



Article

FAS-Res2net: An Improved Res2net-Based Script Identification Method for Natural Scenes

Zhiyun Zhang ¹, Hornisa Mamat ¹, Xuebin Xu ¹, Alimjan Aysa ^{1,2,*}  and Kurban Ubul ^{1,2,*} ¹ School of Information Science and Engineering, Xinjiang University, Urumqi 830046, China² Xinjiang Key Laboratory of Multilingual Information Technology, Xinjiang University, Urumqi 830046, China

* Correspondence: alim@xju.edu.cn (A.A.); kurbanu@xju.edu.cn (K.U.)

Abstract: Problems such as complex image backgrounds, low image quality, diverse text forms, and similar or common character layouts in different script categories in natural scenes pose great challenges to scene script identification. This paper proposes a new Res2Net-based improved script identification method, namely FAS-Res2Net. In the feature extraction part, the feature pyramid network (FPN) module is introduced, which is beneficial to aggregate the geometric feature information extracted by the shallow network and the semantic feature information extracted by the deep network. Integrating the Adaptive Spatial Feature Fusion (ASFF) module is beneficial to obtain local feature information for optimal weight fusion. In addition, the global feature information of the image is extracted by introducing the swin transformer coding block, which makes the extracted feature information more abundant. In the classification part, the convolutional classifier is used to replace the traditional Linear classification, and the classification confidence of each category is output, which improves the identification efficiency. The improved algorithm achieved identification rates of 94.7% and 96.0% on public script identification datasets SIW-13 and CVSI-2015, respectively, which verified the superiority of the method.

Keywords: script identification; feature pyramid; adaptive spatial feature fusion; global feature; convolutional classifier



Citation: Zhang, Z.; Mamat, H.; Xu, X.; Aysa, A.; Ubul, K. FAS-Res2net: An Improved Res2net-Based Script Identification Method for Natural Scenes. *Appl. Sci.* **2023**, *13*, 4434. <https://doi.org/10.3390/app13074434>

Academic Editor: Byung-Gyu Kim

Received: 10 March 2023

Revised: 26 March 2023

Accepted: 26 March 2023

Published: 31 March 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Identifying the script is a crucial aspect of scene text recognition and serves as a prerequisite step for text recognition. The goal is not to determine the specific text content but rather to identify the language of the text, enabling the selection of the appropriate text recognition model. Therefore, script identification can be viewed as an image classification problem, which has been thoroughly researched by an increasing number of experts over the years. Different script languages include Chinese, English, Arabic, Greek, etc. As shown in Figure 1, different character stroke structures from different script languages can be seen. Wide applications of script identification include online archiving of multilingual scene images, product image search, multilingual machine translation, scene image understanding, etc.

Existing approaches for multi-script documents have demonstrated promising outcomes in script identification for printed, handwritten, or mixed documents [1]. However, integrating text recognition information into script identification, as proposed in [2], can be challenging, since extracting text recognition information is typically difficult. In [3], researchers utilized the VGG16 convolutional neural network to extract feature information, along with fine-grained and attention modules to identify fine-grained features at crucial positions. Ref. [4] provides a comprehensive overview of recent research progress in this field, highlighting the valuable achievements made in recent years. While the aforementioned methods have delivered satisfactory results, they primarily focus on script identification for documents.



Figure 1. Sample dataset example.

The extensive research and application of script identification is not only satisfied with document content, but script identification in natural scenarios is also becoming more and more important. Scene text images are different from printed document images, which is a big challenge in identification, mainly for the following reasons:

- (1) Text and image styles in natural scenes are changeable. They can come from various constructions, such as outdoor billboards, traffic signs, user instructions, etc. Text images contain different fonts, colors, multiple artistic styles, and there are great variations.
- (2) Low resolution, noise, and light changes will cause distortion, reduce image quality, and affect the accuracy of identification. Since it is difficult to get rid of the influence of objective factors such as weather, brightness, and equipment when capturing images, the image quality needs to be improved.
- (3) Languages belonging to the same family may have minimal differences, and distinct languages may share a common subset of characters. For instance, Greek, English, and Russian share a set of characters, and the arrangement of these characters is nearly identical. However, accurately distinguishing between them requires identifying the unique characters and components specific to each language, which presents a fine-grained classification challenge.
- (4) Background interference has a direct impact on identification accuracy. When the background of the image overlaps with the text, the background may be mistakenly regarded as part of the text, thereby identifying the wrong script.

The main work results are as follows:

- (1) This paper proposes an improved scene script identification method based on convolutional neural network Res2Net, namely FAS-Res2net.
- (2) Feature Pyramid Network (FPN) was proposed to preserve the deep semantic feature information and shallow geometric feature information of text images.
- (3) An adaptive spatial feature fusion module is proposed to calculate the spatial weights of feature maps at different levels, and the weight fusion feature information solves the feature conflict between positive and negative samples.
- (4) The two block-encoding modules of swin transformer were used to extract the global features of the image, which enriches the feature information of the script image and aggregates the self-applicable local feature information and the global feature information.
- (5) The full convolution classifier was used instead of the traditional Linear fully connected layer to output the classification confidence of each category, and then the script category was determined, which improves the classification efficiency.

This article is organized as follows:

Section 2 of this paper presents an overview of the current research status on script identification in natural scenes. Section 3 outlines the proposed script identification network's overall framework, emphasizing the Res2Net network module, Feature Pyramid, Adaptive Spatial Feature Fusion (ASFF) modules, swin transformer-encoding block, and fully convolutional classifier module. In Section 4, the script identification results are detailed and the experimental data are evaluated. Section 4 analyzes the strengths and

weaknesses of the experimental outcomes. Finally, Section 5 provides a summary of the completed work in this paper while offering insights into areas that need improvement for future research.

2. Related Works

There are two main categories of methods used for identifying scripts in natural scene images: traditional machine learning methods that involve manually designed features for script identification, and deep learning-based methods that automatically extract script features using techniques such as convolutional neural networks and transformer-encoding blocks. In general, automatic feature extraction based on deep learning has better performance than traditional handcrafting [5]. Traditional methods can achieve high accuracy for script identification of document images but are not suitable for scene text because they usually contain less text and are more difficult to image.

Over the past few years, there have been significant advancements in script identification technology due to the development of convolutional neural networks [6]. These networks offer an end-to-end solution that is more efficient and less time-consuming than traditional methods, and they continue to evolve and improve over time. The classic CNN network such as first AlexNet [7] to VGG [8], ResNet [9], and then to MobileNet [10], EfficientNet [11], and so on. Various pooling methods, such as overlapping pooling [12], global average pooling [13], and feature pyramid pooling [14], have been employed. Loss functions, including cross-entropy loss [15], focal loss [16], and dice loss [17], have also been used. In addition, commonly used optimization techniques include SGD [18], Adam [19], and AdamW [20]. To overcome the issue of fixed-size input images, fully convolutional networks and global pooling layers have been adopted. The ResNet residual learning framework solves the problems of degradation and precision saturation caused by deep networks and improves the identification accuracy by paying attention to the residual structure. Res2Net is an advanced version of ResNet that enhances the fine-grained multi-scale feature and receptive field size capabilities by introducing deep residual blocks.

Several approaches have been proposed for script identification. Shi et al. [21] designed a shallow CNN based on AlexNet, which introduced the horizontal position invariant. To handle irregular text, Luo et al. [22] proposed an attention-based sequence recognition and correction network. Ankan Kumar Bhunia et al. [23] proposed a CNN-LSTM network that extracts global and local features and employs a softmax layer to calculate patch weights for attention. Karim et al. [24] proposed a multivariate time series classification method composed of LSTM and FCN, which showed better results with the addition of an attention mechanism. Cheng et al. [25] processed similar sequence inputs with a patch aggregator and converted the number of channels to the number of script classes. Finally, Fujii et al. [26] used an Encoder and Summarizer to extract local features and fused them using an attention mechanism to reflect the importance of different patches.

3. Methods

3.1. Overview of Network Structure

This paper proposes an innovative approach to identify scripts in scene text images using an end-to-end network architecture consisting of feature extraction and classification layers. With the network structure shown in Figure 2, firstly, CNN feature extraction and feature pyramid structure were used to combine multi-layer spatial feature information. The attention mechanism calculates the weight of spatial features and concatenates the weight feature information of each layer to generate a fine-grained multi-scale feature map. In addition, the encoding block of Transformer extracts global feature information while serially aggregating local information. The second part of the classification layer uses a new convolutional classifier, which converts feature information into a class-level classification map through the superposition of convolutional layers. An adaptive pooling layer compresses the feature map into each class's classification confidence to determine the script

category. Additionally, the input images were rescaled to a fixed square size of 384×384 , which removes the need to preprocess the images in the dataset to a constant size.

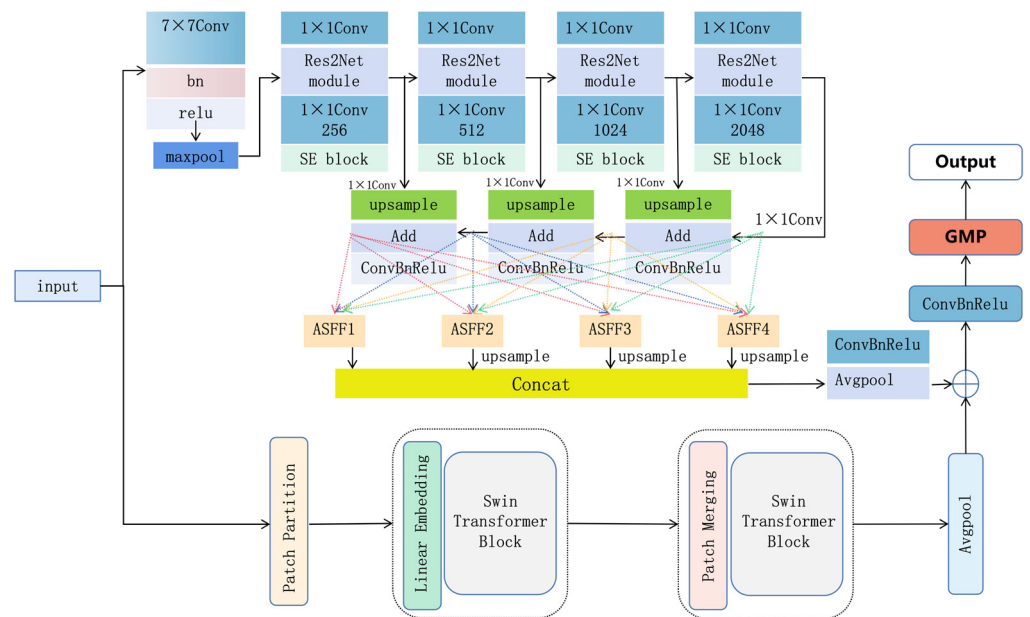


Figure 2. Network structure.

3.2. Feature Extraction Module

The Res2Net [27] convolutional neural network was used as the backbone network in the feature extraction module. As shown in Figure 3, it is an improved version of the ResNet50 bottleneck structure. The 3×3 convolution in the bottleneck structure was replaced by the 3×3 convolution group of the multi-level residual structure. This enhances the receptive field of the network, enabling it to obtain different levels of fine-grained scale feature information of objects. The multi-scale here does not refer to the combination of levels, but to the combination of multiple receptive fields at a finer granularity.

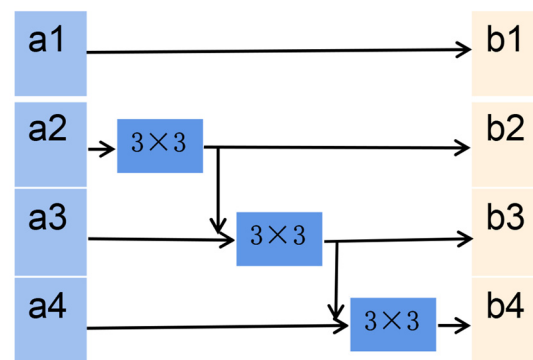


Figure 3. Res2Net module structure diagram.

To begin with, the input feature maps were subjected to a 1×1 convolution operation. Following this, they were divided equally into s map subsets in the channel dimension. These subsets have the same scale size, but the number of channels was reduced to $1/s$ of the input channels. Subsequently, each 3×3 convolution operation contains all the feature information from the previous stage and increases the receptive field. Res2Net can obtain combined feature information of varying numbers and receptive field sizes, while the hierarchical residual connections in individual residual blocks can capture fine-grained changes both for details and globally. Finally, the concat function was employed to parallelize the output features of each stage and obtain the final feature information.

3.3. Adaptive Multi-Layer Feature Fusion Module

The module for fusing features from text images adopts the Feature Pyramid Network (FPN) architecture [28]. Different text instances are mapped onto feature maps with different resolutions. While key feature details are extracted in one layer, other layers are considered as regions with background noise. Any incorrect evaluation can hinder the gradient propagation during backpropagation. The feature information of each layer in the pyramid structure has its unique use, and its information weight may hinder the gradient transfer and affect the identification results.

Figure 4 illustrates the feature fusion process. Initially, the feature information of each layer enters the pyramid, and the p1 feature map undergoes 1×1 convolution and upsampling. The fusion feature information is obtained by adding C_i of the same size to p1, p2, p3, and p4, and the resulting feature map sizes are 1/4, 1/8, 1/16, and 1/32 of the original image, respectively. Next, the Adaptive Spatial Feature Fusion (ASFF) algorithm [29] was applied to adjust the size and weight fusion of feature information. Feature information resizing involves downsampling by pooling and convolution and upsampling using interpolation. Adjust the feature map resolution in each layer to match the scale size of the specified layer. When upsampling, 1×1 convolution modifies the feature information channel size, and nearest neighbor interpolation adjusts the spatial size ratio. During downsampling, feature maps can be compressed to half their original size using 3×3 convolutions with a stride of 2, and the channels resized. For example, when downsampling by 1/4 to reduce resolution, pooling is performed first, followed by 3×3 convolution with stride 2. After resizing, perform adaptive blending is based on size. The fusion calculation formula is shown in formula (1):

$$S_{mn}^i = a_{mn}^i v_{mn}^{1 \rightarrow i} + b_{mn}^i v_{mn}^{2 \rightarrow i} + c_{mn}^i v_{mn}^{3 \rightarrow i} + d_{mn}^i v_{mn}^{4 \rightarrow i} \quad (1)$$

where $v_{mn}^{x \rightarrow i}$ represents the feature vector at (i, j) in the feature map adjusted from layer x to layer i . $v_{mn}^{1 \rightarrow i}, v_{mn}^{2 \rightarrow i}, v_{mn}^{3 \rightarrow i}, v_{mn}^{4 \rightarrow i}$ represent the scaled feature maps of layers 1, 2, 3, and 4, respectively, a, b, c , and d are the weight parameters of the four layers. The resized feature map enters the convolution calculation to obtain each weight parameter, and after concat, softmax, and normalization, the weight ratio between 0 and 1 is obtained, and the total weight sum is 1. The normalized calculation formula is as follows (2):

$$a_{mn}^i = \frac{e^{l_{a_{mn}}^i}}{e^{l_{a_{mn}}^i} + e^{l_{b_{mn}}^i} + e^{l_{c_{mn}}^i} + e^{l_{d_{mn}}^i}} \quad (2)$$

where $l_{a_{mn}}^i, l_{b_{mn}}^i, l_{c_{mn}}^i, l_{d_{mn}}^i$ is the core parameter. Adaptive aggregation of features at each layer was achieved by computing the weights of each layer at each scale. This adaptive multi-layer feature fusion strategy can effectively eliminate the influence of background noise, accurately identify positive samples, and retain aggregated effective information, enhancing the scale invariance of features. Finally, adjust the weight-fused p1_fu, p2_fu, p3_fu, and p4_fu feature information so that the dimensions, width, and height are all equal, and then perform concat parallel connection to obtain the best local feature information extracted by the entire CNN part.

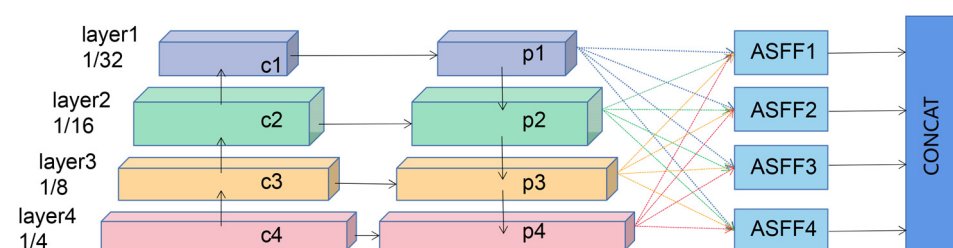


Figure 4. Adaptive multi-layer feature fusion module diagram.

3.4. Swin Transformer-Encoding Block

Although convolutional neural networks (CNNs) are good at extracting local information from text images, the depth of the network leads to loss of specific object details. To address this issue, this study incorporates the concept of the swin transformer [30] and introduces a transformer-encoding block. This block employs a window-based multi-head attention mechanism to capture extensive global contextual information while minimizing the computational power required for low-resolution mapping.

The network model diagram of the Swin Transformer is shown in Figure 5, illustrating the following steps to extract feature information from the input image. First, the RGB image was divided into small and non-overlapping feature map modules, and feature information was extracted from these sequential elements. Next, the global modeling semantic information was extracted, and the mapping was linearly embedded. A two-layer Swin Transformer block was used to obtain semantic feature information, and finally the input dimension was reduced by patch merging. The receptive fields of patches and windows are enhanced to extract feature information from each layer.

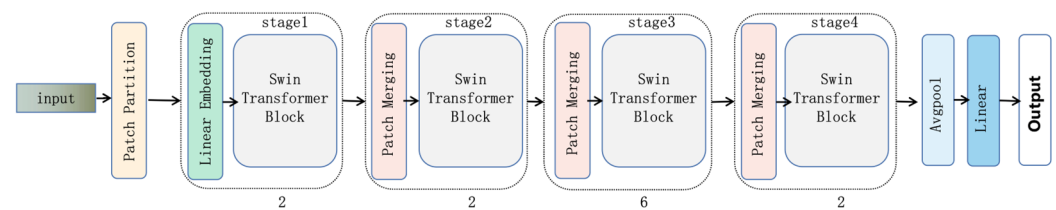


Figure 5. Swin transformer module diagram.

The structure of the Swin transformer block is shown in Figure 6 and consists of three main components: layer norm (LN), window-based multi-head self-attention (W-MSA), and shifted window multi-head self-attention (SW-MSA). This module builds windowed multi-head self-attention and multi-layer perceptrons using residual connections. It adopts local windows as basic units and focuses on windows to preserve image features and minimize computational complexity. Furthermore, feature maps are shifted in various directions based on half the window length to facilitate cross-window feature interactions and improve feature extraction.

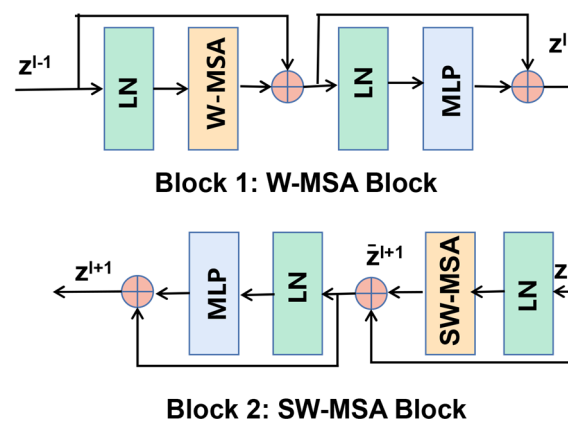


Figure 6. Swin transformer Block structure.

The complete calculation formula for the Swin transformer module is presented below:

$$\bar{Z}^l = W - MSA(LN(Z^{l-1})) + Z^{l-1} \quad (3)$$

$$Z^l = MLP(LN(\bar{Z}^l)) + \bar{Z}^l \quad (4)$$

$$\bar{Z}^{l+1} = SW - MSA(LN(Z^l)) + Z^l \quad (5)$$

$$Z^{l+1} = MLP(LN(\bar{Z}^{l+1})) + \bar{Z}^{l+1} \quad (6)$$

where Z^l and \bar{Z}^l are the output of the MLP and self-attention in the first block, Z^{l-1} is the input image sequence, Z^{l+1} is output image feature sequences, LN is a layer normalization operation, W -MSA is a windowed multi-head self-attention module, MLP is a multi-layer perceptron, and SW -MSA is a shifted window self-attention module.

3.5. Fully Convolutional Classifier

We replaced the traditional linear fully connected layer with a fully convolutional classifier [31], as shown in Figure 7. This replacement significantly improves the efficiency of classification training. First, the combined local and global feature information was forwarded to stacked convolutional layers, whose detailed structural specification is shown in Table 1. The number of channels is adjusted by 3×3 convolution and then by 1×1 convolution to equal the number of script classes. The class classification map was further compressed using global max pooling (GMP). Each channel number corresponds to a score map, and only the most important values are collected as follows:

$$a_i = \max f(i, x, y) \quad i = 1, 2 \dots T. \quad (7)$$

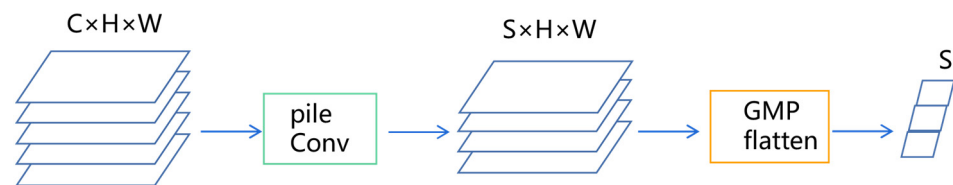


Figure 7. Full volume classification diagram.

Table 1. Convolution classifier configuration information.

Type	Configuration
Input	$512 \times H \times W$
Conv2d	Kernel: 3, Stride: 1, padding: 1, output: $384 \times H \times W$
BatchNorm	Channels: 384
Relu	
Conv2d	Kernel: 3, Stride: 1, padding: 1, output: $384 \times H \times W$
BatchNorm	Channels: 384
Relu	
Conv2d	Kernel: 1, Stride: 1, padding: 0, output: $C \times H \times W$

Term (i, x, y) represents the score map of the i -th channel, corresponding to the i -th script class. Therefore, the output size is the number of classes T , and each output a_i corresponds to the score of the i -th class. For each script category, output the score for each category. During training, it is converted into a probability through the softmax layer, which can be used directly during testing.

4. Experiments

The effectiveness of the proposed method was evaluated on two publicly available datasets for scene script identification, namely SIW-13 [23] and CVSI-2015 [31]. This section first introduces the two public datasets, then describes the specific implementation details, and finally presents and analyzes the experimental results.

4.1. Dataset

Table 2 summarizes the SIW-13 and CVSI-2015 datasets. The SIW-13 dataset contains 16,291 images, with 9791 training samples and 6500 testing samples. It includes 13 scripts: Arabic, Cambodian, Chinese, English, Greek, Hebrew, Japanese, Kannada, Korean, Mongolian, Russian, Thai, and Tibetan. Images were captured from natural scene regions and then corrected for horizontal orientation. Script images contain various font styles, complex backgrounds, and may be affected by different lighting conditions during capture. Furthermore, different script categories may share the same or similar subsets of characters, making this dataset very challenging.

Table 2. Image category and number in the dataset.

Dataset	Category	Train	Validate	Test
SIW-13	13	9791	-	6500
CVSI-2015	10	6412	1069	3207

The ICDAR 2015 Video Script Identification Competition released the CVSI-2015 dataset, which includes English, Hindi, Bengali, Oriya, Gujarati, Punjabi, Kannada, Tamil, and ten script categories for Telugu and Arabic. The dataset contains 6412 scene text pictures in the training set, 1069 in the verification set, and 3207 in the test set. Each script class accounts for about 10% of the data, and the dataset is relatively balanced. Compared with the SIW-13 dataset, this dataset is characterized as tidier, more uniform, and clearer. However, due to the limited number of script categories and the small amount of data, deep learning models may encounter learning and convergence difficulties.

4.2. Implementation Details

During the experiments, we used a computer equipped with an NVIDIA GeForce RTX 3090 graphics card with 24G of video memory, an Intel(R) Xeon(R) Platinum 8255C CPU running at 2.50 GHz, and 45 GB of RAM. The system is Ubuntu20.04, compiled with Python 3.8, PyTorch 1.10.0, and Cuda 11.3, and we use Pycharm 2021.1 professional version as the compiler. During the experiment, we initialized the feature extraction network with pre-trained models, specifically the pre-trained models res2net50_26w_4s and res2net101_26w_4s of the Res2Net50 and Res2Net101 networks on the ImageNet dataset. The input image size was resized to 384×384 before training, and the batch size is set to 32. We used stochastic gradient descent (SGD) as the optimizer with momentum set to 0.9 and weight decay set to e^{-4} . Finally, the initial learning rate was set to 10^{-2} .

4.3. Results of the Method in This Paper

To evaluate the impact of each newly added module, we conducted a series of ablation experiments on the SIW-13 dataset, which is widely regarded as the authoritative benchmark. These experiments allow us to analyze the effectiveness of each module and its contribution to the overall performance. The results of these experiments are listed in Table 3.

Table 3. Ablation experiment results of SIW-13.

Method	FPN	ASFF	Transformer	GMP	Accu. (%)
Res2Net					93.3
1	✓				94.0
2	✓	✓			94.2
3	✓	✓	✓		94.7
4	✓	✓	✓	✓	94.1

4.3.1. Benchmark Results

As detailed in the second row of Table 3, we only used the Res2Net module to extract features for script identification. This module is good at multi-scale representation at fine-grained level, and gradually enlarges the receptive field by multi-convolution to capture image details. For classification, we adopted traditional linear layers. This produced an identification accuracy of 93.3%, which was lower than the other variants. This is largely attributed to the lack of feature fusion and attention modules, as well as an inability to extract rich global feature information.

4.3.2. Feature Pyramid Results

We combined the Feature Pyramid Network (FPN) module with Res2Net to prevent information loss in shallow networks. This enabled us to simultaneously obtain semantic information from deep networks and geometric information from shallow networks, resulting in richer feature representations. This operation has also been verified in the experiment. At this time, the script identification rate reached 94.0%, which is 0.7% higher than that of the simple Res2Net network classification identification rate.

4.3.3. Adaptive Multi-Layer Feature Fusion Results

In the experiments detailed in the fourth row of Table 3, the addition of the Adaptive Spatial Feature Fusion (ASFF) module resulted in a performance improvement of 0.2%. This result indicates that the ASFF module can successfully filter out positive and negative sample conflicts in the spatial dimension during training and suppress the inconsistency of different features across network layers in multi-scale targets.

4.3.4. Transformer-Encoding Block

For the fourth experiment, we selected two coding blocks of the Swin Transformer to extract global feature information from the image. The local feature fusion information extracted by CNN was aggregated in series with the global information extracted by transformer, and the obtained information was classified. The identification rate reached 94.7%. It can be seen that the transformer block has a very good effect.

In order to determine the number of swin transformer-encoding blocks that achieve the best experimental results in script identification, we used 1, 2, 3, and 4 encoding blocks to conduct experiments on the dataset. The experimental results are shown in Table 4. When 3 and 4 coding blocks are used to extract the global feature information of the image, the script identification rate reaches the lowest 93.9%. When adding one encoding block, the effect is better, reaching an identification rate of 94.1%. The best identification rate is to use two encoding blocks to extract image feature information, and the identification rate reaches 94.7%.

Table 4. Comparison of identification results with different swin transformer-coding blocks.

Method	Transformer Block	Accu. (%)
Res2Net50 + FPN + ASFF	1	94.1
Res2Net50 + FPN + ASFF	2	94.7
Res2Net50 + FPN + ASFF	3	93.9
Res2Net50 + FPN + ASFF	4	93.9

4.3.5. GMP Results

The classification layer uses the convolutional classifier instead of the traditional Linear classification. During training, changes in the behavior of the classification layer led to changes in the top convolutional layer to accommodate the convolutional classifier. Although the experimental results show that the identification accuracy is 0.6% lower than the traditional classification, it has a lower number of parameters, which greatly improves the efficiency of training classification. In the experiment, the running time of each epoch is

more than 22 s faster than that of the traditional classification layer. It is a good choice to improve the efficiency of training classification within a limited tolerable accuracy range.

4.3.6. Experimental Results under Different Parameters

In order to further confirm the superiority of the method proposed in this paper and find out the best experimental results, we carried out related experiments on the SIW-13 dataset under different parameter conditions. The experimental results are shown in Table 5. The basic network structure of Res2Net50 and Res2Net101 was used for improvement, and experiments were carried out under different batch sizes. In general, the effect of extracting features based on the network structure of Res2Net50 was better than that of Res2Net101. Because the Res2Net101 network has too many layers, the extracted feature information is relatively redundant, and the problem of over-fitting has occurred. It can be seen from the experimental results that the identification rate does not necessarily strictly follow the linear growth relationship with the increase of the batch size, but may decrease. When the value of Batchsize is 32, the effect of improving the network based on Res2Net50 is the best to achieve an identification rate of 94.7%.

Table 5. The identification results of the improved method based on Res2Net50 and Res2Net101 under different batch sizes (%).

Method	16	32	64	128
Res2Net50 + FPN + ASFF + Swin	94.3	94.7	94.5	94.6
Res2Net101 + FPN + ASFF + Swin	94.1	94.4	94.4	94.3

To demonstrate the effectiveness and generalization of our proposed method, we also conducted experiments on the publicly available CVSI-2015 dataset. Compared with the SIW-13 dataset, the CVSI-2015 dataset contains fewer samples, which may hinder the training convergence of deep learning networks. To address this issue, we employed a novel data augmentation technique to increase the number of training sets by generating edge flow images from raw images. Our experimental results show that the script identification rate increases to 96.0%, which is better after applying edge flow data augmentation.

4.3.7. Error Analysis



Although this paper uses an innovative script identification method, there are still some scripts that cannot be identified correctly. As shown in Figure 8, for the first and sixth images (id:327_1 and id:188_1), the identification results are English and Chinese, respectively, while the groundtruths are Greek and Japanese, respectively. The identification error is because there are no special characters of Greek and Japanese in the image, only English letters and Chinese characters. The distinction between the two scripts cannot be made, so this task is extremely difficult. For the second image (id:139_2), due to the irregular orientation of the script symbols and the short length of the script, the feature information is biased towards Japanese, and Arabic is misidentified as a Japanese script. The third and fourth images are both due to the complex background, and the extracted feature information cannot be matched well, resulting in identification errors. The identification of the fifth image as English should be the correct identification result, but it was wrongly marked as the Russian label when making the dataset.



Figure 8. Samples of script identification errors in the SIW-13 dataset. The three lines of labels below each script are image id, identification result, and groundtruth.

As shown in Table 6, it shows the gradual change from the basic Res2Net network to the added modules, from the identification error to the correct one. These two script images are from the public dataset SIW-13 test set. The first script image is Arabic, but because its background is complex and runs through the entire image, neither the pure Res2Net or the FPN structure can extract effective features for correct identification. After adding the Adaptive Spatial Feature Fusion (ASFF) module, the effective features are weighted, and the script is identified correctly. The groundtruth of the second image is Japanese. It can be seen that the script in the image is composed of some Japanese characters and Chinese characters. Without the transformer module in the network, the identified script is Chinese. After the block extracts the global feature information of the image, it is correctly identified as Japanese, thus verifying the effectiveness of each module.

Table 6. Two script identification examples illustrating the gradual change from wrong to correct. These five methods correspond to the five methods in the ablation experiments in Table 3, respectively.

Image	Id	Groundtruth	Res2Net	Res2Net + FPN	Res2Net + FPN + ASFF	Res2Net + FPN + ASFF + Transformer	Res2Net + FPN + ASFF + Transformer + GMP
	141_2	Arabic	English	English	Arabic	Arabic	Arabic
	460_1	Japanese	Chinese	Chinese	Chinese	Japanese	Japanese

4.4. Comparison with State-of-the-Art Results

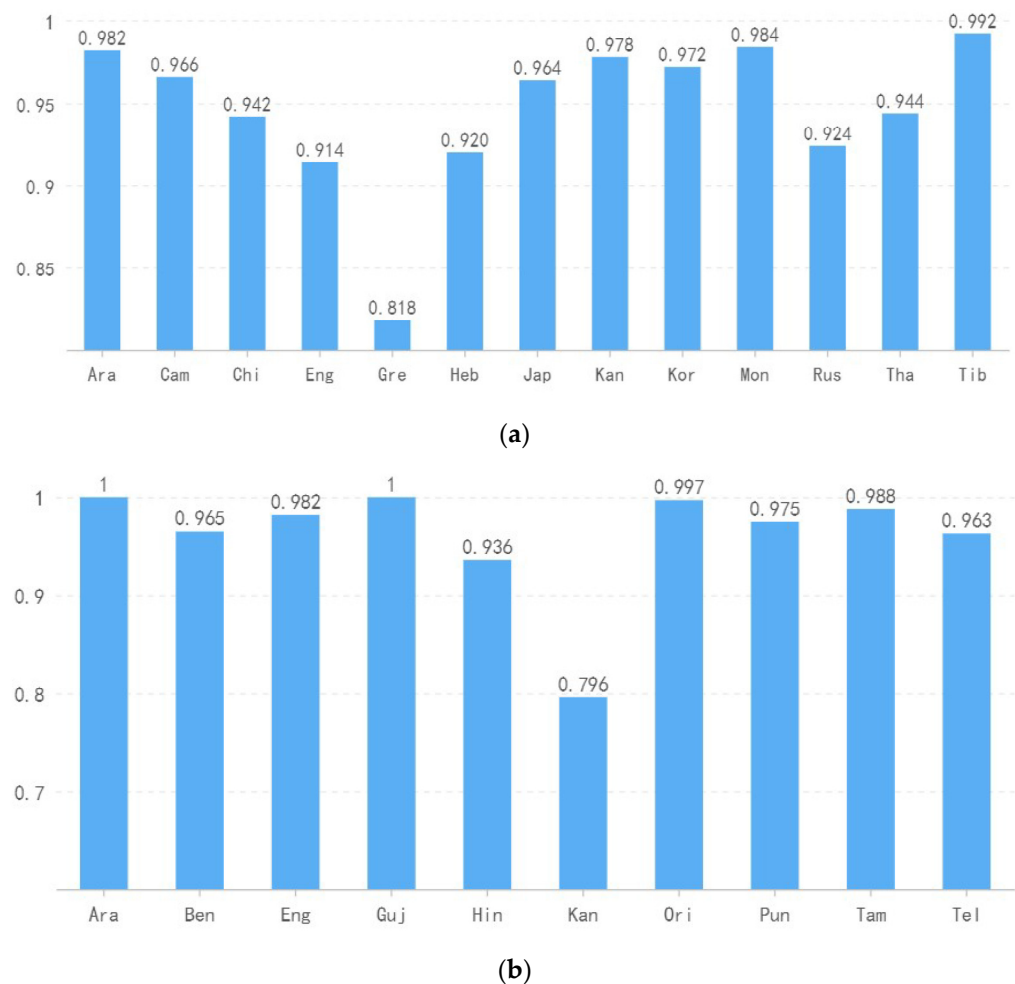
We compared the performance of our proposed method with other existing methods on the SIW-13 and CVSI-2015 datasets, and the results are shown in Table 7. Since this research direction is relatively small, the existing comparative experiments are limited. Our proposed method outperforms DisCNN, CNN + LSTM, and CNN + BoVW methods on the SIW-13 dataset. Furthermore, we achieved satisfactory results on the CVSI-2015 dataset, with slightly lower performance than CNN + BoVW and 2-stage CNN methods, but better than others. Perhaps these methods design a special softmax loss function specifically for this dataset.

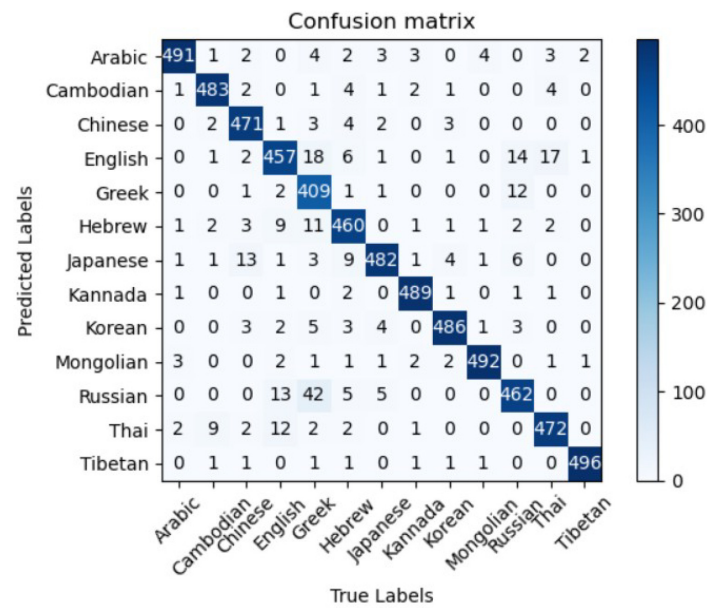
Table 7. Comparison accuracy (%) on two public script datasets.

Method	SIW-13	CVSI-2015
Our	94.7	96.0
DisCNN [21]	88.0	94.0
CNN + LSTM [32]	92.0	94.0
SRS + LBP + KNN [33]	-	94.0
Conv.feature + NBNN [34]	-	95.0
CNN + BoVW [35]	92.0	97.0
CNN + LSTM, VGG16 [31]	-	88.3
2-stage CNN [36]	-	97.4

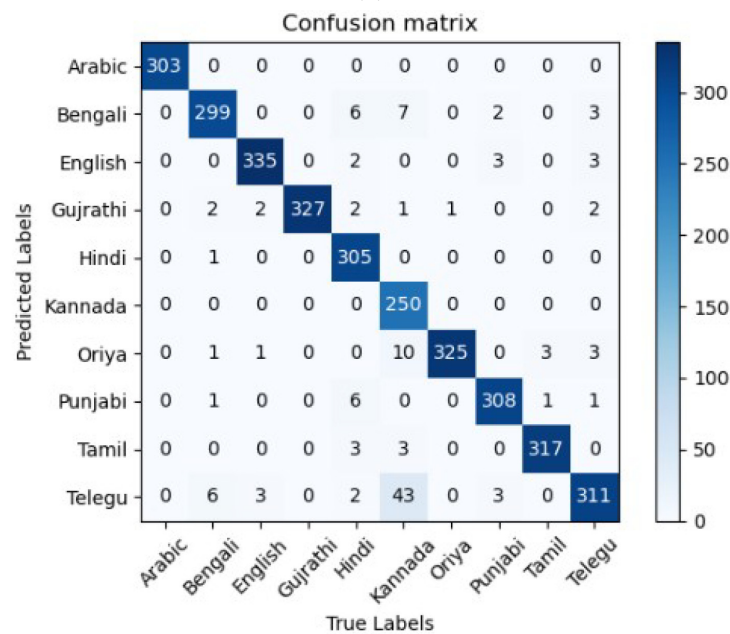
4.5. Experimental Analysis

To verify the effectiveness of our method, we further predicted the identification accuracy of each language in each dataset. The prediction results are shown in Figure 9, and a confusion matrix was generated for each dataset prediction, as shown in Figure 10.

**Figure 9.** Classification performance on two public datasets. (a) SIW-13, (b) CVSI-2015.



(a)



(b)

Figure 10. Confusion matrix on two publicly available datasets. The x-coordinates and y-coordinates represent the real label and the predicted category, respectively, and the darker the color, the higher the proportion of correct predictions. (a) SIW-13, (b) CVSI-2015.

As shown in Figure 9a, it shows that the identification performance of Tibetan in the SIW-13 dataset is the best, reaching an identification rate of 99.2%. Greek had the lowest accuracy rate of 81.8%. From the confusion matrix in Figure 10a, it can be seen that 42 Greek images are misclassified as Russian. The reason for this is a language similarity issue, and that there are shared similar characters. English has a relatively low accuracy rate and is mainly misidentified as Russian and Thai. Most other languages have an identification rate of 94% or above, and our overall identification rate on this dataset is 94.7%.

To illustrate the accuracy of the proposed method on the CVSI-2015 dataset, Figure 9b shows the identification accuracy for each category and Figure 10b shows the corresponding confusion matrix. Since Arabic and Gujarati have their own independent and special character forms, the identification rate has reached 100%. Except for Hindi and Kannada,

the identification rate of other languages has reached over 96%. Except for Kannada, none of the scripts had more than seven misclassifications. Most of the misclassifications of Kannada were identified as Telugu because they all have relatively curved font textures, and a small part were identified as Oriya and Bengali. We achieved an overall identification rate of 96.0% on this dataset.

5. Conclusions

Aiming at the problems of complex background, variable text style, and shared characters of script images in natural scenes, this paper proposes a script multi-classification identification method based on convolutional neural network Res2Net, namely FAS-Res2Net. A convolutional neural network Res2Net with improved residual blocks is used as the backbone of feature extraction, and improvements are mainly made in the feature extraction module and classification layer. For the problem of missing feature information in the deep network, the feature pyramid module was used to aggregate the deep and shallow semantic and geometric feature information of the image. An adaptive spatial feature fusion module was also integrated to calculate the weight of feature information of each layer, which solves the feature conflict between positive and negative samples and obtains the best adaptive local feature information. For the global feature of the image, two encoding blocks of the swin transformer were proposed to extract the image feature information and aggregated with the local information extracted by CNN. Finally, using the full convolution classifier instead of the traditional Linear classification reduces the network parameters and improves the training efficiency. In the future, the script identification model can be embedded into the text detection model. For script images in natural scenes, text detection directly performs script identification after determining the text position to realize the automation of script identification in natural scenes.

Author Contributions: Conceptualization, Z.Z.; methodology, Z.Z.; software, Z.Z.; investigation, Z.Z.; validation, Z.Z.; resources, H.M., X.X. and K.U.; writing—original draft preparation, Z.Z. and A.A.; writing—review and editing, Z.Z. and A.A.; visualization, Z.Z.; project administration, H.M., X.X. and K.U.; funding acquisition, A.A. and K.U. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Natural Science Foundation of China (No. 62266044, 61363064 and 61862061) and Natural Science Foundation of Science and Technology Department of Xinjiang Uygur Autonomous Region (No. 2021D01C119).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: We The datasets used in this paper are publicly available, including SIW-13 and CVSI-2015.

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

References

1. Ubul, K.; Tursun, G.; Aysa, A.; Impedovo, D.; Pirlo, G.; Yibulayin, I. Script Identification of Multi-Script Documents: A Survey. *IEEE Access* **2017**, *5*, 6546–6559. [\[CrossRef\]](#)
2. Cao, Y.; Li, J.; Wang, Q.F.; Huang, K.; Zhang, R. Improving Script Identification by Integrating Text Recognition Information. *Aust. J. Intell. Inf. Process. Syst.* **2019**, *16*, 67–75.
3. Ma, M.; Wang, Q.-F.; Huang, S.; Goulermas, Y.; Huang, K. Residual attention-based multi-scale script identification in scene text images. *Neurocomputing* **2020**, *421*, 222–233. [\[CrossRef\]](#)
4. Naosekpam, V.; Sahu, N. Text detection, recognition, and script identification in natural scene images: A Review. *Int. J. Multimed. Inf. Retr.* **2022**, *11*, 291–314. [\[CrossRef\]](#)
5. Gomez, L.; Nicolaou, A.; Karatzas, D. Improving patch-based scene text script identification with ensembles of conjoined networks. *Pattern Recognit.* **2017**, *67*, 85–96. [\[CrossRef\]](#)
6. Huang, K.; Hussain, A.; Wang, Q.F.; Zhang, R. (Eds.) *Deep Learning: Fundamentals, Theory and Applications*; Springer: Berlin, Germany, 2019.

7. Hosny, K.M.; Kassem, M.A.; Fouad, M.M. Classification of skin lesions into seven classes using transfer learning with AlexNet. *J. Digit. Imaging* **2020**, *33*, 1325–1334. [\[CrossRef\]](#)
8. Sitaula, C.; Hossain, M.B. Attention-based VGG-16 model for COVID-19 chest X-ray image classification. *Appl. Intell.* **2021**, *51*, 2850–2863. [\[CrossRef\]](#)
9. Roy, S.K.; Manna, S.; Song, T.; Bruzzone, L. Attention-Based Adaptive Spectral–Spatial Kernel ResNet for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2020**, *59*, 7831–7843. [\[CrossRef\]](#)
10. Srinivasu, P.N.; SivaSai, J.G.; Ijaz, M.F.; Bhoi, A.K.; Kim, W.; Kang, J.J. Classification of skin disease using deep learning neural networks with MobileNet V2 and LSTM. *Sensors* **2021**, *21*, 2852. [\[CrossRef\]](#)
11. Marques, G.; Agarwal, D.; Diez, I.D.L.T. Automated medical diagnosis of COVID-19 through EfficientNet convolutional neural network. *Appl. Soft Comput.* **2020**, *96*, 106691. [\[CrossRef\]](#)
12. Akhtar, N.; Ragavendran, U. Interpretation of intelligence in CNN-pooling processes: A methodological survey. *Neural Comput. Appl.* **2019**, *32*, 879–898. [\[CrossRef\]](#)
13. Kumar, R.L.; Kakarla, J.; Isunuri, B.V.; Singh, M. Multi-class brain tumor classification using residual network and global average pooling. *Multimed. Tools Appl.* **2021**, *80*, 13429–13438. [\[CrossRef\]](#)
14. Zhu, Y.; Wan, L.; Xu, W.; Wang, S. ASPP-DF-PVNet: Atrous Spatial Pyramid Pooling and Distance-Filtered PVNet for occlusion resistant 6D object pose estimation. *Signal Process. Image Commun.* **2021**, *95*, 116268. [\[CrossRef\]](#)
15. Dong, Y.; Shen, X.; Jiang, Z.; Wang, H. Recognition of imbalanced underwater acoustic datasets with exponentially weighted cross-entropy loss. *Appl. Acoust.* **2020**, *174*, 107740. [\[CrossRef\]](#)
16. Yeung, M.; Sala, E.; Schönlieb, C.-B.; Rundo, L. Unified Focal loss: Generalising Dice and cross entropy-based losses to handle class imbalanced medical image segmentation. *Comput. Med. Imaging Graph.* **2022**, *95*, 102026. [\[CrossRef\]](#) [\[PubMed\]](#)
17. Zhao, R.; Qian, B.; Zhang, X.; Li, Y.; Wei, R.; Liu, Y.; Pan, Y. Rethinking dice loss for medical image segmentation. In Proceedings of the 2020 IEEE International Conference on Data Mining (ICDM), Sorrento, Italy, 17–20 November 2020; pp. 851–860.
18. Woodworth, B.E.; Patel, K.K.; Srebro, N. Minibatch vs local sgd for heterogeneous distributed learning. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 6281–6292.
19. Liu, Z.; Shen, Z.; Li, S.; Helwegen, K.; Huang, D.; Cheng, K.T. How do adam and training strategies help bnns optimization. In Proceedings of the International Conference on Machine Learning, Virtual, 18–24 July 2021; pp. 6936–6946.
20. Kalfaoglu, M.; Kalkan, S.; Alatan, A.A. Late temporal modeling in 3d cnn architectures with bert for action recognition. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; Springer: Cham, Switzerland, 2020; pp. 731–747.
21. Shi, B.; Bai, X.; Yao, C. Script identification in the wild via discriminative convolutional neural network. *Pattern Recognit.* **2016**, *52*, 448–458. [\[CrossRef\]](#)
22. Luo, C.; Jin, L.; Sun, Z. MORAN: A Multi-Object Rectified Attention Network for scene text recognition. *Pattern Recognit.* **2019**, *90*, 109–118. [\[CrossRef\]](#)
23. Bhunia, A.K.; Konwer, A.; Bhunia, A.K.; Bhowmick, A.; Roy, P.P.; Pal, U. Script identification in natural scene image and video frames using an attention based Convolutional-LSTM network. *Pattern Recognit.* **2019**, *85*, 172–184. [\[CrossRef\]](#)
24. Karim, F.; Majumdar, S.; Darabi, H.; Harford, S. Multivariate LSTM-FCNs for time series classification. *Neural Netw.* **2019**, *116*, 237–245. [\[CrossRef\]](#)
25. Cheng, C.; Huang, Q.; Bai, X.; Feng, B.; Liu, W. Patch aggregator for scene text script identification. In Proceedings of the 2019 International Conference on Document Analysis and Recognition (ICDAR), Sydney, Australia, 20–25 September 2019; pp. 1077–1083.
26. Fujii, Y.; Driesen, K.; Baccash, J.; Hurst, A.; Popat, A.C. Sequence-to-label script identification for multilingual ocr. In Proceedings of the 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), Kyoto, Japan, 9–15 November 2017; IEEE: Piscataway, NJ, USA, 2017; Volume 1, pp. 161–168.
27. Gao, S.H.; Cheng, M.M.; Zhao, K.; Zhang, X.Y.; Yang, M.H.; Torr, P. Res2net: A new multi-scale backbone architecture. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *43*, 652–662. [\[CrossRef\]](#) [\[PubMed\]](#)
28. Peng, F.; Miao, Z.; Li, F.; Li, Z. S-FPN: A shortcut feature pyramid network for sea cucumber detection in underwater images. *Expert Syst. Appl.* **2021**, *182*, 115306. [\[CrossRef\]](#)
29. Cheng, X.; Yu, J. RetinaNet with difference channel attention and adaptively spatial feature fusion for steel surface defect detection. *IEEE Trans. Instrum. Meas.* **2020**, *70*, 1–11. [\[CrossRef\]](#)
30. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 10012–10022.
31. Dastidar, S.G.; Dutta, K.; Das, N.; Kundu, M.; Nasipuri, M. Exploring knowledge distillation of a deep neural network for multi-script identification. In Proceedings of the International Conference on Computational Intelligence in Communications and Business Analytics, Santiniketan, India, 7–8 January 2021; Springer: Cham, Switzerland, 2021; pp. 150–162.
32. Mei, J.; Dai, L.; Shi, B.; Bai, X. Scene text script identification with convolutional recurrent neural networks. In Proceedings of the 2016 23rd International Conference on Pattern Recognition (ICPR), Cancun, Mexico, 4–8 December 2016; IEEE: Piscataway, NJ, USA, 2016; pp. 4053–4058.

33. Nicolaou, A.; Bagdanov, A. D.; Liwicki, M.; Karatzas, D. Sparse radial sampling LBP for writer identification. In Proceedings of the 2015 13th International Conference on Document Analysis and Recognition (ICDAR), Tunis, Tunisia, 23–26 August 2015; IEEE: Piscataway, NJ, USA, 2015; pp. 716–720.
34. Gomez, L.; Karatzas, D. A fine-grained approach to scene text script identification. In Proceedings of the 2016 12th IAPR Workshop on Document Analysis Systems (DAS), Santorini, Greece, 11–14 April 2016; IEEE: Piscataway, NJ, USA, 2016; pp. 192–197.
35. Zdenek, J.; Nakayama, H. Bag of local convolutional triplets for script identification in scene text. In Proceedings of the 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), Kyoto, Japan, 9–15 November 2017; IEEE: Piscataway, NJ, USA, 2017; Volume 1, pp. 369–375.
36. Mahajan, S.; Rani, R. Word Level Script Identification Using Convolutional Neural Network Enhancement for Scenic Images. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* **2022**, *21*, 1–29. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.