LEARNING IMAGE AESTHETICS BY LEARNING INPAINTING

June Hao Ching *John See* *Lai-Kuan Wong*

Visual Processing Lab, Faculty of Computing and Informatics,
Multimedia University, Malaysia

ABSTRACT

Due to the high capability of learning robust features, convolutional neural networks (CNN) are becoming a mainstay solution for many computer vision problems, including aesthetic quality assessment (AQA). However, there remains the issue that learning with CNN requires time-consuming and expensive data annotations especially for a task like AQA. In this paper, we present a novel approach to AQA that incorporates self-supervised learning (SSL) by learning how to inpaint images according to photographic rules such as rules-of-thirds and visual saliency. We conduct extensive quantitative experiments on a variety of pretext tasks and also different ways of masking patches for inpainting, reporting fairer distribution-based metrics. We also show the suitability and practicality of the inpainting task which yielded comparably good benchmark results with much lighter model complexity.

Index Terms— Aesthetic quality assessment, CNN, self-supervised learning, image inpainting, photographic rules

1. INTRODUCTION

With the advancement of mobile camera technology and the growth of social media, online photo sharing has become an increasingly popular phenomenon. As such, personal galleries or media retrieval systems are also inundated with a massive deluge of images; many which could be of poor quality or lack in appeal. The growing interest in *aesthetic quality assessment* (AQA) in recent years [1, 2, 3] is testament of the need to automate the process of selecting or sorting out images from the perspective of aesthetic appeal.

In the early days, most of the works that proposed for AQA were designing hand-crafted features that correspond to known aesthetic principles such as low-level features that are based on photographic rules [4], and SIFT or color descriptors [5]. With the success of deep learning, researchers started to use CNN-based models in their works [6, 7, 3, 1], and these methods easily outperform the handcrafted methods by a significant margin.

Although work on AQA using deep learning techniques outperformed most traditional feature extraction methods, the initial data collection and annotation works are most essential to the success of using a heavily supervised method like CNN.

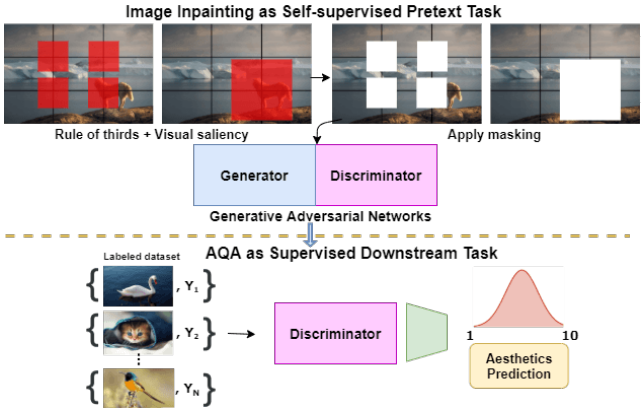

Fig. 1. Image inpainting according to photographic rules as a self-supervised learning (SSL) pretext task for AQA.

This is particularly challenging and expensive for a subjective task like AQA as opinions need to be collected from many professional photographers to provide useful ratings of the aesthetics of an image. Self-supervised learning (SSL) offers a new paradigm towards learning visual features from an unlabeled dataset on a pretext task (with pseudo labels) before transferring to (the actual) downstream supervised prediction task. In recent years, works like [8, 9, 10, 11, 12] proposed different SSL pretext tasks trained on ImageNet [13], which reported strong capabilities at various downstream tasks. Hence, we are motivated to design a viable SSL pretext task that can incorporate photographic rules for to better understand image aesthetics. By teaching the machine to inpaint portions of the image that corresponds closely to aesthetic concepts, we hypothesize that the model will also learn intrinsic knowledge and features of these concepts, which in turn, can perform the AQA task well.

In this paper, we propose a novel approach to AQA by incorporating SSL based on image inpainting. The main contributions in this work are as follows:

1. We propose new ways of performing image inpainting based on compositional rules (rule of thirds, visual saliency) as a self-supervising pretext task for the CNN before transferring to the downstream supervised AQA task.

2. We demonstrate that using SSL can work reasonably well for AQA, matching close to the performance of most state-of-the-art deep learning methods, with a smaller dispersion in correlation.

3. We provide a comprehensive benchmark (including network complexity) for a variety of pretext tasks, highlighting the capability of our lightweight model, which garnered comparable results against heavier methods.

2. RELATED WORK

AQA. Works in AQA can generally be divided into two categories – (1) classical handcrafted low-level features based on photography rules or using image descriptors, and (2) methods that leverage deep learning models (e.g. CNN).

In the first category, AQA involved designing handcrafted features that correspond to known aesthetic principles such as the composition of an image, sharpness, saturation and the contrast levels of an image, and the use of basic photographic rules e.g. rule-of-thirds, etc. Well-known early works are like [4, 5] extracted low level features or color/shape based descriptors and train with standard machine learning classifiers or regressors.

In the second category, deep learning models, particularly CNNs have made a huge impact in many image processing tasks such as image classification and object detection, largely due to their powerful ability to learn different hierarchies of spatial features from images. Works like [6, 14] even proposed to use more than one CNN to learn from different types/scales of visual features from images for AQA. A number of recent works [7, 3] attempt to change the structure of InceptionNet [15] in particular to learn local and global features of images. [16] trained a Siamese network to encode images into a visual aesthetics space. Besides, [17] incorporated adaptive spatial pyramid pooling to input images of any resolution to make better predictions. Our work is inspired by the recent work of [1], which modified the standard classification loss function to Earth Mover Distance (EMD) loss and achieved relatively good performance without the need for sophisticated architectures.

SSL. Self-supervised learning (SSL) comes across as a viable learning technique which leverages on *pretext tasks* trained when labels are scarce. Across recent literature [18], there are four known categories of pretext tasks for SSL: generation-based, context-based, free semantic label-based and cross modal-based. Of the four, the generation-based and context-based methods are the most relevant to this work as they focus on the generation of pixels and the structure of objects in image, tasks which have affinity with the concept of aesthetics in images. Context-based methods [11, 10, 12] allow CNNs to learn useful visual features by learning how to identify angles and arrangement (spatial structure) of objects or patches in images. Meanwhile, generation-based methods like [19, 8, 9] enable CNNs to learn visual features from unlabeled images via synthesis, inpainting and colorization tasks.

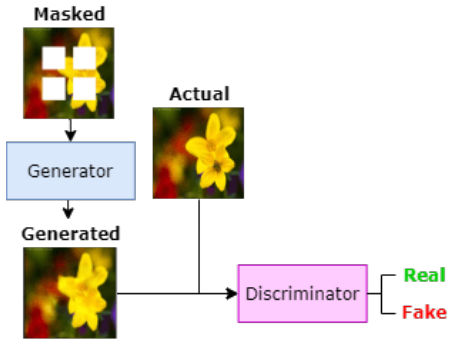

Fig. 2. Illustration of the architecture for image inpainting.

3. METHODS

Motivated by the work of Pathak et al. [8], we introduce a number of new ways of performing image inpainting pretext task based on Generative Adversarial Networks (GAN) by incorporating compositional photographic rules to enable the learning of context and structure in images. This forms the main intuition behind the idea of "learning aesthetics by learning inpainting" – the computer basically learns the subjective concept of aesthetics by learning to fill in removed patches. During the pretext task, the generator network is to inpaint the image on the patches masked based on compositional rules while the discriminator network acts to differentiate between the actual and generated patches. Hence, upon transfer to the downstream AQA task, the discriminator inherently understands the areas (where the patches are) to focus on when predicting the aesthetic quality level of images.

3.1. Compositional Image Inpainting as Pretext Task

To enable the CNN to learn visual features from with a sense of photographic rules, we designed several feasible image inpainting methods as pretext tasks for the SSL process. Specifically, we designed 3 new ways of performing image inpainting based on photographic compositional rules such as rule-of-thirds and visual saliency by altering the masking area.

Masking Area for Image Inpainting. Fig. 3 shows the different ways of masking patches for the inpainting pretext task. The white square areas indicate the masking areas for the generator to generate and for the discriminator to differentiate in the learning process. The leftmost image is the original image. The center cropped (CC) area is the original masked patch used in [8]. We propose to mask off parts of the image that corresponds closely to some photographic rules. Four Power Points (FPP) uses rule-of-thirds to obtain the four power points of the image, and we apply masking on 32×32 square areas on each power point. For Highest Salient

Fig. 3. Different ways of masking patches for image inpainting: rule-of-thirds (FPP), visual saliency (HSA) or both (HSP).

Area (HSA), we take the point with the highest saliency value (using the method by [20]) as center point and apply masking on a 64×64 square area. Lastly, for Highest Salient Power point (HSP), we fixed 4 areas in the image – each a 64×64) square area centered at each power point, and compute the sum of saliency values. Masking is applied on the area with the highest total saliency.

Generative Adversarial Network. The pretext image inpainting task is first trained using GAN before transferring its discriminator network to the downstream AQA task [5]. Typical GAN architectures consists of two networks: G and D, in a two-player minimax game where G tries to generate image patches that look as close as possible to the real image patch while D tries to distinguish between a generated image patch and a real image patch:

$$\min_{G} \max_{D} log(D) + log(1 - D(G)) \tag{1}$$

The GAN architecture that is used in the experiment is motivated by that of [8], but with modifications to its loss functions.

Generator. The generator $\mathcal{G}$ consists of 6 convolutional layers (leaky ReLU as activation function), followed by 4 transposed convolutional layers (ReLU as activation function) and ending with another convolutional layer (Tanh as activation function). We implement the loss function of the generator in our work as a weighted combination of adversarial loss and pixel-wise loss.

The adversarial loss measures the extent of which the generator fools the discriminator. Here, the mean squared error (MSE) is a feasible measure that can be minimized, where I_{mp} indicates a full image with masked patches, $True$ indicates true values (1s) and $False$ indicates false values (0s):

$$\min_{G} \mathcal{L}_{G,adv} = MSE(D(G(I_{mp})), True) \tag{2}$$

In an inpainting task, the contribution of this loss must be set to a low value to allow gradual perturbations to the generated patch.

Meanwhile, the pixel-wise loss measures how close are the real values of a full image X to the generated image G with $L1$ loss:

$$\min_{G} \mathcal{L}_{G,pw} = L1(G(I_{mp}), X) \tag{3}$$

The overall loss function of the generator is given by:

$$\min_{G} \mathcal{L}_{G} = \lambda_{adv} * \mathcal{L}_{G,adv} + \lambda_{pw} * \mathcal{L}_{G,pw} \tag{4}$$

Discriminator. The discriminator D network consists of 5 convolutional layers, first 4 convolutional layers attached with leaky ReLU as activation function. The loss function for D measures how much can the discriminator differentiate between the true one as true and the fake one is fake with adversarial loss:

$$\mathcal{L}_{D,real} = MSE(D(X), True) \tag{5}$$

$$\mathcal{L}_{D,fake} = MSE(D(G(I_{mp})), False) \tag{6}$$

The overall loss function of the discriminator:

$$\mathcal{L}_{D} = \lambda_{dis} \cdot (L_{D,real} + L_{D,fake}) \tag{7}$$

In this set up, both G and D are jointly optimized concurrently using Adam [21]. The balancing parameters are set to $\lambda_{adv} = 0.001$, $\lambda_{pw} = 0.999$ as in [8] while $\lambda_{dis} = 0.5$.

3.2. Downstream Supervised AQA Task Training

After training the GAN on the pretext task (image inpainting) to good convergence, the discriminator network is transferred over to the downstream supervised AQA task. In this second training, we use the large-scale AVA dataset AVA [5] which is popularly used as a AQA benchmark. The transferred network is appended with fully connected layers. We proceed to fine-tune the CNN (discriminator) on AVA to predict the probability distribution score of the images using EMD loss proposed by [1].

4. EXPERIMENTAL RESULTS

We train our pretext tasks using the large-scale AVA dataset [5] which contains around 250K images which were crawled from a popular photography community DPChallenge[1] where each image was rated by an average of 210 photographers with a score ranging from 1 to 10.

For the AQA task, a fully-connected layer (randomly initialized) is added at the end of the SSL-trained CNN with a dropout rate of 0.75. Training was performed using Adam [21] optimizer with the learning rates of the convolutional layers and the last FC layer set to 10^{-5} and 10^{-4}, respectively. The models presented in this paper are all implemented using PyTorch. Code and models are available[2] for public use.

[1] www.dpchallenge.com

[2] https://github.com/chingjunehao/SSL-Inpainting-AQA

Architecture	No. Params.	Pretext task	pre-trained on	Accuracy↑ (2 classes)	LCC↑ (mean)	SRCC↑ (mean)	LCC (std.dev)	SRCC (std.dev)	EMD↓
NIMA-AlexNet	2.56M	-	ImageNet	79.89%	0.522	0.498	0.191	0.185	0.0509
NIMA-VGG16*	14.96M	-	ImageNet	81.13%	0.598	0.576	0.219	0.209	0.0477
AlexNet	57.05M	RotNet	ImageNet	79.00%	0.423	0.408	0.166	0.160	0.0535
VGG16	14.97M	Colorization	ImageNet	80.93%	0.501	0.476	0.179	0.168	0.0506
Resnet152	58.16M	Colorization	ImageNet	80.72%	0.543	0.514	0.203	0.194	0.0498
VGG16	14.97M	Colorization	AVA	80.55%	0.510	0.484	0.185	0.175	0.0507
Generator-FPP	3.07M	Inpainting	AVA	79.10%	0.271	0.263	0.104	0.099	0.0570
Generator-HSP	3.07M	Inpainting	AVA	79.43%	0.404	0.390	0.165	0.157	0.0536
Discriminator-CC	**1.56M**	Inpainting	AVA	80.31%	0.428	0.402	0.109	0.100	0.0527
Discriminator-HSP	**1.56M**	Inpainting	AVA	79.77%	0.428	0.404	0.150	0.141	0.0533
Discriminator-HSA	**1.56M**	Inpainting	AVA	80.23%	0.452	0.426	0.139	0.128	0.0523
Discriminator-FPP+CC	3.12M	Inpainting	AVA	80.26%	0.506	0.482	0.161	0.149	0.0510
Discriminator-FPP	**1.56M**	Inpainting	AVA	**80.43%**	**0.520**	**0.494**	0.171	0.160	**0.0504**

Table 1. Performance comparison with methods pre-trained on different SSL pretext task (2nd and 3rd section) before transferring to the supervised downstream task, or direct supervised learning (1st section). The AQA task is trained using the method proposed by [1]. Other notations: +: concatenation of features after 5th conv layer. M: million.

Model	Accuracy (2 classes)	LCC (mean)	SRCC (mean)	LCC (std.dev)	SRCC (std.dev)	EMD
Murray et al. [5]	66.70%	-	-	-	-	-
Lu et al. [6]	74.46%	-	-	-	-	-
Hii et al. [7]	75.76%	-	-	-	-	-
Schwarz et al. [16]	75.83%	-	-	-	-	-
Wang et al. [14]	76.80%	-	-	-	-	-
Murray et al. [17]	80.30%	-	0.709	-	-	-
Talebi et al. [1]	81.51%	0.636	0.612	0.233	0.218	0.0500
Hosu et al. [22]	81.72%	0.757	0.756	-	-	-
Zhang et al. [2]	81.81%	0.704	0.690	-	-	0.045
Jin et al. [3]	82.66%	-	-	-	-	-
Ours (SSL-D-FPP)	80.43%	0.520	0.494	0.171	0.160	0.0504

Table 2. Performance comparison against state-of-the-art methods on AVA. We report the performance of the best methods that are reported in their respective works.

4.1. Results and Discussion

Table 1 compares the performance of various pretext tasks, particularly different configurations of our proposed inpainting scheme. Results indicate the viability and practicality of learning an inpainting task to learn robust features for AQA.

We achieved the best all-rounded result on Discriminator-FPP setting with an EMD that is comparable to a standard NIMA [1] on AlexNet. The performance of Discriminator-FPP is also comparable to other SSL pretext tasks evaluated (RotNet [10], Colorization [9]) but with much lesser complexity in terms of parameters, *i.e.* Resnet152 (37x larger), VGG16 (9.5x larger). Given that the size of ImageNet is around 60x larger than AVA, SSL models pre-trained on it performed marginally better expectedly, but our results show the sufficiency of AVA for the purpose of AQA. Table 2 summarizes the performance of our preliminary attempt at SSL in comparison with several benchmark methods on AVA. We hope this work spurs future directions towards SSL methods.

It is worth noting that transferring the generator network also did not work as well as with the discriminator network. Note also that we had attempted to fine-tune the ImageNet pre-trained inpainting model which was released by [8] for the purpose of benchmarking but the AQA training did not converge properly using the EMD loss function.

On evaluation metrics. The evaluation metric that is most commonly reported by the AQA community on AVA is a standard 2-class accuracy, which has been found to be sorely lacking [17, 1] due to class imbalance issues, *i.e.* AVA has over 180k of positive labels (good images) and only 74k of negative labels (bad images) if rating 5 is considered as the boundary. Therefore, [1] proposed to include distribution-based metrics such as linear correlation coefficient (LCC) between the mean and standard deviation of distributions, Spearman's rank correlation coefficient (SRCC) between the mean and standard deviation of distributions, and the EMD between distributions. In this paper, we reported all 6 evaluation metrics initiated by [1] for all conducted experiments.

Pre-training on other AQA datasets. Besides comparing the performance of models that were pre-trained on ImageNet and AVA, we also ran an experiment by pre-training Discriminator-FPP with image inpainting as SSL on a bigger dataset called AROD[16] before transferring back to AQA on AVA. The result was not as good as pre-training on AVA, with an accuracy of 80.47%, LCC (mean) of 0.449, SRCC (mean) of 0.412 and EMD of 0.526.

5. CONCLUSION

In this paper, we propose new image inpainting methods that abide by common photographic rules (rule-of-thirds and visual saliency) as self-supervisory signal for aesthetic quality assessment. We demonstrate the notion that we can teach a machine to understand aesthetic appeal by learning how to fill up salient areas of a photograph. Our comprehensive experiments show the feasibility of exploiting an unlabeled dataset to achieve comparable results at lesser complexity.

6. REFERENCES

[1] Hossein Talebi and Peyman Milanfar, "Nima: Neural image assessment," *IEEE Transactions on Image Processing*, vol. 27, no. 8, pp. 3998–4011, 2018.

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

[3] Xin Jin, Le Wu, Xiaodong Li, Xiaokun Zhang, Jingying Chi, Siwei Peng, Shiming Ge, Geng Zhao, and Shuying Li, "Ilgnet: inception modules with connected local and global features for efficient image aesthetic quality classification using domain adaptation," *IET Computer Vision*, vol. 13, no. 2, pp. 206–212, 2018.

[4] Ritendra Datta, Dhiraj Joshi, Jia Li, and James Z Wang, "Studying aesthetics in photographic images using a computational approach," in *ECCV*. Springer, 2006, pp. 288–301.

[5] Naila Murray, Luca Marchesotti, and Florent Perronnin, "Ava: A large-scale database for aesthetic visual analysis," in *IEEE CVPR*, 2012, pp. 2408–2415.

[6] Xin Lu, Zhe Lin, Hailin Jin, Jianchao Yang, and James Z Wang, "Rapid: Rating pictorial aesthetics using deep learning," in *Proc. of the 22nd ACM Multimedia*, 2014, pp. 457–466.

[7] Yong-Lian Hii, John See, Magzhan Kairanbay, and Lai-Kuan Wong, "Multigap: Multi-pooled inception network with text augmentation for aesthetic prediction of photographs," in *ICIP*. IEEE, 2017, pp. 1722–1726.

[8] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros, "Context encoders: Feature learning by inpainting," in *IEEE CVPR*, 2016, pp. 2536–2544.

[9] Richard Zhang, Phillip Isola, and Alexei A Efros, "Colorful image colorization," in *ECCV*. Springer, 2016, pp. 649–666.

[10] Spyros Gidaris, Praveer Singh, and Nikos Komodakis, "Unsupervised representation learning by predicting image rotations," *arXiv preprint arXiv:1803.07728*, 2018.

[11] Carl Doersch, Abhinav Gupta, and Alexei A Efros, "Unsupervised visual representation learning by context prediction," in *IEEE ICCV*, 2015, pp. 1422–1430.

[12] Mehdi Noroozi and Paolo Favaro, "Unsupervised learning of visual representations by solving jigsaw puzzles," in *ECCV*. Springer, 2016, pp. 69–84.

[13] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *IEEE CVPR*, 2009, pp. 248–255.

[14] Zhangyang Wang, Shiyu Chang, Florin Dolcos, Diane Beck, Ding Liu, and Thomas S Huang, "Brain-inspired deep networks for image aesthetics assessment," *arXiv preprint arXiv:1601.04155*, 2016.

[15] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich, "Going deeper with convolutions," in *IEEE CVPR*, 2015, pp. 1–9.

[16] Katharina Schwarz, Patrick Wieschollek, and Hendrik P. A. Lensch, "Will people like your image?," *CoRR*, vol. abs/1611.05203, 2016.

[17] Naila Murray and Albert Gordo, "A deep architecture for unified aesthetic prediction," *arXiv preprint arXiv:1708.04890*, 2017.

[18] Longlong Jing and Yingli Tian, "Self-supervised visual feature learning with deep neural networks: A survey," *CoRR*, vol. abs/1902.06162, 2019.

[19] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative adversarial nets," in *NIPS*, 2014, pp. 2672–2680.

[20] Xiaodi Hou and Liqing Zhang, "Saliency detection: A spectral residual approach," in *IEEE CVPR*, 2007, pp. 1–8.

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

[22] Vlad Hosu, Bastian Goldlucke, and Dietmar Saupe, "Effective aesthetics prediction with multi-level spatially pooled features," in *IEEE CVPR*, 2019, pp. 9375–9383.