Hindawi
Discrete Dynamics in Nature and Society
Volume 2022, Article ID 9188356, 10 pages
https://doi.org/10.1155/2022/9188356

Research Article

Multilevel Attention and Multiscale Feature Fusion Network for Author Classification of Chinese Ink-Wash Paintings

Wei Jiang, Xianglian Meng, and Ji Xi

School of Computer Information and Engineering, Changzhou Institute of Technology, Changzhou, China

Correspondence should be addressed to Wei Jiang; jiangweitju@163.com

Received 20 December 2021; Accepted 15 February 2022; Published 10 March 2022

Academic Editor: Jinchang Ren

Copyright © 2022 Wei Jiang et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to effectively extract features with high representation ability has always been a research topic and a challenge for classification tasks. Most of the existing methods mainly solve the problem by using deep convolutional neural networks as feature extractors. Although a series of excellent network structures have been successful in the field of Chinese ink-wash painting classification, but most of them adopted the methods of only simple augmentation of the network structures and direct fusion of different scale features, which limit the network to further extract semantically rich and scale-invariant feature information, thus hindering the improvement of classification performance. In this paper, a novel model based on multi-level attention and multi-scale feature fusion is proposed. The model extracts three types of feature maps from the low-level, middle-level and high-level layers of the pretrained deep neural network firstly. Then, the low-level and middle-level feature maps are processed by the spatial attention module, nevertheless the high-level feature maps are processed by the scale invariance module to increase the scale-invariance properties. Moreover, the conditional random field module is adopted to fuse the optimized three-scale feature maps, and the channel attention module is followed to refine the features. Finally, the multi-level deep supervision strategy is utilized to optimize the model for better performance. To verify the effectiveness of the model, extensive experimental results on the Chinese ink-wash painting dataset created in this work show that the classification performance of the model is better than other mainstream research methods.

1. Introduction

In recent years, the most effective visual recognition tasks are based on the complex and deep convolutional neural networks (CNNs) which stack multiple convolution and pooling layers to generate the high-level semantic features [1–9] or other technology [10, 11]. Specially, the high-level semantic information is widely used to achieve competitive performance in many research areas, for example, the classification on Chinese ink-wash paintings (IWPs) [12–14] which attracts increasing attention. To further improve the classification performance, the handcraft features extracted from Chinese IWPs are integrated with those high-level semantic features in some works. However, instinctively, the high-level features from CNNs and handcraft features are heterogeneous essentially. It is difficult to make full use of these different types of features which are extracted by separately methods.

Not to be overlooked, low-level spatial structural features from deep CNN architectures also play a crucial role in those visual recognition tasks. Some works introduced that the low-level and high-level features extracted from different layers [15–17] were integrated to improve the classification performance. Whereas, in those methods they ignored the middle-level features which were complementary and could contribute to the final performance. Therefore, how to extract the middle-level features from the network and apply them to the recognition task is a very worthy topic.

In addition, low-level, middle-level, and high-level features extracted from different layers of deep neural networks play different roles in image classification tasks. The simple cascading or weighted average fusion method for these multi-scale features can not well mine the complementarity among features to obtain the more discriminative representation [18–20]. Therefore, it is still a

difficult problem how to effectively integrate these complementary multi-scale feature information to obtain the better performance.

On the other hand, the feature maps extracted from the shallow structure of the network mainly contain spatial structured information such as background, while the feature maps extracted from the deep structure have more abstract semantic information. However, many research works have not fully considered their contributions before integrating these features. Therefore, to solve this problem, the attention mechanisms were proposed to learn more discriminative fusion features [21–23]. In a recognition task, spatial and channel attention mechanisms are introduced to enhance the useful information units and restrain the redundant information units, so that different scales of feature information can be utilized efficiently.

To address the problems mentioned above on the classification of Chinese IWPs, our work explores how to make the best use of the low-level, middle-level, and high-level image features obtained from different layers, and how to take full advantage of the attention mechanism to fuse the multi-scale information for achieving better classification performance. In comparison with the existing researches on author classification of Chinese IWPs dataset, our key contributions can be highlighted below and detailed in the following sections.

(i) Different from directly using low-level and high-level features for image classification, we propose a hybrid model which extracts the low-level, middle-level, and high-level feature representations from different layers in a deep CNN architecture. As for low-level and middle-level features, a spatial attention module is adopted to filter out some irrelevant details. While, the high-level feature is processed by a scale invariant module to increase the scale-invariant properties of advanced features.

(ii) The multi-scale feature fusion-based conditional random field is utilized to integrate the processed three-scale features. After that, a channel attention module is introduced to assign different weights to different feature channels with their contributions. Moreover, we design a multi-level deep supervision module to optimize the model by using three types of processed features.

(iii) We conduct a series of experiments to compare the proposed model with other prominent approaches for author classification of Chinese IWPs dataset. Extensive experimental results show that our model achieves promising performance, which demonstrates the effectiveness and superiority of our model.

The remainder of this paper is organized as follows. Some important and related works are reviewed in Section 2. In Section 3, we present the proposed hybrid model for author classification and describe it in detail. Section 4 represents the experimental results and analysis. The conclusions are given in Section 5.

2. Related Work

As mentioned in the previous section, deep CNN architecture, attention mechanisms, and multi-scale feature fusion are essential to improve the classification performance of Chinese IWPs. Therefore, this section will introduce these mainstream concepts and related technologies.

2.1. Deeper Architecture Design. Many researchers have focused on the classification of Chinese IWPs by utilizing deeper neural networks due to the success of achieving state-of-the-art performance with better representations [13, 14]. As a result, deeper architecture design plays a critical role in computer vision tasks. For example, as a success CNN model, the AlexNet [3] significantly improved performance compared to traditional visual recognition methods by stacking filters sequentially. However, AlexNet architecture should be changed and improved due to its limited network depth and filter kernel size. VGGNet [24] and ResNet [2] showed that increasing the depth of neural networks could significantly improve the quality of the representation and the performance of the classification. To deal with the problem of the slow convergence of training caused by the increase of network depth, Batch Normalization (BN) [25] was introduced to regulate the distribution of the inputs to each layer in deeper networks. Furthermore, the ResNet introduced short connections to neural networks and obtained much deeper network architectures by solving the problem of gradient vanishing.

With the rapid development of deep learning technology, there are increasing excellent deep neural network models and exciting architectures. Deep learning models with pre-trained and some strategies have also emerged to perform image classification and recognition tasks. Based on the above discussion, these pre-trained deep network models, such as VGGNet, will also be chosen to perform the classification task of Chinese IWPs in our work.

2.2. Attention Mechanism. In recent years, the attention mechanism has been proved to be applied successfully in visual recognition tasks such as image/video captioning [22, 26, 27], image classification [28, 29], saliency detection [8, 30], Visual Question Answering (VQA) [31, 32]. Chen et al. [22] introduced a novel CNN dubbed SCA-CNN, in which spatial and channel-wise attentions in the task of image captioning. They evaluated the proposed architecture on some open image captioning datasets and the results demonstrated the effectiveness of SCA-CNN. Sun et al. [30] proposed a novel and efficient video eye fixation detection model to improve the saliency detection performance by utilizing the memory mechanism and visual attention mechanism. Through hierarchical training, the proposed model achieved improved performance compared to other state-of-the-art methods. In paper [31], Xu et al. proposed and applied a novel question-guided spatial attention architecture named the Spatial Memory Network to the VQA task. They evaluated Spatial Memory Network on some

available visual question answering datasets and obtained competitive results.

As mentioned above, because the attention mechanism has the powerful ability to select discriminative features, it is very suitable for visual recognition tasks. Nevertheless, most current approaches just extract low-level and high-level features from different layers in deep neural networks. Those methods only simply integrated these features and ignored the middle-level features. As such, instinctively, the middle-level feature is a crucial factor affecting the performance for a recognition task. Therefore, these multi-scale features should be adopted for image classification. Due to the differences between the multi-scale features, spatial attention and channel-wise attention strategies are employed to assign large weights to important information, which has a greater impact on the classification results.

3. Proposed Architecture

Based on the above discussion, a novel model based on multi-level attention mechanism and multi-scale fusion is proposed in this paper. The model mainly uses the attention mechanism to process the multi-scale feature maps and designs an elaborate feature fusion strategy to learn more discriminative feature representations for obtaining better classification performance.

The overall architecture of the proposed model is shown in Figure 1. This model will be introduced detailedly in the following section.

3.1. Multiscale Features Extraction. As mentioned above, many researches focused on how to design effective network models to extract low-level and high-level feature representations, and then fusion the two types of features for current vision tasks. In the field of IWPs classification, low-level features from image in general contain specific strokes, textures, and other information which play an important role in the final classification task. On the other hand, high-level features relatively represent a large receptive field and contain the global semantic features of IWPs, which can be further used for accurate classification.

In previous research work, only low-level and high-level image features were extracted for recognition tasks. Therefore, to use low-level, middle-level, and high-level features at the same time, a multi-scale feature extraction and fusion strategy is proposed in this paper. Specifically, the pre-trained VGG16 network is utilized to extract low-level, middle-level features and high-level feature representations from the IWPs images.

The details are shown in Figure 1. Separately, the conv1-2 feature in the VGG16 network is used as the low-level feature and the conv2-2 feature is used as the middle-level feature. And the conv5-3 feature in the VGG16 network is adopted as the high-level feature. Generally speaking, features at different layers often contain semantic information with different levels of abstraction, which affects the final classification performance. A lot of research work just fuses multi-scale features directly without considering their differences. However, there are many redundant information in different scales of features, which introduces a burden of calculation and leads to a decrease in the accuracy of classification tasks.

To solve this problem, make full use of low-level, middle-level, and high-level features, some strategies should be adopted, such as attention mechanism, multi-scale feature fusion.

3.2. Spatial Attention Module. It can be seen from the analysis of the above sections that due to the different contributions of different scales of features for the classification task, the proposed model in this paper has a spatial attention module which filters irrelevant information after extracting the low-level features and the middle-level features, as shown in Figure 1.

For example, there are some textures, stroke features, and other information in the low-level and middle-level features which may promote performance, and the background noise contained in the low-level and middle-level feature maps may also disturb the classification, so these feature maps need to be filtered at the pixel level to retain the most valuable pixel area. Nevertheless, the high-level features have a large receptive field, re-weighting the spatial attention of high-level features will have a greater impact on the low-level features, so the high-level features are not processed by the spatial attention mechanism.

Based on this, the low-level and middle-level feature map extracted from Chinese IWPs are processed by the spatial attention mechanism to generate more discriminative information. The proposed model first obtains low-level and middle-level feature maps from the pre-trained VGG16 network and undergoes max pooling and 1×1 convolution operations, and then serves as the input of the respective spatial attention operation.

Assuming that a Chinese IWP image is given, the extracted low-level features are expressed as $f^{\text{low}} \in R^{W \times H \times C}$ by using pre-trained VGG16 network, where $W \times H$ represents the size of the low-level feature map, C is the channel number of the low-level feature maps.

Similar to that in [8], to increase the receptive field to obtain global information, two convolutional layers are mainly applied. The size of the convolution kernel of the first layer is $1 \times k$, and the size of the convolution kernel of the second layer is $k \times 1$, so the specific definitions are described as follows:

$$s_1 = g_2\left(g_1\left(f^{\text{low}}, W_1\right), W_2\right). \tag{1}$$

$$s_2 = g_1\left(g_2\left(f^{\text{low}}, W_3\right), W_4\right), \tag{2}$$

where g_1 and g_2 in (1) and (2) and are convolution operations of size $1 \times k$ and $k \times 1$, respectively. Then, by using the normalization processing method to encode the feature maps to $[0, 1]$, the final spatial attention feature maps are obtained.

$$\text{Sp} = F\left(f^{\text{low}}, W\right) = \sigma_1\left(s_1 + s_2\right), \tag{3}$$

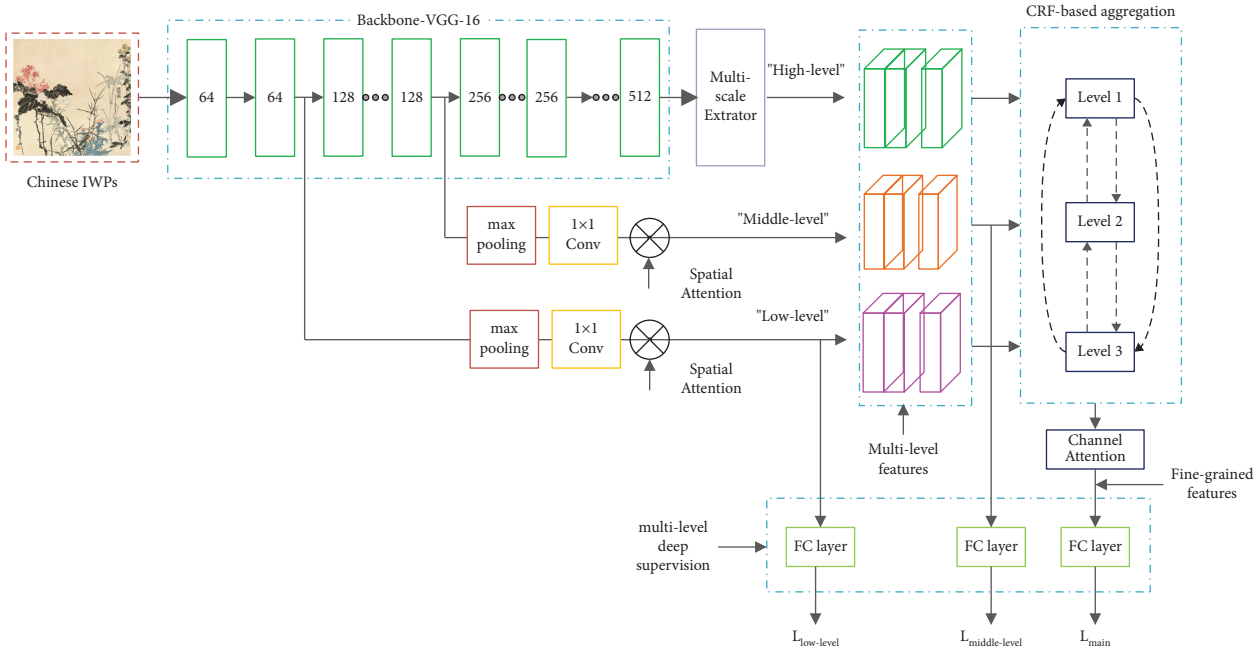

Figure 1: The proposed model for author classification of Chinese IWPs by multi-level attention and multi-scale fusion.

where $W = \{W_1, W_2, W_3, W_4\}$ in (3) is the parameters of the spatial attention module. σ_1 in (3) is a sigmoid function. The refined low-level features are obtained by the spatial attention maps re-weighting, which are defined as:

$$\widehat{f^{\text{low}}} = \text{Sp} \cdot f^{\text{low}}. \tag{4}$$

For the middle-level feature $f^{\text{mid}} \in R^{W \times H \times C}$, a similar spatial attention processing method is utilized to generate a refined middle-level feature map $\widehat{f^{\text{mid}}}$. The specific operation will not be described. Based on the above operations, the proposed model in this paper filters the low-level and middle-level feature maps to complete the spatial attention mechanism processing.

3.3. Scale Invariance Processing. Different from the low-level and middle-level features, the model extracts the conv5-3 layer features from the VGG16 network as the high-level semantic feature. Instead of directly processing the high-level features by the spatial attention mechanism, the model sends features to a scale invariance module to increase the scale-invariant nature of advanced features.

The scale invariance module proposed in this paper is inspired by the method in [33] to solve the over-fitting problem that may be caused by the high-level semantic features. To obtain a more generalized feature representation, the high-level feature maps enter the scale invariance module to undergo multiple branch stochastic affine operations.

As shown in Figure 2, the initial high-level feature f^{high} enters two branches of the scale invariance processing module as input data, respectively. The corresponding scale parameters are $(\alpha, 1-\alpha)$ and $(\beta, 1-\beta)$, where $\alpha, \beta \in [0, 1]$. These scale parameters in the above modules are uniformly distribution during the training process of the network and are randomly reset in each epoch. Specially, these parameters are set to the desired value 0.5 during the testing period.

In Branch-1 of the scale invariance processing module, DR = 1 indicates the dilated the convolution operation in which convolution kernel is 3×3, and the value 1 after DR represents the dilation rate. And other DR branches in the module represent the dilated convolution operations in which the dilation rates are 2, 3, and 4, respectively. After processed in Branch-1, high-level semantic features are weighted, and the new features $f1$ are formed through the element-wise add operation.

Similarly, the corresponding features $f2$ are formed through the processing of Branch-2, and then the new features $[f1, f2]$ are obtained by the two-branch networks through the concat operation. In order to keep the same number of channels and sizes corresponding to the processed low-level and middle-level features, the features $[f1, f2]$ are processed by a 1×1 convolution operation to obtain the features $f^{\text{high}'}$.

3.4. Multiscale Feature Fusion Based on Conditional Random Field. From the above discussion, the different scales of feature maps have different representational abilities and are highly complementary. Therefore, multi-scale feature maps are extracted through pre-trained VGG16 in the proposed model. Nevertheless, how to make the best of these multi-scale features extracted from different receptive fields to improve the recognition performance is a very worthy research topic. At present, weighted average or cascade processing on multi-scale features are commonly used feature fusion strategies, but these strategies are too simple to fuse features well.

Conditional Random Field (CRF) [34] has been widely used in natural language processing. In recent years, CRF was often used in the field of computer vision as a message passing mechanism to refine the features of convolutional neural networks.

In this paper, CRF is proposed to perform fusion operation to achieve feature learning and obtain richer representation information. The multi-scale feature fusion module based on CRF in our model refines the extracted features of different scales with each other. Specifically, the module dynamically transfers complementary information from different scale (low-level, middle-level and high-level) feature information for fusion operations to enhance the representation capabilities of specific scales.

Assuming that the given multi-scale feature maps are expressed as: $\mathrm{Fe} = \{f^{\mathrm{low}}, f^{\mathrm{middle}}, f^{\mathrm{high}}\}$, and the optimized multi-scale feature maps based on CRF fusion processing are expressed as: $\widehat{\mathrm{Fe}} = \left\{\widehat{f}^{\mathrm{low}}, \widehat{f}^{\mathrm{middle}}, \widehat{f}^{\mathrm{high}}\right\}$, where the optimized feature maps $\widehat{f}^{i}$ are obtained from the initial feature maps f^{i} by using CRF model fusion processing.

Specifically, the conditional probability distribution of the original multi-scale feature maps set and the optimized multi-scale feature maps set is defined as:

$$P(\widehat{\mathrm{Fe}}|\mathrm{Fe}, \Theta) = \frac{\exp\left\{\mathrm{En}(\widehat{\mathrm{Fe}}, \mathrm{Fe}, \Theta)\right\}}{\int_{\widehat{\mathrm{Fe}}} \exp\left\{\mathrm{En}(\widehat{\mathrm{Fe}}, \mathrm{Fe}, \Theta)\right\} d\widehat{\mathrm{Fe}}}. \tag{5}$$

In (5), Θ is the parameter set of formula and the denominator is the normalized partition function. The corresponding energy function $\mathrm{En}(\widehat{\mathrm{Fe}}, \mathrm{Fe}, \Theta)$ is defined as the sum of two potential functions:

$$\mathrm{En}(\widehat{\mathrm{Fe}}, \mathrm{Fe}, \Theta) = \sum_{i,j} \Psi_1\left(\widehat{f}_i, \widehat{f}_j\right) + \sum_{i} \Psi_2\left(\widehat{f}_i, f_i\right). \tag{6}$$

In (6), the potential function $\Psi_2(\widehat{f}_i, f_i)$ represents the similarity between the original feature maps and the optimized feature maps. Here, the *L2* distance is utilized to define the similarity, which is specifically defined as:

$$\Psi_2\left(\widehat{f}_i, f_i\right) = -\frac{1}{2}\left\|\widehat{f}_i - f_i\right\|^2. \tag{7}$$

In addition, in (6), the potential function $\Psi_1(\widehat{f}_i, \widehat{f}_j)$ represents the correlation between two optimized feature information, which is specifically defined as:

$$\Psi_1\left(\widehat{f}_i, \widehat{f}_j\right) = \left(\widehat{f}_i\right)^T \omega_j^i \widehat{f}_j. \tag{8}$$

In (8), ω_j^i represents the parameter used to calculate the correlation between two optimized feature maps $(\widehat{f}_i, \widehat{f}_j)$. The refined feature maps $\widehat{f}_i$ are fused with the original feature maps f_i and the complementary information passed from other optimized feature maps, and finally are formed through multiple iterations.

Through the fusion processing module based on CRF, the multi-scale feature representations (low-level, middle-level, and high-level) are processed into more optimized and more robust feature information.

3.5. Channel Attention Module. In the proposed model, the feature information optimized by the multi-scale feature fusion module based on CRF will be sent into the channel attention module. The channel attention mechanism is used to recorrect all feature channels, and the rich-information feature channels are strengthened, while the low-information feature channels are suppressed. Therefore, it can be seen from Figure 1, a channel attention module is added to the model proposed in this work.

Assuming that the optimized multi-scale feature maps are expressed as $f^{\mathrm{mul}} \in R^{W \times H \times C}$, where $W \times H$ represents the size of the multi-scale feature maps, and C is the number of channels.

Firstly, the feature maps f^{mul} are expanded into a channel set $f^{\mathrm{mul}} = [f_1^{\mathrm{mul}}, f_2^{mul}, f_3^{\mathrm{mul}}, \ldots, f_C^{\mathrm{mul}}]$, where f_j^{mul} is the j-th channel in the multi-scale features. Then, each $f_j^{\mathrm{mul}} \in R^{W \times H}$ is processed to generate a channel feature vector $x^{\mathrm{mul}} \in R^C$ by average pooling operation. Subsequently, two fully connected layers are added to fully capture the correlation of channels. To reduce the complexity of the model, the Relu function is introduced between the two fully connected layers. Afterwards, similar to the spatial attention operation, by using the normalization method to encode the feature maps to $[0, 1]$, the final channel attention weighted feature maps are obtained.

$$g = \delta\left(\mathrm{fc}_1\left(x^{\mathrm{mul}}, W_1\right)\right). \tag{9}$$

In (9), fc_1 represents the first fully connected layer, W_1 represents the parameters of the first fully connected layer, and δ represents the Relu function.

$$\mathrm{Ch} = \sigma_2\left(\mathrm{fc}_2\left(g, W_2\right)\right). \tag{10}$$

In (10), fc_2 represents the second fully connected layer, W_2 is the parameters, and σ_2 represents the sigmoid function. After processing by the above channel attention mechanism, the reweighted multiscale feature maps f^{main} are obtained, which are specifically defined as:

$$f^{\mathrm{main}} = \mathrm{Ch} \cdot f^{\mathrm{mul}}. \tag{11}$$

3.6. Multilevel Deep Supervision. To supervise the model to achieve better performance, unlike the existing work that only adds supervision at the high-level feature, a multi-level deep supervision strategy is proposed to ensure the learning efficiency of the network.

As shown in Figure 1, three types of optimized feature representations: $\widehat{f^{\mathrm{low}}}$, $\widehat{f^{\mathrm{mid}}}$, f^{main} are obtained in the proposed model. Fully considering the impact of the classification loss of the above three types of optimized feature representations, a joint loss function is meticulous designed in this work. Specifically, the ℓ_{low}, ℓ_{middle}, and ℓ_{main} represent the loss corresponding to the features $\widehat{f^{\mathrm{low}}}$, $\widehat{f^{\mathrm{mid}}}$, f^{main}, respectively.

The joint loss function of this model is specifically defined as follows:

$$\ell_{\mathrm{all}} = 0.1\ell_{\mathrm{low}} + 0.1\ell_{\mathrm{middle}} + 0.8\ell_{\mathrm{main}}. \tag{12}$$

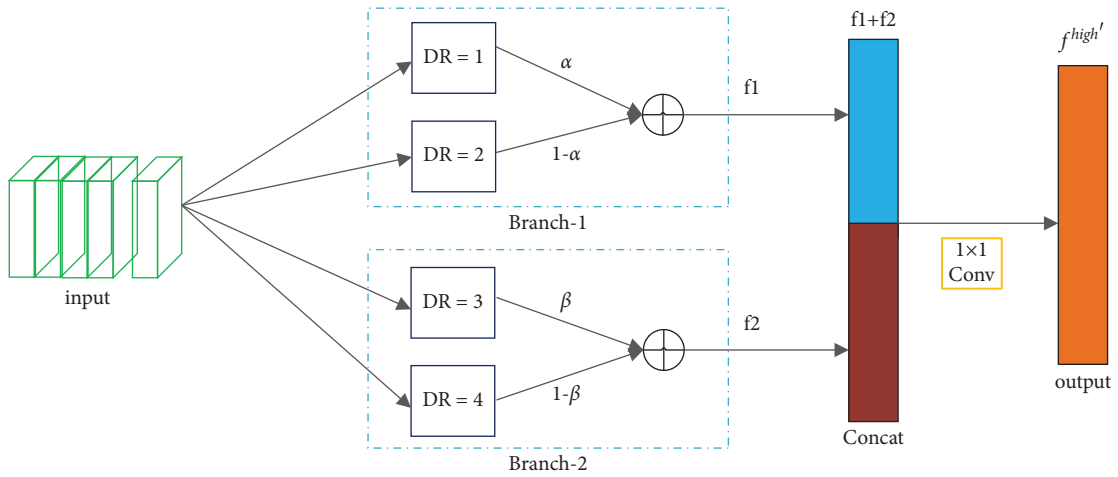

Figure 2: The scale invariance processing module.

It can be seen from (12) that the multilevel deep supervision of the model uses the three-level classification joint loss function to optimize the model. In particular, because the three-level optimized features have different contributions to classification, the three-level losses are assigned different weights in the model.

4. Experimental

4.1. DataSet. To verify the proposed author classification model of Chinese IWPs based on multi-level attention and multiscale feature fusion, a Chinese IWPs dataset is collected and established from the Internet in this work.

In order to ensure rationality in this work, the following requirements should be met in the establishment of the Chinese IWPs dataset: Firstly, in order to achieve better results in the training process of the deep learning model proposed in this experiment, the dataset established in this experiment should have a larger number of Chinese IWPs. Secondly, in order to verify the validity of the author classification of the model, the dataset should contain a sufficient number of artists, and each artist should have a certain number of Chinese IWPs. Thirdly, in order to verify the rationality and accuracy of the experiment, the diversity of artists' ages and styles, as well as the diversity of each artist's IWPs, should be fully taken into account when selecting and collecting artists' IWPs.

The dataset contains 3040 Chinese IWPs from ten artists, including: Cao Jianlou, Fan Zeng, Li Xiaoming, Lu Yanshao, Pan Tianshou, Qi Baishi, Wu Changshuo, Xu Beihong, Zeng Xiaolian and Zhu Da. The dataset is shown in Table 1.

In the experimental setting of this paper, 3/5 of the dataset established above is used as the training set, 1/5 as the verification set, and the remaining 1/5 as the test set. In addition, for the rationality of the experiment, the dataset should be divided into three sets at random, and to ensure that there is no duplication of the paintings of each artist. Figure 3 shows some examples of the dataset used in the experiments.

To facilitate the benchmark test of the model, three different classification scenarios are designed for the test of the above dataset:

Case 1: classify the paintings of three artists: Qi Baishi, Wu Changshuo and Cao Jianlou;

Case 2: classify the paintings of five artists: Fan Zeng, Pan Tianshou, Wu Changshuo, Qi Baishi, and Xu Beihong;

Case 3: classify the paintings of six artists: Cao Jianlou, Li Xiaoming, Lu Yanshao, Zhu Da, Zeng Xiaolian, and Xu Beihong.

4.2. Experiment Results and Discussion. To prove the effectiveness of the model proposed in this work, a series of experiments will be conducted on the author classification dataset introduced above. We will compare and analyze the results of the experiments on the three cases in the dataset.

4.2.1. Performance Analysis of Different Classifiers. For classification tasks in different research fields, selecting an appropriate classifier is one of the most important factors. Therefore, to find the most effective classifier, the classifiers tested in this experiment are carried out, in which includes: K-nearest neighbor (KNN), logistic regression (LR), random forest (RF), and support vector machine (SVM).

As for the SVM classifier, to find the appropriate kernel function, the experiment also tested many types of kernel functions. In the end, the RBF kernel function is adopted in the experiment. In addition, 10-fold cross-validation is performed to find the hyperparameter C' in the range of [0.001, 20], and to find the optimal parameter δ' in the range of [0.001, 10]. For the RF classifier, the number of trees is selected in the range of 50 to 900 with the step size being set to 50, and the depth of the tree is searched from 2 to 18.

In this experiment, the Case-1 category of the author classification dataset is selected for verification. The specific experimental results are shown in Table 2.

Table 1: The dataset of Chinese IWPs by ten artists.

Number	Artists	Age	Number of paintings
1	Cao Jianlou	1913–2005	397
2	Fan Zeng	1938-	113
3	Li Xiaoming	1972-	114
4	Lu Yanshao	1909–1993	295
5	Pan Tianshou	1897–1971	103
6	Qi Baishi	1864–1957	1129
7	Wu Changshuo	1844–1927	168
8	Xu Beihongu	1895–1953	240
9	Zeng Xiaolian	1939-	300
10	Zhu Da	1626–1705	181

From the comparison of the results of the four different classifiers in Table 2, the SVM classifier achieves the best performance on Case-1 category of the classification task than other classifiers, with an average classification accuracy rate of 96.2%. The classification accuracy achieved by SVM is 11.5% higher than that of the KNN classifier (84.7%), which is the largest gap between the classification results of the four classifiers. The results show that the SVM classifier and the RF classifier have achieved the closest results on the author classification dataset, which also reflects the excellent classification effect of the RF classifier (90.5%). Therefore, due to the optimal classification performance on the dataset, the SVM is utilized as the final classifier for the experiments in this work. In addition, it can also be observed from Table 2 that the SVM classifier achieves the highest average accuracy rate in the Case-1 category, and the lowest accuracy rate is 94.6% for the Wu Changshuo's painting category. The main reason is that Wu Changshuo's painting category is the smallest number in this dataset (168 frames), which leads to the worst classification result. Compared with Wu Changshuo's painting category, the classification results of Qi Baishi and Cao Jianlou's painting categories are better. Furthermore, it can be seen from the results that the similar phenomenon happens when other classifiers are used.

4.2.2. Performance Analysis of Different Methods. To benchmark the model proposed in this paper, a series of experiments compared with the existing methods are carried out on the Case-2 category of the dataset. In the experiments, these representative methods include: Sheng and Jiang [12], Sun et al. [13], and Jiang et al. [14]. The specific experimental results are shown in Table 3.

As can be seen from Table 3, the proposed method based on multi-level attention and multi-scale feature fusion in this paper has achieved the performance with 94.8% accuracy, which is significantly better than those three methods above mentioned. Specifically, the classification accuracy of the methods by Sheng and Jiang [12], Sun et al. [13], and Jiang et al. [14] reached 88.1%, 82.0% and 87.4%, respectively.

In the method of Sun et al. [13], CNN network model was proposed to extract features from IWPs, and Sparse Group Lasso was used to perform the final classification task. However, they only considered selecting the most representative ten local subgraphs from the images, while ignoring the global characteristics of the IWPs images. And the proposed method in our work extracts multi-scale feature maps and makes full use of these features for the classification task. In the method of Jiang et al. [14], although the trained deep neural network was also used to obtain features from images, the CNN model which they used was too shallow to extract the discriminating information. However, the model proposed in our work uses pre-trained CNN to extract multi-scale feature maps and makes full use of attention mechanism and CRF to fuse these feature maps.

Analyzing the reasons, the model proposed is a well-designed network architecture, which is very effective in extracting features and feature fusion for obtaining good classification results. The experimental results in Table 3 also verify the superiority of the model proposed on the author classification dataset.

4.2.3. Ablation Study. Specifically, to further study the effectiveness of each module in the model proposed in our work, we also conduct a series of ablation experiments. Table 4 shows the accuracy comparison of the classification results of the ablation experiment on the Case-3 category.

In Table 4, the method named ours-mid means that only the low-level and high-level feature maps are extracted from the model proposed, and the middle-level feature extraction module is removed from the model (see Figure 1). In detail, only the low-level and high-level feature maps are used as the multi-level feature processing data. The method named ours-spatial in Table 4 means that the low-level and middle-level features are not processed by the spatial attention mechanism and are directly sent to the multi-level feature fusion module based on CRF for processing. The third method, named ours-channel means that the channel attention processing module is removed from our model. The fourth method, named ours-fusion means that the multi-scale feature fusion module based on CRF is removed from the model in our work.

In the ours-mid method, an average classification accuracy of 86.5% is achieved, which is 2.7 percentage points lower than the best classification result (89.2%) of the model in this work. This experimental result also verifies the importance of the middle-level feature maps above mentioned for the classification task.

The ours-spatial method obtains an average classification accuracy of 86.2%, which is the worst classification result in the experiment. This result is 3 percentage points lower than the method based on the multi-level attention and multi-scale fusion network proposed (89.2%). This experimental result proves that the spatial attention module can well remove irrelevant and redundant information from the low-level and middle-level feature maps and plays a very essential role in the proposed model.

Similarly, the ours-channel method achieves a classification result of 87.6%, which is 1.6 percentage points lower than the best classification result (89.2%). The experimental result verifies the role of the channel attention processing module in the proposed model, which could also slightly improve the classification performance. This result also shows that the feature maps processed by the multi-scale

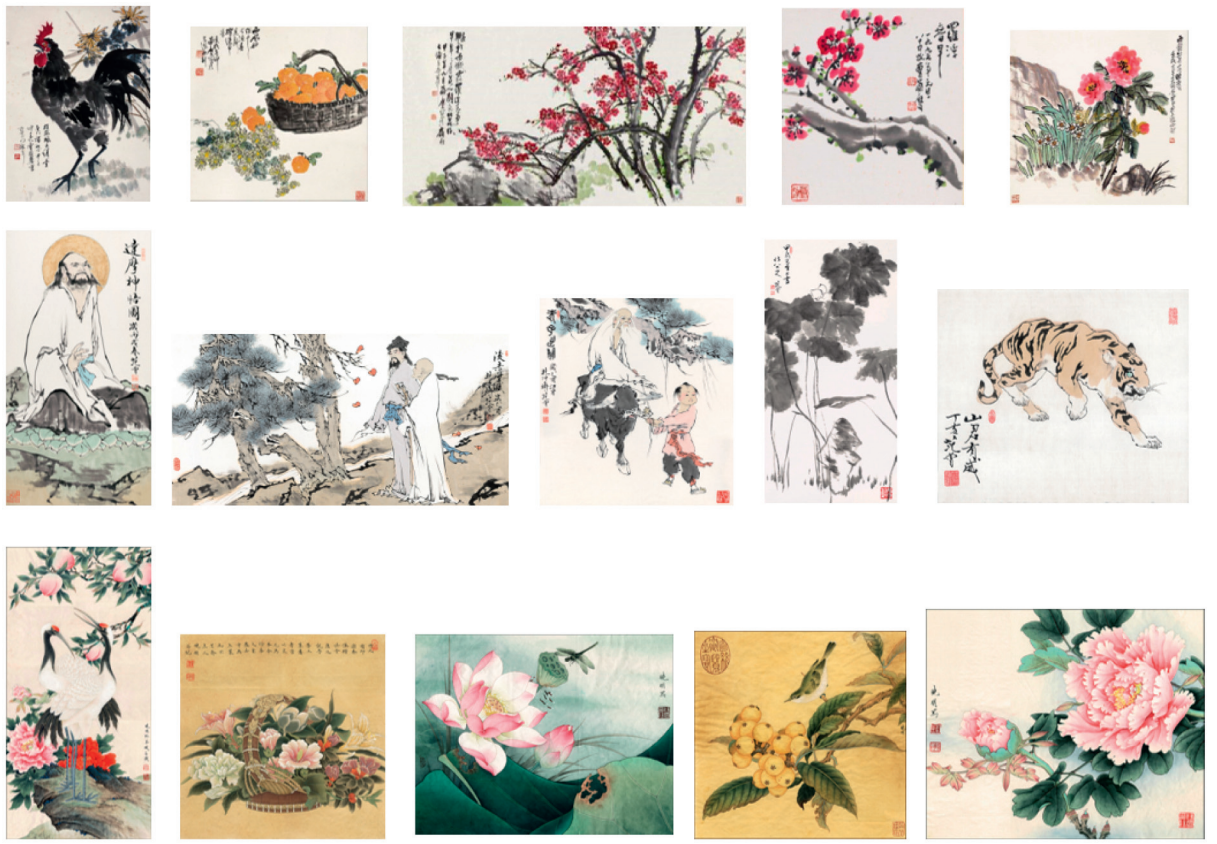

Figure 3: Some samples of the dataset in the experiment. The three rows show some samples of Cao Jianlou's, Fan Zeng's, and Li Xiaoming's paintings, respectively.

Table 2: Comparison of classification results of different classifiers on the Case-1 category of the dataset.

Artists/classifiers	KNN	LR	RF	SVM
Qi Baishi	86.1	89.3	92.7	96.7
Wu Changshuo	82.3	80.9	85.8	94.6
Cao Jianlou	85.6	91.5	93.1	97.4
Average	84.7	87.2	90.5	96.2

Table 3: Comparison of classification results of different methods on the Case-2 category of the dataset.

Artists/methods	Sheng [12]	Sun [13]	Jiang [14]	Proposed
Fan Zeng	88.5	82.7	85.4	94.3
Pan Tianshou	92.3	84.1	89.3	93.6
Wu Changshuo	83.8	76.3	81.6	94.7
Qi Baishi	85.2	81.5	90.2	96.9
Xu Beihong	90.9	85.4	90.7	94.4
Average	88.1	82.0	87.4	94.8

Table 4: Accuracy comparison of the ablation experiment on the Case-3 category of the dataset.

Artists/methods	Ours-mid	Ours-spatial	Ours-channel	Ours-fusion	Proposed
Cao Jianlou	85.2	83.4	87.3	85.2	88.3
Li Xiaoming	84.5	85.1	84.8	82.4	86.7
Lu Yanshao	87.1	85.6	86.2	85.3	88.4
Zhu Da	86.0	85.7	87.5	87.6	88.2
Zeng Xiaolian	86.8	89.3	89.1	88.7	92.4
Xu Beihong	89.2	88.2	90.8	89.1	91.2
Average	86.5	86.2	87.6	86.4	89.2

feature fusion network based on CRF have better representation ability.

The accuracy of the ours-fusion method in Table 4 on the Case-3 category of the dataset reaches 86.4%. Compared with the classification results obtained by the model proposed in this paper, it shows that if the fusion network module is removed, the classification accuracy will be reduced by 2.8%. This experimental result shows that the fusion module designed in the model has also played a better role in improving the final classification performance.

5. Conclusions

To solve the problem of extracting low-level and high-level features from different layers in deep CNN for classification and using simple feature fusion strategies in the previous researches, a novel model based on multi-level attention and multi-scale feature fusion is proposed in this paper. In this model, the pre-trained VGG16 network is used to extract low-level, middle-level and high-level feature representations from Chinese IWPs images. The spatial attention mechanism is adopted to filter out irrelevant information from the low-level and middle-level features, and a scale invariant module is added to increase the scale-invariant properties of high-level features in the model. Then the three optimized features are fused together based on the CRF mechanism. Subsequently, the channel attention mechanism is used to obtain the features with more discriminating ability. The model is trained by designing a multi-level deep supervision module, and a large number of experimental results show that the proposed model can achieve more competitive performance.

Data Availability

The raw/processed data required to reproduce these findings cannot be shared at this time as the data also forms part of an ongoing study.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this article.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant 61901063, Humanities and Social Science Fund of Ministry of Education of China under Grant 19YJCZH120, Science and Technology Plan Project of Changzhou under Grant CE20205042, and Qinglan Project of Jiangsu Province (2020).

References

[1] R. Girshick, J. Donahue, Trevor Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 580–587, IEEE, Columbus, OH, USA, 23 June 2014.

[2] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, IEEE, Las Vegas, NV, USA, 27 July 2016.

[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proceedings of the Advances in neural information processing systems*, pp. 1097–1105, DBLP, Lake Tahoe, Nevada, US, 3 December 2012.

[4] C. Szegedy, W. Liu, Y. Jia et al., "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1–9, IEEE, Boston, MA, USA, 7 June 2015.

[5] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: towards real-time object detection with region proposal networks," in *Proceedings of the Advances in neural information processing systems*, pp. 91–99, MIT Press, Montreal Canada, 7 December 2015.

[6] J. Long, E. Shelhamer, and Trevor Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3431–3440, Boston, MA, USA, 7 June 2015.

[7] J. Fu, H. Zheng, and T. Mei, "Look closer to see better: recurrent attention convolutional neural network for fine-grained image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4438–4446, IEEE, Honolulu, HI, USA, 21 July 2017.

[8] T. Zhao and X. Wu, "Pyramid feature attention network for saliency detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3085–3094, Long Beach, CA, USA, 15 June 2019.

[9] S. Gao, M.-M. Cheng, K. Zhao, X.-Yu Zhang, M.-H. Yang, and P. H. S. Torr, "Res2net: a new multi-scale backbone architecture," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, 2019.

[10] M. Jian, Y. Yin, J. Dong, and W. Zhang, "Comprehensive assessment of non-uniform illumination for 3d heightmap reconstruction in outdoor environments," *Computers in Industry*, vol. 99, pp. 110–118, 2018.

[11] M. Jian, J. Dong, M. Gong et al., "Learning the traditional art of Chinese calligraphy via three-dimensional reconstruction and assessment," *IEEE Transactions on Multimedia*, vol. 22, no. 4, pp. 970–979, 2019.

[12] J. Sheng and J. Jiang, "Recognition of Chinese artists via windowed and entropy balanced fusion in classification of their authored ink and wash paintings (iwps)," *Pattern Recognition*, vol. 47, no. 2, pp. 612–622, 2014.

[13] M. Sun, D. Zhang, J. Ren, Z. Wang, and S. Jesse, "Brushstroke based sparse hybrid convolutional neural networks for author classification of Chinese ink-wash paintings," in *Proceedings of the 2015 IEEE International Conference on Image Processing (ICIP)*, pp. 626–630, IEEE, Quebec City, QC, Canada, 27 September 2015.

[14] W. Jiang, Z. Wang, J. S. Jin, Y. Han, and M. Sun, "DCT-CNN-based classification method for the Gongbi and Xieyi techniques of Chinese ink-wash paintings," *Neurocomputing*, vol. 330, pp. 280–286, 2019.

[15] Q. Hou, M.-M. Cheng, X. Hu, B. Ali, Z. Tu, and P. H. S. Torr, "Deeply supervised salient object detection with short connections," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3203–3212, IEEE, Honolulu, HI, USA, 21 July 2017.

[16] Y. Tang, X. Wu, and W. Bu, "Deeply-supervised recurrent convolutional neural network for saliency detection," in *Proceedings of the 24th ACM international conference on Multimedia*, pp. 397–401, ACM, Amsterdam The Netherlands, 16 October 2016.

[17] P. Zhang, D. Wang, H. Lu, H. Wang, and R. Xiang, "Amulet: aggregating multi-level convolutional features for salient object detection," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 202–211, IEEE, Venice, Italy, 22 October 2017.

[18] G. Sun, X. Zhang, X. Jia et al., "Deep fusion of localized spectral features and multi-scale spatial features for effective classification of hyperspectral images," *International Journal of Applied Earth Observation and Geoinformation*, vol. 91, Article ID 102157, 2020.

[19] J. Ren, S. He, H. Zhao et al., "Effective extraction of ventricles and myocardium objects from cardiac magnetic resonance images with a multi-task learning u-net," *Pattern Recognition Letters*, 2021.

[20] Z. Fang, J. Ren, C. MacLellan et al., "A novel multi-stage residual feature fusion network for detection of covid-19 in chest x-ray images," *IEEE Transactions on Molecular, Biological and Multi-Scale Communications*, 2021.

[21] X. Zhang, X. Zhou, M. Lin, and J. Sun, "Shufflenet: an extremely efficient convolutional neural network for mobile devices," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6848–6856, IEEE, Salt Lake City, UT, USA, 18 June 2018.

[22] L. Chen, H. Zhang, J. Xiao et al., "Sca-cnn: spatial and channel-wise attention in convolutional networks for image captioning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5659–5667, IEEE, Honolulu, HI, USA, 21 July 2017.

[23] G. Zhang, M. Kan, S. Shan, and X. Chen, "Generative adversarial network with spatial attention for face attribute editing," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 417–432, Munich, Germany, 8 September 2018.

[24] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, https://arxiv.org/abs/1409.1556.

[25] Sergey Ioffe and C. Szegedy, "Batch normalization: accelerating deep network training by reducing internal covariate shift," in *Proceedings of the International conference on machine learning*, pp. 448–456, PMLR, Lille, France, 6 July 2015.

[26] Li Yao, A. Torabi, K. Cho et al., "Describing videos by exploiting temporal structure," in *Proceedings of the IEEE international conference on computer vision*, pp. 4507–4515, IEEE, 7 December 2015.

[27] K. Xu, Ba Jimmy, K. Ryan et al., "Show, attend and tell: neural image caption generation with visual attention," in *Proceedings of the International conference on machine learning*, pp. 2048–2057, PMLR, LILLE, France, 6 July 2015.

[28] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 8, pp. 2011–2023, 2020.

[29] S. Woo, J. Park, J.-Y. Lee, and K. So, "Cbam: convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, pp. 3–19, Munich, Germany, 8 September 2018.

[30] M. Sun, Z. Zhou, Q. Hu, Z. Wang, and J. Jiang, "Sg-fcn: a motion and memory-based deep learning model for video saliency detection," *IEEE Transactions on Cybernetics*, vol. 49, no. 8, pp. 2900–2911, 2018.

[31] H. Xu and K. Saenko, "Ask, attend and answer: exploring question-guided spatial attention for visual question answering," in *Proceedings of the European Conference on Computer Vision*, pp. 451–466, Springer, Amsterdam, The Netherlands, 8 October 2016.

[32] Z. Yang, X. He, J. Gao, Li Deng, and Alex Smola, "Stacked attention networks for image question answering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 21–29, IEEE, Las Vegas, NV, USA, 27 June 2016.

[33] X. Gastaldi, "Shake-shake regularization," 2017, https://arxiv.org/abs/1705.07485.

[34] X. He, R. S. Zemel, and M. A. Carreira-Perpinán, "Multiscale conditional random fields for image labeling," in *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004*, vol. 2, p. II, IEEE, Washington, DC, USA, 27 June 2004.