LEARNING MULTI-SCALE ATTENTIVE FEATURES FOR SERIES PHOTO SELECTION

Jin Huang[†], *Chaoran Cui*[⋆], *Chunyun Zhang*[⋆], *Zhen Shen*[‡], *Jun Yu*[†], *Yilong Yin*[‡]

[†] School of Computer Science and Technology, Shandong University
[⋆] School of Computer Science and Technology, Shandong University of Finance and Economics
[‡] School of Software, Shandong University

ABSTRACT

People used to take a series of nearly identical photos about the same subject, but it is usually a tedious chore to select the reversed ones from them. Despite the remarkable progress, most existing studies on image aesthetics assessment fail to fulfill the task of series photo selection. In this paper, we develop a novel deep CNN architecture that aggregates multi-scale features from different network layers, in order to capture the subtle differences between series photos. To reduce the risk of redundant or even interfering features, we introduce the spatial-channel self-attention mechanism to adaptively recalibrate the features at each layer, so that informative features can be selectively emphasized and less useful ones suppressed. Extensive experiments on a benchmark dataset well demonstrate the potential of our approach for series photo selection.

Index Terms— Aesthetics assessment, series photo selection, multi-scale, self-attention mechanism

1. INTRODUCTION

Photography in recent years has become ubiquitous in our daily life, where people are keen to record every memorable moment via a photo. In a real scenario, users often take a series of photos about the same object or scene to ensure that the best appearance or expression can be captured [1]. However, what then follows is that users have to manually decide the reserved ones from these nearly identical images for cost-effective storage, which is a cumbersome and time-consuming process.

Nowadays, there emerges increasing research efforts [2, 3] on assessing image quality from an aesthetic perspective. These methods could guide photo selection, but are hardly applied to users' series photos depicting roughly the same contents. Typically, they are trained on a large general corpus of images with diverse contents, and tend to yield close ratings for similar images [4]. Nevertheless, the visual differences between series photos are very subtle, so existing methods for image aesthetics assessment may be less capable of series photo selection. To illustrate this point, we display several examples of series photos in Fig. 1, with the photo ratings assigned by a known aesthetics model [5]. As can be seen, the gap between photo ratings is marginal in most series, and the one considered of the highest aesthetic quality and that preferred by majority users are usually inconsistent.

Fig. 1. Some examples of series photos. The rating predicted by a known aesthetics model [5] is listed below each photo, and the one preferred by majority users in each series is indicated by an orange box.

For this reason, several methods have been recently proposed to facilitate series photo selection. Kuzovkin et al. [6, 7] defined the multi-level contexts of each photo with hierarchical clustering, and adapted the photo quality score based on its contexts within the series. Chang et al. [8] collected the first large public dataset comprised of photo series from personal photo albums, and presented an end-to-end deep learning method with the Siamese network [9]. Despite the initial breakthrough in the above studies, it still remains a great challenge to extract discriminative and robust features to identify the subtle differences between series photos.

In this paper, we develop a novel deep CNN architecture that aggregates multi-scale features with self-attention mechanism [10] for series photo selection. Intuitively, the features

Chaoran Cui and Yilong Yin are the corresponding authors.

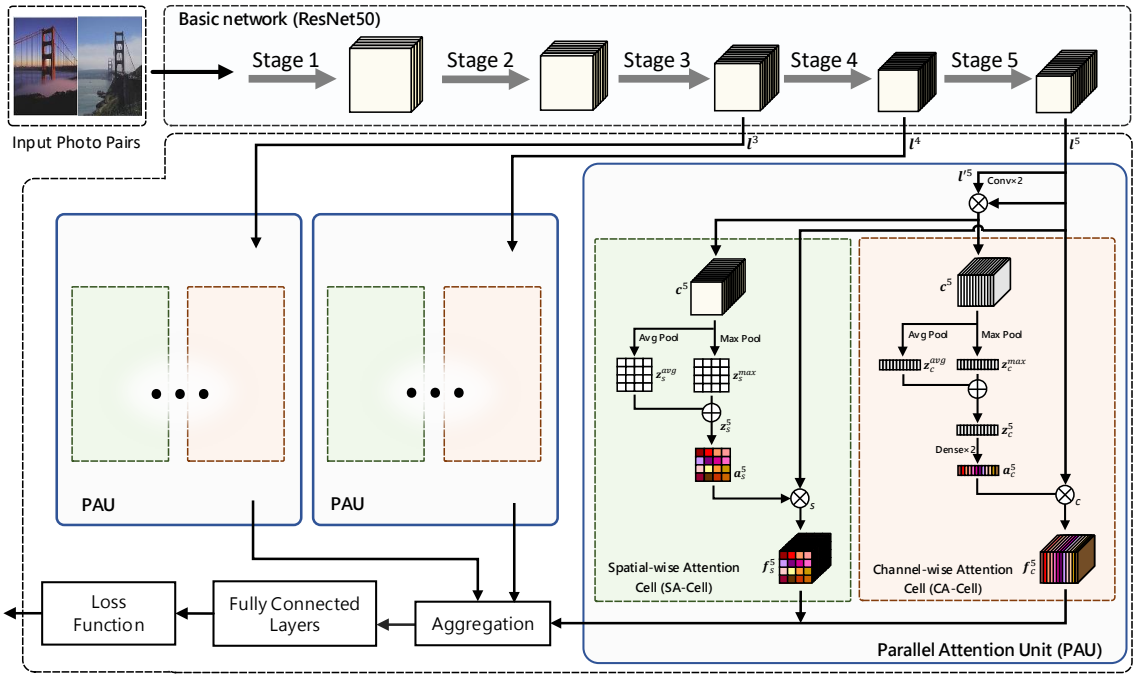

Fig. 2. Overview of the proposed deep network architecture for series photo selection.

produced by top convolutional layers has rich semantics, but low resolution that lose the fine details of images [11]. To capture the nuances in visual contents, our approach additionally exploits the features with high resolution from the intermediate layers, and automatically learns to combine the complementary information of different layers. However, a key disadvantage of fusing multi-scale features lies in that it may introduce redundant or even interfering features, which inevitably degrades the representation capability of our model. To address this issue, we append two types of attention modules at each layer to adaptively recalibrate the features in spatial and channel dimensions respectively. In this way, our approach selectively emphasizes informative features and suppresses less useful ones.

2. FRAMEWORK

2.1. Problem Formulation

The fundamental challenge of series photo selection is to determine the relative orders of photos. Therefore, we cast the problem as a pairwise ranking task. Formally, denote by $\mathcal{D} = \left\{\left(\boldsymbol{x}_i^k, \boldsymbol{x}_j^k, y^k\right) | k = 1, 2, \ldots, n\right\}$ a training set consisting of n pairwise comparisons, where $\boldsymbol{x}_i^k$ and $\boldsymbol{x}_j^k$ are two photos from the same series, and y^k is a binary label indicating the preference relation of the image pair, i.e., $y^k = 1$ if users prefer $\boldsymbol{x}_i^k$ over $\boldsymbol{x}_j^k$, and zero otherwise. We aim to learn a mapping function $f(\boldsymbol{x}_i^k, \boldsymbol{x}_j^k)$ that predicts the probability of the preference of $\boldsymbol{x}_i^k$ over $\boldsymbol{x}_j^k$. The desired mapping function can be obtained by minimizing the cross-entropy loss as follows:

$$L = \sum_{k=1}^{n} -y^k \log f(\boldsymbol{x}_i^k, \boldsymbol{x}_j^k) - (1 - y^k)(1 - f(\boldsymbol{x}_i^k, \boldsymbol{x}_j^k)) \quad (1)$$

In the following, we shall omit the subscript k for notational simplicity.

2.2. Overall Network Architecture

The backbone of our learning algorithm is an end-to-end deep CNN architecture. In our study, features from different network layers are jointly leveraged to help capture the subtle differences between series photos [12]. Specifically, we embark on the ResNet50 network [13], and the major modifications in our network can be summarized as follows:

- Similar to the Siamese network [9], our network contains two disjoint identical streams of ResNet50 with tied weights for feature extraction of $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ from a photo pair.
- The features from each of the last three layer stages of ResNet50 are exported and fed into a Parallel Attention Unit (PAU), in which the features are adaptively recalibrated to enable that informative features are emphasized and less useful ones are suppressed. The details of PAU will be described next.
- The outputs of different PAUs are aggregated by concatenation, resulting in the final representations of $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$. We then compute their distance and append three fully-connected layers to produce $f(\boldsymbol{x}_i, \boldsymbol{x}_j)$.

Table 1. Results of different methods in terms of classification accuracy on photo pairs.

Method	[8]	ResNet50	SENet	Ours
Accuracy (%)	73.01	72.32	68.86	**76.46**

The overall architecture of the proposed deep network is illustrated in Fig. 2.

2.3. Parallel Attention Unit

The workflow of the developed PAU component is shown in the bottom right of Fig. 2. It takes the feature maps $\boldsymbol{l} \in \mathbb{R}^{H \times W \times C}$ at each layer stage as the input, and outputs the refined feature maps with the same size along spatial $\boldsymbol{f}_s$ and channel $\boldsymbol{f}_c$ dimensions respectively, where H, W, and C are the height, width, and number of channels of feature maps. Inspired by the self-attention mechanism [10], $\boldsymbol{l}$ is first fed into two 1×1 convolutional layers [14] to generate the pseudo-query feature maps $\boldsymbol{l}'$. Then, we compute the compatibility between $\boldsymbol{l}$ and $\boldsymbol{l}'$ via element-wise dot product [15]:

$$\boldsymbol{c} = \boldsymbol{l} \otimes \boldsymbol{l}' \tag{2}$$

Based on $\boldsymbol{c}$, we further design two branches to obtain two attention maps, i.e., one is for the spatial attention effect, while the other is for the channel attention effect.

Spatial-wise Attention Cell (SA-Cell). The SA-Cell module is used to model the intra-spatial relationships of $\boldsymbol{l}$. Specifically, the context at each spatial location of the feature maps is aggregated by shrinking $\boldsymbol{c}$ using both average-pooling and maximum-pooling along the channel dimension, yielding the statistics $\boldsymbol{z}_s^{avg} \in \mathbb{R}^{H \times W}$ and $\boldsymbol{z}_s^{max} \in \mathbb{R}^{H \times W}$, respectively. Different from previous studies relying only on average-pooling, we consider that max-pooling gathers complementary clues to infer a finer attention map [16]. Therefore, we simultaneously use average-pooled and maximum-pooled features here. We empirically confirmed that exploiting both pooling operations greatly improves the performance rather than using each independently (see Section 3.3). The final spatial context descriptor $\boldsymbol{z}_s$ is attained by adding $\boldsymbol{z}_s^{avg}$ and $\boldsymbol{z}_s^{max}$. We use $\boldsymbol{z}_s$ as the guidance to generate the attention map $\boldsymbol{a}_s$ over the spatial locations of $\boldsymbol{l}$. This is achieved by a simple sigmoid function, i.e.,

$$\boldsymbol{a}_s = \sigma(\boldsymbol{z}_s) = \frac{1}{1 + exp(-\boldsymbol{z}_s)} \tag{3}$$

Finally, we obtain the recalibrated feature maps $\boldsymbol{f}_s$ by the spatial-wise multiplication [17] between $\boldsymbol{l}$ and $\boldsymbol{a}_s$:

$$\boldsymbol{f}_s = \boldsymbol{l} \otimes_s \boldsymbol{a}_s \tag{4}$$

Channel-wise Attention Cell (CA-Cell). The CA-Cell module is used to model the intra-channel relationships of $\boldsymbol{l}$. In analogy to SA-Cell, CA-Cell calculates the context of each channel of the feature maps by shrinking $\boldsymbol{c}$ with average-pooling and maximum-pooling along the spatial dimensions, yielding the statistics $\boldsymbol{z}_c^{avg} \in \mathbb{R}^C$ and $\boldsymbol{z}_c^{max} \in \mathbb{R}^C$, respectively. The final channel context descriptor $\boldsymbol{z}_c$ is the sum of $\boldsymbol{z}_c^{avg}$ and $\boldsymbol{z}_c^{max}$. To generate the channel attention map $\boldsymbol{a}_c$, we feed $\boldsymbol{z}_c$ into a two-layered fully-connected network with a bottleneck structure, i.e.,

$$\boldsymbol{a}_c = \sigma(\boldsymbol{W}_2 \delta(\boldsymbol{W}_1 \boldsymbol{z}_c)) \tag{5}$$

where δ refers to the ReLU function. $\boldsymbol{W}_1 \in \mathbb{R}^{\frac{C}{r} \times C}$ is the parameters of the first dimensionality reduction layer with reduction ration r, while $\boldsymbol{W}_2 \in \mathbb{R}^{C \times \frac{C}{r}}$ is the parameters of the second dimensionality increasing layer. We empirically set $r = 16$ according to the results presented in [18]. Intuitively, $\mathbf{a}_c$ encodes the importance of each feature map of $\boldsymbol{l}$. Therefore, the recalibrated feature maps $\boldsymbol{f}_c$ can be produced by the channel-wise multiplication [18] between $\boldsymbol{l}$ and $\boldsymbol{a}_c$:

$$\boldsymbol{f}_c = \boldsymbol{l} \otimes_c \boldsymbol{a}_c \tag{6}$$

Table 2. Comparison between our approach and variant methods using features from different layer stages.

Method	L_3	L_4	L_5	Ours
Accuracy (%)	66.78	75.09	73.36	**76.46**

3. EXPERIMENTS

3.1. Dataset

To ensure the comparability of the empirical results, the experiments were carried out on the benchmark dataset collected in [8] for series photo selection. The dataset contains 15,545 photos organized in 5,953 series. In each series, there are about 2 to 8 photos, and the pairwise preference on two photos have been manually labeled as ground-truth. Out of all 15,143 photo pairs, we randomly sampled 12,075 pairs for training, 483 pairs for validation, and the remaining 2,585 pairs for testing.

3.2. Performance Comparison

We compared our approach against the state-of-the-art method proposed in [8] for series photo selection. ResNet50 [13] and SENet [18] were also introduced as baselines, since they are commonly used in previous studies of image aesthetics assessment [2, 19].

The performance of different competitors were evaluated in terms of classification accuracy on photo pairs. Table 1 summarizes the comparison results. As can be seen, our approach substantially outperforms the method presented in [8], yielding around 4.73% relative improvement. This demonstrates the effectiveness of our approach for series photo selection. Besides, both the two methods are much superior to ResNet50 and SENet, suggesting that the classic deep learning models may be uncompetitive for series photo selection.

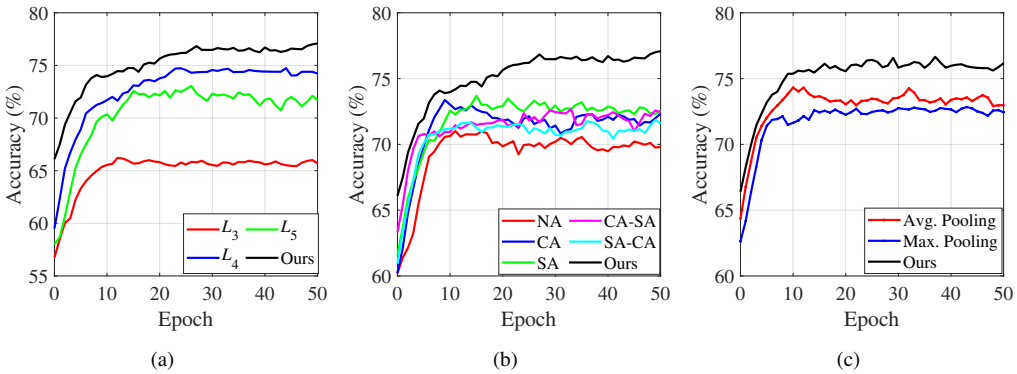

Fig. 3. Accuracy curves over the validation set during training. (a) Our approach versus variant methods with features from different layer stages. (b) Our approach versus variant methods with different attention modules. (c) Our approach versus variant methods with different pooling strategies.

Table 3. Comparison between our approach and variant methods using different attention modules.

Method	Accuracy (%)	Method	Accuracy (%)
NA	71.97	CA-SA	72.32
CA	74.05	SA-CA	72.66
SA	74.39	Ours	**76.46**

3.3. Ablation Study

Multi-scale Feature Fusion. We implemented several variants of our approach that use the features from different layer stages. Table 2 reports their performance results, in which L_3, L_4, and L_5 represent the variant methods using only the features outputted by the third, fourth, and fifth layer stage, respectively. It can be clearly seen that all variant methods fall behind our approach. Such results highlight the benefits of fusing multi-scale features in our approach.

Attention Module Aggregation. We ran our approach with different attention modules at each layer stage. Here, NA denotes the variant method without any attention modules. CA and SA indicate the variant methods separately adopting CA-Cell and SA-Cell, while CA-SA and SA-CA represent the ones simultaneously appending the two attention modules in a sequential manner. As shown in Table 3, NA is considerably worse than the other competitors, revealing the importance of incorporating the self-attention mechanism into our approach. Besides, the sequential use of CA-Cell and SA-Cell fails to provide better results compared to utilizing one of them alone; instead, our approach achieves the best performance when composing them in parallel.

Pooling Strategy Selection. We equipped the attention modules in our approach with three pooling strategies, namely, the average pooling, the maximum pooling, and the union of them as our original design. From Table 4, we can observe that the last exceeds its counterparts significantly. This confirms our belief that the average pooling and maximum pooling complement with each other, and it is advisable to combine them together for improved accuracy.

Table 4. Comparison between our approach and variant methods using different pooling strategies.

Method	Avg. Pooling	Max. Pooling	Ours
Accuracy (%)	73.01	72.32	**76.46**

Model Stability Analysis. Fig. 3 displays the accuracy curves over the validation set of our approach and the variant methods during training in the above ablation studies. As can be observed, our approach leads to a monotonic improvement in the performance of the learning process, and it consistently outperforms the variant methods in all cases. The results further verify the stability of our approach.

4. CONCLUSIONS

We have presented a novel deep CNN architecture to improve feature representation power for series photo selection. Our network jointly leverages multi-scale features from different layers, and introduces attention-based feature refinement with spatial-wise and channel-wise modules. Extensive experiments have confirmed that our network achieves outstanding performance on the benchmark dataset. In the future, we will further learn more discriminative features to capture the subtle differences between series photos.

5. ACKNOWLEDGEMENTS

This work is supported by the National Natural Science Foundation of China (61876098, 61701281, 61703234, 61573219), the Fostering Project of Dominant Discipline and Talent Team of Shandong Province Higher Education Institutions, and the National Key R&D Program of China (2018YFC0830100, 2018YFC0830102).

6. REFERENCES

[1] Baoyuan Wang, Noranart Vesdapunt, and Utkarsh Sinha, "Real-time burst photo selection using a light-head adversarial network," *IEEE Transactions on Image Processing*, 2019.

[2] Yubin Deng, Chen Change Loy, and Xiaoou Tang, "Image aesthetic assessment: An experimental survey," *IEEE Signal Processing Magazine*, vol. 34, no. 4, pp. 80–106, 2017.

[3] C. Cui, H. Liu, T. Lian, L. Nie, L. Zhu, and Y. Yin, "Distribution-oriented aesthetics assessment with semantic-aware hybrid network," *IEEE Transactions on Multimedia*, vol. 21, no. 5, pp. 1209–1220, 2019.

[4] Naila Murray, Luca Marchesotti, and Florent Perronnin, "Ava: A large-scale database for aesthetic visual analysis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 2408–2415.

[5] Xin Lu, Zhe Lin, Hailin Jin, Jianchao Yang, and James Z Wang, "Rating image aesthetics using deep learning," *IEEE Transactions on Multimedia*, vol. 17, no. 11, pp. 2021–2034, 2015.

[6] Dmitry Kuzovkin, Tania Pouli, Rémi Cozot, Olivier Le Meur, Jonathan Kervec, and Kadi Bouatouch, "Context-aware clustering and assessment of photo collections," in *Proceedings of the Symposium on Computational Aesthetics*. ACM, 2017, p. 6.

[7] Dmitry Kuzovkin, Tania Pouli, Rémi Cozot, Olivier Le Meur, Jonathan Kervec, and Kadi Bouatouch, "Image selection in photo albums," in *Proceedings of the ACM International Conference on Multimedia Retrieval*. ACM, 2018, pp. 397–404.

[8] Huiwen Chang, Fisher Yu, Jue Wang, Douglas Ashley, and Adam Finkelstein, "Automatic triage for a photo series," *ACM Transactions on Graphics*, vol. 35, no. 4, pp. 148, 2016.

[9] Luca Bertinetto, Jack Valmadre, Joao F Henriques, Andrea Vedaldi, and Philip HS Torr, "Fully-convolutional siamese networks for object tracking," in *Proceedings of the European Conference on Computer Vision*. Springer, 2016, pp. 850–865.

[10] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *Proceedings of the Advances in Neural Information Processing Systems*, 2017, pp. 5998–6008.

[11] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2117–2125.

[12] Yun Liu, Ming-Ming Cheng, Xiaowei Hu, Kai Wang, and Xiang Bai, "Richer convolutional features for edge detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 3000–3009.

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

[14] Min Lin, Qiang Chen, and Shuicheng Yan, "Network in network," *arXiv preprint arXiv:1312.4400*, 2013.

[15] Saumya Jetley, Nicholas A. Lord, Namhoon Lee, and Philip Torr, "Learn to pay attention," in *International Conference on Learning Representations*, 2018.

[16] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 3–19.

[17] Long Chen, Hanwang Zhang, Jun Xiao, Liqiang Nie, Jian Shao, Wei Liu, and Tat-Seng Chua, "Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5659–5667.

[18] Jie Hu, Li Shen, and Gang Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7132–7141.

[19] Xiaodan Zhang, Xinbo Gao, Wen Lu, and Lihuo He, "A gated peripheral-foveal convolutional neural network for unified image aesthetic prediction," *IEEE Transactions on Multimedia*, vol. 21, no. 11, pp. 2815–2826, 2019.