PEAR: PHOTOGRAPHIC EMBEDDING FOR AESTHETIC RATING

Hao Wu, Jiangchao Yao

Alibaba Group, China

ABSTRACT

Image aesthetic quality assessment has gained the enormous interest in recent years with the advancement of deep learning and the large-scale datasets. Current state-of-the-art methods generally leverage deep features, while the hand-crafted photographic attributes which are also useful but not always available have not drawn the sufficient attention. In this paper, we propose the Photographic Embedding for Aesthetic Rating (PEAR) framework to assimilate their advantages. In PEAR, a prior network is constructed to smoothly inject the photographic attributes to the latent space, and handle the attribute missing case via knowledge transfer. Simultaneously, aesthetic quality assessment is formulated as multi-task learning of aesthetic rating and image reconstruction to regularize the learning of latent codes. The extensive experiments on two challenging datasets demonstrate PEAR outperforms the state-of-the-art methods.

***Index Terms*—** Photographic embedding, aesthetic rating

1. INTRODUCTION

Recent years have witnessed the bursting of digital photography on the Internet, which leads to the increasing demand of automatically rating images based on human aesthetic experience for applications, e.g., image search and retrieval. Essentially, the task of image aesthetic quality assessment aims to simulate the human rating on image aesthetics. In the early works [1, 2, 3, 4, 5], handcrafted features like photographic attributes are common exploited for model decision. As can be seen in the left panel of Figure 1, photographic attributes, such as rules of thirds and depth of field, are commonly established rules for human visual perception of images. Each specific rule is evaluated with a binary annotator. The normalized score is the mean of multiple votes.

Recent works greatly enjoy the benefits from the advancement of deep neural networks (DNNs) and the large-scale aesthetic datasets [6, 7]. Hand-crafted features used in early researches have long been substituted by the flexible high-level features. Thus, in the era of deep learning, little attention is paid to the photographic attributes. A few recent works [8, 9] generally train a model that predicts both aesthetic rating and photographic attributes in the manner of multi-task learning. We argue that learning photographic attributes as an extra discriminative task limits the commitment of the photographic information. Moreover, annotating photographic attributes are expensive in practice. For the current largest aesthetic dataset AVA [6], the photographic attributes are mostly missing and current methods lack the solution to infer the photographic attributes , thus the performance may degenerate. These leave us with two questions—how to effectively incorporate the information of photographic attributes and how to handle the attribute missing case?

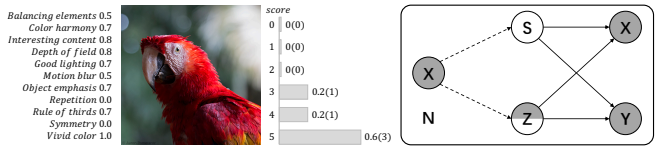

Fig. 1. **Left:** An example from the AADB dataset. For score distribution, each bar represents a score from 0 to 5. The number inside the brackets denotes the number of voters and the number outside denotes its proportion. **Right:** The PEAR graphical model. The shaded nodes are observed instances X and aesthetic score Y. Photographic embedding Z is partially observed. Solid and dashed lines represent generative and inference procedures.

To tackle the issues above, we propose the Photographic Embedding for Aesthetic Rating (PEAR) framework. As shown in the right panel of Figure 1, we introduce the autoencoding latent photographic embedding and deep feature codes in two different subspaces, on which both aesthetic rating and image reconstruction are conditional dependent in a generative fashion. The intuition is that the human judgement on image aesthetic relies on both photographic attributes and deep contents. A prior network is constructed to smoothly inject the photographic attributes to the latent space, and handle the attribute missing case via knowledge transfer. Besides. we formulate aesthetic quality assessment as the multi-task learning of aesthetic rating and image reconstruction. Here, image reconstruction as a second task in an autoencoding style serves as a regularizer to make the latent codes more meaningful. We conduct a range of experiments on two challenging benchmarks and extensive results show that our PEAR outperforms the state-of-the-art methods.

2. THE PROPOSED METHOD

2.1. Preliminaries

Given an aesthetic dataset $\mathcal{D} = \{(x_n, y_n, z_n)\}_{n=1}^{N}$, where N is the total sample number, x_n denotes an image instance and $y_n \in [0,1]^d$ is the corresponding aesthetic rating that serves as the primary supervision. In practice, each instance x_n is rated by multiple raters to provide scores within an ordinal value set $v = \{v_1, \ldots, v_d\}$, and y_n is the normalized rating distribution with each element y_{ni} represents the proportion of raters that gave the score v_i. The overall score can be calculated by $\sum_{i=1}^{d} y_i v_i$. For the case of binary classification, an empirical threshold is given to decide an image is whether good or bad in the sense of aesthetics. Thus different learning objectives can be deduced by the score distribution y_n. $z_n = [z_{n1}, \ldots, z_{nk}]$ is the photographic attribute vector containing k elements that indicate the existence of k individual photographic attributes. Each element of z_n is either binary or a mean of a set of binary values. However, z_n might not be given in real-world settings thus we need to infer such latent variable. Our goal is to train a model that can predict aesthetic ratings close to the human ground-truth.

2.2. The Photographic Embedding Framework

In this section, we give the general presentation of our PEAR framework. We formulate the aesthetic quality assessment task as multi-task learning. Different from previous methods, the image reconstruction task in the upper branch is jointly performed with aesthetic predicting in the lower branch. Except for commonly used deep latent code S extracted by DNNs, we introduce the photographic embedding Z parallelly in the latent space which jointly transits to aesthetic rating and contributes to image reconstruction. In inference process, both the distributions of photographic embedding Z and deep feature S are modeled based on image X, we denote these two distributions with $q(Z|X)$ and $q(S|X)$ respectively in terms of posterior approximation. In the generative process, aesthetic rating Y and reconstructed image $\hat{X}$ depend on both latent deep code S and photographic embedding Z modeled with $p(Y|Z,S)$ and $p(\hat{X}|Z,S)$.

According to the graphical model presented in Figure 1, we deduce the log-likelihood for the aesthetic rating branch:

$$\begin{aligned}\log p(Y|X) &= \sum_{n=1}^{N} \log p(y_n|x_n) \\ &= \sum_{n=1}^{N} \log \int_{s_n} \sum_{z_n} p(y_n|z_n, s_n) p(z_n|x_n) p(s_n|x_n) ds_n \\ &= \sum_{n=1}^{N} \log E_{p(z_n|x_n), p(s_n|x_n)} \left[p(y_n|z_n, s_n) \right] \end{aligned} \tag{1}$$

Due to the intractability in calculating the log-likelihood function, we instead choose to optimize the evidence lower bound (ELBO) [10, 11, 12]. To formulate the ELBO, we introduce two variational distributions $q(z_n|x_n)$ and $q(s_n|x_n)$ and deduce the log-likelihood function in the following. We neglect subscripts with shorter notations for simplicity:

$$\begin{aligned}\log p(Y|X) \geq\ & E_{q(z|x), q(s|x)} \log p(y|z,s) \\ & - D_{KL}(q(z|x)||p(z|x)) - D_{KL}(q(s|x)||p(s|x)).\end{aligned}$$

Correspondingly, for the image reconstruction branch, we yield the following log-likelihood function:

$$\begin{aligned}\log p(\hat{X}|X) \geq\ & E_{q(z|x), q(s|x)} \log p(\hat{x}|z,s) \\ & - D_{KL}(q(z|x)||p(z|x)) - D_{KL}(q(s|x)||p(s|x)).\end{aligned}$$

Finally, we present the the optimization object in the manner of multi-task learning by combining the above functions with a balancing factor λ:

$$\begin{aligned}\min L = & \underbrace{(-E_{z,s} \log p(y|z,s))}_{aesthetic\ rating\ loss} + \lambda \underbrace{(-E_{z,s} \log p(\hat{x}|z,s))}_{reconstruction\ loss} \\ & + (1+\lambda) \underbrace{D_{KL}(q(z|x)||p(z|x))}_{photographic\ info\ injection} \\ & + \beta D_{KL}(q(s|x)||p(s|x)).\end{aligned} \tag{2}$$

The first term is the aesthetic rating loss directly supervised by the score distribution y. The second term is the image reconstruction term that poses self-supervision in learning the latent codes. We tune this term with a weight λ to explicitly control its influence on the latent codes in joint learning. The third term is the key to inject the photographic attribute knowledge into the latent codes. The photographic attribute embedding variable z is regularized by the prior generated from the prior network. The knowledge of photographic attributes are smoothly injected or transferred into PEAR through this KL divergence term. For the forth term, the Gaussian prior regularization is applied on the deep embeddings. Originally, the weight of this term should be $1 + \lambda$. However, since the Gaussian prior is too simple and non-informative, we follow [13] to modify this weight with a relaxed hyperparameter β.

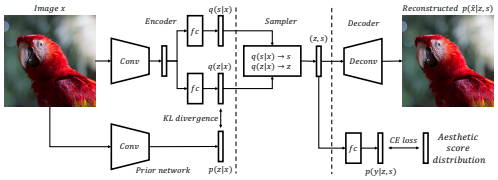

Fig. 2. The illustration of PEAR model structure. The three module are separated by dashed lines. The single-headed arrows represent the data flow while the double-headed arrows are indicator for losses.

2.3. The Model Structure

In this section, we systematically demonstrate the instantiation of PEAR.As illustrated in Figure 2, the whole model can

be decomposed into the following three modules.

The encoder module models the variational distributions $q(z|x)$ and $q(s|x)$. Since x_n is an image instance, we deploy a convolutional neural network (CNN). The structure of CNNs do not have to be specified thus our framework is compatible with any current networks. To inject the information of photographic attributes, we utilize an auxiliary prior network, which is simply pre-trained on the AADB photographic attributes, to generate the flexible prior $p(z|x)$. In this way, even if the photographic attributes are not given, we can transfer the knowledge of photographic attributes.Thus it naturally handles the attribute missing case. Moreover, the autoencoding style mitigates the noisy supervision [14, 15, 16, 17] problem caused by unreliable attributes. We set the prior $p(s) \sim N(0, 1)$ like [18] to ease the calculation of the forth term in Eq. (2) which can be deduced by:

$$D_{KL}(q(s|x)||p(s|x)) = -\frac{1}{2}(1 + \log \sigma^2(x) - \mu(x)^2 - \sigma^2(x)).$$

The sampler module is utilized to perform Monte Carlo sampling for $q(z|x)$ and $q(s|x)$. We sample from multivariate Bernoulli distribution for the photographic embedding variable while we sample from Gaussian distribution for the deep feature variable. Practically DNNs are unable to backpropagate through samples and we resort to the reparameterization tricks [18, 19] for stochastic optimization. For discrete variable z, we apply the modified reparameterization trick from [19] with temperature ignored:

$$z = \text{sigmoid}(\log q(z|x) - \log(1 - q(z|x)) + g),$$

where $g = \log u - \log(1 - u)$ and $u \sim Uniform(0, 1)$. g is the difference of two Gumbel random variables. For continuous variable s, we deploy the reparameterization trick [18]: $s = \mu(x) + \sigma^2(x)\epsilon$, where $\epsilon \sim N(0, 1)$ is the Gaussian noise. With the help of the above reparameterization tricks, the first and the second term in Eq. (2) can be optimized by SGD.

The decoder module plays the part in generating aesthetic rating Y and reconstructed image $\hat{X}$ with the concatenation of sampled z and s from the sampler module. As the two tasks are simultaneously performed, we deploy two heads respectively. For aesthetic rating, the simple MLP structure is adopted and we optimize with cross-entropy loss to learn the score distribution. For image reconstruction, the deconvolution network is deployed and we use the MSE loss for reconstruction. The detailed structures are explained in Section 3.2.

3. EXPERIMENTS

3.1. Datasets and Baselines

AVA [6] is the current largest aesthetic quality assessment benchmark that consists of around 250k images. The scores are collected from multiple viewers while the photographic attributes are poorly labeled. To evaluate all methods in three different perspectives of,we report classification accuracy for binary classification, Spearman's rank correlation coefficient (SRCC) for aesthetic overall score regression (ranking) and the Earth Mover's Distance (EMD) [20] for distribution learning. The above criterion cover all aspects in previous studies. **AADB** [7] contains around 10,000 images which are annotated by 5 individual workers and the votes range from 0 to 5. All images are provided with photographic attributes. As AADB is specifically designed for aesthetic ranking, only SRCC is reported following previous studies [7, 21, 9]. We compare our PEAR with the following methods. **PAR** [7] trains a model through a Siamese network with score regression. **AAL** [9] leverages multi-task learning to jointly regresses aesthetic scores and predict photographic attributes. **CE** trains the ResNet50 with cross-entropy loss for distribution learning. **NIMA** [22] deploys the squared Earth Mover's Distance for distribution learning. **UIAA** [21] reshapes the original supervision into Gaussian for distribution learning.

3.2. Implementation Details

For convolutions layers, we leverage ResNet50 [23] as backbone to align with the previous works [22, 9, 21]. The prior network adopts the same architecture and is initiated by the photographic attributes on AADB and then shared in experiments on both datasets. We feed the flattened output of ResNet50 to photographic encoder $(2048 \rightarrow k)$, μ encoder $(2048 \rightarrow j)$ and σ encoder $(2048 \rightarrow j)$. The dimension of the photographic attribute k is 11. We empirically set j to 100. The concatenation of sampled z and s is fed to the two decoding heads. For the aesthetic head, a fc layer $(k + j \rightarrow d)$ is deployed. The dimension d is 10 on the AVA dataset and 6 on the AADB dataset. For the image reconstruction head, the concatenation of sampled z and s are expanded with a fully-connected layer $(k + j \rightarrow 8192)$ and then resized to $(128, 8, 8)$. Five deconvolutional layers implemented with $ConvTranspose2d$ via Pytorch [24] are followed to reconstruct the image. The kernel size, stride and padding are set to $4 \times 4, 2, 1$ for all deconvolutional layers and the output channel numbers are set to $64, 32, 16, 8, 3$ respectively. BatchNorm and ReLU are applied succeedingly to each layer, except that Sigmoid is used for the final deconvolutional layer. In the default setting, we set λ of the reconstruction branch to 0.1 to relatively emphasize the other aesthetic rating branch. We set β to 0.1 as we do not expect the learning of the latent deep feature to be restricted by the simple Gaussian prior. We adopt the Adam optimizer [25] and set batch size to 128. The learning rate of the backbone and the prior network is set to 0.0001, while the learning rate for the rest of PEAR is set to 0.001.

3.3. Results on AVA and AADB

In Table 1, we summarize the results of all methods on the AVA and AADB datasets. For the AVA dataset, we report

Method	AVA Accuracy (%)	AVA SRCC	AVA EMD	AADB SRCC
PAR	77.33	0.558	-	0.6782
AAL	-	0.631	-	0.7041
CE	79.56	0.696	0.067	0.7089
NIMA	79.33	0.690	0.067	0.7081
UIAA	80.35	0.714	0.066	0.7264
PEAR	**80.52**	**0.719**	**0.065**	**0.7281**

Table 1. Results on AVA and AADB.

Fig. 3. **Left:** The absolute difference of correlation matrices. Lighter is better. **Middle:** The inferred photographic attributes of an AVA sample. **Right:** The t-SNE visualizations with binary labels superimposed.

classification accuracy for binary classification, SRCC for aesthetic score regression (ranking) and EMD [20] for score distribution learning. Generally, score distribution learning methods achieve better results than score regression methods. CE yields slightly better results than NIMA, which demonstrates cross-entropy loss is more suitable in learning score distribution than EMD loss. Among all baselines, UIAA achieves the best result as it utilizes cross-entropy loss to learn from a stabilized Gaussian distribution to address the unreliability of the origin score distribution. Nonetheless, our PEAR achieves better results than UIAA in accuracy, SRCC and EMD. The benefits are from the incorporation of photographic attributes in our PEAR. For the AADB dataset, we report SRCC in Table 1. The results are similar with those on the AVA dataset that UIAA achieves the best results in all baselines while our PEAR surpasses UIAA. The reason of the superiority of PEAR is that PEAR innovatively injects the photographic attributes into latent space as well as the image reconstruction branch helps the regularization of the latent codes. In the subsequent sections, we will show the effectiveness of such two aspects.

3.4. Effectiveness of Photographic Attribute Embedding

In this section, we demonstrate the effectiveness of the photographic attribute embedding of our PEAR. For AADB which contains the ground-truth photographic attributes, we analyse the joint distributions of the estimated photographic attributes along with the aesthetic score versus the ground-truth quantitatively as done in [9]. The correlation matrix is computed on the test set of AADB. We compute the correlation matrix of PEAR model predictions and the ground-truth respectively and illustrate the absolute difference of the two correlation matrices. As can be seen in left of Figure 3, the left panel denotes the absolute difference matrix of AAL [9] and the right panel denotes the absolute difference matrix of PEAR. For AAL, the average of difference matrix is 0.243, while the average of PEAR matrix is 0.131 which indicates our PEAR models the photographic attributes more closely. Among all photographic attributes, balancing elements and rule of thirds are the two hardest attributes that account for high-level aesthetics. For AVA, we present the inferred photographic attributes by our PEAR in the middle of Figure 3. The most confident predicted attributes are object emphasis, depth of field and interesting content, which precisely capture the essential aesthetics in the image. PEAR is also relatively confident about the color harmony. For the two difficult attributes we mentioned before, PEAR is weakly confident about rule of thirds (present) while relatively not confident about balancing elements (not present). The predictions with lowest scores are also correct as there is neither repetition nor symmetry in the original image. However, we admit that PEAR needs to be further improved, since it gives a neutral score for motion blur which is actually not caused by motion.

3.5. Effectiveness of Joint Image Reconstruction

To verify the effectiveness of the image reconstruction branch in our PEAR, we visually analyze its regularization effects on the learned latent code s via t-SNE [26]. Specifically, we consider the situation when the reconstruction is applied ($\lambda = 0.1$) and is canceled ($\lambda = 0$). As illustrated in the right of Figure 3, we superimpose the binary label with label 1 indicates a good image. The clusters are intertwined in both cases due to the learning of score distribution rather than binary labels. When image reconstruction is canceled, although we can discriminate some blue dots in the lower part, the reds dots are generally mixed up with the blue ones. However, we can still find in the right panel that the blue and red dots are distinguishable in the upper and lower areas respectively when the image reconstruction is applied. This demonstrates the effectiveness of the image reconstruction branch in PEAR.

4. CONCLUSION AND FUTURE WORK

In this paper, we propose the PEAR framework. In PEAR, a prior network is constructed to smoothly inject the photographic attributes to the latent space. When photographic attributes are unavailable, a flexible prior can be obtained by transferring the knowledge of existing data. Simultaneously, the aesthetic quality assessment is formulated as multi-task learning of aesthetic rating and image reconstruction to regularize the learning of latent codes. A range of experiments have been conducted on two challenging benchmarks to demonstrate the effectiveness of aesthetic rating and photographic attributes modeling. For future directions of research, our PEAR can be extended to incorporate the content information encoded in the latent space, and the architecture of the image reconstruction decoder can be further explored.

5. REFERENCES

[1] Ritendra Datta, Dhiraj Joshi, Jia Li, and James Z Wang, "Studying aesthetics in photographic images using a computational approach," in *ECCV*, 2006.

[2] Yan Ke, Xiaoou Tang, and Feng Jing, "The design of high-level features for photo quality assessment," in *CVPR*, 2006.

[3] Yiwen Luo and Xiaoou Tang, "Photo and video quality evaluation: Focusing on the subject," in *ECCV*, 2008.

[4] Luca Marchesotti, Florent Perronnin, Diane Larlus, and Gabriela Csurka, "Assessing the aesthetic quality of photographs using generic image descriptors," in *CVPR*, 2011.

[5] Ou Wu, Weiming Hu, and Jun Gao, "Learning to predict the perceived visual quality of photos," in *ICCV*, 2011.

[6] Naila Murray, Luca Marchesotti, and Florent Perronnin, "Ava: A large-scale database for aesthetic visual analysis," in *CVPR*, 2012.

[7] Shu Kong, Xiaohui Shen, Zhe Lin, Radomir Mech, and Charless Fowlkes, "Photo aesthetics ranking network with attributes and content adaptation," in *ECCV*, 2016.

[8] Gautam Malu, Raju S Bapi, and Bipin Indurkhya, "Learning photography aesthetics with deep cnns," *arXiv:1707.03981*, 2017.

[9] Bowen Pan, Shangfei Wang, and Qisheng Jiang, "Image aesthetic assessment assisted by attributes through adversarial learning," in *AAAI*, 2019.

[10] Martin J Wainwright and Michael Irwin Jordan, *Graphical models, exponential families, and variational inference*, Now Publishers Inc, 2008.

[11] John Paisley, David Blei, and Michael Jordan, "Variational bayesian inference with stochastic search," *arXiv:1206.6430*, 2012.

[12] David M Blei, Alp Kucukelbir, and Jon D McAuliffe, "Variational inference: A review for statisticians," *JASA*, 2017.

[13] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner, "beta-vae: Learning basic visual concepts with a constrained variational framework," 2016.

[14] Jiangchao Yao, Jiajie Wang, Ivor W Tsang, Ya Zhang, Jun Sun, Chengqi Zhang, and Rui Zhang, "Deep learning from noisy image labels with quality embedding," *IEEE Transactions on Image Processing*, vol. 28, no. 4, pp. 1909–1922, 2018.

[15] Hao Wu, Jiangchao Yao, Jiajie Wang, Yinru Chen, Ya Zhang, and Yanfeng Wang, "Collaborative label correction via entropy thresholding," in *2019 IEEE International Conference on Data Mining (ICDM)*, 2019, pp. 1390–1395.

[16] Hao Wu, Jiangchao Yao, Ya Zhang, and Yanfeng Wang, "Cooperative learning for noisy supervision," in *2021 IEEE International Conference on Multimedia and Expo (ICME)*, 2021.

[17] Hao Wu, Jiajie Wang, Yuanzhe Gu, Peisen Zhao, and Zhonglin Zu, *A Solution to Multi-Modal Ads Video Tagging Challenge*, Association for Computing Machinery, 2021.

[18] Diederik P Kingma and Max Welling, "Auto-encoding variational bayes," *arXiv:1312.6114*, 2013.

[19] Eric Jang, Shixiang Gu, and Ben Poole, "Categorical reparameterization with gumbel-softmax," *arXiv:1611.01144*, 2016.

[20] Le Hou, Chen-Ping Yu, and Dimitris Samaras, "Squared earth mover's distance-based loss for training deep neural networks," *arXiv:1611.05916*, 2016.

[21] Hui Zeng, Zisheng Cao, Lei Zhang, and Alan C Bovik, "A unified probabilistic formulation of image aesthetic assessment," *TIP*, 2019.

[22] Hossein Talebi and Peyman Milanfar, "Nima: Neural image assessment," *TIP*, 2018.

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *CVPR*, 2016.

[24] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al., "Pytorch: An imperative style, high-performance deep learning library," in *NeurIPS*, 2019.

[25] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv:1412.6980*, 2014.

[26] Laurens van der Maaten and Geoffrey Hinton, "Visualizing data using t-sne," *JMLR*, 2008.