mathematics

MDPI

Article

A Multi-Scale Hybrid Attention Network for Sentence Segmentation Line Detection in Dongba Scripture

Junyao Xing [1], Xiaojun Bi [2,*,†] and Yu Weng [2,†]

[1] College of Information and Communication Engineering, Harbin Engineering University, Harbin 150009, China; 595587572@hrbeu.edu.cn
[2] School of Information and Engineerin, Minzu University of China, Beijing 100081, China; wengyu@muc.edu.cn
* Correspondence: bixiaojun@hrbeu.edu.cn
† Current address: Key Laboratory of Ethnic Language Intelligent Analysis and Security Governance of MOE, Beijing 100081, China.

Abstract: Dongba scripture sentence segmentation is an important and basic work in the digitization and machine translation of Dongba scripture. Dongba scripture sentence segmentation line detection (DS-SSLD) as a core technology of Dongba scripture sentence segmentation is a challenging task due to its own distinctiveness, such as high inherent noise interference and nonstandard sentence segmentation lines. Recently, projection-based methods have been adopted. However, these methods are difficult when dealing with the following two problems. The first is the noisy problem, where a large number of noise in the Dongba scripture image interference detection results. The second is the Dongba scripture inherent characteristics, where many vertical lines in Dongba hieroglyphs are easily confused with the vertical sentence segmentation lines. Therefore, this paper aims to propose a module based on the convolutional neural network (CNN) to improve the accuracy of DS-SSLD. To achieve this, we first construct a tagged dataset for training and testing DS-SSLD, including 2504 real images collected from Dongba scripture books and sentence segmentation targets. Then, we propose a multi-scale hybrid attention network (Multi-HAN) based on YOLOv5s, where a multiple hybrid attention unit (MHAU) is used to enhance the distinction between important features and redundant features, and the multi-scale cross-stage partial unit (Multi-CSPU) is used to realize multi-scale and richer feature representation. The experiment is carried out on the Dongba scripture sentence segmentation dataset we built. The experimental results show that the proposed method exhibits excellent detection performance and outperforms several state-of-the-art methods.

Keywords: hybrid attention mechanism; multi-scale depthwise convolution; multi-scale features; sentence segmentation line detection; Dongba scripture sentence segmentation line detection dataset

MSC: 68T45

Citation: Xing, J.; Bi, X.; Weng, Y. A Multi-Scale Hybrid Attention Network for Sentence Segmentation Line Detection in Dongba Scripture. *Mathematics* **2023**, *11*, 3392. https://doi.org/10.3390/math11153392

Academic Editor: Shuai Liu

Received: 11 July 2023
Revised: 29 July 2023
Accepted: 31 July 2023
Published: 3 August 2023

1. Introduction

The Dongba hieroglyph is one of the oldest hieroglyphs in China [1]. It was invented by the ancestors of the Naxi nationality in China. It has been handed down in the form of Dongba scripture and has been collected in many countries around the world. These scriptures record the development of the Naxi minority civilization, and have special academic value in studying the origin of religion, characters, and culture. In 2003, Dongba scripture was listed as "World Memory Heritage" by UNESCO, which shows the important position and great influence of the Dongba language in the world language research. Figure 1 shows a piece of a Dongba scripture image. However, due to the lack of accumulation of relevant expertise in Dongba scripture, it is difficult for nonprofessionals to accurately translate Dongba scripture. For this reason, the research on the digitalization and machine translation of Dongba scripture is of great significance for inheriting and studying Dongba

culture and realizing the salvage protection of Dongba classics. Among them, the automatic sentence segmentation of Dongba scripture is an important basic work of Dongba scripture digitization and machine translation research, as its accuracy may directly affect the results of machine translation. Therefore, as one of its core tasks, Dongba scripture sentence segmentation line detection (DS-SSLD) should improve detection accuracy as much as possible.

Figure 1. A Dongba scripture image containing several sentences, where the horizontal line represents the row segmentation line and the vertical line represents the single sentence segmentation line.

Generally, the main text of the Dongba scripture is composed of three lines of characters separated by a horizontal segmentation line, in which each line of characters is separated into several single sentences by a single vertical segmentation line or double vertical segmentation [2]. In addition to horizontal and vertical lines, some scripture have outer borders or decorative patterns. According to such writing characteristics, finding the position of horizontal and vertical lines in the image is an important basis for sentence segmentation. In recent years, the task of sentence segmentation of Dongba scripture has attracted attention due to the urgency of the digital protection of Dongba scripture. Several methods based on projection are adopted [3–5] according to the feature that the sentence segmentation line runs through the Dongba scripture image. However, these segmentation lines are bent due to manual writing, and there are breakpoints due to the long retention time, which makes the projection value of some segmentation lines not prominent enough or even lower than the projection value of the text area after projection as shown in Figure 2. These issues may cause errors in the judgment of the segmentation lines. Therefore, it is necessary to further improve the detection accuracy of segmentation lines by means of deep learning.

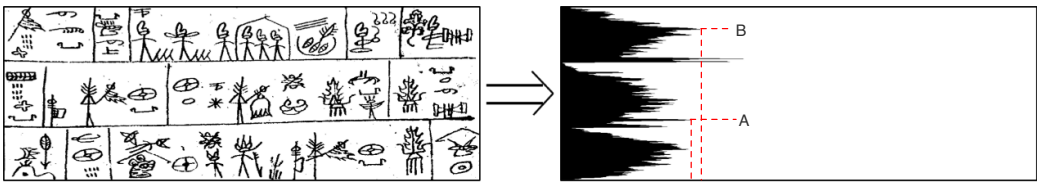
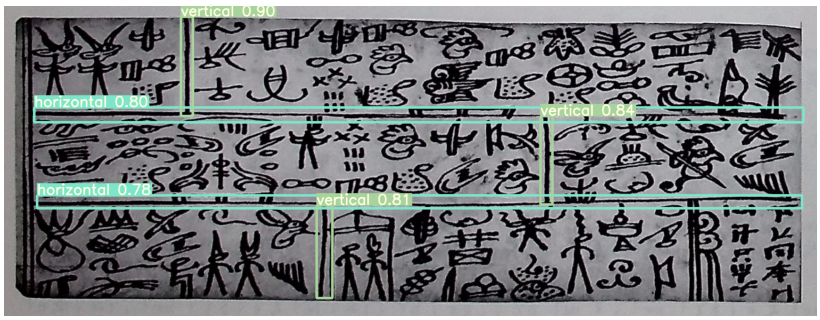


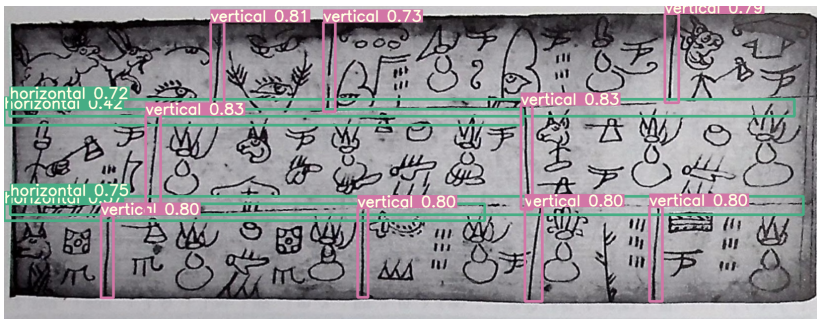
Figure 2. The left is a binary image of Dongba scripture, and the right is the result of horizontal projection on this binary image. It can be seen that the projection value of the horizontal segmentation line (point A) is less than the projection value of the text area (point B).

With the rapid development of deep learning algorithms [6], the method based on deep learning has penetrated into most related tasks of computer vision [7–9], including image restoration, pattern recognition and object detection. Therefore, the method based on deep learning for DS-SSLD is promising to be worthy of further exploration and research. However, at present, there is no public dataset available for DS-SSLD, which makes it impossible to conduct fair evaluation and comparison and is not conducive to more scholars' research. As a result, we begin to build a new tagged dataset for Dongba scripture sentence segmentation line detection.

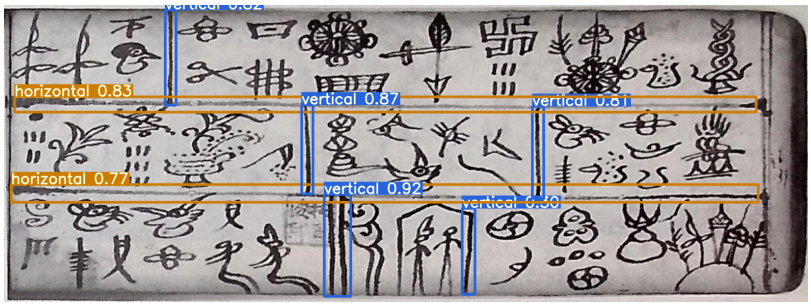
Due to the large image noise, different writing styles, different scales between horizontal lines and vertical lines, and the confusion between the text area and the segmentation lines, the object detection based on deep learning methods of sentence segmentation lines in Dongba scripture is more challenging. It mainly faces three problems: (1) The horizontal lines used for branches and the vertical lines used for sentence segmentation are different in scale, and it is difficult to include the feature information of the two segmentation lines with the feature information of a single scale. (2) The images of Dongba scripture have plenty of noise, and the long horizontal lines of the segmentation lines often have a lot of breakpoints, which lead to duplicate detection or missed detection. (3) Dongba hieroglyphs contain long vertical lines, which are easily confused with the vertical lines used to break sentences as shown in Figure 3.



(a) missed detecction



(b) duplicate detecction



(c) error detecction

Figure 3. Examples of some failed detection from YOLOv5s network. (**a**) The vertical segmentation line at the lower right corner is not detected; (**b**) horizontal segmentation line is detected in duplicate due to breakpoint; (**c**) the vertical lines in the Dongba characters below are incorrectly detected.

In order to enhance the network's attention to important regions, we propose a new multiple hybrid attention unit (MHAU), which can handle the spatial information of the multi-scale input feature map and can establish dependence among the multi-scale channel attention with effect. It can improve the ability of the multi-scale hybrid attention network (Multi-HAN) to utilize valuable features of different depths and scales from the channel and spatial dimensions of feature maps, and enhance the ability of perception to

the target area. At the same time, the cross-stage connection mode is used to strengthen and constrain the deep global information with shallow details so as to enhance the network's attention to important regions. Through this method, Multi-HAN can obtain more important aggregation features and model Dongba scripture images with different feature scales.

It is experimentally validated that rich global and semantic information is frequently expressed through deeper level feature maps, while shallower features contain more spatial and detailed information. In order to consider the two situations at the same time, shallow features are gradually integrated with deep features in the process of information transmission. The feature pyramid network (FPN) + path aggregation network (PAN) structure in YOLOv5s is possible to aggregate details and spatial information with deep global and semantic information. At the same time, in order to enhance the extraction of multi-scale features in the process of fusion, we propose a multi-scale cross-stage partial unit (Multi-CSPU), which replaces the convolution layer in the original cross-stage partial network (CSPNet) structure with a multi-scale depthwise convolution (MDconv) module [10]. It can provide the gradient combination with richer scales and improve the learning ability of the network. Thus, the MDconv [10] can reduce the computational complexity while providing multi-scale information.

In conclusion, we make the following contributions as follows:

(1) A new Dongba scripture sentence segmentation line detection dataset named DBS2022 is constructed, which consists of 2504 standard images and their annotations. It supports Dongba scripture sentence segmentation line detection (DS-SSLDD). As far as we know, this is the only dataset from the real Dongba scriptures, including images of Dongba scripture such as sacrifices and ceremonies.
(2) A novel structure named multi-scale hybrid attention network (Multi-HAN) is proposed, where both novel multiple hybrid attention unit (MHAU) and multi-scale cross-stage partial unit (Multi-CSPU) are designed. MHAU is designed for establishing dependence between multi-scale channel attention. Multi-CSPU is proposed for multi-scale and richer feature extraction , and implements a gradient combination with richer scales, thus improving the learning ability of the network.
(3) As an application, a new object detection framework named Multi-HAN for DS-SSLD is proposed. We conducted experiments on our constructed DBS2022, which validated the effectiveness of our methods, and the quantity of the experimental results demonstrates that our proposed Multi-HAN is superior to the state-of-the-art methods.

2. Related Work

2.1. Traditional Approaches

The Dongba scripture sentence segmentation task has been of interest since 2020, and it only stays in the row segmentation stage. The traditional methods for Dongba scripture sentence segmentation include two independent tasks, preprocessing and segmentation line detection. Image denoising is an important problem in image preprocessing, which can reduce intra class changes and expand inter class changes. Researchers have invented numerous utility image denoising algorithms, such as linear denoising [11], median denoising [12], adaptive denoising [13] and wavelet denoising [14]. Other preprocessing methods, such as binary processing, are also widely used. In the row segmentation stage, they usually use a combination of projection [15] and K-means clustering [16]. The segmentation point is determined by finding the maximum projection value and setting the threshold value [3–5].

However, the real images of Dongba scripture images are written by hand or by knife carving and have a long history. They are vulnerable to the influence of acquisition equipment and acquisition methods, and often appear blurry with large-area shadows. The pixel values of these noises are close to the text area and cannot be completely removed by filtering or binarization. In addition, manual writing may make the segmentation line between sentences curved or inclined at a large angle, as a result of which many

segmentation lines presumably have no sharp peak after projection, and may also even be lower than the projection value of the text area. The threshold value is difficult to determine, leading to errors in the judgment of segmentation lines. Therefore, it is necessary to further improve the detection accuracy of segmentation lines by means of deep learning.

Although, there are few research works related to the segmentation of Dongba scripture, there are a lot of research studies on other aspects related to it. For example, deep learning is applied to table segmentation [17,18], handwritten text line segmentation [19], etc. The characteristics of these tasks are similar to DS-SSLD; it can prove that deep learning is indeed an effective method in DS-SSLD.

2.2. Deep Learning Network Approaches

Object detection is one of the three major tasks in the field of computer vision, and it is also the most basic and challenging hot topic. The methods of object detection algorithm can be divided into two categories.

Firstly, the object detection algorithm based on the convolutional neural network (CNN) can be divided into two technical development routes, according to whether there is an anchor frame or not: anchor-based and anchor-free methods. The anchor-based method first sets the size of the object frame through prior knowledge, and then returns and classifies the anchors in the form of a sliding window based on the feature map. The anchor-based method includes one-stage and two-stage detection algorithms. The core idea of the two-stage approaches is to first generate a series of candidate boxes from the first phase network as samples, which are named RoIs (regions of interest), and then classify the RoIs through the convolutional neural network. For example, the representative two-stage object detection model Faster-RCNN [20] is the first end-to-end deep learning detection algorithm that is closest to real-time performance. Then, based on Faster-RCNN, the feature pyramid networks (FPNs) [21] technology is further proposed. It is adopted to build high-level semantic information at all high and low levels with different scales to improve the accuracy of the detection network (especially for some datasets with large-scale changes of objects to be detected). Cascade-RCNN [22] further stacks several cascaded modules behind the detectors and uses different IOU threshold training. For different problems in object detection, other meaningful work is proposed, such as [23,24], focusing on architecture design, and [25,26] focusing on multi-scale unification. Although the two-stage model has high accuracy, its calculation speed is slow. The one-stage object detection algorithm does not need a region recommendation stage. It directly generates the category probability and location coordinate value of the object. After a stage, the final test results can be obtained directly. Therefore, it has a faster detection speed. For example, SDD [27] uses the anchor mechanism in Faster-RCNN and uses multi-scale feature maps for prediction, which can consider objects of different sizes in the image. YOLOv3 [28] uses DarkNet53 as the feature extraction network, and uses three branches (three feature maps with different scales/receptive fields) to detect objects of different sizes. YOLOv5 [29], as the best and most used object detection model at present, is a new object detection model developed on the basis of YOLOv3 [28] and YOLOv4 [30] models. CornerNet [31] is the pioneering work of the anchor-free technology route. This work proposes a new object detection framework, which transforms the detection of target bounding box by the network into the detection of a pair of key points (i.e., upper left corner and lower right corner). By detecting objects into pairs of key points, it is unnecessary to design the anchor box as an a priori box. Different from the CornerNet detection algorithm, CenterNet [32] has a very simple structure. It discards the idea of two key points by directly detecting the center point of the target and realizes real anchor-free detection.

Secondly, with the great success of the transformer [33] framework in the field of natural language processing (NLP), researchers have tried to migrate it to the field of computer vision. Ref. [33] firstly presents a complete end-to-end DETR object detection framework by combining CNN with transformer. Subsequently, similar algorithms emerge in large numbers. The deformable DETR [34] model is proposed that is based on a variable

convolution neural network. The ACT algorithm [35] is proposed to reduce the computational complexity of the self-attention module. The ViT-FRCNN [36] model uses the vision transformer (ViT) [37] model as a feature extraction network for object detection.

2.3. Attention Mechanism in CNN

The attention mechanism is used to strengthen the allocation of feature expressions with the largest amount of information, while suppressing less useful feature expressions so that the model can adaptively focus on important areas in the context [38–41]. Multiple attention mechanism models have been carried out to optimize the modeling of the deep learning network, which mainly include two aspects: channel attention and spatial attention.

For squeeze and excitation (SE) [42], note that the channel correlation can be captured by selectively modulating the channel size.Efficient channel attention (ECA) [43] further uses 1-D convolution to efficiently achieve local cross-channel interactions, improve channel dependencies, and avoid the adverse effects of dimension reduction by compression. Efficient pyramid split attention (ESPA) [44] can effectively obtain and utilize the spatial information of feature maps at different scales, and effectively establish long-term dependence of multi-scale channel attention. Convolutional block attention module (CBAM) [45] innovatively proposes the attention mechanism of the fusion of channel attention and spatial attention serialization. Triplet [46] uses a three-branched structure to capture cross-dimensional interactions to calculate attention weights.

Both approaches focus on designing more complex attention modules, which will inevitably result in higher computational costs and obtain only a single scale of attention information without establishing a remote channel dependency. Therefore, in order to further improve the efficiency of the model and reduce the complexity of the model, a new attention module multi-hybrid attention unit (MHAU) is proposed, which can effectively establish multi-scale attention information and learn attention weights with low model complexity in a low-cost way to achieve the accurate detection of sentence segmentation lines in Dongba scripture.

3. Method

In this section, we first give the proposed overview of multi-scale hybrid attention network (Multi-HAN), then propose the multiple hybrid attention unit (MHAU), and at last introduce the important components of multi-scale cross-stage partial unit (Multi-CSPU).

3.1. Overview Structure of Multi-HAN

Based on the proposed multiple hybrid attention unit (MHAU) and multi-scale cross-stage partial unit (Multi-CSPU), a multi-scale hybrid attention network (Multi-HAN) based on the YOLOv5s [29] structure is designed. The structure is shown in Figure 4. It contains three parts, a backbone which is used for feature extraction, a neck which is used for feature enhancement, and a head which is used to predict classes and bounding boxes of objects.

Based on the original structure of YOLOv5s [29], we first propose MHAU, which can fully extract the spatial information of multi-scale feature maps of different channels to realize the interaction of multi-scale important features in cross-dimensional channels. At the same time, it adopts a cross-stage transmission mode. That is, the two attention maps generated by shallow features are multiplied by deep features to establish a dependency among the multi-scale channels. It can gather shallow details and deep global semantic information to strengthen and constrain the deep global information with shallow details. Through such operations, the network can effectively focus on important features and enhance effective features to improve the detection capability of target features. In the neck stage, a new multi-scale cross-stage partial unit (Multi-CSPU) is proposed, that is, the multi-scale depthwise convolution (MDconv) [10] is embedded into the CSP2_X structure (which is in the original structure of YOLOv5s [29]) to further fully learn the multi-scale features

without increasing the complexity of the model. The following Sections 3.2 and 3.3 will detail the MHAU and Multi-CSPU.

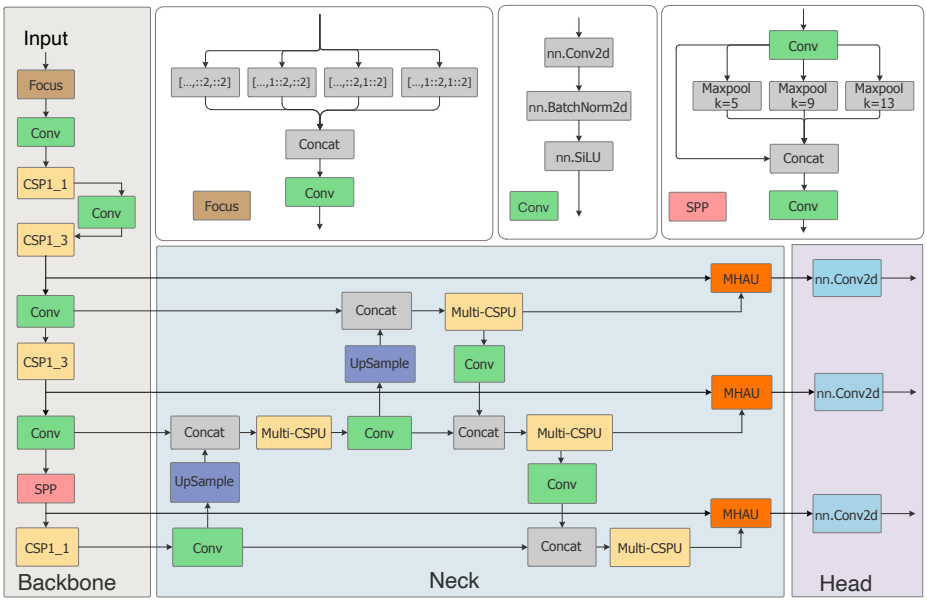


Figure 4. Overview of the proposed multi-scale hybrid attention network (Multi-HAN). "MHAU" and "Multi-CSPU" respectively represent the multiple hybrid attention unit in Section 3.2 and multi-scale cross-stage partial unit in Section 3.3. Wherein, feature maps of different scales in the backbone stage and feature maps of the same scales in the neck stage are implemented cross-stage transmission through MHAU.

3.2. Multiple Hybrid Attention Unit

Considering that the vertical sentence segmentation lines in Dongba scripture are similar to the strokes of individual Dongba characters, it is difficult to distinguish details with a single scale, so it is necessary to consider extracting multi-scale features that are effective for multi-granularity networks at different stages. Based on the above ideas and inspired by the creative expression of the attention mechanism efficient pyramid split attention (ESPA) [44], we improved it, added a parallel spatial attention mechanism, and proposed a multiple hybrid attention unit (MHAU). This module is located in the neck part of the network, which consists of the squeeze and concat (SPC) module [44], split multi-scale channel attention mechanism (sMCA) [44] and split maximum spatial attention mechanism (sMSA), which is used to obtain multi-scale attention information of different scale feature maps. The detailed settings of MHAU in Multi-HAN are shown in the Figure 5.

MHAU proposed in this paper is a parallel cross-feature fusion method of the split multi-scale channel attention mechanism (sMCA) [44] and split maximum spatial attention mechanism (sMSA). First, use the squeeze and concat (SPC) module [44] to extract the multi-scale feature information on each channel feature map from shallow layers, and then divide it into two branches. One branch uses sMCA [44] to extract the channel attention of the multi-scale feature map and uses softmax to recalibrate the multi-scale channel attention vector to obtain a new multi-scale channel attention weight with a sum of 1; another branch uses sMSA to extract multi-scale spatial attention of feature maps with different scales. Finally, the multi-scale channel attention weight and multi-scale spatial attention weight are point multiplied with the deep corresponding scale feature map generated from $\mathbf{X}_2$. Through this operation, MHAU can strengthen the feature cross-fusion ability of Multi-HAN, and enhance the ability to distinguish between effective and redundant features. Therefore, MHAU can obtain fusion features with more meaningful information, providing valuable guidance for the output of subsequent detection results. In addition, due to the significant difference between the foreground and background pixel values of the Dongba

scripture images, average pooling may smooth semantic information in the foreground, which is adverse to the extraction and processing of valuable features. Therefore, in the structure of sMCA [44] and sMSA, this paper only uses the maximum pool operation to fuse the feature map information, which can preserve the beneficial semantic information of the foreground to the greatest extent possible. At the same time, serially extracting spatial and channel attention will lose important feature information, so this paper uses parallel processing to extract two kinds of attention to ensure that important semantic information is not ignored.

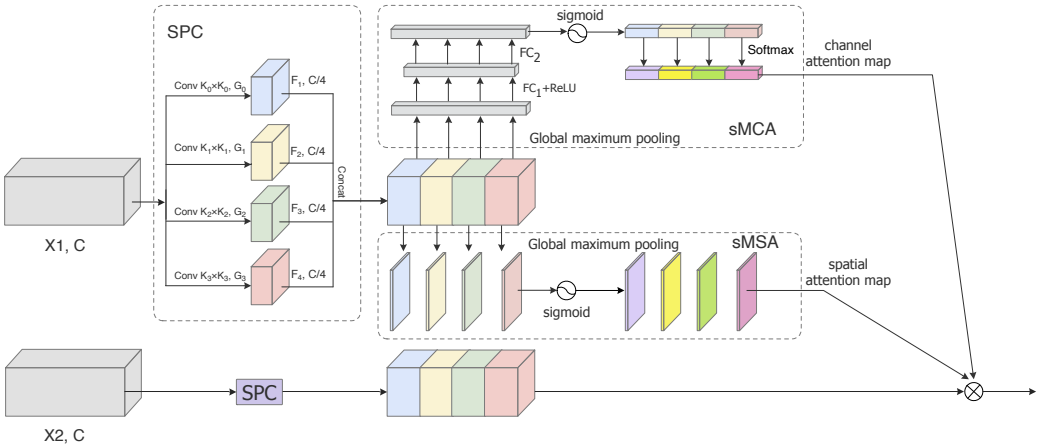


Figure 5. The detailed settings of our proposed multiple hybrid attention unit (MHAU). Respectively, X_1 represents the input from different scale feature maps in the backbone stage, X_2 represents the input from corresponding scale feature maps in the neck stage, and $\otimes$ represents element-wise multiplication.

3.2.1. Squeeze and Concat Module

The squeeze and concat (SPC) module [44] is an important module to extract multi-scale information. Details are shown in the left side of Figure 5. We extract multi-scale feature information from multiple branches in a parallel way. Each branch uses the convolution kernel of different size in the pyramid structure to independently learn multi-scale feature information, and compress the channel dimension of the input tensor from C to C/4. Through splicing, we can effectively extract the feature information of different scales on each channel feature map. However, as the size of the convolution kernel increases, the number of parameters will increase. In order to reduce the calculation pressure, the corresponding group parameter G is introduced at different scales, and its size is transformed with the size of the convolution kernel:

$$G_i = 2^{\frac{k_i - 1}{2}} \qquad i = 0, 1, 2, \ldots, S-1 \tag{1}$$

where $k_i = 2 \times (i+1) + 1$ represents the i-th kernel size, and G_i is the i-th group size. In particular, when $i = 0$, the default of G_i is 1. S is the number of branches. Finally, the multi-scale feature map generation function is given by

$$F_{l-i} = Conv(k_i \times k_i, G_i)(X_l) \qquad i = 0, 1, 2, \ldots, S-1 \qquad l = 1, 2 \tag{2}$$

where $F_{l-i} \in \mathbb{R}^{C/4 \times H \times W}$ denotes the feature map with different scales. We use the SPC [44] module to extract multi-scale features for feature maps of different sizes in the backbone stage and corresponding sizes in the end of the neck stage so as to calculate the weight of attention on each scale subsequently. The experimental results show that this method has better results than the method of weighting on the original input feature map and passing it down.

3.2.2. Attention Extraction Module

In order to simultaneously obtain the channel and spatial attention weights on feature maps of different scales, the channel and spatial attention weights are correspondingly extracted in the following two ways. The detailed settings of the split maximum channel attention mechanism (sMCA) [44] and split maximum spatial attention mechanism (sMSA) are shown in the right side of Figure 5.

One of the branches is sMCA [44]. sMCA [44] first uses the global maximum pooling [47] in the spatial dimension (GMP_s) to fuse the entire spatial information of the corresponding scale feature map in the input so as to enhance the channel weight of important information. The structure of the sMCA [44] is shown in the upper right part of Figure 5. sMCA [44] is mainly composed of (GMP_s) layer, two fully connected (FC) layers, ReLU [48] and Sigmoid [49] functions. The attention weight vector of i-th obtained can be expressed as

$$A_{Ci} = \sigma(FC_2(Relu(FC_1(GMP_S(F_{1-i}))))) \qquad i = 0, 1, 2, \ldots, S-1 \tag{3}$$

where GMP_S represents global maximum pooling in the spatial dimension; F_{C1} represents the fully connected layer, which reduces the channel dimension C to C/4; and F_{C2} restores the channel dimension back to the original. σ denotes the Sigmoid [49] activation function.

All the multi-scale attention vectors are concatenated in a cascading manner. Finally, in order to establish channel attention dependence and realize the information interaction among multi-scale channel attention, we further use Softmax [50] to recalibrate the channel attention weights, which are learned to explicitly model the correlation between the feature channels.

The sMSA module is used in another branch to independently obtain multi-scale spatial attention information. The detailed settings of sMSA are shown in the lower right part of Figure 5. First, the feature mapping of multiple channels at each scale is compressed into a single channel using the global maximum pooling operation in the channel dimension. Secondly, use the Sigmoid [49] function to activate the single channel feature map so that the part with more semantics and detailed information on the feature map can obtain higher weight so as to obtain the spatial attention weight information on the feature map with different scales:

$$A_{Si} = \sigma(GMP_C(F_{1-i})) \qquad i = 0, 1, 2, \ldots, S-1 \tag{4}$$

where GMP_C represents global maximum pooling in the channel dimension.

Finally, we weight the multi-scale channels and spatial attention with the deep corresponding scale feature map, namely,

$$M_i = F_{2-i} \odot A_{Ci} \odot A_{Si} \qquad i = 0, 1, 2, \ldots, S-1 \tag{5}$$

where $\odot$ represents the channel-wise multiplication, and $\mathbf{M}_i$ denotes the feature map with the obtained multi-scale channel-wise attention weight. Finally, the weighted feature maps are dimension connected, and the dimension connection is more efficacious than summing up the feature maps after dimension ascension because it can completely reserve the feature representation without destroying the original feature map information.

Integrating multi-scale spatial information and cross-channel attention into each feature block can preferably model the local and global information jointly, improving the network's attention to the target area.

3.3. Multi-Scale Cross-Stage Partial Unit

Considering the different shapes of hieroglyphs in the images of Dongba scripture, and the existence of characters similar to the vertical line of sentence segmentation lines, it is necessary to improve the feature extraction capability by increasing the network depth, but it will increase the complexity of the model. At the same time, in YOLOv5s [29] model, the cross stage partial (CSP) [51] module, which is shown in Figure 6a, uses more

standard convolutions, which brings the problem of large computation, and using only single-scale convolution kernel is not conducive to the network's extraction of multi-scale features. Therefore, aiming to improve the feature extraction ability of the Multi-HAN, this paper learned from Mixconv's [10] ideas to introduce a multi-scale depthwise convolution (MDconv) module to replace the convolutional layer of the CSP2_X structure in the original YOLOv5s [29] neck network, called multi-scale cross-stage partial unit (Multi-CSPU) as shown in Figure 6b.

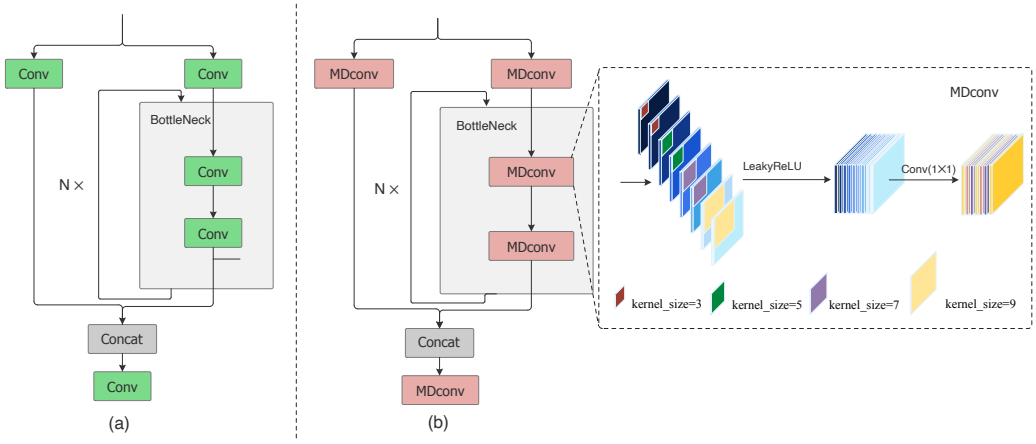


Figure 6. Overview of the CSP module (**a**) and Multi-CSPU (**b**), where a detailed illustration of the MDconv [10] module is shown on the right side.

The Multi-CSPU introduces a multi-scale convolution kernel (3×3, 5×5, 7×7, 9×9) to fully extract the input multi-scale features. The small size convolution kernel pays more attention to the edge structure information, while the large size convolution kernel is more conducive to extract the global information. It is worth noting that we do not use the dilated convolution [52] with different dilation rates to replace the large-scale convolution kernel in the module design because the dilated convolution [52] has a certain grid effect and loses part of the local information of the image. However, the introduction of a large convolution kernel may undoubtedly increase the total parameters of the object detection model, cause the redundancy of model parameters, and increase the risk of overfitting. This phenomenon is not conducive to the improvement of the generalization ability. To this end, multi-scale depthwise convolution (MDconv) [10] is proposed so that each convolution kernel only acts on the characteristic graph of per channel, and through a non-linear activation of LeakyReLU and a 1×1 convolution to complete the interaction and fusion of information of each independent channel. Among them, the negative region of LeakyReLU activation function has a small positive slope, so even for negative input values, it can also carry out back propagation, which can effectively solve the problem of neuron death. Experiments [53] proved that using LeakyRelu in this structure can provide more effective training for the network. The specific settings of MDconv [10] module are shown in the right side of Figure 6.

The above MDconv [10] can greatly improve the feature extraction ability of the model, and ensure that the parameter quantity is consistent with the conventional 3×3. On the premise that the convolution is almost the same, the segmentation line in the Dongba scripture image can be detected very accurately.

4. The Construction of Dongba Scripture Sentence Segmentation Dataset

Since there is no public standard Dongba scripture sentence segmentation line detection dataset at present, it is not conducive to more scholars' research and fair evaluation. Therefore, we established a dataset for Dongba scripture sentence segmentation line detection (DBS2022), which contains 2504 images and their annotations. The detailed data of the DBS2022 are shown in Table 1. All these images are from the real Dongba scripture collected in "An ANNOTATED COLLECTION OF NAXI DONGBA MANUSCRIPTS" [2],

which includes sacrifice funerals, divination, etc. The position of the sentence segmentation line is marked according to the "horizontal" line and the "vertical" line (both single vertical line and double vertical line). We do not consider the decorative pattern at the border because it has no effect on the sentence segmentation. As far as we know, this is the only deep learning dataset used to train the DS-SSLD task, which contains a large number of authentic samples. All images and annotations in DBS2022 are manually collected and checked, so the quality of DBS2022 is reliable.

Table 1. The detailed data of the DBS2022.

Name	DBS2022
Object categories	2 (horizontal, vertical)
Format	TXT
Image	2504
Bounding boxes of horizontal	5250
Bounding boxes of vertical	15,776

Quality Control and Annotations

In order to ensure the authenticity of the data and the generalization ability of the deep learning model, we all cut out the original picture of the book by scanning the complete set of translations and annotations of "An ANNOTATED COLLECTION OF NAXI DONGBA MANUSCRIPTS" [2], without any image-processing steps, such as resize, denoising and blur removal, so as to ensure the authenticity of the data and enable the network to break sentences on any collected real images of Dongba scripture after training.

5. Experiments

In order to evaluate the effectiveness and superiority of Multi-HAN in dealing with the Dongba scripture sentence segmentation line detection (DS-SSLD) task, we conducted many comparative experiments on the Dongba scripture sentence segmentation line detection dataset (DBS2022) we built. The experimental setups, evaluation metrics, experimental results, and analysis are presented in the following subsections.

5.1. Experimental Setup

We use the PyTorch [54] framework to perform the proposed Multi-HAN. In Section 4, we finally obtain 2504 handwritten Dongba scripture images. To ensure the authenticity of the data and the robustness of the network to different noise conditions, we do not make any other image processing operations. We use the 2-image Mosaic [30] during training instead of a single image. The batch size is set as 32, and the epoch number is 300. All models are trained by the optimizer Adam [55] from scratch. The learning rate is set as 0.02 at first, and decreases to 0.001 after 300 epochs.

In various comparative experiments, 80% of the total number of images in the DBS2022 is randomly selected as the training set, and 10% is used as the validation set. After each training period, the training results are verified with the validation set, and the remaining 10% of the images is used as the testing set. After the network training, the detection accuracy of Dongba scripture sentence segmentation line is tested on the testing set.

Since the dataset is built by ourselves and randomly split, we conduct the cross validation on the DBS2022 to verify the stability of the proposed model. In cross validation, we randomly extract one sixth of the dataset as the final evaluation data, and then randomly divide the remaining dataset into five equal parts. One part is used as the test set, while the other four parts are selected as the training set. We take turns training and testing the model five times until each part of the data is used as a test set. Finally, we conduct the final evaluation on the pre-extracted test set.

5.2. Evaluation Metrics

To compare the detection performance of different methods, we introduce mean average precision (mAP) as the measurements. mAP measures the quality of models in all categories; it is one of the most important indicators in object detection. Since the main purpose of the DS-SSLD task is to find the position of the sentence segmentation line accurately to achieve sentence segmentation, according to the particularity of the task, we believe that when the value of IOU reaches 0.5, the segmentation line position can be determined through the current detection box. Therefore, this paper chooses mAP@0.5 as an important measurement to judge the quality of the model. It is formulated as follows:

$$mAP = \frac{\sum_{i=1}^{k} AP_i}{k} \tag{6}$$

where AP_i is the area enclosed below the PR curve of each category, and the PR curve is the curve obtained by taking the precision as the ordinate and the recall as the abscissa when IOU is set to 0.5. k represents the number of categories.

5.3. Training Result

The training process of our model can be seen in Figure 7, which shows different training indicators of the training set and verification set.

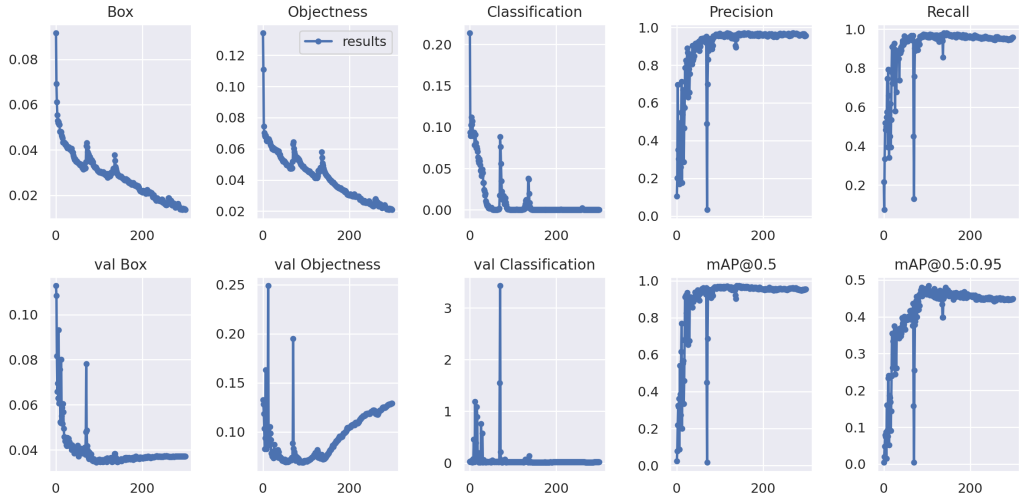


Figure 7. Plots of box loss, objectness loss, classification loss, precision, recall and mean average precision (mAP@0.5 and mAP@0.5:0.95) over the training epochs for the training and validation set.

Figure 7 shows three different types of losses: box loss, object loss, and classification loss. All losses converge well. Box loss represents the ability of the algorithm to locate the center of the object and predict the extent to which the bounding box covers the object. Objectness essentially represents the probability that an object exists in the extracted region of interest. Classification loss measures the ability to predict the correct classification of a given object. Our model rapidly improved the precision, recall and mean average precision indicators, and stabilized after 100 epochs. The box loss, objectness loss and classification loss of the validation set also show a rapid downward trend.

After training our model, we predicted the new pictures in the test set. Figure 8 shows some samples of the test results. The test results show that our Multi-HAN can determine the position of the sentence segmentation line to a high degree, and can accurately detect, even in the case of high interference noise and low foreground pixels. It can successfully deal with the breakpoint of the segmentation line, and there is no problem of missed detection and duplicate detection. At the same time, it can accurately distinguish the short vertical segmentation line from the vertical line in the text, whether it is a single vertical line or double vertical line, without any error.

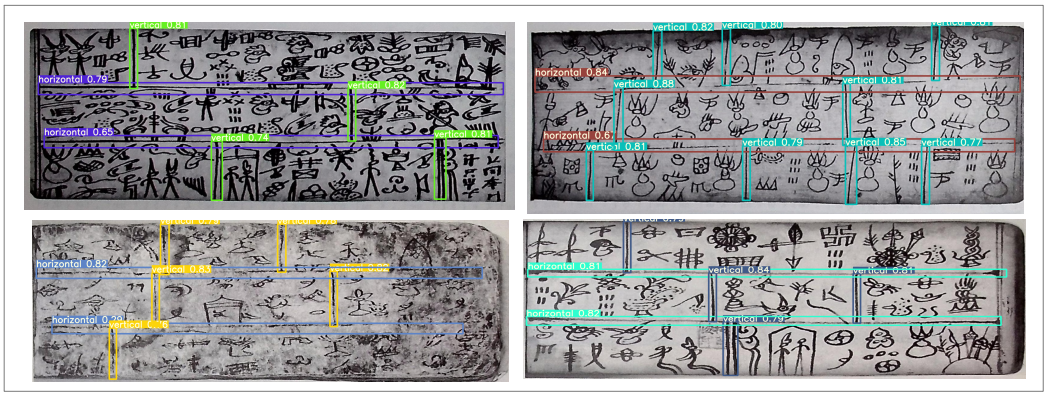


Figure 8. Images form the test dataset showing the performance for detecting the two classes "horizontal" and "vertical" under different conditions.

5.4. Superiority Studies

Currently, a large number of remarkable deep neural network models have been proposed in the field of object detection. To demonstrate the superiority of our method, we conduct extensive experiments to compare these remarkable models with our proposed Multi HAN on the DBS2022. In addition, we also execute several comparative experiments to make a comparison with different representative attention mechanisms to estimate the superiority of our proposed attention mechanism.

5.4.1. Comparison with Deep Neural Network Models

Our proposed method is aimed at the DS-SSLD task. To illustrate the superiority of the proposed Multi-HAN, we conduct extensive comparative experiments on our DBS2022 dataset with one-stage and two-stage deep neural network models, which have been highlighted in the field of object detection in recent years. For the sake of fairness in comparison, we retrain such comparative models on the DBS2022 by ourselves. The experimental settings and dataset partitioning of these models are consistent with the values we mentioned in Section 5.1, and we make certain that these outstanding models can converge to optimal values. The experimental results are shown in Table 2.

Table 2. Evaluation results of excellent object detection models on the proposed dataset. Our methods are shown in **bold**.

Models	Anchor-Based (Free)	Backbone	mAP@.5
Two-stage			
Faster-RCNN [20]	Anchor-based	ResNet101	87.7
	Anchor-based	Vgg16	82.4
One-stage			
SSD [27]	Anchor-based	ResNet50	79.7
YOLOv3 [28]	Anchor-based	DarkNet53	88.8
YOLOv5s [29]	Anchor-based	Focus + CSPDarkNet + SPP	91.5
CornerNet [31]	Anchor-free	Hourglass-104	79.9
CenterNet [32]	Anchor-free	ResNet50	89.2
CenterNet [32]	Anchor-free	Hourglass-104	91.6
Multi-HAN (Ours)	Anchor-based	Focus + CSPDarkNet + SPP	**94.7**

It is obvious from Table 2 that the Multi-HAN proposed in this paper achieves the most advanced detection results, reaching 94.7 of mAP@0.5. For instance, compared with the excellent anchor-free model, CenterNet [32] based on Hourglass-104 [56], Multi-HAN has achieved 3.4% improvement. These results can also prove that, although CornerNet [31] outperforms other models, such as Yolov3, in the MS COCO [57] dataset, it does not obtain excellent results in the DS-SSLD task, which indicates that the DS-SSLD task is special; it is different from conventional object detection tasks. Our approach is specifically designed to address the characteristics of the DS-SSLD task, and MHAU and Multi-HAN are both

designed based on the characteristics of Dongba scripture images. Therefore, the above experimental results can verify that the Multi-HAN designed by us can better deal with the DS-SSLD task consequently.

5.4.2. Comparison with Recently Proposed Attention Mechanism

Since most attention modules are plug-and-play modules, they can extract attention maps from any shallow feature map and apply them to subsequent feature maps. In order to prove the superiority of our attention module in Multi-HAN, without changing other network structures, we replace MHAU in Multi-HAN with other excellent attention modules, Triplet [46], CBAM [45], SE [42], EPSA [44] and ECA [43]. We retrain and test the comparative models that replace the attention module using the same training settings and dataset partitioning on the DBS2022. The evaluation indicators are parameters and mAP@0.5. The experimental results are listed in Table 3.

Table 3. Evaluation results of different attention mechanisms acting on our proposed network framework on the proposed DBS2022 dataset. Our methods are shown in **bold**.

Models	Parameters	mAP@.5
YOLOv5s + Multi-CSPU + Triplet [46]	16.1 M	90.5
YOLOv5s + Multi-CSPU + CBAM [45]	16.3 M	92.8
YOLOv5s + Multi-CSPU + SE [42]	16.2 M	93.2
YOLOv5s + Multi-CSPU + EPSA [44]	25.5 M	93.8
YOLOv5s + Multi-CSPU + ECA [43]	16.2 M	94.2
YOLOv5s + Multi-CSPU + MHAU (Ours)	25.5 M	**94.7**

From the results, it can be seen that except for the Triplet [46] attention module, all other detection results are improved, indicating that designing an appropriate attention mechanism can indeed improve the performance of the model. Overall, it can be summarized that the proposed MHAU is superior to any other attention model. Although our model sacrifices a part of the parameter quantity, it is 0.53% higher than ECA and 0.96% higher than EPSA with the same parameter quantity. It can be concluded that the MHAU proposed by us can obtain the most advanced detection effect.

5.5. Ablation Studies

Adequate ablation experiments are conducted on DBS2022 to evaluate the effectiveness of all components of Multi-HAN, including MHAU, Multi-CSPU and transmission of MHAU. The results of the ablation studies are illustrated in Table 4. mAP@0.5 is used as the measurement. These results effectively demonstrate the excellent performance of our proposed Multi-HAN. We evaluate and analyze the performance of the model from the following aspects based on the results from Table 4:

(1) We find that adding the MHAU with the cross-stage connection mode can improve DS-SSLD performance on our DBS2022 according to the results of Baseline and Baseline + MHAU (cross-stage) (the first and third row). It may be concluded that the proposed MHAU is more beneficial for Multi-HAN to focus on effective features, However, the degree of improvement is not significant because the role of the attention mechanism is to increase weights for certain effective features. If the ability to enhance features is insufficient, the improvement brought by the attention mechanism is limited.

(2) According to the results of Baseline + MHAU (downward) and Baseline + MHAU (cross-stage) (the second and third row), it can be found that the cross-stage connection mode can improve the performance of DS-SSLD more than the downward transmission connection mode. It can be concluded that when extracting attention maps from shallow feature maps of the network, the spatial feature map is large, and the number of channels is small; the extracted weights, especially the channel weights, are too

general and not applied to some specific features. The extracted spatial attention is sensitive and difficult to capture due to limited channels. Downward transmission may directly and fundamentally affect the ability of the backbone network to extract features, and may have a negative impact, on the contrary. The cross-stage MHAU can fully aggregate shallow detailed features and deep semantic information. Without affecting feature extraction, attention maps captured with corresponding sizes of shallow detail information can enhance and constrain deep global information, helping to distinguish important features from redundant features at the head stage.

(3) According to the results of Baseline and Baseline + Multi-CSPU (the first and forth row), we can know that Multi-CSPU-based MDconv [10] achieves more improvement. It benefits by the powerful ability of MDconv [10] to extract abundant multi-scale features, and also proves that rich feature expression is more important for improving performance.

(4) Baseline + MHAU (downward) (the second row) decreased by 0.1% compared with Baseline (the first row), while Baseline + MHAU (cross-stage) + Multi-CSPU (the fifth row) increased by 1.2% compared with Baseline + Multi-CSPU (the forth row). It can further prove that MDconv [10] has a strong ability to enhance effective feature extraction.

(5) Combining all the above components, the Multi-HAN model achieves the highest detection results, demonstrating further improvement in the combination of MHAU and Multi-CSPU. Moreover, on the basis of adding MDconv [10], increasing MHAU significantly improves the performance of the model, proving that strong feature extraction ability is a prerequisite for the attention mechanism to function.

Table 4. Results of ablation studies for the proposed Multi-HAN model; "downward" means the attention map extracted by MHAU is directly applied to the next layer, and "cross-stage" denotes that MHAU uses cross-stage connection to aggregate the deep information. Our methods are shown in **bold**.

Baseline	MHAU		Multi-CSPU	mAP@0.5
	Downward	Cross-Stage		
YOLOv5s				91.5
	√			91.4
		√		92.9
			√	92.7
	√		√	93.8
Multi-HAN (Ours)		√	√	94.7

5.6. Cross Validation

Since the DBS2022 dataset is proposed by ourselves and split randomly, it is necessary to conduct a cross validation to measure the stability and performance of the proposed model. The detailed settings for data partitioning are explained in Section 5.1. The experimental results of the cross validation are shown in the Table 5. The mAP@.5 values on the grouped test set fluctuate between −1.8 and +1.3, while the test results on the final evaluation dataset fluctuate between −0.3 and +0.4. The fluctuation value is within the normal range. The results and average values of the five tests tend to stabilize, It can prove the stability of our proposed model.

Table 5. The experimental results of cross validation. Average of five experimental results shown in **bold**.

k-th	Test Set		Final Evaluation	
	mAP@.5	**mAP@.5:.95**	**mAP@.5**	**mAP@.5:.95**
1-th	92.4	46.5	94.5	46.5
2-th	92.7	47.1	94.7	47.5
3-th	94.2	47.3	95.0	46.8
4-th	94.0	48.8	95.1	49.9
5-th	91.1	47.1	94.4	47.0
Average	92.9	47.4	94.7	47.5

6. Conclusions

The Dongba scripture sentence segmentation line detection (DS-SSLD) task, as the core technology of Dongba scripture sentence segmentation processing, is one of the basic tasks of Dongba scripture digital processing as machine translation. In order to improve the detection accuracy through using the deep learning-based methods, we first build a dataset (DBS2022) for the detection of sentence segmentation lines in Dongba scripture. DBS2022 can be used as a benchmark dataset to transform into datasets in various formats, such as coco and voc to, train any deep learning model. As far as we know, this is the only dataset for Dongba scripture sentence segmentation line detection, including the adequate authentic samples and the complete segmentation line labeling. Then, a multi-scale hybrid attention network (Multi-HAN) based on YOLOv5s is proposed to accurately detect sentence segmentation lines with large size differences. In Multi-HAN, MHAU is proposed to enhance the distinction between important features and redundant features, and cross-stage weighting is used to obtain fusion output with more meaningful features. MDconv is introduced to enhance multi-scale feature extraction without increasing the number of parameters. Through sufficient experiments, it is shown that Multi-HAN is a very effective sentence segmentation line detection framework for Dongba scripture. The experiments on the dataset we built verify the effectiveness and superiority of our method. A large number of experimental results show that our Multi-HAN method is superior to the most advanced method. In the future, we will focus on better handling the distinction of similar features between vertical lines and vertical sentence segmentation lines in Dongba characters, and consider carrying out research on the text segmentation of Dongba scriptures so as to lay a foundation for the research on digital processing and machine translation of Dongba scriptures.

Author Contributions: J.X.: Investigation, Formal analysis, Methodology, Resources, Data Curation, Writing Original Draft. X.B.: Conceptualization, Project administration, Funding acquisition, Writing—Review and Editing. Y.W.: Validation, Supervision, Writing—Review and Editing. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded by National Natural Science Foundation of China [grant number 62236011]; The National Social Science Fund of China [grant number 20&ZD279].

Data Availability Statement: We created a new datasets for Dongba scripture sentence segmentation line detection; we can submit the datasets, codes and other electronic materials if necessary.

Acknowledgments: We acknowledge all support given by Key Laboratory of Ethnic Language Intelligent Analysis and Security Governance of MOE.

Conflicts of Interest: The authors declare that they have no known competing financial interest or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Zheng, F. *Analysis and Segmentation Algorithm of Dongba Pictograph Document*; Nationalities Publishing House: Beijing, China, 2005; pp. 1–230.
2. Institute, D.C.R. *An Annotated Collection of Naxi Dongba Manuscripts*; Yunnan People's Publishing House: Kunming, China, 1999.
3. Yang, Y.T.; Kang, H.L. Analysis and Segmentation Algorithm of Dongba Pictograph Document. In Proceedings of the 2020 4th Annual International Conference on Data Science and Business Analytics (ICDSBA), Changsha, China, 5–6 September 2020; pp. 91–93.
4. Yang, Y.; Kang, H. Dongba Scripture Segmentation Algorithm Based on Discrete Curve Evolution. In Proceedings of the 2021 14th International Symposium on Computational Intelligence and Design (ISCID), Hangzhou, China, 11–12 December 2021; pp. 416–419.
5. Yang, Y.; Kang, H. Text Line Segmentation Algorithm for Dongba Pictograph Document. In Proceedings of the 2021 14th International Symposium on Computational Intelligence and Design (ISCID), Hangzhou, China, 11–12 December 2021; pp. 412–415.
6. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. In Proceedings of the Advances in Neural Information Processing Systems, Nevada, CA, USA, 3–6 December 2012; pp. 1097–1105.
7. Liu, S.; Huang, S.; Wang, S.; Muhammad, K.; Bellavista, P.; Del Ser, J. Visual tracking in complex scenes: A location fusion mechanism based on the combination of multiple visual cognition flows. *Inf. Fusion* **2023**, *96*, 281–296. [CrossRef]
8. Liu, S.; Gao, P.; Li, Y.; Fu, W.; Ding, W. Multi-modal fusion network with complementarity and importance for emotion recognition. *Inf. Sci.* **2023**, *619*, 679–694. [CrossRef]
9. Chen, Z.; Sun, Y.; Bi, X.; Yue, J. Lightweight image de-snowing: A better trade-off between network capacity and performance. *Neural Netw.* **2023**, *165*, 896–908. [CrossRef] [PubMed]
10. Tan, M.; Le, Q.V. Mixconv: Mixed depthwise convolutional kernels. *arXiv* **2019**, arXiv:1907.09595.
11. Gong, D.; Sha, F.; Medioni, G. Locally linear denoising on image manifolds. In Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics. JMLR Workshop and Conference Proceedings, Sardinia, Italy, 13–15 May 2010; pp. 265–272.
12. Chen, T.; Ma, K.K.; Chen, L.H. Tri-state median filter for image denoising. *IEEE Trans. Image Process.* **1999**, *8*, 1834–1838. [CrossRef]
13. Zhang, X.P.; Desai, M.D. Adaptive denoising based on SURE risk. *IEEE Signal Process. Lett.* **1998**, *5*, 265–267. [CrossRef]
14. Pan, Q.; Zhang, L.; Dai, G.; Zhang, H. Two denoising methods by wavelet transform. *IEEE Trans. Signal Process.* **1999**, *47*, 3401–3406. [CrossRef]
15. Zhou, S.F.; Liu, C.P.; Liu, G.; Gong, S.R. Multi-step segmentation method based on minimum weight segmentation path for ancient handwritten Chinese character. *J. Chin. Comput. Syst.* **2012**, *33*, 614–620.
16. Likas, A.; Vlassis, N.; Verbeek, J.J. The global k-means clustering algorithm. *Pattern Recognit.* **2003**, *36*, 451–461. [CrossRef]
17. Paliwal, S.S.; Vishwanath, D.; Rahul, R.; Sharma, M.; Vig, L. Tablenet: Deep learning model for end-to-end table detection and tabular data extraction from scanned document images. In Proceedings of the 2019 International Conference on Document Analysis and Recognition (ICDAR), Sydney, NSW, Australia, 20–25 September 2019; pp. 128–133.
18. Siddiqui, S.A.; Fateh, I.A.; Rizvi, S.T.R.; Dengel, A.; Ahmed, S. Deeptabstr: Deep learning based table structure recognition. In Proceedings of the 2019 International Conference on Document Analysis and Recognition (ICDAR), Sydney, NSW, Australia, 20–25 September 2019; pp. 1403–1409.
19. Renton, G.; Soullard, Y.; Chatelain, C.; Adam, S.; Kermorvant, C.; Paquet, T. Fully convolutional network with dilated convolutions for handwritten text line segmentation. *Int. J. Doc. Anal. Recognit. (IJDAR)* **2018**, *21*, 177–186. [CrossRef]
20. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *Adv. Neural Inf. Process. Syst.* **2015**, *28*, 1137–1149. [CrossRef] [PubMed]
21. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
22. Cai, Z.; Vasconcelos, N. Cascade r-cnn: Delving into high quality object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 6154–6162.
23. Zhu, Y.; Zhao, C.; Wang, J.; Zhao, X.; Wu, Y.; Lu, H. Couplenet: Coupling global structure with local parts for object detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 4126–4134.
24. Lee, H.; Eum, S.; Kwon, H. Me r-cnn: Multi-expert r-cnn for object detection. *IEEE Trans. Image Process.* **2019**, *29*, 1030–1044. [CrossRef] [PubMed]
25. Li, Y.; Chen, Y.; Wang, N.; Zhang, Z. Scale-aware trident networks for object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 6054–6063.
26. Cai, Z.; Fan, Q.; Feris, R.S.; Vasconcelos, N. A unified multi-scale deep convolutional neural network for fast object detection. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 354–370.
27. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. Ssd: Single shot multibox detector. In Proceedings of the Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; Proceedings, Part I 14; Springer: Amsterdam, The Netherlands, 2016; pp. 21–37.
28. Redmon, J.; Farhadi, A. YOLOv3: An Incremental Improvement. *arXiv* **2018**, arXiv:1804.02767.

29. Jocher, G.; Stoken, A.; Borovec, J.; Chaurasia, A.; Liu, C.; Laughing, A.; Hogan, A.; Hajek, J.; Diaconu, L.; Marc, Y.; et al. ultralytics/yolov5: V5. 0-YOLOv5-P6 1280 models AWS Supervise. ly and YouTube integrations. *Zenodo* **2021**, *11*. [CrossRef]
30. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. Yolov4: Optimal speed and accuracy of object detection. *arXiv* **2020**, arXiv:2004.10934.
31. Law, H.; Deng, J. CornerNet: Detecting Objects as Paired Keypoints. *Int. J. Comput. Vis.* **2020**, *128*, 642–656. [CrossRef]
32. Duan, K.; Bai, S.; Xie, L.; Qi, H.; Huang, Q.; Tian, Q. Centernet: Keypoint triplets for object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 6569–6578.
33. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. *arXiv* **2017**, arXiv:1706.03762.
34. Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; Dai, J. DD Deformable transformers for end-to-end object detection. In Proceedings of the 9th International Conference on Learning Representations, Virtual Event, 3–7 May 2021; pp. 3–7.
35. Zheng, M.; Gao, P.; Zhang, R.; Li, K.; Wang, X.; Li, H.; Dong, H. End-to-end object detection with adaptive clustering transformer. *arXiv* **2020**, arXiv:2011.09315.
36. Beal, J.; Kim, E.; Tzeng, E.; Park, D.H.; Zhai, A.; Kislyuk, D. Toward transformer-based object detection. *arXiv* **2020**, arXiv:2012.09958.
37. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
38. Qin, Z.; Zhang, P.; Wu, F.; Li, X. Fcanet: Frequency channel attention networks. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10-17 October 2021; pp. 783–792.
39. Sang, H.; Zhou, Q.; Zhao, Y. Pcanet: Pyramid convolutional attention network for semantic segmentation. *Image Vis. Comput.* **2020**, *103*, 103997. [CrossRef]
40. Chen, Y.; Kalantidis, Y.; Li, J.; Yan, S.; Feng, J. A^2-nets: Double attention networks. *arXiv* **2018**, arXiv:1810.11579.
41. Zhang, H.; Wu, C.; Zhang, Z.; Zhu, Y.; Lin, H.; Zhang, Z.; Sun, Y.; He, T.; Mueller, J.; Manmatha, R.; et al. Resnest: Split-attention networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 2736–2746.
42. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7132–7141.
43. Wang, Q.; Wu, B.; Zhu, P.; Li, P.; Zuo, W.; ECA-Net, Q.H. Efficient Channel Attention for Deep Convolutional Neural Networks. *arXiv* **2019**, arXiv:1910.03151.
44. Zhang, H.; Zu, K.; Lu, J.; Zou, Y.; Meng, D. EPSANet: An efficient pyramid squeeze attention block on convolutional neural network. In Proceedings of the Asian Conference on Computer Vision, Macao, China, 4–8 December 2022; pp. 1161–1177.
45. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
46. Misra, D.; Nalamada, T.; Arasanipalai, A.U.; Hou, Q. Rotate to attend: Convolutional triplet attention module. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Virtual Conference, 5–9 January 2021; pp. 3139–3148.
47. Lin, M.; Chen, Q.; Yan, S. Network in network. *arXiv* **2013**, arXiv:1312.4400.
48. Nair, V.; Hinton, G.E. Rectified linear units improve restricted boltzmann machines. In Proceedings of the 27th International Conference on Machine Learning (ICML-10), Haifa, Israel, 21–24 June 2010; pp. 807–814.
49. Nielsen, M.A. *Neural Networks and Deep Learning*; Determination Press: San Francisco, CA, USA, 2015; Volume 25.
50. Luce, R.D. *Individual Choice Behavior: A Theoretical Analysis*; Courier Corporation: North Chelmsford, MA, USA, 2012.
51. Wang, C.Y.; Liao, H.Y.M.; Wu, Y.H.; Chen, P.Y.; Hsieh, J.W.; Yeh, I.H. CSPNet: A new backbone that can enhance learning capability of CNN. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Seattle, WA, USA, 14–19 June 2020; pp. 390–391.
52. Yu, F.; Koltun, V. Multi-scale context aggregation by dilated convolutions. *arXiv* **2015**, arXiv:1511.07122.
53. Chen, Z.; Bi, X.; Zhang, Y.; Yue, J.; Wang, H. LightweightDeRain: Learning a lightweight multi-scale high-order feedback network for single image de-raining. *Neural Comput. Appl.* **2022**, *34*, 5431–5448. [CrossRef]
54. Paszke, A.; Gross, S.; Chintala, S.; Chanan, G.; Yang, E.; DeVito, Z.; Lin, Z.; Desmaison, A.; Antiga, L.; Lerer, A. Automatic differentiation in pytorch. In *NIPS 2017 Workshop*; NIPS: Long Beach, CA, USA, 2017.
55. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
56. Newell, A.; Yang, K.; Deng, J. Stacked hourglass networks for human pose estimation. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2016, pp. 483–499.
57. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2014, pp. 740–755.