This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2018.2801718, IEEE Transactions on Multimedia

Image Style Classification based on Learnt Deep Correlation Features

Wei-Ta Chu, *Senior Member, IEEE,* Yi-Ling Wu

***Abstract*—This paper presents a comprehensive study of deep correlation features on image style classification. Inspired by that correlation between feature maps can effectively describe image texture, we design various correlations and transform them into style vectors, and investigate classification performance brought by different variants. In addition to intra-layer correlation, inter-layer correlation is proposed as well, and its effectiveness is verified. After showing the effectiveness of deep correlation features, we further propose a learning framework to automatically learn correlations between feature maps. Through extensive experiments on image style classification and artist classification, we demonstrate that the proposed learnt deep correlation features outperform several variants of CNN features by a large margin, and achieve the state-of-the-art performance.**

***Index Terms*—Painting images, convolutional neural network, Gram matrix, deep correlation features, learnt correlation features.**

I. Introduction

DESPITE various studies on visual attributes and semantic concept detection, some image properties are implicit and difficult to extract, but may be quite useful in image management and retrieval. Some bio-inspired properties, like sentiment [1] and emotion, can be easily perceived by human beings, but are hard to be modeled in a computational way. In this work, we focus on the *image style* that emerges recently and is believed to be a potential extension of current classification/retrieval works. We will mainly take oil painting images as the targeted domain because image styles such as Academicism, Baroque, and Cubism are well defined. We will extract or learn correlations between feature maps obtained based on deep neural networks, and achieve image style classification on various image datasets.

Foreseeing the potential of visual style analysis, several inspiring works have been proposed. Focusing on branded handbag recognition, Wang et al. [2] proposed a random forest-based strategy to determine discriminative patches on handbags, in order to construct handbag style representation. Tarvainen et al. [3] proposed a movie dataset to facilitate affective movie content analysis, which is largely determined by movie style and aesthetics. Karayev et al. [4] proposed two image datasets respectively consisting of photos from Flickr and painting images from Wikiart.org, and investigated various visual features on image style classification. Specific to painting images, Khan et al. [5] constructed a large-scale painting image dataset consisting of paintings from 91 different artists. They studied how local and global features perform in three applications, i.e., artist categorization, style classification, and saliency detection. Most recently, Tseng et al. [6] proposed a ranking model for style identification based on random forests. Based on visual features like Lab color histogram and GIST, they concentrate on mitigating the overfitting problem and the ambiguity problem by using random forests.

To do image style classification, image representation is obviously the key. In [4], Karayev et al. reported that deep features, which have been demonstrated to achieve promising performance in various fields, also yield performance much better than hand-crafted features like color histogram, GIST, and visual saliency. However, the complex interplay between visual appearance and perceived image style is still not clear. Recently, Gatys et al. [7] proposed a feature space that was originally designed for texture synthesis [8] on top of the filter responses of each layer in a convolutional neural network. Particularly, the correlations between different filter responses are calculated, as the important clues to transfer a photograph into a painting of some artist's style. Since then, a large number of studies and commercial software are developed for style transfer [9][10][11][12][13][14][15][16][17] and image style analysis [18] [19]. In fact, style transfer is viewed as one of the major achievements in deep learning in year 2016 [20].

We focus on image style analysis and take painting image classification as the main target. Figure 1 shows sample images of styles from Academicism to Rococo. Given a painting image, style descriptor is extracted, and then the image is classified into one of the style classes. In our previous work [19], we transformed correlations between feature maps into image style descriptors. Descriptors derived from different layers with different settings were comprehensively evaluated. In addition to correlations between feature maps at the same layer (intra correlations), we further proposed descriptors from correlations across multiple layers (inter correlations). Benefits of jointly considering various correlations were verified. After verifying the effectiveness of deep correlation features, in this work we attempt to discover correlations by a learning framework and demonstrate the learnt correlations can significantly improve classification performance. In addition to painting styles, we also show the effectiveness of the proposed descriptors on photo style classification, illustration style classification, and artist classification.

Notice that taking correlations between different features as texture description is not a new idea. In early years, Jain and Healey [21] proposed correlation between different Gabor filter bands to describe texture information in multiple scales. In our work, we try to discover effectiveness of correlation-based

W.-T. Chu and Y.-L. Wu are with National Chung Cheng University, Taiwan.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2018.2801718, IEEE Transactions on Multimedia

description in the context of deep learning. Recently, Lin et al. [22] proposed bilinear CNN models that take outer product between outputs derived from two CNN models to be image representation. They claimed that the outer product captures pairwise correlations between feature channels. Bilinear CNN models provide a generic framework to consider correlation between outputs of multiple CNN models, while our work considers correlation between feature maps at different layers of a single CNN model.

The rest of this paper is organized as follows. In Section II, most recent related works on image style transfer and image style analysis are described. Section III describes basic deep correlation features derived from common statistical techniques, and Section IV describes the learning framework to automatically discover implicit correlation between feature maps. Comprehensive experimental results on the collected oil painting dataset are provided in Section V, and performance comparison on other painting datasets as well as generalization to other types of datasets are given in Section VI. Finally, summary and discussion are given in Section VII.

II. Related Works

A. Style Transfer

Image style refers to how an image is perceived. It is hardly to be described literally but can be easily perceived visually. Difference between various styles may come from texture, color distribution, semantics, etc. One research topic highly related to image style is thus texture synthesis, which has been studied for years. Approaches based on physical simulation, Markov random field, Gibbs sampling, or feature matching were proposed [23]. For example, Zhang et al. [24] formulated style transfer as an optimization problem by using Markov random fields. They jointly considered the content and style so that the generated results not only synthesize the style, but also preserve the image content well. Recently, the power of convolutional neural networks (CNN) on texture synthesis has been revealed [8]. Gatys et al. found that textures can be well represented by the correlations between feature maps of different layers. This finding motivates the pioneering work [7] that transfers a photo into an image with a style similar to a given targeted painting. Given a photo P and a reference image I with the targeted style, the idea of [7] is to adjust P, such that correlations between feature maps (obtained by inputting the adjusted image $\hat{P}$ to a CNN) are similar to that of I, while CNN features of $\hat{P}$ are similar to the original P. With this breakthrough, this research topic becomes quite flourishing in year 2016.

The adjustment process in [7] is computationally expensive, and thus a few works were proposed to speed up the process. Johnson et al. [10] constructed transformation networks using perceptual loss functions, which are defined based on high-level features from a loss network. The perceptual loss functions more robustly measure image similarities, and the transformation networks achieve style transfer much more efficiently. Ulyanov et al. [9] proposed a feed-forward convolutional network to move the adjustment process into a learning stage, and made style transfer much more light-weight. They improved the proposed Texture Networks by designing a new normalization scheme and a learning formulation to further improve quality and diversity of stylization [17]. Although the learnt networks mentioned above are efficient, they are tied to a single style. To transfer an image into a specific style, a separate network should be constructed. Dumoulin et al. [12] proposed conditional instance normalization, and developed a generic network that can flexibly capture properties of different styles. Ruder et al. [13] extended style transfer to videos. In addition to do style transfer for each video frame, they introduced temporal consistency loss and a multi-pass algorithm to make transformation results smooth. Tanno et al. [14] extended the style transfer network in [10] to learn multiple styles at the same time, and further reduced computational requirement to develop a real-time style transfer application on mobile phones. In addition to this, other style transfer mobile applications like Prisma [15] are also available.

Style transfer achieved by the convolutional neural network is basically semantics-unaware. Champandard [11] empowered this approach by further considering a semantic map corresponding to the input image. A user can sketch a spatial layout associated with semantic meanings, and then the proposed system is able to synthesize a fine artwork with specified style conforming to the sketched layout and semantics. Despite the amazing results of style transfer, why correlations between feature maps can well catch style representation is unclear. Li et al. [16] proposed a novel interpretation to show that matching the Gram matrices of feature maps is equivalent to minimize the maximum mean discrepancy with the second order polynomial kernel.

B. Image Style Classification

In addition to style transfer, some works have been proposed to utilize style representation in image style analysis. Wang et al. [2] constructed handbag style representation and color representation based on discriminative patch discovery and dominant color features, respectively. Handbags are first classified into different style classes, and within each class handbags of different colors are further discriminated. These aforementioned representations are respectively used to measure inter-class style similarity and intra-class color variations. Based on the Gestalt theory, Shen and Cheng [25] proposed Gestalt feature points to improve repeatability of local features in images of the same content but in different styles. They demonstrated superior performance yielded by the proposed Gestalt feature points over existing local features.

Karayev et al. [4] showed that deep features yield performance much better than conventional hand-crafted features in recognizing image styles. Lu et al. [26] proposed a multi-patch aggregation neural network to integrate feature learning and aggregation. This network was shown to yield promising performance in image style classification, aesthetics categorization, and quality estimation. In [27], the same research group jointly considered global and local characteristics by a deep learning framework to predict image aesthetics. They further proposed to utilize image styles and semantic attributes to boost aesthetics estimation performance. Folego et al. [28]

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2018.2801718, IEEE Transactions on Multimedia

Fig. 1. Sample painting images of different styles. Left to right, top to bottom: Academicism, Baroque, Expressionism, High Renaissance, Low Renaissance, Impressionism, Neoclassicism, Primitivism, Realism, and Rococo.

divided a painting image into non-overlapping patches, and extracted CNN features from each patch. Each patch was classified individually, and results of different patches are fused to get final results, i.e., whether a painting was produced by van Gogh or not. Motivated by the finding of [7], Matsuo and Yanai [18] transformed the Gram matrices of feature maps into style vectors, and then used them to do style image retrieval and artist retrieval.

Deep features have been demonstrated to yield good performance in image style analysis. However, the potential of deep features on image style analysis, especially the correlation between deep features mentioned in [7], was far from well explored. Karayev et al. [4] showed deep features extracted from pre-trained neural networks work well, and Matsuo and Yanai [18] showed that correlation between feature maps yield even better performance. In our previous work [19], in addition to the Gram-based style representation, we investigated several more types of correlations and demonstrated performance variations. In this work, in contrast to calculate correlations defined in statistics textbooks, we further propose to automatically learn correlations by neural networks, and demonstrate that the learnt correlations can provide significantly better performance. The main contribution of this work is thus the proposal of automatically learnt deep correlation features on image style classification.

III. Gram-based Deep Correlation Features

Motivated by the inspiring work [7], we attempt to transform correlations between feature maps into style representation of images, and then construct classifiers based on such representations to do image style classification. In addition to the Gram-based representation [7][18], we further investigate performance of various correlations well defined in statistics.

In this work we utilize the VGG-19 network [29] trained based on the ImageNet dataset to obtain filter responses at different layers. This network consists of sixteen convolutional layers and three fully-connected layers. At each convolutional layer, the receptive field is fixed to 3×3 with convolution stride 1 pixel. Spatial pooling is carried out by five max-pooling (or average-pooling) layers, which respectively follow the 2nd, the 4th, the 8th, the 12th, and the 16th convolutional layers (note that not every convolutional layer is followed by a pooling layer). Max-pooling (or average-pooling) is performed over 2×2 pixel window, with stride 2. Because of the pooling layers, convolutional layers in this framework can be divided into five groups. The work in [7] named convolutional layers as 'conv1_1', 'conv1_2', 'conv2_1', 'conv2_2', and so on. The 'conv2_1' layer, for example, are the 3rd convolutional layer that just follows the first pooling layer. In this work, we use the imagenet-vgg-verydeep-19 model trained by the MatConvNet toolbox [30], based on the ImageNet dataset, to conduct the following studies.

Gram-based Deep Correlation Features. Gatys et al. [8] built a style representation based on the correlations between filter responses (feature maps), in order to transfer a photo into an image with a targeted style. In [7], the correlations are represented by the Gram matrix $G^l \in R^{N_l \times N_l}$, where G^l_{ij} is the inner product between the vectorized feature map i and j in layer l, i.e.,

$$G^l_{ij} = \sum_k F^l_{ik} F^l_{jk}, \tag{1}$$

where F^l_{ik} is the activation of the ith filter at position k in layer l.

Taking the 'conv5_1' layer of the VGG-19 model as the example, there are 512 feature maps[1], and the width and height of each feature map are both 14. Each feature map is therefore vectorized into a $14 \times 14 = 196$-dimensional vector. We can calculate the inner product between any pair of 196-dim vectors, and thus a (symmetric) 512×512 Gram matrix can be constructed.

In order to achieve image style classification, we traverse the Gram matrix G^l by raster scan and transform the matrix into a *style vector*, which is then classified by an SVM classifier (support vector machine) constructed for image styles. With the example mentioned above, A 512×512 Gram matrix is flattened as a 262,144-dim deep correlation feature vector. In the following, we will describe such style vector as Gram-based deep learning features.

Figure 2 illustrates the system framework. In the evaluation section, we will especially investigate performance

[1]Detailed configurations of the VGG-19 model please refer to [29].

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2018.2801718, IEEE Transactions on Multimedia

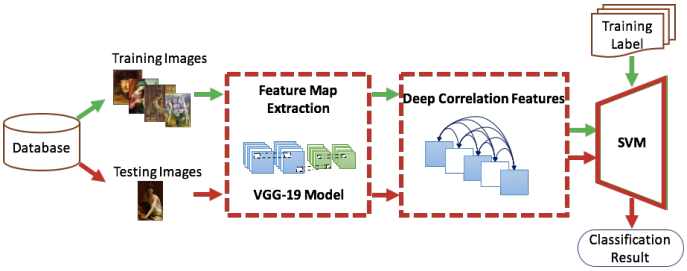

Fig. 2. Illustration of the system framework. Given an image, feature maps are extracted based on the VGG-19 model, and then deep correlation features are calculated. A support vector machine is constructed for style classification.

obtained based on style representation derived from 'conv1_1', 'conv2_1', 'conv3_1', 'conv4_1', and 'conv5_1', respectively.

Other Deep Correlation Features. In addition to Gram-based features, which is derived from inner product, we would like to further investigate the possibility of other correlation features. In statistics, many similarity measures are designed to describe relationships between random variables. We conceptually think that each feature map is generated according to a random variable, and then evaluate correlations between feature maps based on several well-adopted measurements. In addition to Gram matrix, we calculate the following correlations between feature maps: (1) Pearson correlation, (2) Spearman correlation, (3) covariance, (4) Chebychev distance, (5) Euclidean distance, and (6) Cosine similarity. We will respectively transform each of these correlations into style representation, and evaluate its performance on image style classification. Combinations of some of them will also be extensively evaluated.

To clearly describe these correlation metrics in mathematics, we take feature maps of 'conv5_1' as the main example. The 196 values of a feature map can be viewed as values generated based on a random variable, and the ith feature map and the jth feature map can be represented by two 196-dimensional vectors $F^{(i)} = \{f_1^{(i)}, ..., f_{196}^{(i)}\}$ and $F^{(j)} = \{f_1^{(j)}, ..., f_{196}^{(j)}\}$, respectively.

- Pearson correlation: The Pearson correlation coefficient $\rho_{i,j}$ between the ith feature map and the jth feature map is defined as

$$\rho_{i,j} = \frac{E[(f^{(i)} - \bar{f}^{(i)})(f^{(j)} - \bar{f}^{(j)})]}{\sigma^{(i)}\sigma^{(j)}}, \tag{2}$$

  where $\sigma^{(i)}$ and $\sigma^{(j)}$ are standard deviations of $F^{(i)}$ and $F^{(j)}$, respectively, $\bar{f}^{(i)}$ and $\bar{f}^{(j)}$ are means of $F^{(i)}$ and $F^{(j)}$, respectively, and $E(\cdot)$ is the expectation function.
- Spearman correlation: The Spearman correlation coefficient $r_{i,j}$ is actually the Pearson correlation coefficient between the ranked variables. Therefore, the value $r_{i,j}$ between the ith feature map and the jth feature map is defined as

$$r_{i,j} = \frac{E[(g^{(i)} - \bar{g}^{(i)})(g^{(j)} - \bar{g}^{(j)})]}{\sigma_g^{(i)}\sigma_g^{(j)}}, \tag{3}$$

  where $g^{(i)}$ is the sorted $f^{(i)}$, $\bar{g}^{(i)}$ is the mean of sorted $F^{(i)}$, and $\sigma_g^{(i)}$ is the standard deviation of sorted $F^{(i)}$.
- Covariance: The covariance $cov_{i,j}$ between the ith feature map and the jth feature map is defined as

$$cov_{ij} = E[(f^{(i)} - \bar{f}^{(i)})(f^{(j)} - \bar{f}^{(j)})]. \tag{4}$$

- Chebychev distance: The Chebychev distance $v_{i,j}$ between the ith feature map and the jth feature map is defined as

$$v_{ij} = \lim_{k\to\infty}\left(\sum_{i=1}^{196}|f_k^{(i)} - f_k^{(j)}|^k\right)^{1/k}, \tag{5}$$

  which is also known as the L_∞ norm between $F^{(i)}$ and $F^{(j)}$.
- Euclidean distance: The Euclidean distance $d_{i,j}$ between the ith feature map and the jth feature map is defined as

$$d_{ij} = \sqrt{\sum_{k=1}^{196}(f_k^{(i)} - f_k^{(j)})^2}. \tag{6}$$

- Cosine similarity: The Cosine similarity $s_{i,j}$ between the ith feature map and the jth feature map is defined as

$$s_{ij} = \frac{F^{(i)} \cdot F^{(j)}}{\|F^{(i)}\|\|F^{(j)}\|}, \tag{7}$$

  where $F^{(i)} \cdot F^{(j)}$ denotes the inner product of $F^{(i)}$ and $F^{(j)}$.

We can calculate a specific correlation value between any pair of feature map. The 'conv5_1' layer outputs 512 feature maps, and thus totally $512 \times 512 = 262,144$ Spearman correlation values, for example, can be obtained.

Inter-layer Deep Correlation Features. The correlations mentioned above are calculated based on feature maps coming from the same layer. They are 'intra-layer' correlations. We are wondering if *correlations between feature maps across layers* also benefit style classification. To verify this, we calculate Gram matrices of feature maps at each convolutional layer, and then calculate inner products between intra-layer Gram matrices (after dimension reduction) to measure the inter-layer correlation, i.e., the Gram matrix of Gram matrices. For example, the correlation between Gram matrices derived from 'conv4_1' and 'conv5_1' can be calculated. Other forms of inter-layer correlations will also be studied.

IV. Learnt Deep Correlation Features

In [19], we verified the effectiveness of deep correlation features on image style classification. With the success of correlations defined in statistics textbooks, we are wondering if the correlations between feature maps can be automatically learnt by a neural network, such that the transformed style representation can yield even better performance. Therefore, given a set of feature maps, we would like to construct a neural network that automatically learns correlations between feature maps. This neural network replaces the roles of inner product or Pearson correlation mentioned above. Figure 3 illustrates the framework to automatically learn deep correlations as well as style classification.

The idea of automatically learning correlation between features or modalities was already proposed before. Wang et

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2018.2801718, IEEE Transactions on Multimedia

al. [31] proposed a hashing-based orthogonal deep model to learn multimodal representation. This method captures intra-modality and inter-modality correlations, and yields substantially better performance on cross-modal retrieval. Also for cross-modal retrieval, Hua et al. [32] put more focus on semantic coherence and proposed a correlation learning method by adaptive hierarchical semantic aggregation. In contrast to learning correlation between different modalities, our work attempts to learn correlation between feature maps derived from the deep learning framework. Different feature maps conceptually represent responses of different object parts at different scales.

The essential idea of building a neural network to automatically learn correlation between random variables is described as follows. Calculating the correlation between two random variables X and Y can be viewed as a real-valued function f such that $C = f(X, Y)$, where C means the correlation between X and Y. A neural network can be viewed as a function set, and a specific set of parameters in the network indicates a specific function. By finding the best parameters to build the network (with respect to the defined loss function), conceptually we find the best function that transforms two inputs X and Y into C so that C most appropriately describes image styles and thus yields better classification results.

Particularly, to learn correlation between feature maps, we first flatten each feature map into a vector, and then stack all vectors as a matrix. Taking feature maps from 'conv5_1' as the example, there are 512 feature maps, and each feature map is 14×14. Each feature map is flatten into a 196-dimensional vector, and thus all vectors are stacked to be a 196×512 aggregated feature map, which is resized into 224×224 afterwards. To especially capture feature deviation, it is subtracted from the mean aggregated map obtained from the training data. After this process, this aggregated feature map is fed into a neural network to learn correlation features.

With the same idea, we can also learn "correlation between correlations (especially in terms of Gram matrix)". Taking the Gram matrix derived from 'conv5_1' as the example again, the Gram matrix is 512×512, and we resize it into 224×224. The resized Gram matrix is fed into the learning model to derive correlation between correlations.

Note that resizing the aggregated feature map or Gram matrix into smaller size can largely reduce the required training time, but may incur information loss or distortion. Fortunately, according to our experiments, performance remains similar when the input is resized.

Figure 4 shows detailed structure and settings of the learning model. A given aggregated feature map is convolved with a 19×19 window, with zero padding and the ReLU activation function, and 32 filter responses are obtained. After max pooling, these filter responses are further sequentially processed by two fully-connected layers, consisting of 512 nodes and 256 nodes, respectively. These fully-connected layers act as a classifier, and the probabilities of the input being each image style are output by the final softmax layer. This model is trained by minimizing the cross entropy loss function. Parameters of this model are updated by using the ADADELTA method [33], with the learning rate $lr = 1.0$,

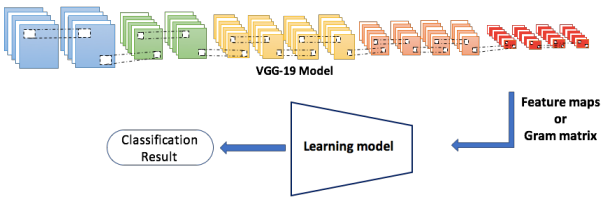

Fig. 3. Illustration of learnt deep correlation features.

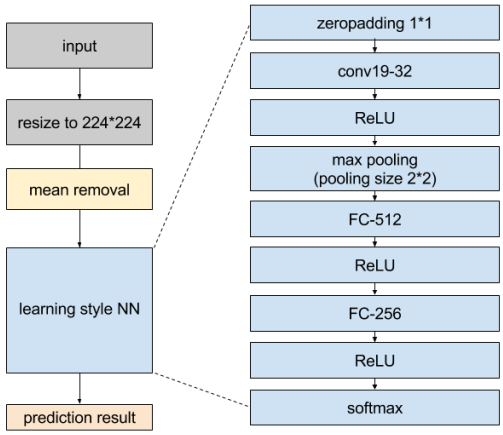

Fig. 4. Detailed structure of the correlation learning model.

decay constant $\rho = 0.95$, and fuzz factor $\epsilon = 1 \times 10^{-8}$. The mini-batch size is 100, and the numbers of epochs for training are around 20 to 50, depending on different datasets.

The main difference between frameworks in Figure 2 and Figure 3 is how to extract the deep correlation features. In Figure 2, correlation features are extracted by computing the Gram matrix between feature maps. In Figure 3, correlation features are automatically learnt by a neural network. One another difference is the classification method. In Figure 2, an SVM classifier is constructed based on the Gram-based correlation features. In Figure 3, classification is done by the neural network same as feature learning.

V. Performance Evaluation

We conduct comprehensive performance evaluation and divide it into two sections. In Sec. V, we focus on verifying that (1) considering correlation between feature maps (Gram-based correlation) is better than commonly-used CNN features (usually derived from the first or the second fully-connected layers); and (2) learnt correlation features further work better than the Gram-based correlation features. Results shown in Sec. V are obtained based on the OilPainting dataset. In Section VI, we experiment on more datasets and demonstrate the state-of-the-art performance comparing to existing works.

A. Painting Image Datasets

From WikiArt.org, we collected totally 19,787 oil painting images belonging to 17 image styles for the following evaluation. Table I shows detailed information of the collected

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2018.2801718, IEEE Transactions on Multimedia

dataset, named OilPainting, where each style class contains at least 200 images. To fairly do performance comparison, we also evaluate performance on the Wikipaintings dataset [4] and the Painting-91 dataset [5]. The former consists of 82,442 images belonging to 25 styles, and the latter consists of 2,338 painting images belonging to 13 styles. Note that our OilPainting dataset contains only oil paintings, while the other two datasets may contain other types of images, like comics and watercolor paintings. We have shown sample images of the OilPainting dataset in Figure 1. Figure 5 shows five sample images of the WikiPaintings dataset, and Figure 6 shows five sample images of the Painting-91 dataset. We see that some styles are common in these three datasets. In this section, we will mainly use the OilPainting dataset to demonstrate the effectiveness of deep correlation features.

Different artists have their unique styles in producing artworks. Several previous works thus also study classifying images according to artists. In this paper, we also study this issue based on two datasets. We select the artists who produced more than 50 images from the OilPainting dataset, and construct the OilPainting Artist dataset that includes totally 15,357 images produced by 104 artists. Another dataset is from [5], called the Painting-91 Artist dataset, which contains 4,266 images produced by 91 artists. Table II shows overall information of all evaluated datasets.

In all the following experiments, the five-fold cross validation scheme was adopted. Taking the OilPainting dataset as the example, at each run, from each style 80% of the data is randomly selected as the training data, and the remaining 20% is used for testing. We conduct the same process for five runs, and the report the average classification accuracy.

B. Performance of Gram-based Deep Correlation Features from Different Layers

We first investigate performance variations yielded by deep correlation features computed from different layers. According to [7], we especially focus on the Gram matrices derived from 'conv1_1', 'conv2_1', 'conv3_1', 'conv4_1', and 'conv5_1', respectively, which are all the first convolutional layer after the pool layer (except for 'conv1_1'). The five Gram matrices are of different dimensions, and so do the transformed style vectors. To fairly compare performance of style vectors from different layers, we adopt principal component analysis (PCA) to reduce dimensions of all style vectors into 4096.

Table III and Table IV show performance variations obtained by Gram-based deep correlation features from different layers, when average pooling and max pooling are applied, respectively. The experiments are conducted based on the OilPainting dataset with the five-fold cross validation scheme, and the average classification accuracies are reported. As can be seen from both tables, we see that Gram-based features derived from the 16th convolutional layer, i.e., 'conv5_1', perform the best. The 'conv5_1' layer is thus widely used in the following experiments. The 'fc7' row shows the performance obtained by vectors coming from the second fully-connected layer (other than convolutional layers, before this layer there are five max pooling layers and one fully-connected layer, and this is why it is called fc7), which was commonly used in many classification tasks. Comparing fc7 with others, fc7 outperforms most except for 'conv5_1'. This shows that output of the fully-connected is quite effective. However, more performance gain can be obtained if we extract deep correlation features from an appropriate layer, e.g., 'conv5_1'. In Table IV, we also show even if the VGG-19 model is fine-tuned with the painting images, performance similar to that without fine-tuning is obtained.

Previous studies showed that features extracted from convolutional layers may give better generalization abilities than the output of fully-connected layers. Therefore, we also evaluate performance obtained by directly taking feature maps as features. The 'FM of conv5_1' shows the performance yielded by the feature maps of the 'conv5_1' layer. As can be seen, comparing with the Gram-based features derived from 'conv5_1', much worse performance is obtained. Finally, by comparing Table III and Table IV, we found that the network with max pooling performs better.

C. Performance of Various Deep Correlation Features

The aforementioned results show effectiveness of Gram-based correlation features. We would like to investigate if other correlations commonly seen in statistics also yield effective style representation. Table V shows performance variations of various deep correlation features. This table can be divided into four parts. The first part is just the subset of Table IV, showing the best performance obtained by Gram-based features. The second part shows average accuracies obtained by six different style vectors derived from six correlations, respectively. Note that each individual style vector is reduced to 4096-dimensional by PCA. By comparing the first two parts, we see no other correlation works better than Gram-based features. This verifies that the choice in [7] is really good. Among the correlations other than Gram matrix, Euclidean distances and Cosine similarity are relatively better.

The third part of Table V tests the conjecture: will better performance be obtained if we jointly consider multiple style representations derived from different correlations? For example, the cell 'Gram-Cos.' means that we concatenate the style representation derived from Gram matrices with that from Cosine similarity. Note that in order to make fair comparison, we reduce dimensionality of each kind of style vector into 2048, so that concatenation of two different style vectors form a 4096-dimensional vector. The third part of Table V shows that, by concatenating style vectors derived from Gram matrices and covariance, performance better than other combinations can be achieved (accuracy=60.56%). This verifies that combining two different style vectors outperforms the best individual one (Gram matrix, accuracy=58.13%).

Since considering multiple correlations yields performance gain, how about calculating *correlation between multiple correlations* and viewing it as a "meta style representation"? The fourth part of Table V shows performances obtained by style representations derived from correlation (measured by inner product) between Euclidean distances and Cosine similarity (denoted by 'Eud. dot Cos.'), and correlation between Gram

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2018.2801718, IEEE Transactions on Multimedia

Fig. 5. Five sample images from the WikiPaintings dataset. From the left to the right, the styles are Color Field, Cubism, Post Impressionism, Realism, and Rococo.

Fig. 6. Five sample images from the Painting-91 dataset. From the left to the right, the styles are Abstract Expressionism, Baroque, Constructivism, Cubism, and Impressionism.

TABLE I
STYLE CLASSES AND THE NUMBERS OF IMAGES IN EACH CLASS IN THE OILPAINTING DATASET.

Style	Academicism	Art Nouveau	Baroque	Cubism	Expressionism	High Renaissance
#img	342	263	1892	349	1127	408
Style	Impressionism	Mannerism	Naive Art	Neoclassicism	Northern Renaissance	Post-Impressionism
#img	4557	607	373	442	549	2183
Style	Realism	Rococo	Romanticism	Surrealism	Symbolism	
#img	2766	1097	1532	794	506	

TABLE II
OVERALL INFORMATION OF THE EVALUATED DATASETS.

Datasets	#Styles	#Artists	#Images
OilPainting	17	–	19,787
Wikipaintings	25	–	82,442
Painting-91	13	–	2,338
OilPainting Artist	–	105	15,357
Painting-91 Artist	–	91	4,266

TABLE III
PERFORMANCE VARIATIONS OF GRAM-BASED DEEP CORRELATION FEATURES FROM DIFFERENT LAYERS, BASED ON THE OILPAINTING DATASET (AVERAGE POOLING WAS APPLIED IN THE FRAMEWORK).

Layer	Original dim.	Reduced dim.	Avg. Accuracy
fc7	4096	4096	52.86%
FM of conv5_1	100352	4096	31.98%
conv1_1	4096	4096	32.71%
conv2_1	16384	4096	34.24%
conv3_1	65536	4096	40.77%
conv4_1	262144	4096	47.87%
conv5_1	262144	4096	**57.19%**

matrices and Cosine similarity (denoted by 'Gram dot Cos.'). Surprisingly, we obtain further performance gain (61.28% vs. 60.56%), by comparing the 'Gram dot Cos.' with 'Gram-Cov.'. Other meta correlations are also experimented, but performance gains are not significant and are not shown here. We can thus push the idea proposed in [7] one step further: *correlation between deep correlation features even works better*.

TABLE IV
PERFORMANCE VARIATIONS OF GRAM-BASED DEEP CORRELATION FEATURES FROM DIFFERENT LAYERS, BASED ON THE OILPAINTING DATASET (MAX POOLING WAS APPLIED IN THE FRAMEWORK).

Layer	Original dim.	Reduced dim.	Avg. Accuracy
fc7	4096	4096	56.83%
fc7 (fine tuned)	4096	4096	56.59%
FM of conv5_1	100352	4096	31.81%
conv1_1	4096	4096	30.08%
conv2_1	16384	4096	35.05%
conv3_1	65536	4096	44.70%
conv4_1	262144	4096	50.60%
conv5_1	262144	4096	**58.13%**

D. Intra-Layer and Inter-Layer Correlations

The Gram matrices mentioned above are calculated based on feature maps of the 'conv5_1' layer. They are 'intra-layer' correlations because only information within the 'conv5_1' layer is considered. We are wondering if *correlations between feature maps across layers* also benefit style classification. To verify this, we calculate Gram matrices of feature maps at each convolutional layer, and then calculate inner products between intra-layer Gram matrices (after dimension reduction) to measure the inter-layer correlation, i.e., the Gram matrix of Gram matrices.

Table VI shows performances obtained by style representation derived from 'conv5_1' only, and by the concatenation of style vectors from 'conv5_1' and the Gram matrix of Gram matrices, respectively. As can be seen, by further considering inter-layer correlation, performance gain can be obtained (59.91% vs. 58.13%). There may be many ways to jointly

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2018.2801718, IEEE Transactions on Multimedia

TABLE V
PERFORMANCE VARIATIONS OF VARIOUS DEEP CORRELATION FEATURES, BASED ON THE OILPAINTING DATASET.

Correlation	fc7	fc7 (fine tune)	**Gram matrix**			
Avg. Acc.	56.83%	56.59%	**58.13%**			
Correlation	Pearson	Spearman	Covariance	Chebychev dist.	Euclidean dist.	**Cosine Sim.**
Avg. Acc.	44.96%	44.92%	45.51%	46.05%	51.33%	**53.34%**
Correlation	Pear.-Spear.	Pear.-Cos.	Gram-Pear.	Gram-Cos.	Gram.-Eud.	**Gram-Cov.**
Avg. Acc.	47.36%	55.68%	60.22%	60.36%	60.42%	**60.56%**
Correlation	Eud. dot Cos.	**Gram dot Cos.**				
Avg. Acc.	51.17%	**61.28%**				

TABLE VI
PERFORMANCE VARIATIONS OF STYLE VECTORS DERIVED FROM INTRA-LAYER CORRELATION ONLY AND INTRA-INTER CORRELATION.

Correlation	Average accuracy
fc7	56.83%
Gram matrix (from 'conv5_1')	58.13%
Gram matrix + Gram of Gram	**59.91%**

consider intra-layer and inter-layer correlations, which are left for future exploration.

E. Performance of Learnt Correlation Features

In this section, we evaluate deep correlation features learnt from information derived from 'conv5_1'. Inputs of the learning models include Gram matrix of feature maps (FM), and inner products between the Gram matrix and Cosine similarity, respectively. Table VII shows performance comparison between Gram-based deep correlation features (GDCF) and learnt deep correlation features (LDCF). Table VII(1) shows accuracy obtained based on the Gram-based representation derived from FM, while Table VII(3) shows accuracy obtained based on features learnt from FM. They are both correlations between feature maps. Comparison between them clearly shows that the learnt transformation (Table VII(3)) significantly outperforms inner products (Table VII(1)). Table VII(4) shows the performance obtained by the "meta correlation" transformed from the Gram matrix. The "learnt meta representation" (Table VII(4)) significantly outperforms "handcrafted meta representation" (Table VII(2)). These results verify that correlation between feature maps can be automatically learnt, and such learnt features can more effectively describe image styles. As can be seen, LDCF yields much better performance no matter which input is given. The best performance achieved by GDCF is 61.28%, while the best performance achieved by LDCF is 71.99%.

Fig. 7 shows the confusion matrix of classification results for the OilPainting dataset, obtained based on learnt deep correlation features. We obtain the best performance for Impressionism paintings. On the other hand, more confusions exist between High-Renaissance/Northern-Renaissance/Mannerism (Late Renaissance), between Realism/Romanticism/Symbolism, and between Academicism/Neoclassicism. More elegant features or external knowledge may be utilized to improve performance in the future.

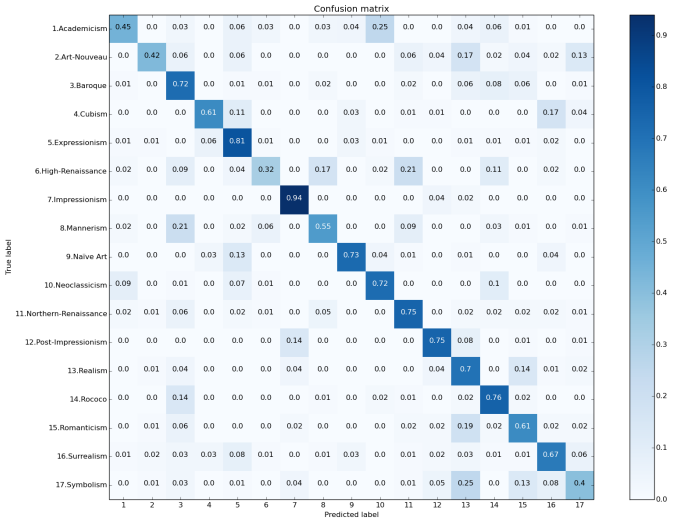

Fig. 7. The confusion matrix of classification results for the OilPainting dataset.

VI. PERFORMANCE COMPARISON

A. Other Painting Datasets

After showing effectiveness of GDCF and LDCF, in this section we will extensively evaluate the proposed features based on more datasets and compare with other state-of-the-art methods.

Table VIII shows performance comparison between our methods and [4] [18], based on the Wikipaintings dataset. In [4], deep features from a fully-connected layer are used as image representation (DeCAF6). They also utilized class confidences of high-level attribute classifiers [34] as image presentation, by further considering the inter-correlation of four aggregated classifier confidence (Fusion $\times$ Content). The work [18] also used Gram matrix of feature maps (with the VGG-16 framework, while we use the VGG-19 framework) as image representation. However, they did not thoroughly study the influence of different types of intra-layer correlations and the inter-layer correlation.

As can be seen from Table VIII, in GDCF, correlation between Gram matrices and Cosine similarity again yields the best performance, which surpasses the most recent results reported in [18]. The last row shows that the learnt deep correlation features further improve classification accuracy significantly. With learnt correlation between correlation, around 71% accuracy can be achieved.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2018.2801718, IEEE Transactions on Multimedia

TABLE VII
AVERAGE ACCURACIES OBTAINED BY GRAM-BASED DEEP CORRELATION FEATURES AND LEARNT DEEP CORRELATION FEATURES. THESE EXPERIMENTS ARE CONDUCTED BASED ON THE OILPAINTING DATASET.

	Correlation btw feature maps	*Correlation btw Correlation*
	Gram matrix of FM	Gram dot Cos.
Gram-based deep corr. features (GDCF)	(1) 58.13%	(2) 61.28%
	learnt from FM	learnt from Gram
Learnt deep corr. features (LDCF)	(3) **65.37%**	(4) **71.99%**

TABLE VIII
AVERAGE ACCURACIES OBTAINED BY DIFFERENT STYLE REPRESENTATIONS, BASED ON THE WIKIPAINTINGS DATASET.

	Fusion × Content [4]	DeCAF6 [4]	[18]
Avg. accuracy	47.30%	35.60%	57.00%
GDCF	Gram matrix of FM	Gram dot Cos.	
Avg. accuracy	56.58%	58.19%	
LDCF	learnt from FM	learnt from Gram	
Avg. accuracy	63.96%	**70.99%**	

TABLE IX
AVERAGE ACCURACIES OBTAINED BY DIFFERENT STYLE REPRESENTATIONS, BASED ON THE PAINTING-91 DATASET.

	[5]	[35]
Avg. accuracy	62.20%	69.21%
GDCF	Gram of FM	Gram dot Cos.
Avg. accuracy	71.86%	73.59%
LDCF	learnt from FM	learnt from Gram
Avg. accuracy	73.20%	**78.27%**

Table IX shows average accuracies obtained based on different style representations, based on the Painting-91 dataset. Khan et al. [5] integrated handcrafted local and global features as image representation, which is surpassed by [35] that considered features extracted from multiple layers of CNN. From the second row, we see that the Gram-based deep correlation features provide clear improvement over [35] (71.86% vs. 69.21%). If correlation between correlations is considered (Gram dot Cos), more improvement (73.59% vs. 69.21%) can be made.

In LDCF, the features learnt from feature maps have similar performance to 'Gram dot Cos'. Unlike performance improvement seen from Table VIII (63.96% vs. 58.19%), in Table IX the learnt features are not significantly better. This may be because the number of training images in the Painting-91 dataset is much fewer (see Table II). Even so, the features learnt from Gram matrices still clearly outperform other settings (78.27%).

To visualize the effectiveness of style classification, we randomly select 100 paintings from each of the five picked styles in the Painting-91 dataset. The styles are Abstract Expressionism, Constructivism, Popart, Post Impressionism, and Surrealism. For each painting image, we take the result of the first fully-connected layer of the model shown in Figure 4 as image representation. This representation is 512-dimensional, and we adopt the t-SNE technique [36] to map it into a 2-dimensional vector. Figure 8 shows the 2-dimensional embedding of the selected painting images. From this figure we clearly see that paintings of the same styles tend to be embedded into a neighboring area. This visualizes effectiveness of the learnt correlation features.

B. Other Types of Datasets

With the success of deep correlation features, we are wondering if these features can also be used in databases other than painting images. In [4], in addition to painting images, they also collected 80K photos from Flickr groups and classified them into 20 styles. These styles may be related to optical techniques (e.g., Macro and Bokeh), atmosphere (e.g., Hazy and Sunny), Mood (e.g., Serene and Melancholy), and so on. We can expect that these styles may not be mutually exclusive, i.e., one image may be related to both Sunny and Romantic. However, because of the collection mechanism of [4], only one label is associated with each image. This is viewed as an unfortunate but acceptable reality of working with a large-scale dataset.

TABLE X
AVERAGE ACCURACIES OBTAINED BY DIFFERENT STYLE REPRESENTATIONS, BASED ON THE FLICKR DATASET.

	Fusion × Content [4]	DeCAF6 [4]
Avg. accuracy	36.80%	33.60%
GDCF	Gram matrix of FM	Gram dot Cos.
Avg. accuracy	41.76%	43.22%
LDCF	learnt from FM	learnt from Gram
Avg. accuracy	48.57%	**60.04%**

With the program provided by the authors of [4], we collected 72,440 photos belonging to 20 styles in total, because some photos have been unlinked from Flickr. Figure 9 shows five sample photos with the styles Bokeh, Bright, Noir, Romantic, and Sunny, respectively. This dataset is adopted to evaluate the proposed deep features.

Table X shows average classification accuracies based on the Flickr dataset. As can be seen, the GDCF outperforms the deep features proposed in [4]. With the proposed learning method, the LDCF significantly improves classification performance. This verifies that the proposed deep correlation features are also effective in discriminating styles of natural images. This may be because texture (which is the information deep correlation features tend to describe) is also an important clue to present different photography styles.

Another type of images we want to evaluate is illustration images, in particular clip art images. There is a vast range of visual styles in clip art images, such as sketches, woodcuts, cartoon, and gradient-shading. We adopt the illustration dataset collected in [37], where 4,591 illustrations of 220 styles are included. Of the 4,591 illustrations, 1,000 illustrations were

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2018.2801718, IEEE Transactions on Multimedia

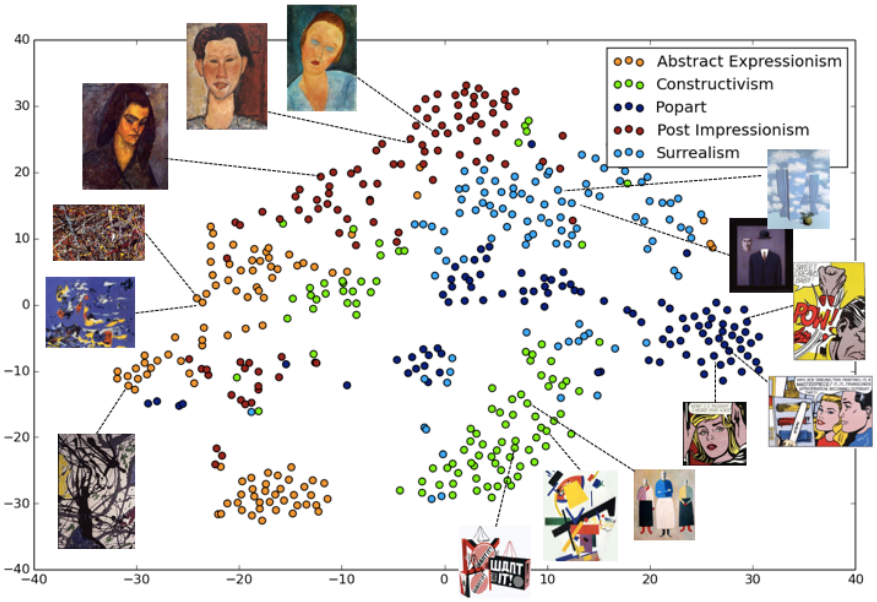

Fig. 8. An illustration showing the two-dimensional distribution of selected painting images in the Painting-91 dataset.

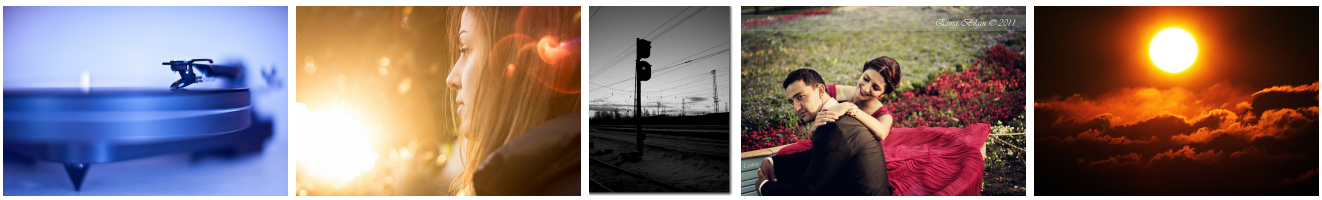

Fig. 9. Five sample photos from the Flickr datasets. From the left to the right, the styles are Bokeh, Bright, Noir, Romantic, and Sunny.

TABLE XI
AVERAGE ACCURACIES OBTAINED BY DIFFERENT STYLE REPRESENTATIONS, BASED ON THE ILLUSTRATION DATASET.

GDCF	Gram matrix of FM	Gram dot Cos.
Avg. accuracy	66.73%	68.57%
LDCF	learnt from FM	learnt from Gram
Avg. accuracy	71.83%	**72.44%**

collected from the Art Explosion dataset[2], and the remaining 3,591 illustrations were from the clipart included in Microsoft Office. Figure 10 shows sample illustrations of five different styles. We can clearly perceive style difference between different sample illustrations, though the difference or the definition of style cannot be verbally stated.

Table XI shows average classification accuracies based on the illustration dataset. We again see the superiority of LDCF over GDCF. Overall, the classification accuracy is quite promising (over 72% accuracy for a dataset of 220 styles).

C. Artist Classification

Different artists have different skills and preferences to make their artworks. Classifying paintings based on style representation can thus be applied to do artist classification. Here we evaluate the proposed style representations based on two artist datasets, i.e., the OilPainting Artist dataset and Painting-91 Artist dataset mentioned in Table II.

Figure 11 shows some paintings produced by two artists, Paul Cezanne and Vincent van Gogh. Their paintings are famous artworks categorized into Expressionism. However, from Figure 11 we can implicitly perceive difference in painting styles and topics.

Table XII shows average classification accuracies for the OilPainting Artist dataset, which consists of 105 artists. The reported average accuracy is obtained based on the five-fold cross validation scheme. Table XII again shows the superiority of deep correlation features. Correlation between Gram matrices and Cosine similarity yields 63.17% accuracy that significantly outperforms the fully-connected layer (55.59%). The learnt deep correlation features further improve performance to 64.65% and 69.74% based on the schemes of learning from feature maps and learning from Gram matrices, respectively.

The Painting-91 Artist dataset consists of 91 artists, and each artist has 30 to 50 painting images. Table XIII shows performance comparison between ours (GDCF and LDCF) and [5][35]. By considering the correlation between Gram matrices and Cosine similarity, average accuracy 63.17% can be obtained, which is better than prior studies [5] and [35].

[2]http://www.novadevelopment.com

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2018.2801718, IEEE Transactions on Multimedia

Fig. 10. Sample illustrations of five different styles.

TABLE XII
AVERAGE ARTIST CLASSIFICATION ACCURACIES OBTAINED BY DIFFERENT STYLE REPRESENTATIONS, BASED ON THE OILPAINTING ARTIST DATASET.

GDCF	fc7	Gram matrix of FM	Gram dot Cos
Avg. Accuracy	55.59%	60.90%	63.17%
LDCF	–	learnt from FM	learnt from Gram
Avg. Accuracy	–	64.65%	**69.74%**

TABLE XIII
AVERAGE ARTIST CLASSIFICATION ACCURACIES OBTAINED BY DIFFERENT STYLE REPRESENTATIONS, BASED ON THE PAINTING-91 ARTIST DATASET.

	[5]	[35]	
Avg. Accuracy	53.10%	56.40%	
GDCF	fc7	Gram matrices of FM	Gram dot Cos
Avg. Accuracy	55.59%	60.90%	63.17%
LDCF	–	learnt from FM	learnt from Gram
Avg. Accuracy	–	61.28%	**64.32%**

The best average accuracy 64.32% is obtained based on LDCF (learnt from Gram) for this challenging dataset. Comparing with other experiments, the extent of improvement brought by LDCF is smaller. This again is due to the small volume of training data.

D. Influence of Volume of Training Data

From Table IX and Table XIII, we see that performance improvement tends to be smaller when training data are fewer. To quantitatively verify this, we combine the collected OilPainting dataset and the WikiPaintings dataset [4]. Only images of the 16 styles common in both datasets are combined, and finally the combined dataset contains 90,572 images. From each style, 80% of images is randomly selected in the set X and the remaining 20% is put into the set Y. We intentionally select 10%, 20%, ..., and 100% of X to respectively construct ten learning models for LDCF extraction (learnt from Gram matrices), which are then used to do style classification for the images in Y. Figure 12 shows the relationship between classification accuracy and the number of training data (in percentage). As can be seen, this curve clearly shows the influence of the number of training data on the performance of learnt deep correlation features.

VII. CONCLUSION

Inspired by the interesting work [7] that showed the effectiveness of correlation between feature maps, we transform such correlations into style vectors, and utilize them to achieve image style classification. We comprehensively study performance variations brought by correlations in different layers, performance variations of different correlations, and the idea of inter-layer correlation. In addition to correlations commonly defined in statistics, we further explore the possibility of learning correlation between feature maps automatically based on a neural network. Based on experimental results on various datasets, including the ones other than painting images, we demonstrate that in most cases the learnt deep correlation features yield better performance by a large margin.

In the future, in addition to deep correlation features that mainly capture texture information, we will explore the impact of semantics or image attributes on image style classification. Furthermore, as shown in Fig. 7, some painting styles are more confused and hardly to be discriminated. To combat this issue, hierarchical approaches, i.e., doing rough classification first and then finer classification, may be helpful.

ACKNOWLEDGMENT

This work was partially supported by the Ministry of Science and Technology of Taiwan under the grant MOST 105-2628-E-194-001-MY2 and MOST 106-3114-E-002-009.

REFERENCES

[1] D. Borth, R. Ji, T. Chen, T. Breuel, and S.-F. Chang, "Large-scale visual sentiment ontology and detectors using adjective noun pairs," in *Proceedings of ACM Multimedia Conference*, 2013.
[2] Y. Wang, S. Li, and A. C. Kot, "On branded handbag recognition," *IEEE Transactions on Multimedia*, vol. 18, no. 9, pp. 1869–1881, 2016.
[3] J. Tarvainen, M. Sjoberg, S. Westman, J. Laaksonen, and P. Oittinen, "Content-based prediction of movie style, aesthetics, and affect: Data set and baseline experiments," *IEEE Transactions on Multimedia*, vol. 16, no. 8, pp. 2085–2098, 2014.
[4] S. Karayev, M. Trentacoste, H. Han, A. Agarwala, T. Darrell, A. Hertzmann, and H. Winnemoeller, "Recognizing image style," in *Proceedings of British Machine Vision Conference*, 2014.
[5] F. Khan, S. Beigpour, J. van de Weijer, and M. Felsberg, "Painting-91: A large scale database for computational painting categorization," *Machine Vision and Application*, vol. 25, no. 6, pp. 1385–1397, 2014.
[6] T.-E. Tseng, W.-Y. Chang, C.-S. Chen, and Y.-C. F. Wang, "Style retrieval from natural images," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2016.
[7] L. A. Gatys, A. S. Ecker, and M. Bethge, "A neural algorithm of artistic style," Aug 2015. [Online]. Available: http://arxiv.org/abs/1508.06576
[8] L. Gatys, A. S. Ecker, and M. Bethge, "Texture synthesis using convolutional neural networks," in *Proceedings of Neural Information Processing Systems*, 2015.
[9] D. Ulyanov, V. Lebedev, A. Vedaldi, and V. Lempitsky, "Texture networks: Feed-forward synthesis of textures and stylized images," in *Proceedings of International Conference on Machine Learning*, 2016.
[10] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proceedings of European Conference on Computer Vision*, 2016.
[11] A. J. Champandard, "Semantic style transfer and turning two-bit doodles into fine artworks," Mar 2016. [Online]. Available: https://arxiv.org/abs/1603.01768
[12] V. Dumoulin, J. Shlens, and M. Kudlur, "A learned representation for artistic style," Dec 2016. [Online]. Available: https://arxiv.org/abs/1610.07629

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2018.2801718, IEEE Transactions on Multimedia

Fig. 11. Five sample images of two different artists, Paul Cezanne (top) and Vincent van Gogh (bottom).

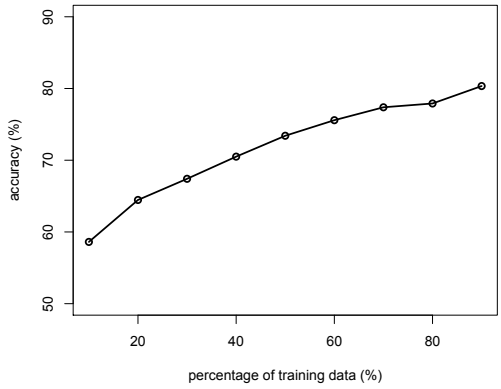

Fig. 12. The relationship between classification accuracy and the number of training data.

[13] M. Ruder, A. Dosovitskiy, and T. Brox, "Artistic style transfer for videos," in *Proceedings of German Conference on Pattern Recognition*, 2016, pp. 26–36.
[14] R. Tanno, S. Matsuo, W. Shimoda, and K. Yanai, "Deepstylecam: A real-time style transfer app on ios," in *Proceedings of International Conference on Multimedia Modeling*, 2017, pp. 446–449.
[15] PRISMA. (2016) Prisma. [Online]. Available: http://prisma-ai.com
[16] Y. Li, N. Wang, J. Liu, and X. Hou, "Demystifying neural style transfer," Jan 2017. [Online]. Available: https://arxiv.org/abs/1701.01036
[17] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis," Jan 2017. [Online]. Available: https://arxiv.org/abs/1701.02096
[18] S. Matsuo and K. Yanai, "Cnn-based style vector for style image retrieval," in *Proceedings of ACM International Conference on Multimedia Retrieval*, 2016, pp. 309–312.
[19] W.-T. Chu and Y.-L. Wu, "Deep correlation features for image style classification," in *Proceedings of ACM International Conference on Multimedia*, 2016, pp. 402–406.
[20] T. Blog. (2016) The major advancements in deep learning in 2016. [Online]. Available: https://tryolabs.com/blog/2016/12/06/major-advancements-deep-learning-2016/
[21] A. Jain and G. Healey, "A multiscale representation including opponent color features for texture recognition," *IEEE Transactions on Image Processing*, vol. 7, no. 1, pp. 124–128, 1998.
[22] T.-Y. Lin, A. RoyChowdhury, and S. Maji, "Bilinear cnn models for fine-grained visual recognition," in *Proceedings of IEEE International Conference on Computer Vision*, 2015, pp. 1449–1457.
[23] L.-Y. Wei and M. Levoy, "Fast texture synthesis using tree-structured vector quantization," in *Proceedings of ACM SIGGRAPH*, 2000, pp. 479–488.
[24] W. Zhang, C. Cao, S. Chen, J. Liu, and X. Tang, "Style transfer via image component analysis," *IEEE Transactions on Multimedia*, vol. 15, no. 7, pp. 1594–1601, 2013.
[25] I.-C. Shen and W.-H. Cheng, "Gestalt rule feature points," *IEEE Transactions on Multimedia*, vol. 17, no. 4, pp. 526–537, 2015.
[26] X. Lu, Z. Lin, X. Shen, R. Mech, and J. Z. Wang, "Deep multi-patch aggregation network for image style, aesthetics, and quality estimation," in *Proceedings of IEEE International Conference on Computer Vision*, 2015.
[27] X. Lu, Z. Lin, H. Jin, J. Yang, and J. Z. Wang, "Rating image aesthetics using deep learning," *IEEE Transactions on Multimedia*, vol. 17, no. 11, pp. 2021–2034, 2015.
[28] G. Folego, O. Gomes, and A. Rocha, "From impressionism to expressionism: Automatically identifying van gogh's paintings," in *Proceedings of IEEE International Conference on Image Processing*, 2016.
[29] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proceedings of International Conference on Learning Representation*, 2015.
[30] A. Vedaldi and K. Lenc, "Matconvnet: Convolutional neural networks for matlab," in *Proceedings of ACM International Conference on Multimedia*, 2015, pp. 689–692.
[31] D. Wang, P. Cui, M. Ou, and W. Zhu, "Learning compact hash codes for multimodal representations using orthogonal deep structure," *IEEE Transactions on Multimedia*, vol. 17, no. 9, pp. 1404–1416, 2015.
[32] Y. Hua, S. Wang, S. Liu, A. Cai, and Q. Huang, "Cross-modal correlation learning by adaptive hierarchical semantic aggregation," *IEEE Transactions on Multimedia*, vol. 18, no. 6, pp. 1201–1216, 2016.
[33] M. D. Zeiler, "Adadelta: An adaptive learning rate method," Dec 2012. [Online]. Available: https://arxiv.org/abs/1212.5701
[34] S. Dhar, V. Ordonez, and T. L. Berg, "High level describable attributes for predicting aesthetics and interestingness," in *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, 2011.
[35] K.-C. Peng and T. Chen, "Cross-layer features in convolutional neural networks for generic classification tasks," in *Proceedings of IEEE International Conference on Image Processing*, 2015.
[36] L. van der Maaten and G. Hinton, "Visualizing high-dimensional data using t-sne," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008.
[37] E. Garces, A. Agarwala, D. Gutierrez, and A. Hertzmann, "A similarity measure for illustration style," *ACM Transactions on Graphics*, vol. 33, no. 4, p. Article No. 93, 2014.

Wei-Ta Chu received the B.S. and M.S. degrees in Computer Science from National Chi Nan University, Taiwan, in 2000 and 2002, and received the Ph.D. degree in Computer Science from National Taiwan University, Taiwan, in 2006. He is now a Professor in the Department of Computer Science and Information Engineering, National Chung Cheng University, Taiwan. His research interests include digital content analysis, multimedia indexing, deep learning, and pattern recognition.

He won the Best Full Technical Paper Award in ACM Multimedia 2006. He was awarded Outstanding Youth Electrical Engineer Award by the Chinese Institute of Electrical Engineering in 2017, the Distinguished Alumni Award presented by National Chi Nan University in 2014, Best GOLD Member Award presented by IEEE Tainan Section in 2013, the K. T. Li Young Researcher Award presented by Institute of Information & Computing Machinery in 2012, and the Young Faculty Awards presented by National Chung Cheng University in 2011. He was a visiting professor at Nagoya University from January to March 2017, and a visiting scholar at Digital Video & Multimedia Laboratory, Columbia University, from July to August 2008. He is an associate editor of IEICE Transactions on Information and Systems since 2016. He serves as Publication Co-Chair of International Computer Symposium 2016, Publication Co-Chair of IEEE International Conference on Visual Communications and Image Processing 2018, and Program Co-Chair of International Conference on Multimedia Modelling 2020.

Yi-Ling Wu received her B.S. degree in Computer Science from Tamkang University, Taiwan, in 2015, and received M.S. degree in Computer Science from National Chung Cheng University, Taiwan, in 2017. Her research interests include image processing, computer vision, and machine learning.