid-level detailed features through multi-

 doi:10.20944/preprints202602.1167.v1

scale feature extraction with progressive up-sampling mechanism by cavity convolution, while cross-layer connectivity constructs a hierarchical feature pyramid through feature splicing operations, which enables the model to capture both global context and local boundary information. It is worth noting that the performance gain shows a nonlinear growth trend with the deepening of decoder layers, and the introduction of Layer3 brings a 3.57% jump in mIoU, which is significantly higher than the gain magnitude of Layer2, verifying the necessity of the multilevel feature refinement mechanism in complex scene parsing. The experimental results fully demonstrate that the layer decoder designed in this paper realizes the progressive reconstruction of semantic information from coarse-grained to fine-grained through the refined feature recalibration of the DA module, the multi-scale context fusion of the DC module, and the progressive feature up-sampling strategy, which provides an effective decoding scheme to improve the segmentation accuracy.

4.5. Comparative Experiment

4.5.1. Comparative Experiments of Multiscale Feature Fusion Modules

The lightweight multi-scale feature fusion module proposed in this paper improves the model's ability to represent complex visual patterns by hierarchically integrating features from different receptive fields and attention mechanisms.

As shown in Figure 7 (c), a three-level feature fusion mechanism is adopted to capture multi-scale spatial context information through dilated convolution, whose differential expansion rate can effectively cover the feature distribution from local details to global structure; a dual-attention mechanism is introduced to perform adaptive recalibration of channel and spatial dimensions to strengthen the response strength of important feature regions; and the original input features (X) are deeply spliced to the above-optimized features through cross-layer connections, forming a model with both detail preserving and semantic features. The original input features (X) and the above-optimized features are deeply spliced through a cross-layer connection to form a hybrid feature representation with both detail preservation and semantic enhancement. Compared with the traditional baseline model (e.g., Figure 7 (a) using only linear superposition), this design avoids the problem of detail loss caused by the traditional cascade operation (e.g., Figure 7 (b) simply splicing DA and DC) by introducing the hopping connection to retain the underlying high-resolution information. During the feature fusion process, the inflated sampling strategy of dilated convolution complements the feature selection property of the attention mechanism, where the former explicitly constructs a multi-scale feature pyramid by parameterization and the latter implicitly mines long-distance dependencies between features in a data-driven manner.

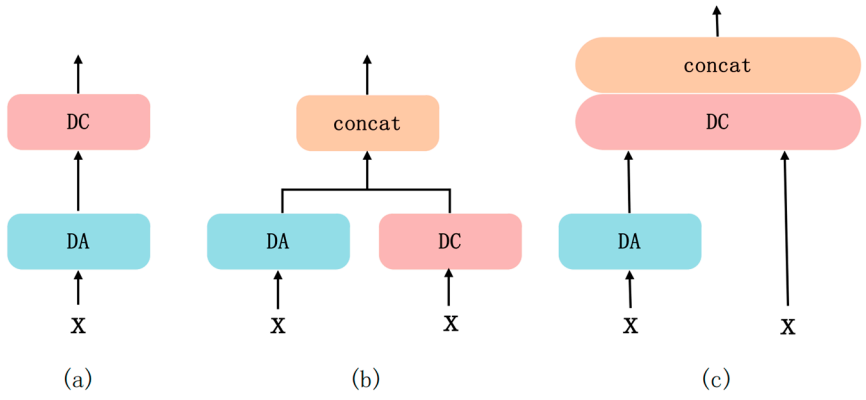

Figure 7. Comparison of multi-scale feature fusion modules.

 doi:10.20944/preprints202602.1167.v1

This section validates the superiority of the proposed hierarchical fusion architecture by systematically comparing the performance differences of different multi-scale feature fusion strategies in semantic segmentation tasks as shown in Table 3.

Table 3. Comparative experimental results of multi-scale feature fusion modules.

Backbone	Description	mIoU	PA
ResNet101	SERIES	77.89%	92.32%
ResNet101	PARALLEL	75.79%	91.44%
ResNet101	INTEGRATION	79.22%	92.47%

The experimental data show that the baseline model (a) with serial structure achieves 92.32% in pixel accuracy (PA), but its mean intersection and merger ratio (mIoU) is only 77.89%, reflecting the significant deficiencies of the model in category boundary delineation and complex scene adaptation; while the model (b) with simple parallel splicing suffers from insufficient feature interactions resulting in a comprehensive degradation of performance (mIoU 75.79%, PA 91.44%). In contrast, model (c) achieves a breakthrough improvement of 79.22% in mIoU through a three-level feature fusion mechanism, which is statistically significant ($p<0.01$) in terms of performance gain of 1.33% and 3.43% over models (a) and (b), respectively. The significant advantage of mIoU of model (c) confirms that it is more suitable for the core requirement of semantic segmentation - the metric integrates the accuracy of category prediction with the completeness of region coverage, whereas model (c) effectively mitigates the intra-class discrepancy and inter-class confusion through the multi-scale context capture by dilated convolution and the dual-attention feature recalibration Problems.

4.5.2. Dominant Model Comparison Experiment

In the segmentation task of the batik dataset, the quantitative comparison in Table 4 shows that the mainstream model and the proposed method present significant performance stratification in terms of parameter size (Params), mean intersection, and merger ratio (mIoU) and pixel accuracy (PA).

Table 4. Validation results of the dominant model on the batik dataset.

Methods	Params	mIoU	PA
FCN8	134.28M	64.69%	86.72%
SegNet	72.55M	67.70%	87.99%
DucHdc	69.17M	70.58%	89.64%
UperNet	126.08M	71.46%	89.89%
DeepLabV3+	59.34M	75.92%	91.52%
Proposed	76.65M	**79.22%**$^{+3.3\%}$	**92.47%**$^{+0.95\%}$

Among the traditional models, DeepLabV3+ achieves the optimal baseline performance (mIoU=75.92%, PA=91.52%) by the 59.34M parameter count and the context fusion capability of space pyramid pooling (ASPP), with an improvement of 8.22% compared to SegNet (72.55M, mIoU=67.70%) with similar parameter counts, verifying the effectiveness of multiscale feature extraction. Scale feature extraction; while FCN8 (134.28M), which has the largest number of parameters, has only 64.69% more due to the lack of dynamic feature weighting mechanism, exposing the imbalance between model capacity and efficiency. The proposed method achieves mIoU=79.22% with PA=92.47% with 76.65M parameters (17.31% increase over DeepLabV3+), which is 3.3% and 0.95% improvement over the optimal baseline.

In the analysis of the validation curves of the batik dataset (shown in Figure 8), the models show significant divergence in data trends, convergence speed, stability, and final performance. Traditional models (e.g., FCN8, SegNet) have a slow mIoU growth rate at the initial stage (epoch≤500) and fall

 doi:10.20944/preprints202602.1167.v1

into performance saturation after epoch≥2000 due to insufficient feature fusion capability, accompanied by periodic oscillations, while DeepLabV3+, although reaching near the peak at epoch=1500 by the dilated convolution, subsequently shows a The DeepLabV3+ reaches the peak near epoch=1500 by the cavity convolution but subsequently shows a slight degradation due to the fixed receptive field limitation. In contrast, the proposed method (Proposed) is enhanced by dual attention features, and the mIoU exceeds 75% at epoch=800 and converges stably to 79.22% after epoch=2000. Stability analysis shows that the traditional model fluctuates at the late stage of training due to the insufficient adaptation ability of few-shot samples. The performance comparison shows that the proposed method improves the mIoU by 3.3% over the optimal baseline (DeepLabV3+) under the condition that the number of parameters (76.65M) only increases by 17.31%, which verifies its superiority in terms of accuracy and lightweight trade-off.

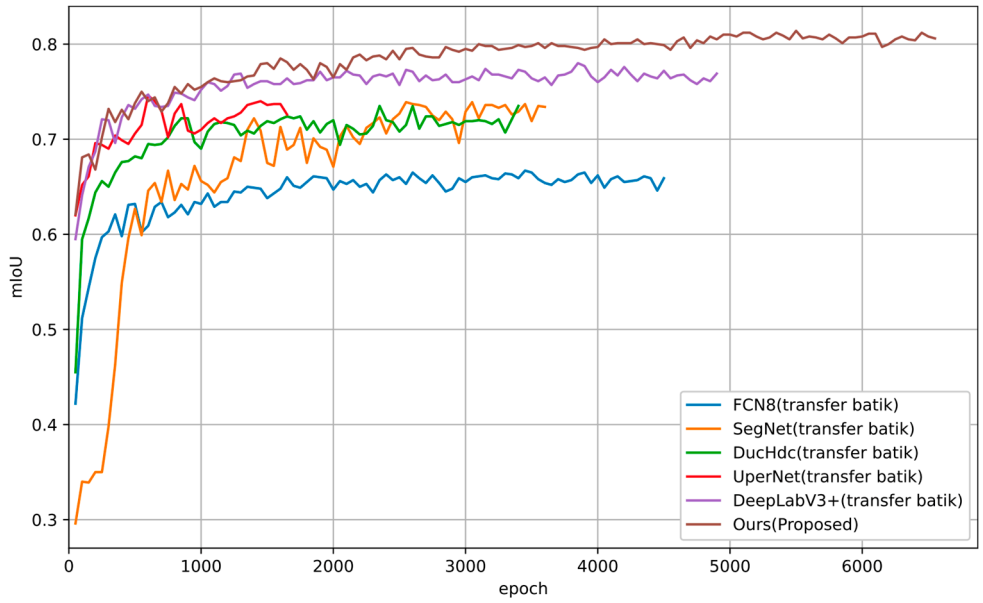

Figure 8. Comparison of mIoU curves of mainstream models on batik dataset.

To evaluate the effectiveness of the proposed method in semantic segmentation and texture recognition of batik patterns, this section compares the visual detection results of UperNet, DeepLabV3+, and the proposed model on the batik dataset, as shown in Figure 9: (a) is the original RGB image, (b) is the manually labeled real labels, (c) and (d) represent the segmentation prediction results of UperNet and DeepLabV3+ segmentation prediction results, and (e) shows the output of the proposed model in this study. The blue boxes label the classification differences in different categories or texture regions, while the red boxes compare the accuracy of the engraving of edges and detailed textures.

 doi:10.20944/preprints202602.1167.v1

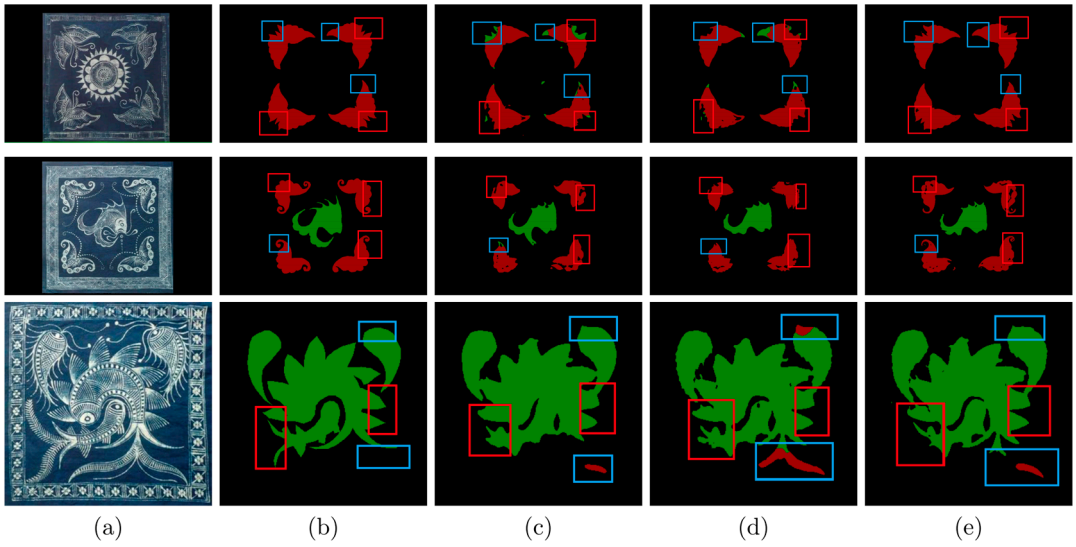

Figure 9. Visualization of the batik dataset (a) RGB Images, (b) Ground truth, (c) UperNet, (d) DeepLabV3+, (e) BMFNet(Proposed).

From the results, it can be seen that UperNet may suffer from texture erosion in localized areas, while DeepLabV3+ has a limited ability to differentiate fine boundaries and misclassified blocks. The proposed model achieves an improvement in the accuracy of overall texture boundary, detail parts, and complex contours, which not only improves the integrity of target category recognition but also effectively suppresses the background noise interference and makes the segmented region more compatible with the real label. Its superior performance in multiscale feature fusion and high-precision texture segmentation is demonstrated.

5. Conclusions

This paper proposed a few-shot sample semantic segmentation algorithm for batik based on attention weighting and hierarchical decoding, and its innovative architecture realizes accurate parsing of complex textures through a multi-level feature interaction mechanism. The network uses ResNet101 as the backbone network to construct the migration learning module, extracts the base features using the pre-training parameters, and combines the dual-attention feature enhancement module to achieve the spatial and channel cocalibration - the spatial attention strengthens the texture contour response through the feature map positional correlation modeling, while the channel attention is based on the global pooling Dynamic recalibration of feature channel weights, and their cascade operation enables the model to effectively focus on key texture regions under few-shot samples. The architecture achieves a balance between texture detail preservation and model generalization capability under very few-shot sample constraints through the synergistic optimization of migration learning initialization, attention-guided feature selection, lightweight multi-scale fusion, and hierarchical decoding, which provides a reliable technical solution for the digital preservation of intangible cultural heritage. The algorithm proposed in this paper demonstrates robustness to noise and localized fuzzy regions in the actual batik pattern segmentation task and maintains good feature discrimination performance under few-shot sample conditions. For batik patterns of different sizes, shapes, and texture styles, the experimental results further demonstrate the model's adaptability to complex scenes.

Future work includes: (1) Aiming at the fine features of batik such as "fine wax cracks and gradient halos", a super-resolution fusion mechanism will be introduced to first improve the detail resolution of images before segmentation, thereby reducing the missed segmentation and missegmentation of small-scale textures. (2) Combining the textual information of intangible cultural

 doi:10.20944/preprints202602.1167.v1

heritage, a multi-modal semantic segmentation method will be designed to enhance the segmentation quality of batik images.

Author Contributions: Conceptualization, Y.M. and W.L.; methodology Y.M. and W.L.; software, H.Q; validation, Y.M., H.Q. and W.L.; data curation, Y.M.; writing—original draft preparation, Y.M.; writing—review and editing, Y.M. and W.L.; visualization, H.Q.; supervision, W.L.; project administration, W.L.; funding acquisition, W.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Key Research Project of Anhui Provincial Department of Education, China (2022AH053089). Outstanding Scientific Research and Innovation Team of Anhui Police College (2023GADT06).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The source code presented in the study are openly available in GitHub https://github.com/TeacherWLee/BaticSemanticSegmentation (accessed on 7 January 2026).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Tian, J.; Xin, P.; Bai, X.; Xiao, Z.; Li, N. An Efficient Semantic Segmentation Framework with Attention-Driven Context Enhancement and Dynamic Fusion for Autonomous Driving. *Applied Sciences* **2025**, *15*, 8373, doi:10.3390/app15158373.
2. Wang, Y.; Zhang, J.; Chen, Y.; Yuan, H.; Wu, C. An Automated Learning Method of Semantic Segmentation for Train Autonomous Driving Environment Understanding. *IEEE Transactions on Industrial Informatics* **2024**, *20*, 6913–6922.
3. Cheng, J.; Deng, C.; Su, Y.; An, Z.; Wang, Q. Methods and Datasets on Semantic Segmentation for Unmanned Aerial Vehicle Remote Sensing Images: A Review. *ISPRS Journal of Photogrammetry and Remote Sensing* **2024**, *211*, 1–34, doi:10.1016/j.isprsjprs.2024.03.012.
4. Yu, A.; Quan, Y.; Yu, R.; Guo, W.; Wang, X.; Hong, D.; Zhang, H.; Chen, J.; Hu, Q.; He, P. Deep Learning Methods for Semantic Segmentation in Remote Sensing with Small Data: A Survey. *Remote Sensing* **2023**, *15*, 4987, doi:10.3390/rs15204987.
5. Azad, R.; Aghdam, E.K.; Rauland, A.; Jia, Y.; Avval, A.H.; Bozorgpour, A.; Karimijafarbigloo, S.; Cohen, J.P.; Adeli, E.; Merhof, D. Medical Image Segmentation Review: The Success of U-Net. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2024**, *46*, 10076–10095, doi:10.1109/TPAMI.2024.3435571.
6. Liu, H.; Gou, S.; Zhou, Y.; Jiao, C.; Liu, W.; Shi, M.; Luo, Z. Object Knowledge-Aware Multiple Instance Learning for Small Tumor Segmentation. *Biomedical Signal Processing and Control* **2026**, *115*, 109400, doi:10.1016/j.bspc.2025.109400.
7. Lu, Y.; Li, W.; Cui, Z.; Zhang, Y. Beyond Low-Dimensional Features: Enhancing Semi-Supervised Medical Image Semantic Segmentation with Advanced Consistency Learning Techniques. *Expert Systems with Applications* **2025**, *261*, 125456, doi:10.1016/j.eswa.2024.125456.
8. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015; Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F., Eds.; Springer International Publishing: Cham, 2015; pp. 234–241.
9. Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Deeplab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected Crfs. *IEEE transactions on pattern analysis and machine intelligence* **2017**, *40*, 834–848.
10. Chen, Z.; Ren, X.; Zhang, Z. Cultural Heritage as Rural Economic Development: Batik Production amongst China's Miao Population. *Journal of Rural Studies* **2021**, *81*, 182–193.

 doi:10.20944/preprints202602.1167.v1

11. Gu, Z.; Zhang, Y.; Xian, S.; Pan, S.; Wang, L. Conservation and Inheritance of Cultural Heritage in Regeneration of Urban Villages: A Tale of Two Villages in Guangzhou, China. *Journal of Rural Studies* **2026**, *122*, 103989, doi:10.1016/j.jrurstud.2025.103989.
12. Yuan, X.; Li, H.; Ota, K.; Dong, M. Building Energy Efficient Semantic Segmentation in Intelligent Edge Computing. *IEEE Transactions on Green Communications and Networking* **2023**, *8*, 572–582.
13. Guo, F.; Zhou, D. Few-Shot Semantic Segmentation Network for Distinguishing Positive and Negative Examples. *Applied Sciences* **2025**, *15*, 3627, doi:10.3390/app15073627.
14. Lee, S.; Kim, S. Semi-Supervised Pointwise VIV Detection via Few-Shot and Sequential Transfer Learning. *Engineering Structures* **2026**, *351*, 122058, doi:10.1016/j.engstruct.2025.122058.
15. He, W.; Zhang, Y.; Zhuo, W.; Shen, L.; Yang, J.; Deng, S.; Sun, L. Apseg: Auto-Prompt Network for Cross-Domain Few-Shot Semantic Segmentation. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2024; pp. 23762–23772.
16. Liu, H.; Peng, P.; Chen, T.; Wang, Q.; Yao, Y.; Hua, X.-S. Fecanet: Boosting Few-Shot Semantic Segmentation with Feature-Enhanced Context-Aware Network. *IEEE Transactions on Multimedia* **2023**, *25*, 8580–8592.
17. Ding, H.; Zhang, H.; Jiang, X. Self-Regularized Prototypical Network for Few-Shot Semantic Segmentation. *Pattern Recognition* **2023**, *133*, 109018.
18. Cao, L.; Guo, Y.; Yuan, Y.; Jin, Q. Prototype as Query for Few Shot Semantic Segmentation. *Complex Intell. Syst.* **2024**, *10*, 7265–7278, doi:10.1007/s40747-024-01539-4.
19. Lu, Z.; He, S.; Li, D.; Song, Y.-Z.; Xiang, T. Prediction Calibration for Generalized Few-Shot Semantic Segmentation. *IEEE transactions on image processing* **2023**, *32*, 3311–3323.
20. Gao, J.; Liao, W.; Nuyttens, D.; Lootens, P.; Xue, W.; Alexandersson, E.; Pieters, J. Cross-Domain Transfer Learning for Weed Segmentation and Mapping in Precision Farming Using Ground and UAV Images. *Expert Systems with applications* **2024**, *246*, 122980.
21. Cuttano, C.; Tavera, A.; Cermelli, F.; Averta, G.; Caputo, B. Cross-Domain Transfer Learning with CoRTe: Consistent and Reliable Transfer from Black-Box to Lightweight Segmentation Model. In Proceedings of the Proceedings of the IEEE/CVF international conference on computer vision; 2023; pp. 1412–1422.
22. Tan, Y.; Zhang, E.; Li, Y.; Huang, S.-L.; Zhang, X.-P. Transferability-Guided Cross-Domain Cross-Task Transfer Learning. *IEEE Transactions on Neural Networks and Learning Systems* **2024**.
23. Li, D.; Li, J.; Zeng, X.; Stankovic, V.; Stankovic, L.; Xiao, C.; Shi, Q. Transfer Learning for Multi-Objective Non-Intrusive Load Monitoring in Smart Building. *Applied Energy* **2023**, *329*, 120223.
24. Gao, G.; Dai, Y.; Zhang, X.; Duan, D.; Guo, F. ADCG: A Cross-Modality Domain Transfer Learning Method for Synthetic Aperture Radar in Ship Automatic Target Recognition. *IEEE transactions on geoscience and remote sensing* **2023**, *61*, 1–14.
25. Fang, J.; Wang, Z.; Liu, W.; Chen, L.; Liu, X. A New Particle-Swarm-Optimization-Assisted Deep Transfer Learning Framework with Applications to Outlier Detection in Additive Manufacturing. *Engineering Applications of Artificial Intelligence* **2024**, *131*, 107700.
26. Wang, Y.; Hu, S.; Liu, J.; Wang, A.; Zhou, G.; Yang, C. Bridging the Gap between Computer Vision and Bioelectrical Signal Analysis. *Information Fusion* **2026**, *129*, 104047, doi:10.1016/j.inffus.2025.104047.