From Part to Whole: Who is Behind the Painting?

Daiqian Ma[1,2], Feng Gao[2], Yan Bai[1,2], Yihang Lou[1,2], Shiqi Wang[3], Tiejun Huang[2], Ling-Yu Duan[2] *
SECE of Shenzhen Graduate School, Peking University, Shenzhen, China[1]
National Engineering Lab for Video Technology, Peking University, Beijing, China[2]
Department of Computer Science, City University of Hong Kong, Hong Kong, China[3]
{madaiqian,gaof,yanbai,yihanglou,tjhuang,lingyu}@pku.edu.cn,shiqwang@cityu.edu.hk

ABSTRACT

Compared with normal modalities, the representations of paintings are much more complex due to its large intra-class and small inter-class variation. This poses more difficulties in the task of authorship identification. In this paper, we propose a multi-task multi-range (MTMR) representation framework and try to resolve this issue in two ways. First, we investigate how to improve the representation through multi-task learning. Specifically, we attempt to optimize authorship identification with subtly correlated identification tasks such as style, genre and date. Second, in order to make the representation more comprehensive and reduce the information loss from image scaling, we propose a multi-range structure which is composed of local, regional and global representations. Experiments on the two most representative large-scale painting datasets, Rijksmuseum Challenge and Wikiart, have shown that our method significantly outperforms the existing methods. To give better understanding and provide more effective predictions, we utilize random forest as the feature ranking method to analyze the importance of different features and apply external knowledge matching to further examine the predictions. Moreover, the framework's effects of identifying the authorship are visualized on the paintings' artist-characteristic regions and t-SNE is further applied to perform artist-based cluster analysis. Extensive validation has demonstrated that the proposed framework yields superior performance in the chanllenging task of painting authorship identification.

KEYWORDS

Painting; multi-task learning; multi-range representation; external knowledge matching; random forest;

1 INTRODUCTION

In the past few years, digitized fine-art collections have been growing rapidly due to the popularity of digital technology. In the evolution of fast growing large art work datasets, fine art categorization problem becomes an emerging research area in computer vision research. In particular, for online galleries, there is an urgent need to analyze, classify and understand the paintings in an automatic way. One of the crucial information is the authorship, which plays an important role in the process of unknown-painting identification.

(a) *Sunrise, Whiting Fishing* (b) *Battle of Krasnaya Gorka*

(c) *The barges in Bezons* (d) *Norman city*

Figure 1: The author of (b)(d) is Alexey Bogolyubov, which shows the large variation within the paintings of the same artist. (a)(c) are drawn by Charles-Francois Daubigny and William Turner, respectively. From (a)(b)(c), we can see the similar content(sky, ship) in different artists' paintings.

Many researchers have been trying to solve painting identification problem by computational methods. Some of them [1, 11, 18, 19, 22, 23, 27] utilized low-level features encoding with color, shadow, boundaries and shapes. Li *et al.* [11] designed a novel extraction method by exploiting an integration of edge detection and clustering-based segmentation to distinguish van Gogh's paintings in different time periods. Tseng *et al.* [27] proposed a ranking method for style identification based on random forests. In addition, Puthenputhussery *et al.* [18, 19] proposed a fusion method of different fisher vector encodings and achieved remarkable performance. These studies have shown low-level features, especially typical local features, are useful in painting identification.

Recently, deep networks [3, 7, 8, 10, 20, 26, 28, 29] have brought about more interesting applications in this topic. Firstly Karayev *et al.* [10] observed that Convolutional Neural Network (CNN) features outperform hand-crafted features like color histogram and

*Ling-Yu Duan is the corresponding author.
Daiqian Ma and Feng Gao are joint first authors.

MM '17, October 23–27, 2017, Mountain View, CA, USA.
 ISBN 978-1-4503-4906-2/17/10…$15.00
DOI: https://doi.org/10.1145/3123266.3123325

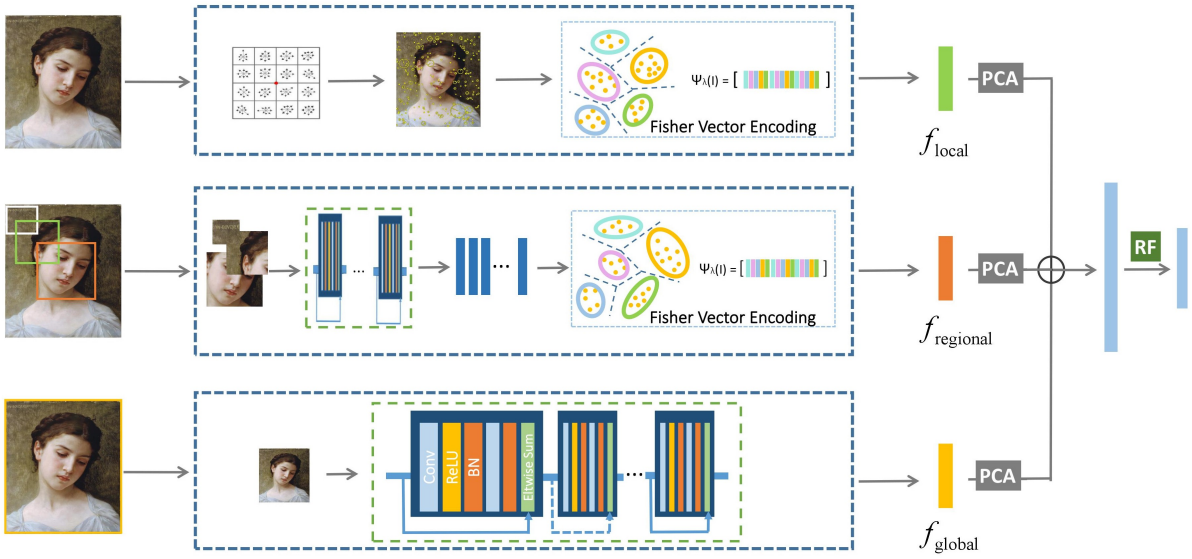


Figure 2: An overview of the proposed MTMR representation framework. Here, the local, regional and global features are extracted from the SIFT-based fisher vector, multi-size region encoding structure and multi-task learning structure, respectively. The network structure in green dashed box employs the residual network. RF represents random forest.

GIST for fine-art classification. Then Chu *et al.* [3] designed and transformed various layer-correlation inside CNN into style vectors and investigated classification performance brought by different variants. More recently, Noord *et al.* [29] designed a multi-scale network to obtain scale-invariant features of paintings. And Jangtjik *et al.* [8] proposed a new weighted scheme to adaptively combine the decision results from different scales. From these studies, we can find that deep networks achieve promising performance in painting identification from the global perspective.

However, a single modality representation is less sufficient to express the entire painting, as an artist may own many different styles of paintings and different artists may create similar contents. For example, Pablo Picasso presented good interest in very rich subjects of every kind and demonstrated a great stylistic versatility that enabled him to work in several styles at once. As shown in Fig. 1, we select several representative paintings from Wikiart. It is extremely difficult to exactly identify the authorship and distinguish paintings of the same artist solely from their styles.

To overcome this problem, some researchers paid more attentions to the combination of local and global features [20, 24]. Saleh *et al.* [20] proposed a unified framework consisting of GIST and CNN features for painting classification. Sheng *et al.* [24] designed a method of combining histogram-based local and global features to characterize different aspects of art styles. Though these methods have achieved promising performance, most of them are still in the circle of traditional local features and global CNN features, ignoring the impact of regional features. Moreover, we find that current works just treat the authorship identification task as a single and independent problem, but actually, authorship identification is a complex procedure involving with many related tasks like painting's style, genre and canvas material. In addition, we find that though the accuracy can be improved by combined features, it is still unclear which kind of features are more effective in the procedure of classification. Therefore, evaluating the importance of features in a technically solid way and visualizing the role of different representation are also meaningful for painting classification.

In this paper, to address these issues, we propose a multi-task multi-range (MTMR) representation framework. On the one hand, we aim to enhance the network's representation ability by joint learning with related tasks (style, material, date and etc.). On the other hand, we are trying to make the paintings' representations richer and more diverse with multi-range features. Our main contributions are summarized as follows:

1) We propose a comprehensive multi-range representation structure, composed of local, regional and global representations to identify paintings at a fine granularity. To the best of our knowledge, this is the first attempt of investigating how painting's authorship identification task can be addressed in an optimal way, together with heterogeneous but subtly correlated tasks.
2) We evaluate the individual role or importance of the proposed three multi-range features in painting identification by random forest, and have revealed their complementarity effects.
3) Extensive experimental results have shown that our framework achieves superior performance on the authorship identification. Moreover, the artist-cluster results further demonstrate that the proposed representation framework is also appropriate to be further applied in painting retrieval.

The rest of this paper is organized as follows. We first introduce our multi-task learning part in Section 2. Section 3 presents the details of multi-range representation part. And appraisal methods are described in Section 4. Then we demonstrate its performance on two large-scale datasets, Rijksmuseum Challenge dataset and

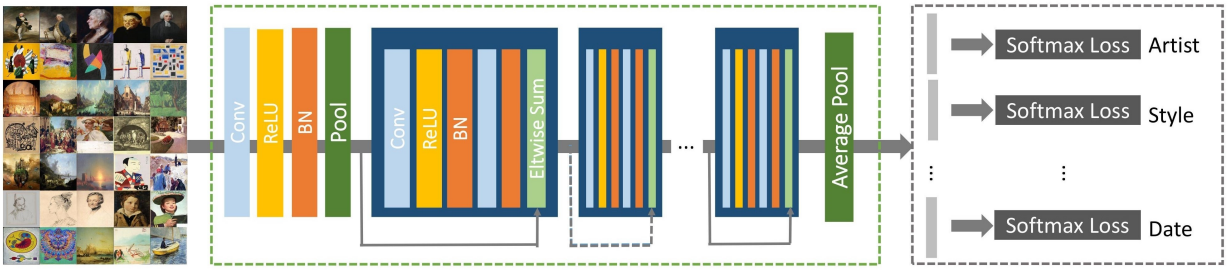


Figure 3: Specification for multi-task learning structure in MTMR representation framework. The part in grey dashed box indicates the multi-task split. The structure in green dashed box is residual network. The grey solid arrows represent identity shortcuts and dashed arrows show a 1x1 convolution with stride 2 to match spatial resolution and feature dimension. The right grey strips are fully connected layers and their lengths are flexible with the class number of related tasks.

Wikiart dataset, and further analyze the influences of this framework on the representation of the paintings. Finally, we conclude this paper in Section 6.

2 MULTI-TASK LEARNING

In this section, as shown in Fig. 3, we investigate how to improve the representation for paintings by multi-task learning.

2.1 Joint Loss Formulation

Here we aim to optimize the main task m, which is on authorship identification, with the assistances of several related/auxiliary tasks $a \in A$. Examples of related tasks include the recognition of paintings' styles, dates, materials and so on. To this end, we design a weighted joint loss L as:

$$L = -\frac{1}{N}\sum_{i=1}^{N}(L^m + \sum_{a\in A}\lambda^a L^a), \tag{1}$$

where N denotes the total number of training images, L^m and L^a represent the loss function of main task and relative auxiliary tasks. λ^a denotes the importance coefficient of task a. Let $\hat{f}_i^m$ and $\hat{f}_i^a$ denote the predicted scores of i-th painting for its ground-truth labels of main task and related tasks in softmax function. Thus the weighted joint loss L can be rewritten as:

$$L = -\frac{1}{N}\sum_{i=1}^{N}(\log(\hat{f}_i^m) + \sum_{a\in A}\lambda^a \log(\hat{f}_i^a)). \tag{2}$$

During training, the errors propagated backwards from these branches are linearly combined and the weights of the shared layers will be updated accordingly.

2.2 Architecture Analysis

There're several typical multi-task learning structures mentioned in previous works [2, 5, 32], such as parallel model, cross-product model, late branching model and early branching model. In this paper, inspired by these structures, we prefer to branch from the last average pool layer of the network and add one target-specific fully connected layer prior to multi-task prediction. The learning procedure is guided by joint softmax loss. Since our main target is to learn a better representation for authorship identification, we take residual network [6] owing to its state-of-the-art performance on several challenging recognition tasks.

Another significant point is the choice of related/auxiliary tasks. Originally we plan to take painting's style, genre, date and material into consideration, but there are not enough labels in existing datasets. With careful comparison, we focus on how to leverage the influence of style and date in authorship identification. In addition, as shown in Fig. 2, there're two residual networks trained by multi-task learning in MTMR framework. One is a 10-layer residual network trained with certain cropped painting patches for extracting regional features, while the other is a 50-layer residual network trained with scaled paintings for extracting global features.

3 MULTI-RANGE REPRESENTATION

In this section, we introduce the multi-range representation structure, as shown in Fig. 2. The typical SIFT-based fisher vector is applied for extracting local representation, and deep residual networks with fisher vector is employed to obtain more efficient regional and global representations.

3.1 SIFT-Based Fisher Vector

As an advanced encoding method, fisher vector(FV) [21] has outperformed the other encoding approches on many image challenge benchmarks. We adopt Fisher vector to encode the local features (i.e. SIFT [13]) to form the local representation of paintings.

The FV encoding starts from extracting dense SIFT descriptors, which have been popularly used for image classification task. We extract dense features from sampled patches (every 8 pixels) with fixed scale and upright orientation. They are first de-correlated by PCA (from 128 to 64) to make the dense features more amenable to the FV description based on the diagonal-covariance GMM. Then we train a Gaussian Mixture Model (GMM) with diagonal covariances, and only the derivatives with respect to the Gaussian mean and variances (64 centroids) are considered. This leads to the representation which captures the average first and second order differences between the features and each GMM centre. Specifically, the

derivation of FV is as follows,

$$\Phi^{(1)} = \frac{1}{N\sqrt{w_k}} \sum_{p=1}^{N} \alpha_p(k) \left(\frac{x_p - \mu_k}{\sigma_k} \right), \tag{3}$$

$$\Phi^{(2)} = \frac{1}{N\sqrt{2w_k}} \sum_{p=1}^{N} \alpha_p(k) \left(\frac{(x_p - \mu_k)^2}{\sigma_k} - 1 \right), \tag{4}$$

where $\{w_k, \mu_k, \sigma_k\}$ are the mixture weights, means, and diagonal covariances of the GMM, and $\alpha_p(k)$ is the soft assignment weight of the p-th feature x_p to the k-th Gaussian. The FV ϕ is obtained by stacking the differences: $\phi = \left[\Phi_1^{(1)}, \Phi_1^{(2)}, ..., \Phi_K^{(1)}, \Phi_K^{(2)}\right]$. Finally, following [17], the performance of an FV can be further improved by passing it through power normalization and $\ell 2$-normalization.

3.2 Regional Feature Encoding

In contrast to dense SIFT, the image patches are expected to be represented from a higher perspective with deep networks, which motivates us to design this structure to extract deep regional features. It is related to the one proposed by Cimpoi *et al.* [4]. Here we apply the global average pooling layer of residual network to extract dense features. It is much more efficient because global average pooling makes it possible to produce fixed-length features.

In addition, the choice of the region size is flexible. The paintings are first split into regions with size s, then we put each region into a pre-trained 10-layer residual network and obtain features f_s from the last *global_pool* layer. With the change of region size, we are able to obtain fixed-length multi-scale regional features. Considering the variation of painting size, the number of patches may vary, we apply fisher vector to produce a single representation. This process is similar as the procedures discussed in Section 2.1, we fit the features into GMM and encode these features by FV. Here it should be noted that this structure is able to get rid of the limitation of image size and produce much more robust regional features.

4 APPRAISAL METHOD

In this section, the importance of each feature is evaluated by random forest and the capacity of the framework for painting identification is analyzed. Moreover, we propose an external knowledge matching method to further examine the predictions.

4.1 Fusion Strategy

The classification ability of our framework is analyzed in this subsection. For fair comparison, we apply linear support vector machine as the classifier. Owing to the promising properties of SVM for dealing with high dimensionality data, we choose concatenation as our main fusion method. Suppose the local, regional and global features as f_l, f_r and f_g, respectively. We first do PCA and whitening operation to reduce the redundancy:

$$f' = diag(1./sqrt(v_1, v_2, ..., v_h)) * U * f, \tag{5}$$

where U is the PCA transformation matrix, h is the number of retained dimensions and v_i is the ith corresponding singular value. Then we perform $\ell 2$ normalization $f'' = \frac{f'}{\|f'\|_2}$, and concatenate

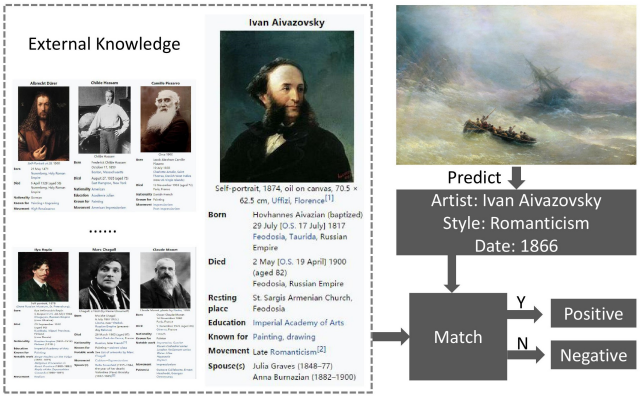


Figure 4: An example of external knowledge matching. The information of these artists are collected from wikipedia. It takes the matched prediction as positive prediction and unmatched prediction as negative prediction.

them directly. It can be represented as:

$$f_{final} = [f_l'', f_r'', f_g'']. \tag{6}$$

Finally, a zero mean-unit variation feature normalization is preformed on f_{final} to obtain the final representation for a painting.

4.2 External Knowledge Matching

Since the painting's authorship identification is a comprehensive subject involving with the judgement of date, style, genre, material and so on, we try to increase the prediction's reliability by matching the predictions with external knowledge. As shown in Fig. 4, the painting's prediction is examined by the matching result. It should be mentioned that this is a posteriori procedure.

Specifically, we design a simplified matching rule. If the painting's predicted date is in the range of predicted artist's lifetime, we judge it matched and take it as a positive prediction. It's hard to take style into consideration because we are not sure whether the predicted artist once tried the predicted style or not. More professional knowledge of art is required to design a more reasonable matching rule. Here we examine the predictions with this simplified rule and the results are listed in the following experiments.

4.3 Feature Ranking

It is necessary to find a way of assessing the importance of these three different kinds of features. Random forest [12] is among the most popular machine learning methods thanks to its relatively good accuracy, robustness and ease of use.

Random forest consists of a number of decision trees. Each node in the decision trees is a condition on a single feature, designed to split the dataset into two so that similar response values end up in the same set. The measure based on which the (locally) optimal condition is chosen is called impurity. For classification, it is typically either Gini impurity or information gain/entropy and for regression trees it is variance. When training a tree, the weighted impurity decrease in a tree for each feature can be computed. For

Table 1: Performance comparisons on Wikiart dataset (evaluated by accuracy, r_x indicates that the size of the region is (x,x)).

Saleh[20]	Tan[26]	ResNet-50	f_{local}	$f_{\text{r\_64}}$	$f_{\text{r\_128}}$	$f_{\text{r\_224}}$	$f_{\text{r\_mix}}$	f_{global}	$f_l + f_r$	$f_l + f_g$	$f_r + f_g$	f_{final}
63.1	76.1	80.6	51.6	63.8	71.4	78.4	80.1	82.2	80.4	82.5	88.3	**88.6**

Table 2: Performance comparisons on Wikiart dataset with different auxiliary tasks (here A, S, D represents Artist, Style and Date, respectively).

Task	A	A+S	A+D	A + S + D
Result	80.6	81.9	81.0	**82.2**

Table 3: Performance comparisons on Rijksmuseum Challenge dataset (evaluated by mean class accuracy, the number of artist refers to the data split in [28]).

Method	Artist			
	34	97	197	958
Noord'15[28]	78.3	74.5	68.2	52.5
ResNet-50	92.3	87.1	80.7	61.7
f_{local}	76.2	73.9	65.4	51.3
$f_{\text{r\_32}}$	79.2	74.3	67.1	50.1
f_{global}	92.9	88.2	81.5	62.7
$f_l + f_r$	86.6	81.4	74.6	56.7
$f_l + f_g$	93.3	88.7	84.4	69.2
$f_r + f_g$	93.7	89.1	85.3	69.1
f_{final}	**94.0**	**89.8**	**85.5**	**69.6**

a forest, the impurity decrease from each feature can be averaged and the features are ranked according to this measure.

In our experiments, we choose 300 decision trees to compose the random forest and finally top-1000 features are treated as the representation of the painting.

5 EXPERIMENTS

5.1 Datasets and Metrics

Rijksmuseum Challenge dataset [15] and Wikiart dataset [1] are the two most representative large-scale online public collections of digitized artworks. The Rijksmuseum Challenge dataset consists of 112,039 digital photographic artworks by 6,629 artists exhibited in Rijksmuseum in Amsterdam, and the Wikiart dataset has images of 81,449 fine-art paintings from 1,119 artists ranging from fifteen centuries to contemporary artists. The splitting of the datasets in our experiments is identical with the settings in [20, 28].

For fair comparison, we apply mean class accuracy as the comparison protocol on Rijksmuseum Challenge dataset, which is the same as [28] and we choose accuracy as the comparison standard on Wikiart dataset, which is the same as [20, 26].

[1] http://www.wikiart.org/

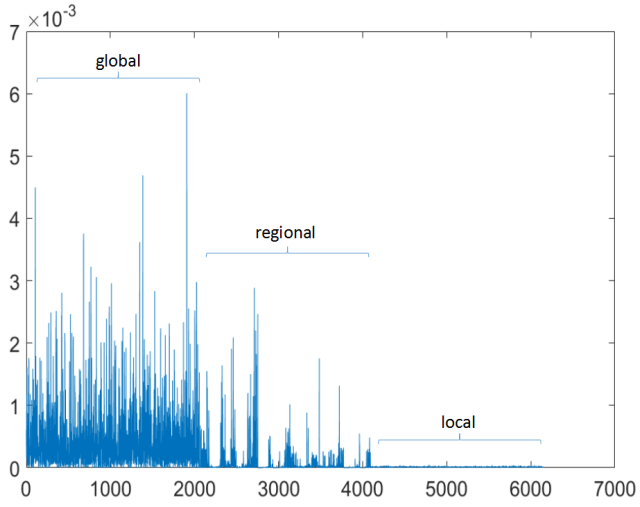


Figure 5: Feature importance evaluated with random forest on Wikiart dataset. From left to right, they are global features, regional features and local features, respectively.

5.2 Implementation Details

Caffe [9] is adopted as our deep learning platform. The mentioned 50-layer and 10-layer residual networks [25] are both trained for 100,000 iterations on 4 Nvidia K80 GPUs with multi-task learning framework. Regional and global features are extracted from the *global_pool* layer in ResNet-10 and ResNet-50. Moreover, we apply the vlfeat toolbox [30] to extract dense SIFT descriptors and perform FV encoding for these descriptors. The linear support vector machine is implemented by scikit-learn [16].

5.3 Results and Analysis

5.3.1 Comparision and evaluation. Our MTMR framework is compared with the representative existing methods on these two datasets. In particular, we set a 50-layer residual network as baseline for better comparison. The experimental results are presented in Tables 1 and 2. The best performance is shown in bold.

It is not surprising to see that deeper structure can yield better performance. Compared with general networks, residual network achieves an obvious performance improvement on both datasets. Moreover, comparing the results of single resnet-50 and multi-task resnet-50 (shown as f_{global}), we can find that multi-task learning is able to further provide better performance with around 1.6% in Wikiart dataset. From the results in Table 2, we can see the contribution of style information is much more obvious than date

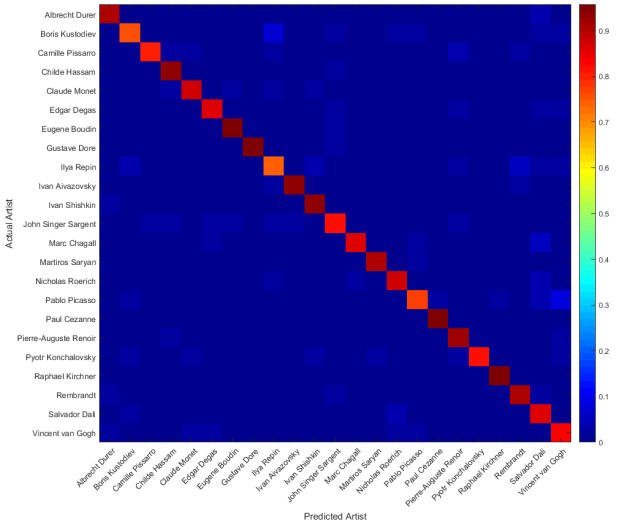


Figure 6: Confusion matrix of Wikiart dataset. The color corresponds to the percent of the artist on vertical axis having been predicted as the artist on horizontal axis. Please zoom in to see details.

information with around 0.9% gain. It should be mentioned that a preferable coefficient setting is 0.6, 0.35 and 0.05 for artist, style and date, respectively. We can also observe the influences of region's size for regional presentation. In Wikiart dataset, the average size of the paintings is around 1500x1500, and therefore we investigate the performance impact of 64x64, 128x128 and 224x224 regions. From the results, we can find that relatively bigger size brings better results and the multi-size feature achieves the best. Here we take the multi-size feature $f_{\text{r\_mix}}$ as f_r for subsequent fusion. As Rijksmuseum Challenge dataset contains many extremely thin paintings (only 42 pixels), we could just take 32x32 regions for test. So the role of regional features is limited in Rijksmuseum Challenge dataset.

Subsequently, we investigate the influences for unbalanced data distribution. As described in [28], we split the Rijksmuseum Challenge dataset according to the number of paintings per artist. Thus the smaller number of artists implies the larger number of the paintings per artist. There is no doubt that unbalanced data distribution may make the training more difficult, which is also exactly what we are facing in reality. From the results in Table 3, we can find the increasing performance with the growth of the number of artist. It shows that our framework is quite effective to deal with this under-training situation. Regional information can make further distinctions on the recognition of the artists who own few paintings. Finally, the complementary effect of fused features is studied. The results on these two datasets show that the combination of different range features obviously boosts the performance further, which achieves nearly 8.0% improvement. All the above-mentioned observations show that, just as the procedure of professional art authentication in reality, observing the painting in multiple views is better than single view.

Besides the numerical results, we also evaluate the importance of the features by random forest. As mentioned, global features are extracted from *global_pool* layer of multi-task resnet-50 with dimensionality of 2048. For fair comparison, local and regional features are also compressed to 2048 with PCA operation. The evaluation results are depicted in Fig. 5. It shows that, from the view of single feature's effect, global features are much more effective than others. But it can not fully represent multiple features' mutual effects because different features' recognizing ranges are different. Regional and local features tend to be more effective for certain kinds of paintings. Recalling the results on Table 1, we can conclude that for large-size painting classification, global features and regional features play dominant roles for authorship identification, while the influence of traditional local features is very limited.

In summary, from above results, we can find that multi-task learning makes the representation more robust and effective. Multi-range representation is able to improve identification capacity on unbalanced datasets and shows remarkable performance improvement for high-resolution paintings.

5.3.2 Results with external knowledge matching.

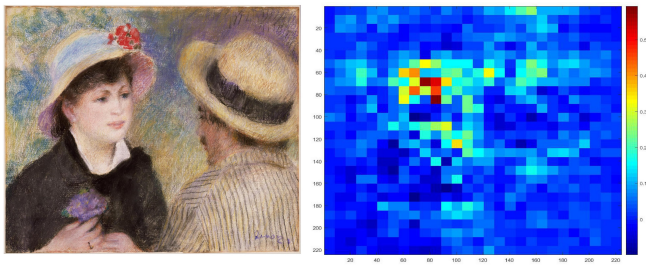
$$P = \frac{TP}{TP + FP} = 90.1\%, \quad R = \frac{TP}{TP + FN} = 75.6\%. \tag{7}$$

The matching is implemented in Wikiart dataset. Here TP, FP, FN represents true positive predictions, false positive predictions and false negative predictions, respectively. P denotes precision and R denotes recall. From the results, we can find that positive predictions (90.1%) show better performance than normal predictions (88.6%) with 1.5%. However, the recall is just 75.6%, indicates the matching rule is insufficient to examine false predictions. In summary, the external knowledge is useful to examine predictions, but the matching rule is too simplified and it still owns a lot of room for improvements.

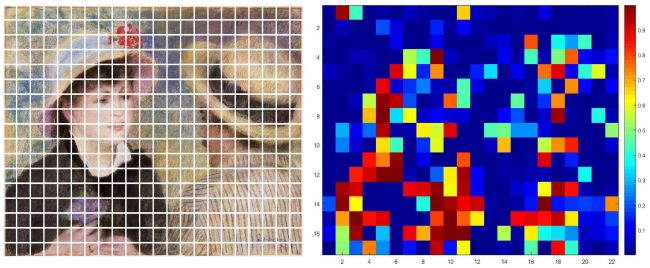
5.3.3 Confusion matrix analysis. The confusion matrix on Wikiart dataset is shown in Fig. 6, and we can observe that there is a relatively obvious misclassification. We have observed that the authorships of Boris Kustodiev's paintings are of considerable possibility to be predicted as Ilya Repin, but Repin's paintings are of smaller possibility to be predicted as Kustodiev's. It seems that Kustodiev once learnt from Repin in a certain period. Thus we search for the background information of these two artists and find out their relationship on wikipedia. It is amazing to see that, from 1896 to 1903, Boris Kustodiev attended Ilya repin's studio at the Imperial Academy of Arts in St. Petersbug. Moreover, when Repin was commissioned to paint a large-scale canvas to commemorate the 100th anniversary of the Stete Council, he ever invited Kustodiev to be his assistant. Therefore, they are in a great relation and their paintings both actually own certain similar drawing skills and styles.

5.4 Further Analysis

5.4.1 Visualization of artist-characteristic regions. In order to gain better understanding of the attribution performed by enhanced deep structure, we adopt the occlusion sensitivity testing method proposed by [31] to visualize the art-characteristic regions in a global view. By systematically occluding a small image region of a painting, the importance of the occluded region is determined by



(a) The artist-characteristic regions in global view.



(b) The artist-characteristic regions in regional view.

Figure 7: The painting is *Boating Couple (Aline Charigot and Renoir)* by Pierre-Auguste Renoir. The heatmap from global view is visualized by occlusion sensitivity testing. The heatmap from regional view is visualized by belief score. In both two heatmaps, red color corresponds to the greater importance in correctly identifying the author.

observing the change in the certainty score for the correct artist. Specifically, the occlusions are performed with a grey block of 8x8 pixels, to indicate approximate regions which are characteristic of the artist. The regions of importance can be visualized using heatmap color coding, as shown in Fig. 7. The region with red color is of greater importance in correctly attributing the painting.

As for the regional view, we pass the split regions into the network in region encoding structure, and take the predicted score at softmax layer of target class as belief score. Then we visualize these belief scores with heatmap. The region with red color is of greater importance in correctly attributing the painting.

For Fig. 7, first it should be noted that both two views assign little weights to the background. It illustrates the importance of the transparency of automatic attribution to allow human experts to interpret and evaluate the visual characteristic. Second, the emphasis of these two views is different. The attention of the global view is centered. By contrast, the attention of the regional view is dispersive, mainly focusing on the outlines of the characters. Therefore, each view in its own way has made important contributions in this task, and multi-range representation is more appropriate.

5.4.2 Artist-based cluster analysis. Since we choose top-1000 features as the representation of the paintings, the visualization techniques are utilized to delineate the space distribution of the whole dataset. Therefore, in Fig. 8, we compute the t-distributed stochastic neighborhood embeddings [14] for the features provided by our MTMR framework. Then we use the embeddings to project each feature into 2-D space, and plot the embedded features by representing them with their corresponding paintings.

Figure 8: t-SNE plot of paintings in the Wikiart test set where spatial distance indicated the similarity as resolved by our proposed analytic framework.

Scrupulous observers may find that many artists' individual drawing styles are unitary, but some may not. It is also worth mentioning that the representation produced by our framework is effective to distinguish paintings of different artists. As can be observed in Fig. 8, each cluster represents an artist, and there are clear margins between different clusters. Moreover, some clusters contain paintings of same style, some are able to contain paintings of different styles. It provides useful evidence that our framework is robust to deal with multi-style artist. In addition, more specific information is shown in Fig. 9.

6 CONCLUSIONS

In this paper, we propose a robust MTMR representation framework composed of multi-task learning and multi-range representation. It obviously outperforms the state-of-the-art methods on the two large-scale datasets. Moreover, we evaluate the importance of individual features and analyze the internal function of our framework. In the future studies, the generative methods will be incorporated to solve the unbalanced data distribution problem. It is a promising choice to pay more attention to the combination of our representation framework and generative adversarial networks.

ACKNOWLEDGMENTS

This work is supported by the National Natural Science Foundation of China (U1611461, 61661146005) and the National Key Research and Development Program of China (No. 2016YFB1001501).

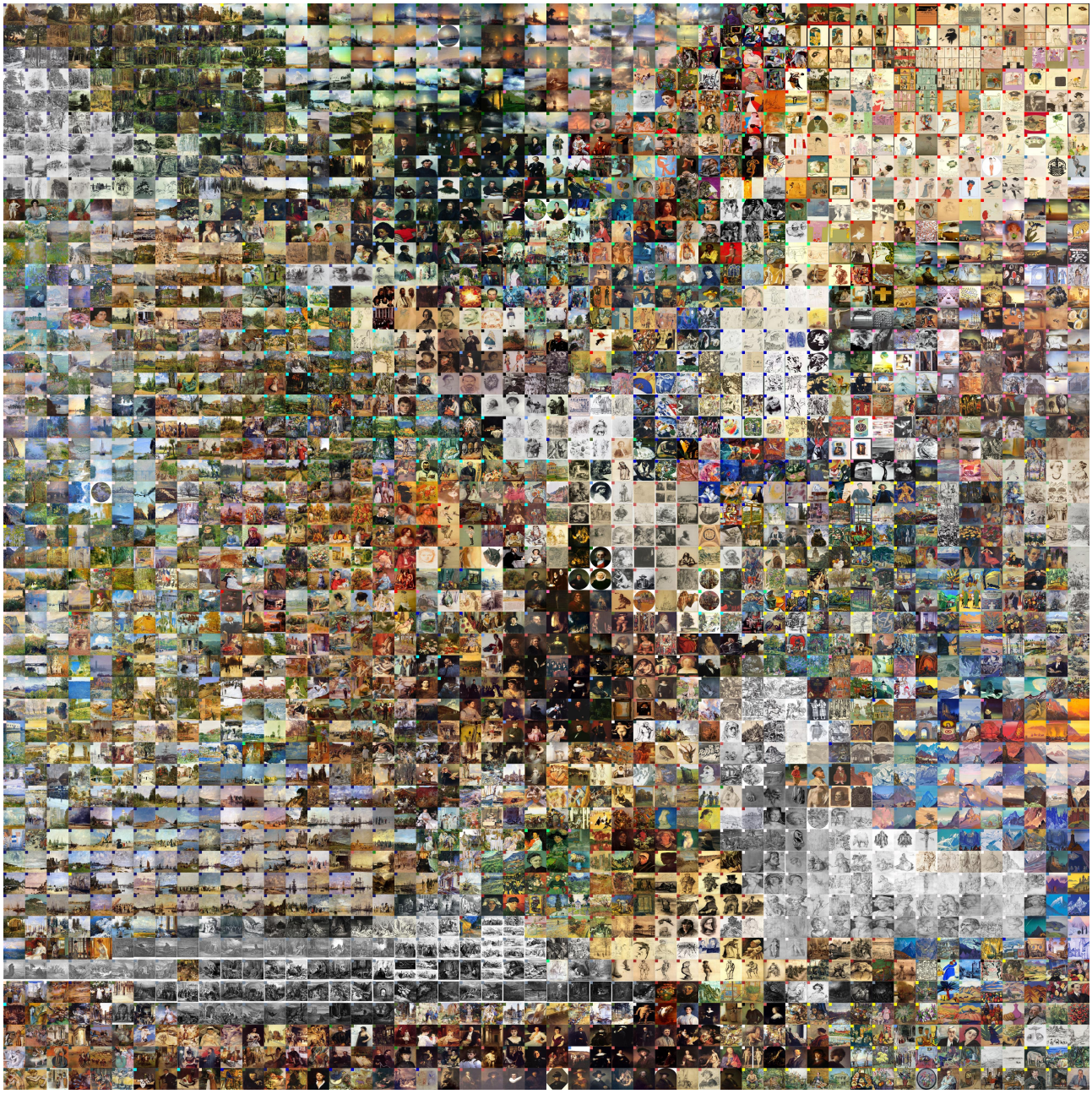


Figure 9: The grid version of Fig. 8, the same color in top-left corner of each painting means the same author. Please zoom in to see more details.

REFERENCES

[1] Ravneet Singh Arora and Ahmed Elgammal. 2012. Towards automated classification of fine-art painting style: A comparative study. In *Pattern Recognition (ICPR), 2012 21st International Conference on*. IEEE, 3541–3544.

[2] Jingjing Chen and Chong-Wah Ngo. 2016. Deep-based ingredient recognition for cooking recipe retrieval. In *Proceedings of the 2016 ACM on Multimedia Conference*. ACM, 32–41.

[3] Wei-Ta Chu and Yi-Ling Wu. 2016. Deep Correlation Features for Image Style Classification. In *Proceedings of the 2016 ACM on Multimedia Conference*. ACM, 402–406.

[4] Mircea Cimpoi, Subhransu Maji, and Andrea Vedaldi. 2015. Deep filter banks for texture recognition and segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3828–3836.

[5] Mohamed Elhoseiny, Tarek El-Gaaly, Amr Bakry, and Ahmed Elgammal. 2016. A Comparative Analysis and Study of Multiview CNN Models for Joint Object Categorization and Pose Estimation. In *Proceedings of The 33rd International Conference on Machine Learning*. 888–897.

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 770–778.

[7] Samet Hicsonmez, Nermin Samet, Fadime Sener, and Pinar Duygulu. 2017. DRAW: Deep networks for Recognizing styles of Artists Who illustrate children's books. *arXiv preprint arXiv:1704.03057* (2017).

[8] Kevin Alfianto Jangtjik, Mei-Chen Yeh, and Kai-Lung Hua. 2016. Artist-based Classification via Deep Learning with Multi-scale Weighted Pooling. In *Proceedings of the 2016 ACM on Multimedia Conference*. ACM, 635–639.

[9] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. 2014. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 675–678.

[10] Sergey Karayev, Matthew Trentacoste, Helen Han, Aseem Agarwala, Trevor Darrell, Aaron Hertzmann, and Holger Winnemoeller. 2014. Recognizing Image Style. In *Proceedings of the British Machine Vision Conference*. BMVA Press. DOI: https://doi.org/10.5244/C.28.122

[11] Jia Li, Lei Yao, Ella Hendriks, and James Z Wang. 2012. Rhythmic brushstrokes distinguish van Gogh from his contemporaries: findings via automated brushstroke extraction. *IEEE transactions on pattern analysis and machine intelligence* 34, 6 (2012), 1159–1176.

[12] Andy Liaw and Matthew Wiener. 2002. Classification and regression by randomForest. *R news* 2, 3 (2002), 18–22.

[13] David G Lowe. 2004. Distinctive image features from scale-invariant keypoints. *International journal of computer vision* 60, 2 (2004), 91–110.

[14] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research* 9, Nov (2008), 2579–2605.

[15] Thomas Mensink and Jan Van Gemert. 2014. The rijksmuseum challenge: Museum-centered visual recognition. In *Proceedings of International Conference on Multimedia Retrieval*. ACM, 451.

[16] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[17] Florent Perronnin, Jorge Sánchez, and Thomas Mensink. 2010. Improving the fisher kernel for large-scale image classification. In *European conference on computer vision*. Springer, 143–156.

[18] Ajit Puthenputhussery, Qingfeng Liu, and Chengjun Liu. 2016. Color multi-fusion fisher vector feature for fine art painting categorization and influence analysis. In *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on*. IEEE, 1–9.

[19] 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*. Springer, 612–627.

[20] Babak Saleh and Ahmed Elgammal. 2015. A unified framework for painting classification. In *Data Mining Workshop (ICDMW), 2015 IEEE International Conference on*. IEEE, 1254–1261.

[21] Jorge Sánchez, Florent Perronnin, Thomas Mensink, and Jakob Verbeek. 2013. Image classification with the fisher vector: Theory and practice. *International journal of computer vision* 105, 3 (2013), 222–245.

[22] Lior Shamir, Tomasz Macura, Nikita Orlov, D Mark Eckley, and Ilya G Goldberg. 2010. Impressionism, expressionism, surrealism: Automated recognition of painters and schools of art. *ACM Transactions on Applied Perception (TAP)* 7, 2 (2010), 8.

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

[24] Jiachuan Sheng and Jianmin Jiang. 2014. Recognition of Chinese artists via windowed and entropy balanced fusion in classification of their authored ink and wash paintings (IWPs). *Pattern Recognition* 47, 2 (2014), 612–622.

[25] Marcel Simon, Erik Rodner, and Joachim Denzler. 2016. ImageNet pre-trained models with batch normalization. *arXiv preprint arXiv:1612.01452* (2016).

[26] Wei Ren Tan, Chee Seng Chan, Hernán E Aguirre, and Kiyoshi Tanaka. 2016. Ceci n'est pas une pipe: A deep convolutional network for fine-art paintings classification. In *Image Processing (ICIP), 2016 IEEE International Conference on*. IEEE, 3703–3707.

[27] Ting-En Tseng, Wei-Yi Chang, Chu-Song Chen, and Yu-Chiang Frank Wang. 2016. Style retrieval from natural images. In *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 1561–1565.

[28] Nanne van Noord, Ella Hendriks, and Eric Postma. 2015. Toward Discovery of the Artist's Style: Learning to recognize artists by their artworks. *IEEE Signal Processing Magazine* 32, 4 (2015), 46–54.

[29] Nanne van Noord and Eric Postma. 2017. Learning scale-variant and scale-invariant features for deep image classification. *Pattern Recognition* 61 (2017), 583–592.

[30] A. Vedaldi and B. Fulkerson. 2008. VLFeat: An Open and Portable Library of Computer Vision Algorithms. (2008).

[31] Matthew D Zeiler and Rob Fergus. 2014. Visualizing and understanding convolutional networks. In *European conference on computer vision*. Springer, 818–833.

[32] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. 2014. Facial landmark detection by deep multi-task learning. In *European Conference on Computer Vision*. Springer, 94–108.