Recognizing the Style of Visual Arts via Adaptive Cross-layer Correlation

Liyi Chen, Jufeng Yang
College of Computer Science, Nankai University, Tianjin, China
chenliyi1995@126.com,yangjufeng@nankai.edu.cn

ABSTRACT

Visual arts consist of various art forms, *e.g.*, painting, sculpture, architecture, *etc.*, which not only enrich our lives but also involve works related to aesthetics, history, and culture. Different eras have different artistic appeals, and the art also has characteristics of each era in terms of expression and spiritual pursuit, in which style is an important attribute to describe visual arts. In order to recognize the style of visual arts more effectively, we present an end-to-end trainable architecture to learn a deep style representation. The main component of our architecture, adaptive cross-layer correlation, is inspired by the Gram matrix based correlation calculation. Our proposed method can adaptively weight features in different spatial locations based on their intrinsic similarity. Extensive experiments on three datasets demonstrate the superiority of the proposed method over several state-of-the-art methods.

KEYWORDS

style classfication; visual arts; deep feature

ACM Reference Format:
Liyi Chen, Jufeng Yang. 2019. Recognizing the Style of Visual Arts via Adaptive Cross-layer Correlation. In *Proceedings of the 27th ACM International Conference on Multimedia (MM '19), October 21–25, 2019, Nice, France.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3343031.3350977

1 INTRODUCTION

As a cluster of art practices, visual arts focus on visual effects, involving various art categories, such as painting, sculpture and decorative art, *etc.* The style of visual arts can reflect strong individual preference as well as unique views of art for some particular subject groups. Recognizing the style of images is the basis of visual art understanding and appreciation, and it includes painting classification, architecture classification, and other abstract tasks of visual arts. In addition, the visual style is closely related to all around our daily life, which can affect the emotion and reflect our aesthetics. For example, we can see various styles of clothing as well as architecture in the street, and sometimes we slow down just for fancy clothes or an attractive building. This has motivated many efforts to recognize the style of images towards a better understanding of visual arts and their high-level appreciations. Although several attempts have been made [18, 20, 22, 42, 43], analyzing visual arts still remains challenging as multiple visual areas of the human brain are involved in this process [14, 41]. Moreover, visual arts of a specific form, *e.g.*, painting, also present high intra-class variations as different artists have different personal habits, experience, *etc.* Therefore, classifying the artist can also be considered as a kind of style classification, *i.e.*, the personal style, which is different from the usual art style (*e.g.*, Renaissance). At the same time, some styles in the same period have a certain degree of similarities on account of a similar background and mutual influence. Furthermore, the emergence of a new style is often an evolution of existing styles. Hence, similar lines, shapes, and themes may appear in multiple styles, which undoubtedly increases the difficulty of classification further.

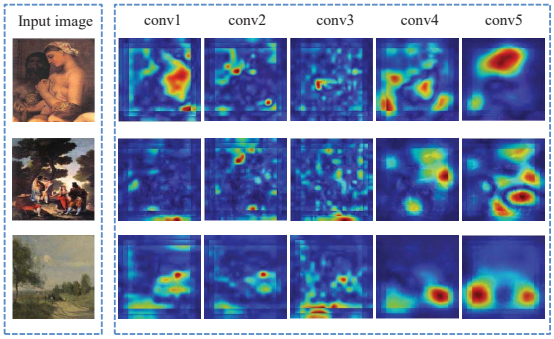

Figure 1: The class activation maps of images using different layers of VGGNet. The activation areas of different layers are different, and the activation areas are more concentrated in higher layers.

Prior studies have investigated various handcrafted features such as GIST [30], LBP [29], SIFT [27], *etc.* for style classification [21, 22]. More recently, features extracted by the convolutional neural network (CNN) have also seen heavy use [1, 31, 32] with better performance. And some researchers [1, 32] consider both holistic and region features to recognize the style of an image, which can encode multi-scale convolutional features and improve the classification accuracy.

Although CNN methods make a great breakthrough on the accuracy, there is still a question that puzzles many researchers: what is the style, and which of the following features can be used to describe the style of an image: color, line, shape, or other things. For the common fine-grained classification tasks (such as food, flowers, *etc.*), researchers usually focus on the object in images. For example, Luan *et al.* [28] recognize objects using Gabor filters on

MM '19, October 21–25, 2019, Nice, France

ACM ISBN 978-1-4503-6889-6/19/10…$15.00
https://doi.org/10.1145/3343031.3350977

each convolutional layer to enhance the resistance of the orientation and scale changes. Different from these tasks, the style of visual art images usually involves the whole picture rather than being limited to objects, especially in art paintings. Moreover, the boundary between various styles is often not clear whether in computer vision or in the field of art. In order to describe style intuitively, it is important to define what is the style. Some studies about artistic style transfer [9, 13, 47] have also demonstrated the feasibility and effectiveness of exploiting texture information for the purpose of artistic style transfer, which means the texture is an important factor to describe the style of visual arts images. This is also in line with our observation. Considering concrete examples in Fig. 1, we show the class activation maps using different layers of VGGNet [40]. The activation areas of different layers are different, and they are more concentrated in higher layers. We further observe that the activation areas of portraits generally focus on the person, whereas for other images, such as landscapes, religious story paintings, *etc*., the activation areas are usually fragmented and tend to respond to the texture information of the painting. Existing deep learning methods can not well describe the texture information of images, especially in high layers the semantic information is abundant while the texture information is decreasing.

To better utilize texture information, an artistic style transfer research [13] proposes a style representation based on the correlation between responses of different filters (feature maps). The correlations of different filter responses are given by the Gram matrix, the inner product of the corresponding feature maps, which can effectively extract the texture information. Chu *et al*. [7] further investigate the effectiveness of the Gram matrix on top of CNN features. However, it limits classification performance since they use fixed several calculations about correlation features and non-end-to-end training methods.

In order to leverage texture information in a more principled manner, we propose a novel adaptive cross-layer CNN model by calculating the inner product of convolution features, in which cross-layer features provide complementary information. It extracts texture information to capture both higher-level abstract information and lower-level detail information in whole images rather than just objects in fine-grained classification tasks, *e.g*., [28]. Our proposed method weight feature descriptors in each spatial location in an adaptive manner towards better discrimination. Experimental results demonstrate the effectiveness of our proposed method in three representative datasets including painting, architecture, and clothing, and our method outperforms state-of-the-art works on all these datasets.

The main contributions of this paper are as follows. First, we calculate inner products between feature responses to extract texture information useful for style classification. Since different layers of the network have different feature responses for the same image, we combine responses from different layers to better cover style information. Second, we present an adaptive cross-layer model, an end-to-end framework, to calculate inner products and extract distinguishable style representations. More specifically, to provide varying importance for different locations, we propose a weight matrix based on Euclidean distance between feature maps, which are automatically updated during the training process.

2 RELATED WORK

2.1 Style Classification

Generally speaking, different characteristics of images, such as color, line, composition, intensity, *etc*., are discriminative for style recognition. This motivates several computer vision based approaches [1, 3, 11, 31, 38, 53].

Some common visual descriptors are used to describe the style [21, 22]. For some images about styles, the background color and idiomatic color tend to be consistent, making the color histogram becomes a good choice for representing the distribution of colors in an image [21]. SIFT descriptors are also proved to be effective for style analysis in [27] due to its robustness to complex deformation and illumination changes. As a holistic representation mainly applied to scene recognition, the GIST features can extract the spatial structure of an image and also shown to be effective in style recognition [30]. The local binary patterns (LBP) is often used to express texture information of images [29]. Guo *et al*. [16] present completed modeling of the LBP operator which captures both image local gray levels and the sign and magnitude features of local differences, which is used for texture classification. LBP operator is also widely used in style classification. For the painting artist and style classification tasks, Khan *et al*. [22] study the performance of multiple local features and global features (*e.g*., LBP, GIST, PHOG [4], SSIM [37], bag-of-words framework, *etc*.), and verify that combining multiple features improve the classification accuracy effectively. In addition, Xu *et al*. [48] extract the morphological features of basic architectural components using deformable part-based models and adopt the multinomial latent logistic regression for the architectural style classification.

Motivated by the success of deep learning for visual tasks [17, 24, 50–52], deep feature representations [2, 8, 10, 19, 44] have also been extensively investigated under the umbrella of visual style analysis. Karayev *et al*. [21] use both visual descriptors features and deep learning features to evaluate single-feature performance and second-stage fusion of multiple features. They focus on the painting style as well as photographic style and define various different types of image styles based on composition styles, photographic techniques, moods, genres, and types of scenes. Lecoutre *et al*. [25] identify the artistic style of the painting using the deep residual neural. Moreover, the multi-scale deep features encode rich feature hierarchies on both global scene layout and part-based details [1, 32]. Fusing these features of various scales can obtain abundant information to represent the image style. Peng *et al*. [32] propose a multi-scale CNN to exponentially generate more training examples for better feature extraction depending on the assumption of the label-inheritable property. And the cross-layer features extracted from multiple layers of CNN also achieve good experimental results on style recognition tasks [31]. Anwer *et al*. [1] also use multi-scale features to extract the regions of interest (ROIs) of images automatically. It describes both holistic and ROIs of images based on multi-scale dense convolutional features and uses Fisher vectors to encode them separately for the painting categorization. In addition, inspired by the context of art history, Yang *et al*. [49] propose a multi-task learning framework based on the label distribution learning for painting style classification. Considering the

historical context information of each art style, including the birthplace, the origin time, and the art movement, they generate a set of label distributions for each style of the art painting to measure the connection between styles. In the training stage, both the style classification task and the label distribution task are optimized, and CNN learns the historical background information related to art and help improve the classification result of the art painting style. Moreover, inspired by the deep learning, Gultepe *et al.* [15] present an unsupervised feature learning method using K-means (UFLK). They extract the style features of paintings and use the Support Vector Machine algorithm to divide the artistic style. Also using unsupervised learning, Wynen *et al.* [46] introduce an approximate technique of sparse coding with the geometric interpretation based on the archetypal analysis for classification. Based on the CNN method as well as three designed visual features, Sethi *et al.* [36] propose an open access online platform in order to classify and analyze the art styles of paintings. However, these above methods still have not answered the question that which visual features could represent the style.

2.2 Texture Representation for Style

In the research of painting style transfer, a deep correlation feature has been proposed as a novel style representation that calculates the correlations between various filter responses, in order to transfer a photo into the specified painting style while preserving its contents [13]. Texture information is considered to be closely related to the style of an image [12, 13, 47]. Motivated by this, deep correlation features, encoded by the Gram matrix, have been proposed as a novel style representation [7] for image style classification. Comprehensive evaluations of different convolutional features are conducted, and a fusion of various correlation strategies is further proposed. They survey the performance distinction of deep correlation features from different layers using the Gram matrix and verify the effectiveness of the Gram matrix in visual arts classification tasks. The remaining six correlation features of these feature maps also used to calculate the style vector, including Spearman correlation, Pearson correlation, covariance, Chebyshev distance, Euclidean distance, and cosine similarity. In addition, they present the style representations that combine many different correlation features to represent the texture information. This work also surveys the performance variations of different correlations and demonstrates the classification performance of combining multiple correlations can outperform using only one correlation. However, this work only uses the fixed several kinds of correlation features to measure the correlation between different feature maps, which is restricted by the existing calculation methods. And it is not an end-to-end method, which means that they only extract features by CNN and use them to calculate texture features for classification. Thereafter, other researchers pay more attention to extract features of multiple scales and locations (such as [1]) or apply historical context information to the style classification of visual art images (such as [49]). In order to solve these problems, we propose an end-to-end network using texture information to classify artistic images, which not only uses CNN features to calculate texture features but also obtains distinguishable features with parameter updates during training. In addition, we propose a weight matrix to measure the importance of different regions in the image.

3 METHODS

Motivated by the effectiveness of using correlations between convolutional features for style recognition, here we propose a novel end-to-end trainable network architecture which encodes the correlations between different feature maps.

3.1 Style Representations

Based on the pioneering work of employing texture information for artistic style transfer [9, 47], Gatys *et al.* [13] introduce the Gram matrix to extract the texture information from the CNN and use it to represent the style of visual art images. The Gram matrix computes the correlations between different feature responses based on the inner product. In this paper, we define the Gram matrix (G^{AB}) as follows:

$$G_{ij}^{AB} = \sum_{k=1\dots d} F_{ij}^{A}(k) F_{ij}^{B}(k), \tag{1}$$

where G_{ij}^{AB} is the inner product of the vectorized feature maps F^A and F^B, the matrix $F_{ij}^{A}(k)$ (resp. $F_{ij}^{B}(k)$) is the activation of the k^{th} filter at the position of the feature map (i, j), d represents the dimension of feature responses in F^A and F^B, and k is the index. Eq. 1 may be sub-optimal as it treats activations at each spatial location as equally important. We improve this by introducing a weight matrix w and formulate it in a more general way as:

$$G_{ij}^{AB} = \sum_{k=1\dots d} w_{ij}(k) F_{ij}^{A}(k) F_{ij}^{B}(k). \tag{2}$$

The size of the weight matrix w_{ij} remains the same as the filter responses. Motivated by the research [7] that investigates six correlation features, we have observed Euclidean distance and Cosine similarity can capture texture information effectively. Considering that Euclidean distance supports backpropagation, we define the weight matrix w using this simple but effective way. During the process of training, we first calculate the distance between features using the Euclidean distance and normalize the distance, and subtract this value from 1 as the weight w_{ij}. Then we scale the weight linearly to $[0, 1]$ using proper normalization. The weight matrix we introduced into the style representation can measure the distance between the different filter responses in every position. Moreover, the style representation G^{AB} is more general and can be applied to any two feature maps F^A and F^B with compatible dimensionality.

3.2 Adaptive Cross-layer Correlation

While it has been shown that the trained representations are transferable between recognition tasks (such as image classification and object detection) to some extent, applicating trained CNN representations as black-box description extractors directly could lead to limited improvements in performance for style classification tasks. As mentioned above, the style of the image has been shown to be closely related to the texture information [13], which may not be well captured by off-the-shelf CNN architectures such as VGGNet, *etc.* This has been partially mitigated by the adoption of the Gram matrix in [7]. Moreover, other distance measures (such as Spearman correlation, Euclidean distance) can also be used to describe the correlation between various features. Interestingly, the choice of the optimal strategy for the distance measure could be

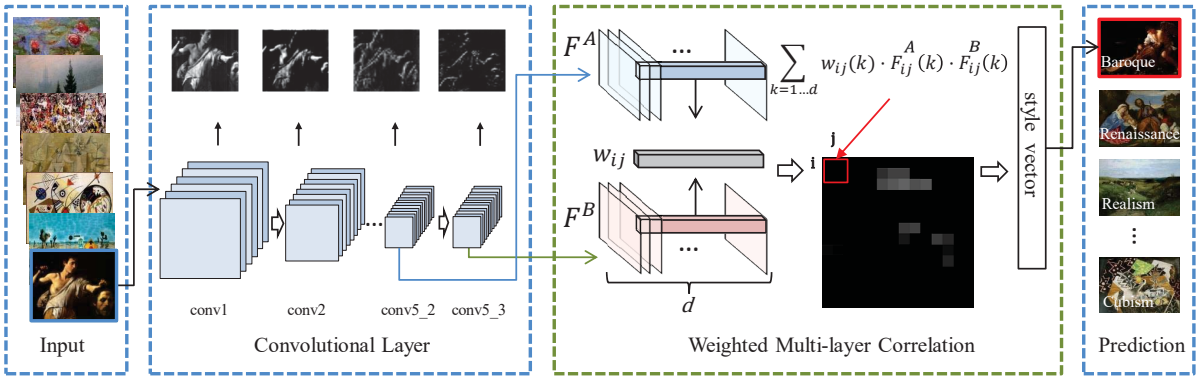

Figure 2: The cross-layer correlation CNN architecture for style classification. We calculate the correlation of the features (F^A and F^B) extracted by the different convolutional layers, here we take the 'conv5_2' layer and the 'conv5_3' layer as an example, and at the same time introduce a similarity weight vector to measure the similarity. The choice of F^A and F^B is generic and can be applied to any layer of feature maps with compatible dimensionality. And d is the dimension of feature responses in F^A and F^B, k is the index.

Algorithm 1 Adaptive Cross-layer Correlation Training

Input: Training examples $I = \{(x_1, y_1), ..., (x_N, y_N)\}$
Initialize model from VGGNet pre-trained on ImageNet
for each $iter \in [1, \max_{iter}]$ **do**
  Train the network on the mini-batch I_{iter} containing M images
  for each $x_m \in [1, M]$ **do**
    Calculate the Euclidean distances between F^A and F^B, and normalize the distance to get L^{AB}
    Initialize the weight by $w = 1 - L^{AB}$
    Normalize weight w to $[0, 1]$ using ℓ_2 normalization
    Calculate the correlation between F^A and F^B to generate G^{AB} by Eq. 2
    Input G^{AB} into the back layers
    Optimize network using the softmax loss
  end for
end for

task-dependent and the effect of various distance measures presents a large gap in the classification accuracy. In the proposed method, the correlation features are well-encapsulated into the network and amenable to training via backpropagation.

As demonstrated in Fig. 1, different layers encode different types of features. Higher layers capture more semantic concepts such as object categories, whereas lower layers encode more detailed discriminative features to capture intra-class variations. Here we argue that learning correlations based on features from different layers could be more effective as in this way, both higher-level abstract and lower-level detailed information could be captured. Some works [31, 33] using various features of multiple layers to provide complementarity information for abstract tasks, *e.g.*, painting classification. Inspired by these methods, we apply a cross-layer strategy to capture more abundant and distinguishable texture information. To this end, we introduce an adaptive cross-layer correlation CNN architecture that calculates the correlation between multiple convolutional layer features on the basis of the VGGNet [40]. As shown in Fig. 2, our proposed method is end-to-end trainable, generic and easily pluggable into any CNN architecture.

As shown in Alg. 1, we calculate the weight matrix for each instance during training. And the Gram-based feature obtained by the correlation calculation $G^{AB} = w \circ AB^T$, where $w = 1 - L^{AB}$, $L^{AB} = (a - b)^2$, and B^T is the transposition of the feature map F^B. We define $\circ$ as the multiplication between elements (Hadamard product), and the size of it depends on the size of the input. The a and b represent the corresponding elements in feature map F^A and feature map F^B respectively. The weight matrix is automatically updated with the training of the network, and the parameters of former layers can also be affected during this process. We use the signed square operation to process the generated bilinear vector. And then, the style vector is normalized using ℓ_2 normalization and passed through a classification layer to predict the style of the visual art image. The feature calculated using the adaptively weighted model can be applied to represent the style.

4 EXPERIMENTS

4.1 Datasets

In this section, we evaluate the performance of our method on three datasets with four style classification tasks involving painting, architecture, and clothing. The artist classification can also be seen as a kind of style classification with personal painting styles.

Painting91 Artists (PA). The Painting91 dataset [22] consists of 4,266 paintings, which contains the paintings of 91 artists from the Renaissance period to modern art. The training/test split provided in the dataset is fixed with 2,275 training images and 1,991 test images.

Painting91 Styles (PS). In the Painting91 dataset, 2,338 paintings from 50 painters are classified into 13 art styles, with style labels as follows: Abstract expressionism, Baroque, Constructivism, Cubism, Impressionism, Neo-classical, Pop art, Post-impressionism,

Realism, Renaissance, Romanticism, Surrealism, and Symbolism. These paintings are divided into 1,250 images for training and remaining for testing.

arcDataset (AD). Xu *et al*. [48] collected an architectural style dataset from Wikimedia, which has a total of 4,786 images, classified into 25 architectural style categories. For each style, 30 images are randomly chosen as the training set and others are used for testing.

Hipster Wars (HW). Hipster Wars dataset [23] contains 1,893 images, and it has five fashion styles: bohemian, goth, hipster, pinup, and preppy. The training set and test set are randomly divided according to 9:1.

In the experiment, we investigate the performance of the intra-layer correlation features and the cross-layer correlation features in the two classification tasks of the Painting91 dataset and compare our method with other existing style recognition methods on all three datasets. All experimental parameters related to datasets and images are consistent with the relevant references [22, 23, 48] for a fair comparison. For example, we use the ten-fold cross-validation on arcDataset following the existing evaluating protocol [48].

4.2 Implementation Details

As shown in Fig. 2, we build an adaptive cross-layer Convolutional Neural Network framework based on the deep framework VGGNet [40] (with 16 layers). And the model is initialized by training with the large-scale image classification task [24]. We calculate the similarity between two feature responses with 512×512 dimension, and before the calculation, we first change them into 262144×1 dimension vectors. The initial learning rate is 0.0001 and multiplied by 0.1 every 10,000 iterations.

Table 1 shows the baseline results and our adaptive correlation methods using Gram-based features on the two tasks of the Painting91 dataset, where the 'VGGNet_ft' represents the baseline. It uses the classification result yielded by the convolutional layer that connects to the classification layer directly and fine-tunes the classification layer again based on the fine-tuned VGGNet.

4.3 Intra-layer Correlation

In the experiment, we first study the intra-layer correlation which calculates the correlation of the same layer features for the style classification. The setting of the classification layer we used in the method of intra-layer correlation is consistent with the stratified experiment in VGGNet for the fair comparison. In the stratified experiment of VGGNet, we directly connect the output of the convolutional layer to the classification layer instead of using the full connection layer. We consider the modified VGGNet as a classifier that uses the fine-tuned VGG model on the corresponding classification task as the initial value. We set the learning rate of the front layers to zero, only train the final classification layer, and calculate the classification results of each convolutional layer. In Table 1, experimental results show that using the proposed correlation method (VGGNet_ft+Gram) can improve the representation performance of convolutional layers, and the proposed adaptive correlation method (VGGNet_ft+WGram) further improves the results, *e.g.*, for painting style classification task, the correlation method achieves 7.98% improvement using 'conv2' compared with baseline (VGGNet_ft), and the adaptive correlation method get a further 2.22% improvement. Furthermore, a deeper convolution layer is more effective for style classification, so we focus on 'conv5', the last set of convolutional layers, to investigate the intra-layer correlation. We especially consider the style features of the correlation from 'conv5_1', 'conv5_2', and 'conv5_3' respectively. In order to clearly compare the intra-layer correlation features with the common deep features, we also use these three convolutional layers of VGGNet to recognize the artists and styles of the Painting91 dataset.

As shown in Table 1, the 'VGGNet_ft' is the baseline, the 'VGGNet_ft+Gram' represents the classification results using the correlation between convolution layers, and the 'VGGNet_ft+WGram' means the classification results using the adaptive correlation between convolutional layers. The number in the parenthesis represents the increment of classification accuracy compared with the left column.

The experimental results show that the classification accuracy of intra-layer correlation features is much better than using the convolutional layer features directly. This shows that the style is closely related to the texture information, and the correlation feature can describe distinguishable texture information in the image, so it is effective for expressing the style of image. According to the classification results, we also found that the 'conv5_2' layer is better than the 'conv5_3' layer, and the 'conv5_1' layer is better than the 'conv5_2' layer, whether using the common deep features or the correlation features. In addition, the classification accuracies of the 'conv5_1' layer and 'conv5_2' layer are relatively close in both two tasks. We consider that the features of 'conv5_1' and 'conv5_2' can express the style appropriately, so we focus on these two layers in the later experiments.

4.4 Cross-layer Correlation

After investigating the correlation features from the same layers, we use the correlation between different layers to represent the style. Considering that the high-level convolution layer features can better respond to the style, we still mainly study the cross-layer correlation around the features of 'conv5'. In Table 1, we can see the classification results using cross-layer correlation features on the two tasks of the Painting91 dataset, where abbreviated expression is used to indicate the correlation between different features. For example, 'conv5_1&2' represents the experimental results of the correlation between the convolutional layer 5_1 and the convolutional layer 5_2.

The experiment results show that the cross-layer correlation result 'conv5_1&2' is better than the results of 'conv5_1&3' and 'conv5_2&3'. From the results of these experiments, we conclude that the correlation features between 'conv5_1' and 'conv5_2' can effectively extract the style information in the image, and the effectiveness of inter-layer correlations of these two features is also consistent with the intra-layer correlation experiments. In addition, compared with the results of intra-layer correlation, the cross-layer features 'conv5_1&2' is better than the intra-layer correlation using 'conv5_1' and 'conv5_2' features individually, so the correlation between these two features provides a good description of the image style. However, the cross-layer correlations 'conv5_1&3' and 'conv5_2&3', only exceed the intra-layer correlation 'conv5_3'. The

Table 1: Classification results on the Painting91 dataset, where 'VGGNet_ft' is the baseline and means the classification results using the convolutional layer directly. 'VGGNet_ft+Gram' means the classification results using the correlation between convolution layers; the 'VGGNet_ft+WGram' means the classification results using the adaptive correlation between convolutional layers. We use 'conv1' to represents the layer after the first convolutional layers as well as the pooling operation, 'conv5_1' refers to the convolutional layer 5_1, and the 'conv5_1&2' means the adaptive correlation between the convolutional layers 5_1 and 5_2. 'Acc' represents the experimental accuracy. The number in the parenthesis represents the increment of classification accuracy compared with the left column.

	Painting Artist (Acc %)			Painting Style (Acc %)		
	VGGNet_ft	VGGNet_ft+Gram	VGGNet_ft+WGram	VGGNet_ft	VGGNet_ft+Gram	VGGNet_ft+WGram
conv1	22.01	24.30 (+2.29)	25.20 (+0.90)	30.97	40.58 (+9.61)	42.29 (+1.71)
conv2	34.77	35.38 (+0.61)	35.88 (+0.50)	41.73	49.71 (+7.98)	51.93 (+2.22)
conv3	47.74	50.46 (+2.72)	55.18 (+4.72)	55.33	66.25 (+10.92)	67.37 (+1.12)
conv4	49.62	65.29 (+15.67)	67.54 (+2.25)	60.05	74.50 (+14.45)	75.42 (+0.92)
conv5 (5_3)	50.80	66.17 (+15.37)	68.69 (+2.52)	63.60	74.58 (+10.98)	75.55 (+0.97)
conv5_1	53.82	67.08 (+13.26)	69.95 (+2.87)	65.72	75.75 (+10.03)	77.67 (+1.92)
conv5_2	52.76	64.88 (+12.12)	69.55 (+4.67)	65.53	75.33 (+9.8)	77.02 (+1.69)
conv5_3	50.80	66.17 (+15.37)	68.69 (+2.52)	63.60	74.58 (+10.98)	75.55 (+0.97)
conv5_1&2	-	68.09	**70.65 (+2.56)**	-	76.33	**78.13 (+1.80)**
conv5_1&3	-	68.96	69.70 (+0.74)	-	75.00	76.56 (+1.56)
conv5_2&3	-	67.17	69.25 (+2.08)	-	74.99	75.74 (+0.75)

Table 2: Comparison of our proposed method with deep learning methods and other popular methods on the Painting91 dataset, where 'PA' and 'PS' represent artist classification and style classification respectively.

Method	PA (%)	PS (%)
Khan *et al.* [22]	53.10	62.20
VGGNet [40]	57.71	69.12
MSCNN1 [32]	58.11	69.67
MSCNN2 [32]	57.91	70.96
CNN F3 [31]	56.40	68.57
CNN F4 [31]	56.35	69.21
Peng *et al.* [33]	57.51	71.05
CMFFV [34]	59.04	67.43
Gram [7]	60.90	71.86
Gram dot Cos [7]	63.17	73.59
SCMFA [35]	65.78	73.16
Anwer *et al.* [1]	64.50	74.80
Yang *et al.* [49]	-	77.76
Ours	**70.65**	**78.13**

accuracy of 'conv5_1&3' is lower than using the intra-layer correlation of 'conv5_1', and the performance of 'conv5_2&3' is lower than that of the intra-layer correlation of 'conv5_2', which means that using cross-layer correlation features with low correlations can suppress the original style features. The results of the adaptive cross-layer correlation experiments show that for both the style and artist classifications, correlation features between 'conv5_1' and 'conv5_2' achieve a very good classification effect. And we consider the intra-layer correlation and cross-layer correlation of the 'conv5_1' feature and 'conv5_2' feature can be a good expression of style, while the texture information extracted using the 'conv5_3' feature is relatively few.

4.5 Comparison to Previous Work

To evaluate the effectiveness of the adaptive cross-layer correlation feature, we compare our proposed method with deep learning methods and other popular methods on three standard datasets. Table 2 shows the experimental results of our method as well as other style classification methods for artist and style classification in the Painting91 dataset. For artist classification, the experimental result using the Gram matrix is 60.90%, and calculating the correlation independently between Gram matrices and Cosine similarity (Gram dot Cos) [7] achieves a classification accuracy of 63.17%. The multi-scale CNN model (MSCNN) from [32] based on the label inheritance strategy achieves a classification accuracy of 58.11%. By combining holistic and part-based deep representations, [1] obtains the classification accuracy of 64.50%. The existing best classification accuracy on the artist task is 65.78% using the SCMFA, a sparse representation based complete kernel marginal Fisher analysis framework [35]. The adaptive cross-layer correlation model we proposed improves the result and achieves 70.65% accuracy. For style classification, the classification accuracy using the SCMFA method can achieve 73.16%, and the method of the deep features combining holistic and part-based information [1] achieves the result of 74.80%. Yang *et al.* [49] use additional information (*e.g.*, the birthplace, the origin time, and the art movement) other than images while conducting experiments on the PS dataset. However, our method obtains a classification accuracy of 78.13%, which still outperforms it under such an unfair comparison. And the method of using convolutional features to calculate the Gram matrix (Gram dot Cos) [7] only achieves an accuracy of 73.59%. Therefore, the proposed adaptive cross-layer correlation architecture is effective.

We also compare our proposed method with other methods on the arcDataset and the Hipster Wars dataset. Table 3 and Table 4 show the comparison of experimental results for the arcDataset and the Hipster Wars dataset, respectively. In the table, 'Our method'

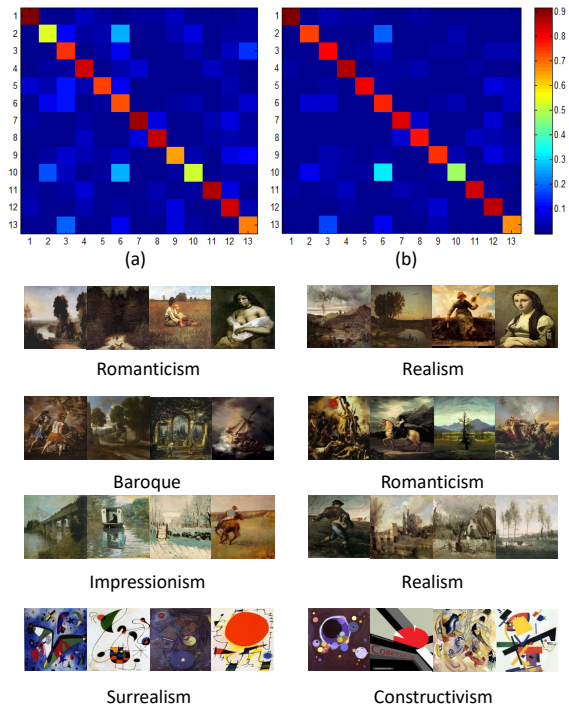

Figure 3: The confusion matrix for 13 painting styles of the Painting91 dataset and some examples of classification results. (a) and (b) represent the confusion matrix of VGGNet and our method respectively. The ordinate is the real label, and the abscissa is the classified label. The numbers 1 to 13 respectively represent Abstract expressionism, Renaissance, Romanticism, Surrealism, Symbolism, Baroque, Constructivism, Cubism, Impressionism, Neo-classical, Pop art, Post-impressionism, Realism. The bottom half is four groups of easy-to-confuse styles that are most noticeably improved using our method. In each row we show the misclassified images in the test set using VGGNet (left), and example images of their misclassified style (right). Using our method can correctly classify these confusing images.

points to the results of our weighted model, which calculates the cross-layer correlation between the convolutional layer 5_1 and the convolutional layer 5_2. An unsupervised image classification method [45] achieves the best accuracy of 59.50% on the arcDataset. Our adaptive cross-layer correlation CNN model obtains a considerable gain on recognition effect, which achieves an accuracy of 73.67%. It is a significant increase because the style of the architectural images is usually reflected in the abundant texture information rather than the semantic information. Specifically, the proposed method can well recognize the architectures with rich texture information. For example, Achaemenid and Ancient Egyptian architectures are often built of stone, and the Russian Revival architecture also has significant texture features in the design structure.

For the Hipster Wars dataset, we similarly compare our method with deep learning methods and other popular methods [5, 6, 39]. The method proposed by Chen *et al.* [6] obtains a classification

Table 3: Comparison of our proposed method with deep learning methods and other popular methods on the arc-Dataset.

Method	Acc. (%)
Xu *et al.* [48]	46.21
Peng *et al.* [33]	55.38
CNN F1 [31]	55.57
Inception V3 [26]	55.67
MSCNN2 [32]	59.13
LDPO-V-PM [45]	59.50
Ours	**73.67**

Table 4: Comparison of our proposed method with deep learning methods and other popular methods on the Hipster Wars dataset.

Method	Acc. (%)
Kiapour *et al.* [23]	70.60
VGG_CNN_M [5]	71.90
[39]	75.90
Chen *et al.* [6]	77.00
Peng *et al.* [33]	77.61
Ours	**80.53**

accuracy of 77.00%. Peng *et al.* [33] present a method based on CNN for various abstract classification tasks and acquire the best classification accuracy of 77.61% on the Hipster Wars dataset. As shown in Table 4, our method achieves an accuracy of 80.53% and outperforms all the existing methods. The proposed method has a satisfactory classification effect for Bohemian, Pinup, and Preppy style apparel. In these styles, the clothing outline (clothing design) is relatively simple and can be easily distinguished according to the texture information. For the Hipster clothing style, the clothing outline is more varied and difficult to observe uniform texture features, so the effect is slightly worse.

We evaluate the proposed method on three widely-used datasets involving painting, architecture, and clothing. Although multiple datasets have different distributions and categories, the proposed method achieves consistent improvement on all the datasets (*e.g.*, about 14% on the arcDataset) illustrating the generalization ability of the model.

4.6 Visualization and Analysis

As shown in Fig. 3, we compare the confusion matrix of VGGNet (a) and our method (b) in the Painting91 dataset, where the numbers 1 to 13 respectively represent Abstract expressionism, Renaissance, Romanticism, Surrealism, Symbolism, Baroque, Constructivism, Cubism, Impressionism, Neo-classical, Pop art, Post-impressionism, Realism. We can observe that our method is very effective in Impressionism, Romanticism, Surrealism, *etc.* And some styles show higher similarity, such as Renaissance, Neo-classical, Baroque, and Realism. Especially, the style Neo-classical has large confusion with the styles Baroque and Renaissance because they flourished in similar periods.

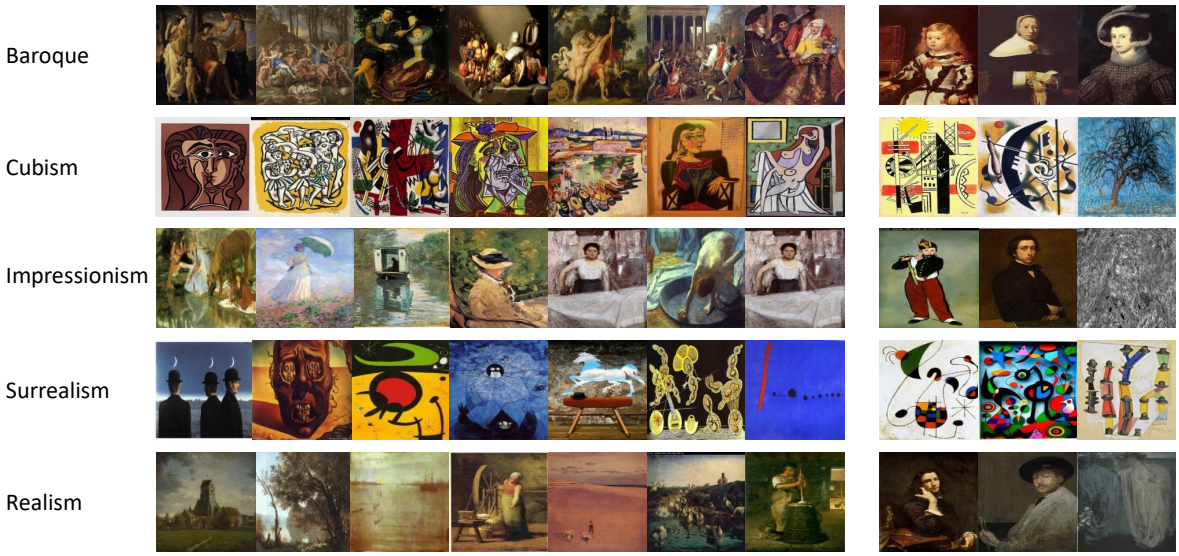

Figure 4: Some images of five classes that can achieve a good recognition rate using the proposed weighted model in style classification tasks of the Painting91 dataset. On the left, we list the example images in the test set, which can be correctly classified based on our model, while incorrectly under the VGGNet. And the right side shows the images distributed into the wrong classes.

The bottom half of Fig. 3 shows four groups of easy-to-confuse styles that are most noticeably improved using our method. On the left of each row, we show the misclassified images of the test set, which are all wrongly assigned to the style on the right side using VGGNet. The images on the right are examples of confusing styles. All the images on the left are correctly classified using our method. Observing these images we can find the four groups of confusing style is relatively similar in the content, *e.g.*, Baroque, Romanticism, Realism all have landscapes and portraits, and Surrealism and Constructivism paintings all have some geometric shapes, lines, *etc.* Moreover, the color of these confusing styles is also very close. Therefore, there are no significant differences between the objects of these paintings. Our method can effectively extract the details and texture information of the painting, and recognize its style.

Fig. 4 shows the images of five style classes that can achieve a satisfactory recognition rate using our weighted model in style classification tasks of the Painting91 dataset. On the left, this example can be correctly classified based on our model, while incorrectly under the VGGNet. And the right side shows the images distributed into the wrong classes. We observe that the proposed weighted model has a good recognition effect for images with strong texture information compared to the VGGNet, especially sensitive for the texture information in the background. In Baroque paintings, our model has a better response to the complex paintings. For example, some images contain many people, but some portraiture with a monotonous background distributed into the wrong classes because the portraiture is common in many style classes. For the Impressionism paintings, the content of some paintings is relatively simple or difficult to identify, so these images are easily identified as the wrong class. Some Surrealism paintings are incorrectly categorized into Constructivism and Cubism styles, because of the fact that several artistic styles that flourished during the same period are mutually influential and have similar characteristics. According to the classification performance, we consider the proposed model also works well for Surrealism paintings.

5 CONCLUSION

For visual art images, texture features have a major impact on their style. The Gram matrix has been explored to encode texture information for style recognition with a commendable performance. However, in principle, this may not be optimal as the Gram matrix construction, and feature learning is completely disjoint. To represent abundant texture information of visual art images, we propose an adaptive cross-layer correlation architecture where the correlation features are spatially weighted and well-encapsulated into the network and amenable to training via backpropagation. Experimental results show that the proposed method can effectively recognize the style of visual arts and significantly outperforms the state-of-the-art methods as well as the traditional CNN architecture in various visual arts recognition tasks.

In the future, we can do some more interesting research on artistic styles, such as processing and analyzing art paintings according to the theme (*e.g.*, Portraits) and other artistic theories, dealing artistic images with foreground/background separately.

ACKNOWLEDGMENTS

This work was supported by the NSFC (NO.61876094), Natural Science Foundation of Tianjin, China (NO.18JCYBJC15400, 18ZXZNGX 00110), the Open Project Program of the National Laboratory of Pattern Recognition (NLPR), and the Fundamental Research Funds for the Central Universities.

REFERENCES

[1] Rao Muhammad Anwer, Fahad Shahbaz Khan, Joost van de Weijer, and Jorma Laaksonen. 2016. Combining holistic and part-based deep representations for computational painting categorization. In *ACM International Conference on Multimedia Retrieval*.

[2] Simone Bianco, Davide Mazzini, and Raimondo Schettini. 2017. Deep multibranch neural network for painting categorization. In *International Conference on Image Analysis and Processing*.

[3] Francesco Bianconi and Raquel Bello-Cerezo. 2018. Evaluation of visual descriptors for painting categorisation. In *IOP Conference Series: Materials Science and Engineering*, Vol. 364. 012037.

[4] Anna Bosch, Andrew Zisserman, and Xavier Munoz. 2007. Representing shape with a spatial pyramid kernel. In *ACM International Conference on Image and Video Retrieval*. 401–408.

[5] Ken Chatfield, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2014. Return of the devil in the details: Delving deep into convolutional nets. *arXiv preprint arXiv:1405.3531* (2014).

[6] Ju-Chin Chen and Chao-Feng Liu. 2015. Visual-based deep learning for clothing from large database. In *Proceedings of the ASE BigData & SocialInformatics*.

[7] Wei-Ta Chu and Yi-Ling Wu. 2016. Deep Correlation Features for Image Style Classification. In *ACM International Conference on Multimedia*.

[8] Wei-Ta Chu and Yi-Ling Wu. 2018. Image style classification based on learnt deep correlation features. *IEEE Transactions on Multimedia* PP, 99 (2018), 1–1.

[9] Alexei A Efros and William T Freeman. 2001. Image quilting for texture synthesis and transfer. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*.

[10] Ahmed Elgammal, Bingchen Liu, Diana Kim, Mohamed Elhoseiny, and Marian Mazzone. 2018. The shape of art history in the eyes of the machine. In *AAAI Conference on Artificial Intelligence*.

[11] Noa Garcia and George Vogiatzis. 2018. How to read paintings: semantic art understanding with multi-modal retrieval. In *Proceedings of the European Conference on Computer Vision*.

[12] Leon Gatys, Alexander S Ecker, and Matthias Bethge. 2015. Texture synthesis using convolutional neural networks. In *Advances in Neural Information Processing Systems*.

[13] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. 2016. Image style transfer using convolutional neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[14] Joseph Goguen. 1999. *Art and the brain*. Vol. 6. Imprint Academic.

[15] Eren Gultepe, Thomas Edward. Conturo, and Masoud Makrehchi. 2017. Predicting and grouping digitized paintings by style using unsupervised feature learning. *Journal of Cultural Heritage* 31 (2017), S1296207417301474.

[16] Zhenhua Guo, Lei Zhang, and David Zhang. 2010. A completed modeling of local binary pattern operator for texture classification. *IEEE Transactions on Image Processing* 19, 6 (2010), 1657–1663.

[17] Junwei Han, Gong Cheng, Zhenpeng Li, and Dingwen Zhang. 2018. A unified metric learning-based framework for co-saliency detection. *IEEE Transactions on Circuits and Systems for Video Technology* 28, 10 (2018), 2473–2483.

[18] Shintami C. Hidayati, Wen Huang Cheng, and Kai Lung Hua. 2012. Clothing genre classification by exploiting the style elements. In *ACM International Conference on Multimedia*.

[19] Yiyu Hong and Jongweon Kim. 2018. Art painting detection and identification based on deep learning and image local features. *Multimedia Tools and Applications* (2018), 1–16.

[20] Juan Xu Hui, Yuan Zou Feng, Wei Jing, and Zhang Ying. 2011. Research on clothing styles classification model based on the MDS and K-Means method. *Advanced Materials Research* 331 (2011), 616–619.

[21] Sergey Karayev, Matthew Trentacoste, Helen Han, Aseem Agarwala, Trevor Darrell, Aaron Hertzmann, and Holger Winnemoeller. 2014. Recognizing image style. In *British Machine Vision Conference*.

[22] Fahad Shahbaz Khan, Shida Beigpour, Joost Weijer, and Michael Felsberg. 2014. Painting-91: A large scale database for computational painting categorization. *Machine Vision & Applications* 25, 6 (2014), 1385–1397.

[23] M Hadi Kiapour, Kota Yamaguchi, Alexander C Berg, and Tamara L Berg. 2014. Hipster wars: Discovering elements of fashion styles. In *European Conference on Computer Vision*.

[24] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*.

[25] Adrian Lecoutre, Benjamin Negrevergne, and Florian Yger. 2017. Recognizing Art Style Automatically in painting with deep learning. In *Asian conference on machine learning*.

[26] Jose Llamas, Pedro M Lerones, Roberto Medina, Eduardo Zalama, and Jaime Gómez-García-Bermejo. 2017. Classification of architectural heritage images using deep learning techniques. *Applied Sciences* 7, 10 (2017), 992.

[27] David G Lowe. 2004. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision* 60, 2 (2004), 91–110.

[28] Shangzhen Luan, Chen Chen, Baochang Zhang, Jungong Han, and Jianzhuang Liu. 2017. Gabor convolutional networks. *IEEE Transactions on Image Processing* PP (2017), 1–1.

[29] Timo Ojala, Matti Pietikainen, and Topi Maenpaa. 2002. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24, 7 (2002), 971–987.

[30] Aude Oliva and Antonio Torralba. 2001. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision* 42, 3 (2001), 145–175.

[31] Kuan-Chuan Peng and Tsuhan Chen. 2015. Cross-layer features in convolutional neural networks for generic classification tasks. In *IEEE International Conference on Image Processing*.

[32] Kuan-Chuan Peng and Tsuhan Chen. 2015. A framework of extracting multi-scale features using multiple convolutional neural networks. In *IEEE International Conference on Multimedia and Expo*.

[33] Kuan-Chuan Peng and Tsuhan Chen. 2016. Toward correlating and solving abstract tasks using convolutional neural networks. In *IEEE Winter Conference on Applications of Computer Vision*.

[34] Ajit Puthenputhussery, Qingfeng Liu, and Chengjun Liu. 2016. Color multi-fusion fisher vector feature for fine art painting categorization and influence analysis. In *IEEE Winter Conference on Applications of Computer Vision*.

[35] Ajit Puthenputhussery, Qingfeng Liu, and Chengjun Liu. 2016. Sparse representation based complete kernel marginal fisher analysis framework for computational art painting categorization. In *European Conference on Computer Vision*.

[36] Ricky J. Sethi, Catherine A. Buell, William P. Seeley, and Swaroop Krothapalli. 2018. An open access platform for analyzing artistic style using semantic workflows. In *International Conference on Web Services*.

[37] Eli Shechtman and Michal Irani. 2007. Matching local self-similarities across images and videos. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[38] Jialie Shen. 2009. Stochastic modeling western paintings for effective classification. *Pattern Recognition* 42, 2 (2009), 293–301.

[39] Edgar Simo-Serra and Hiroshi Ishikawa. 2016. Fashion style in 128 floats: Joint ranking and classification using weak data for feature extraction. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[40] Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*.

[41] Robert L Solso. 1996. *Cognition and the visual arts*. MIT press.

[42] Feifei Sun, Pinghua Xu, and Xuemei Ding. 2018. Multi-core SVM optimized visual word package model for garment style classification. *Cluster Computing* 4 (2018), 1–7.

[43] Guang Lu Sun, Xiao Wu, Hong Han Chen, and Qiang Peng. 2015. Clothing style recognition using fashion attribute detection. In *International Conference on Mobile Multimedia Communications*.

[44] Tiancheng Sun, Yulong Wang, Jian Yang, and Xiaolin Hu. 2017. Convolution neural networks with two pathways for image style recognition. *IEEE Transactions on Image Processing* 26, 9 (2017), 4102–4113.

[45] Xiaosong Wang, Le Lu, Hoo-Chang Shin, Lauren Kim, Mohammadhadi Bagheri, Isabella Nogues, Jianhua Yao, and Ronald M Summers. 2017. Unsupervised joint mining of deep features and image labels for large-scale radiology image categorization and scene recognition. In *IEEE Winter Conference on Applications of Computer Vision*.

[46] Daan Wynen, Cordelia Schmid, and Julien Mairal. 2018. Unsupervised learning of artistic styles with archetypal style analysis. In *Advances in Neural Information Processing Systems*. 6584–6593.

[47] Xuexiang Xie, Feng Tian, and Hock Soon Seah. 2007. Feature guided texture synthesis (fgts) for artistic style transfer. In *Proceedings of the 2nd international conference on Digital interactive media in entertainment and arts*.

[48] Zhe Xu, Dacheng Tao, Ya Zhang, Junjie Wu, and Ah Chung Tsoi. 2014. Architectural style classification using multinomial latent logistic regression. In *European Conference on Computer Vision*.

[49] Jufeng Yang, Liyi Chen, Le Zhang, Xiaoxiao Sun, Dongyu She, Shao-Ping Lu, and Ming-Ming Cheng. 2018. Historical context-based style classification of painting images via label distribution learning. In *ACM International Conference on Multimedia*.

[50] Jufeng Yang, Dongyu She, Ming Sun, Ming-Ming Cheng, Paul L. Rosin, and Liang Wang. 2018. Visual sentiment prediction based on automatic discovery of affective regions. *IEEE Transactions on Multimedia* 20, 9 (2018), 2513–2525.

[51] Dingwen Zhang, Junwei Han, Long Zhao, and Deyu Meng. 2019. Leveraging prior-knowledge for weakly supervised object detection under a collaborative self-paced curriculum learning framework. *International Journal of Computer Vision* 127, 4 (2019), 363–380.

[52] Ning Zhang, Jeff Donahue, Ross Girshick, and Trevor Darrell. 2014. Part-based R-CNNs for fine-grained category detection. In *European Conference on Computer Vision*.

[53] Jana Zujovic, Lisa Gandy, Scott Friedman, Bryan Pardo, and Thrasyvoulos N Pappas. 2009. Classifying paintings by artistic genre: An analysis of features & classifiers. In *IEEE International Workshop on Multimedia Signal Processing*.