Cuteness is Justice: Pawpularity Evaluation Based On Deep Learning

Wu Honghao 30920211154138, Ding Rui 30920211154127, Zhang Kun 30920211154176, Chen Zhipeng 30920211157133, Li Yitong 30920211154134

[1]School of Informatics Xiamen University
Xiamen, China

Abstract

In order to help shelters and rescuers around the world improve the attraction of their pet profiles, many online animal welfare platforms use a basic Cuteness Meter to rank pet photos at present. However, it's still in experimental phases and the algorithm could prospectively be improved with AI strategies. Within the scope of computer vision, image aesthetic evaluation is the most similar task. Several adaptive methods based on CNN have been proposed to receive original images with random sizes to maintain their aesthetic patterns. Transformer as a powerful model with proven effectiveness on computer vision tasks is also prospective. To explore this "cute issue", we apply ResNet model on our "Pawpularity Rating" task, and then reproduce the newest adaptive method so far to verify its effectiveness in the new application scenario. We also make the attempt to apply transformer to our task and confirm its effectiveness, which leads us to propose an assumption of its reason. The interpretation mainly concerns the specific inductive bias determined by the aesthetic task. Further confirmation of the assumption requires more delicate ablation experiments for rigorous comparison and more advanced semantic visualization strategies being designed.

Introduction

Unlike tasks of image classification and object detection, predicting pet picture's attraction to potential adopters is a task concerning much subjective factors, which are usually empirical and ambiguous. Within the scope of computer vision, image aesthetic evaluation is the most similar task but still lacks sufficient research.

The existing strategies focus on quantification of image quality and aesthetics, and are initially conducted through extracting aesthetic features according to both photographic rules (e.g., lighting, contrast) and global image composition (e.g., symmetry, rule of thirds), requiring extensive manual designs (Dhar, Ordonez, and Berg 2011; Ke, Tang, and Jing 2006; Nishiyama et al. 2011; Sun et al. 2009; Tong et al. 2004). However, manual designs for such aesthetic features is not a trivial task even for experienced photographers.

Deep convolutional neural network is consequently considered with its demonstrated effectiveness for various image classification tasks. The RAPID model (Lu et al. 2014) is among the first to apply deep convolutional neural networks (CNN) (Krizhevsky, Sutskever, and Hinton 2012) to the aesthetics rating prediction, where the features are automatically learned. This attempt comes up with the fact that Deep CNN methods are restricted by the structure of linear layers so that the model can only take fixed-size input. Input images need to be transformed via cropping, warping, or padding, which often alter image composition, reduce image resolution, or cause image distortion, result in potential loss of fine grained details and holistic image layout.

Lu et al. propose a deep multi-patch aggregation network to train models with multiple patches generated from one image (Lu et al. 2015) as a substitute of fine grained details. Ma et al. develop an adaptive patch selection strategy A-Lamp to enhance the training efficiency, and use the graph structure to keep holistic information (Ma, Liu, and Wen Chen 2017). In a more intuitive way, Mai et al. present a composition-preserving deep ConvNet with an adaptive spatial pooling layer to directly receive original input images. Chen et al. develop a novel adaptive fractional dilated convolution that is mini-batch compatible and overcomes the aspect ratio restriction of the ConvNet (Chen et al. 2020).

Transformer (Vaswani et al. 2017) was first applied to NLP tasks and achieved great performance (Devlin et al. 2018). Recent work have applied transformer on various vision tasks. Among these, the Vision Transformer (ViT) (Dosovitskiy et al. 2020) employs a pure Transformer architecture to classify images by treating an image as a sequence of patches. ViT has already been introduced to Image Quality tasks(Kaplan et al. 2012), but the work focuses more on image details determined by resolution rather global image composition, let alone being applied to our Pawpularity evaluation task.

One crucial point to mention is that, labels of most Aesthetic Evaluation Dataset (e.t AVA dataset), no matter in binary or continuous form, are evaluated under the guidance of photography and psychological rules. Our dataset's "pawpularity" is derived from each pet profile's page view statistics at the listing pages. Both of them lie in the conditional probability of human's subjective aesthetic and might deviate each other. Effectiviness of methods on aesthetic evaluation tasks are not guaranteed in our project. Besides, effectiveness of methods solving fix-size restriction are also ambiguous since the strategies are not widely applied. Implementation in practice may cause undefined trifles.

So in the project, we introduce Resnet model to pawpularity rating task and then further reproduce the adaptive fractional dilated convolution strategy as a comparison. Transformer is applied as the third step of our attemps. This work is initially carried as a contest on kaggle.com. We evaluate our methods' effect with RMSE according to the evaluation metric of the contest. The best performance of our model is 18.17 applying swin-transfomer (17.60 is the overall best result in the contest). The effectiveness of adaptive dilated convolution strategy is confirmed owing to our reproduction, with a 18.28 RMSE, better than Resnet's 18.38. We also proposed an assumption of rationality of applying transformer to aesthetic evaluation tasks concerning long-range dependencies and images' global composition.

Related Work

Murray's introduction of the AVA dataset on aesthetic assessment and their efforts on manually extracting features for style classification opens up this field (Murray, Marchesotti, and Perronnin 2012) . The following deep learning attempt, conducted by Lu's double-column CNN (Lu et al. 2014) consisting of four convolutional layers and the restricted input size of 224×224 through cropping fully exposes the main challenge in this field: how to maintain the fine grained details, holistic image layout and simultaneously modify images for input. Many efforts have been made to deal with the restriction.

Multi-Patch strategies

From the very beginning, Lu has already made efforts on this fixed-size restriction by designing CNN architectures which simultaneously take multiple versions of the transformed images as input (Lu et al. 2014). This work is further improved with a deep multi-patch aggregation network (DMA-Net) taking multiple randomly cropped patches of fixed size as input (Lu et al. 2015). On top of that, Ma develops a patch selection strategy A-Lamp to take better use of random patches (Ma, Liu, and Wen Chen 2017). These work have shown some promising results, but are still indirect strategies. Fixed aspect ratio cropping does harm to holistic image layout information at the same time.

Adaptive-Layer strategies

Inspired by the success of the SPP-Net for visual recognition (He et al. 2015), strategies to construct adaptive spatial pooling layers have been proposed. The approach of MNA-CNN (Mai, Jin, and Liu 2016), containing multiple sub-networks for different automatically modified pooling sizes, is proposed to preserve image aspect ratios and compositions by feeding the original image, one at a time. Limitation lies on the fact that this strategy can not take images with different aspect ratios for batch process. Then an adaptive fractional dilated convolution (AFDC) is developed which is compatible for mini-batch, using linear interpolation between convolution kernels with different dilation rates.

Transformer in Computer Vision Tasks

Dosovitskiy et al. propose Vision Transformer (ViT) (Dosovitskiy et al. 2020), which is a pure transformer that performs well on image classification task when applied directly to the sequences of image patches. They follow transformer's original design as much as possible. Then plenty of works try to augment a conventional transformer block or self-attention layer with convolution. There has been growing interest in using transformer for high/mid-level computer vision tasks, such as object detection (Beal et al. 2020; Zhu et al. 2020) and segmentation (Wang et al. 2021a,b). Transformer is introduced to low-level vision fields as well. For instance, Image Generation (Parmar et al. 2018), image of processing (Yang et al. 2020). Potential for processing multi-modal tasks with transformer is also under exploration.

Aesthetic Evaluation

The most important goal of the developing objective IQA is to accurately predict the perceived quality by human viewers. The primary criterion of performance measurement is the accuracy of the metrics. Spearman rank order correlation coefficient (SRCC) and the Kendall rank order correlation coefficient (KRCC) are used to estimate the monotonicity and consistency of the quality prediction. Attempts to apply transformer to the area includes evaluating layout quality in UI design, high fidelity prototype (Rahman, Sermuga Pandian, and Jarke 2021) and art price appraisal (Cheon et al. 2021). These works demonstrate the prospects of introducing transformer to aesthetic evaluation fields.

Proposed Solution

baseline-ResNet

Before ResNet is proposed, the depth of deep convolutional neural networks, such as AlexNet, VGG, does not exceed 100, while the depth of ResNet reaches 152 layers, and its complexity is lower than that of VGG. In tasks such as image classification and object detection, ResNet achieves the best performance. Previous work on aesthetic evaluation mainly use AlexNet or VGG as the baseline, or their own network skeletons are AlexNet or VGG (Ma, Liu, and Wen Chen 2017; Lu et al. 2015; Mai, Jin, and Liu 2016). These models are too old to provide enough depth. Therefore, we choose ResNet as our baseline. Although ResNet has been proven to perform well on tasks such as image classifications and object detection, there is still insufficient evidence to show that ResNet can maintain the same performance on image evaluation problems, let alone on our Pawpularity task. Therefore, we will make necessary modifications to ResNet. Since we have less training data, we will fine-tune the existing pretrained ResNet as much as possible to achieve the best performance of the model.

AFDC-Net

In methods before AFDC-Net, the backbone networks are usually adopted from image classification networks. The data augmentation methods, i.e. image cropping and warping, are widely used for preventing overfitting in the image recognition task. A shortcoming is that the compositions and object aspect ratios are altered, which may introduce label noise and harm the task of aesthetic assessment. A simple solution proposed in MNA-CNN (Mai, Jin, and Liu 2016) is to feed one original-size image into the net-work at a time during training and test (bottom stream in Fig.1). A major constraint of the approach is that images with different aspect ratios cannot be concatenated into batches because the aspect ratio of each image should be preserved. Thus it slows down the training and inference.

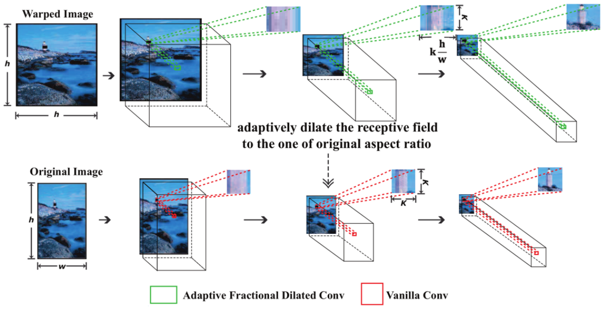

Warped Image
h
h
k
k h/w
adaptively dilate the receptive field
to the one of original aspect ratio
Original Image
h
w
k
k
🟩 Adaptive Fractional Dilated Conv
🟥 Vanilla Conv

Figure 1: Overview of adaptive fractional dilated CNN (above) and the comparison with vanilla CNN (below): Each fractional dilated Conv (above) operated on wrapped input adaptively dilates the same receptive field as the vanilla Conv (below) operated on the original image. It thus helps with the problems: (a) Becomes mini-batch compatible by composition-preserving warping instead of feeding original-size image (b) Preserves aesthetic features related to aspect ratios by adaptive kernel dilation (Chen et al. 2020).

AFDC-Net uses a novel adaptive fractional dilated convolution that is mini-batch compatible. As shown in the top row in figure1, the network adaptively dilates the convolution kernels to the composition-preserving warped images according to the image aspect ratios such that the effective receipt field of each dilated convolution kernel is the same as the regular one. Specifically, as illustrated in figure2, the fractional dilated convolution kernel is adaptively interpolated by the nearest two integer dilated kernels with the same kernel parameters. Thus no extra learning parameters are introduced.

We fully reproduced AFDC-Net and inspected whether it can perform well in our Pawpularity evaluation problem.

Swin-Transformer

Transformer (Vaswani et al. 2017) was initially applied to tasks in the NLP field and achieved sota performance. Subsequently, Transformer was used in the field of computer vision and also achieved good performance (Dosovitskiy et al. 2020). Transformer divides the picture into different patches, and extracts the features on the patches and the features that can measure the information between the patches through the self-attention mechanism. The subsequent Swin Transformer (Liu et al. 2021) introduced Hierarchical Feature Representation and Shifted Window based Multi-head Self-attention, which not only reduces the computational complexity of the model but also improves the performance of the model. Prior to this, there were also many methods based on CNN to divide the picture into patches (Ma, Liu, and Wen Chen 2017; Lu et al. 2015; Mai, Jin, and Liu 2016), trying to extract the composition information of the picture. Although these methods have improved the problem of image evaluation, we think that as a patch-based method, the model mechanism of transformer can better extract the associating information between patches, perceive the layout information of the picture, and then output better evaluation scores that are consistent with subjective rating. Therefore, we investigated the performance of swin transformer on our Pawpularity evaluation problem and analyzed the experimental results.

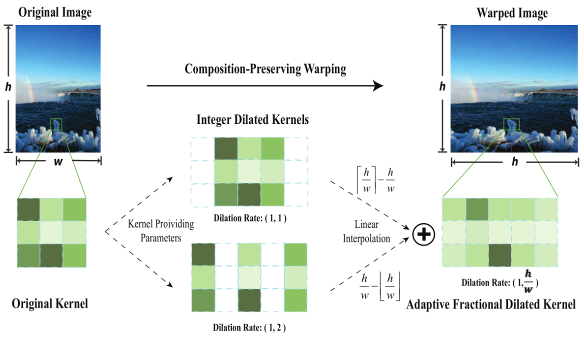

Original Image
h
w
Composition-Preserving Warping
Warped Image
h
h
Integer Dilated Kernels
Kernel Proividing Parameters
Dilation Rate: (1, 1)
$\lceil \frac{h}{w} \rceil - \frac{h}{w}$
Linear Interpolation
$\frac{h}{w} - \lfloor \frac{h}{w} \rfloor$
Dilation Rate: (1, 2)
Original Kernel
Dilation Rate: (1, $\frac{h}{w}$)
Adaptive Fractional Dilated Kernel

Figure 2: Illustration of kernel interpolation: linear interpolation of the nearest two integer dilated kernels shared same kernel parameters are used to tackle the sampling misalignment from fractional dilation rates (Chen et al. 2020).

Experiments

Data set and data exploration

The data set is provided by PetFinder.my, which is Malaysia's leading animal welfare platform, featuring over 180,000 animals with 54,000 happily adopted. The data set includes 9912 photos of cats and dogs, and each photo corresponds to a Pawpularity Score from 1 to 100 (Fig. 3). The Pawpularity Score is derived from each pet profile's page view statistics at the listing pages, using an algorithm that normalizes the traffic data across different pages, platforms and various metrics. We divided the data set into training set, validation set and test set according to the ratio of 6:2:2. We also explored the distribution of Pawpularity Score and found that the distribution showed a clear left-bias. The data at both ends also reveales a higher distribution density (Fig. 4).

Experimental setup

Model details We used ResNet pre-trained on Imagenet as our baseline model. In order to adapt to our image evaluation problem, we modified the last few layers of ResNet and carefully fine-tuned them to utilize its performance to the best. Multi-scale data augmentation were applied to avoid overfitting simultaneously, and the training scale was set to [224,

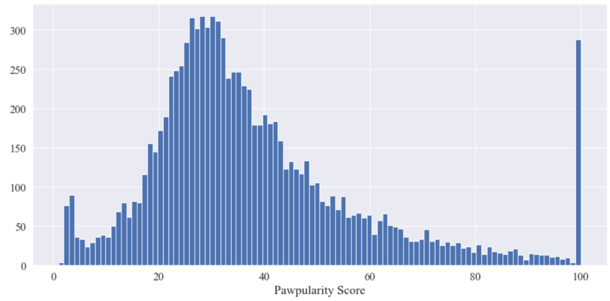


Figure 3: distribution of Pawpularity Score

256, 288, 320, 352, 384]. We reproduced AFDC-Net and modified the model accordingly to adapt it to our specific problems as the second attempt in the experiment. On one hand, we modified the last few layers of the model so that the model can output evaluation scores. In order to speed up the training of AFDC-Net, images with similar aspect ratios are formed into a batch when data is loaded on the other hand. The aspect ratios of the images in the data set are distributed in four intervals. Chen et al.combines the intervals to make the aspect ratios of the images in each batch more diverse (Chen et al. 2020), but this increases the computational complexity of the model. Therefore, we did not adopt interval merging in our experiment. In addition, while training AFDC-Net, we used the same multi-scale data augmentation methods same as what we apply when training ResNet. Like ResNet, the Swin Transformer we used was also pre-trained on Imagenet. We modified the last few layers of the model similarly. The difference is that when training Swin Transformer, we adopt mixup method for data augmentation. During training, the model performs mixup with a probability of 50

Loss function Since the value of Pawpularity Score is between 1 and 100, we scale the value of Pawpularity Score to between 0 and 1 and use the sigmoid function in the last layer of the model in order to limit the range of the output value of the model. As a regression problem, the most common loss function is Mean Square Error(RMSE). However, since we limited the output of the model to between 0 and 1, we can treat it as a probability value. so consequently we can make an attempt to apply Binary Cross Entropy(BCE) as a loss function to train the model. In our experiment, we trained the model separately with RMSE and BCE as the loss function, and compared their differences.

Experimental result

Experimental results (Tab.1) reveal that Swin Transformer has the best performance. AFDC-Net also shows better performance than baseline because it retains the aspect ratio information of the image. This confirms that, for image evaluation problems, a model that can perceive image layout information will indeed perform better. At the same time, we noticed that the model applying BCE as the loss function is better than the model using RMSE.

Model	RMSE	BCE
ResNet	18.41	18.38
AFDC-Net	18.31	18.28
Swin Transformer	18.26	18.17

Table 1: Use RMSE to evaluate the quality of the model, the smaller the better

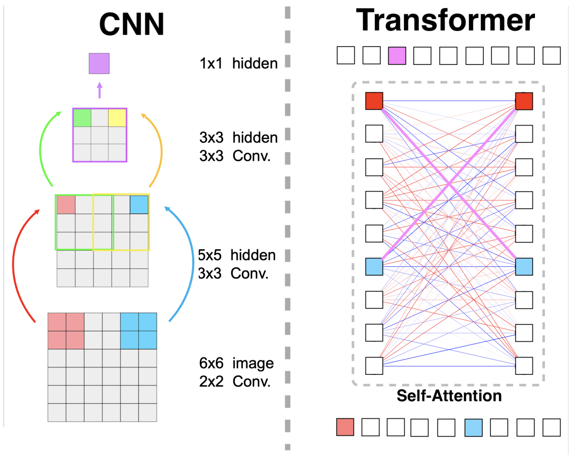


Figure 4: Extracting same features correlating two patches in a long-range, CNN uses more layers than the transformer. The correlation between distant patches can be considered as global composition.

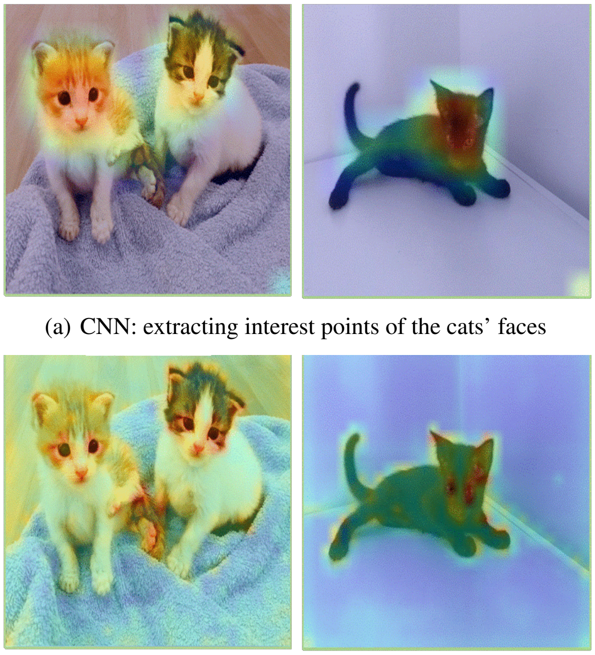
Analysis

Transformer's compatibility for the task Transformer's effectiveness on our Pawpularity task is revealed in some extent according to the result in table1 and table2, which also indicates its prospective power on Aesthetic assessment. The reason behind can be intuitively interpreted by two aspects: the requirements of our specific task and transformer's own characters as a feature extractor.

Although CNN can achieve translation invariance as a positive character for feature extracting, which guarantees its steady effectiveness on image classification tasks, it makes little contribution to our tasks. For instance, if one image under evaluation is split into fixed-size patches and randomly rearranged, features extracted by CNN and the classification result may be invariant, but its aesthetic quality is not even similar to the original.

Unlike the convolution operation in CNNs that has a relatively limited receptive field, self-attention is applied to the whole input sequence, it can therefore effectively capture the image composition information rather than sliding regularly through patches. Consequently, feature maps generated from self-attention models are not constrained in the spatial extent. The most appropriate inductive bias based on the specific task and positions of the layer in the network can be achieved. This can also be interpreted as long-range dependencies among image patches.

Aesthetic evaluation tasks are characterized by its reliance on global image composition information. We propose that



(a) CNN: extracting interest points of the cats' faces

(b) Swin-Transfomer: extracting extra features of cats and including background information

Figure 5: visualization of features extracted by CNN and Swin-transformer

this can be understood as relationship between visual elements in an image. As is shown in figure 4, transformer could extract features with correlations between two relatively distant patches in much efficient ways. The ultimate goal of learning global composition patterns of an image could also be considered as finding its special intuitive bias, for instance, features distributed on appropriate localization. Through visualizing semantic features extracted from the model, we can define that transformer extracted not only the object in the image but also visual guidelines in the scene, compared to simply using RNN. (Fig.5)

BCE vs RMSE In order to find out the reason why BCE is better than RMSE, we examined the curve of BCE when the ground-truth takes different values (Fig.6). Unlike RMSE, which has a symmetrical curve, BCE's curve is asymmetrical. If the difference between the predicted value and the ground-truth is the same, the loss when the predicted value falls in the middle is lower than the loss when it falls on bilateral ends, so the model is prone to predict the value in the middle.

In this way, the distribution of the predicted value of the model will be more similar to the distribution of the true value, but this does not mean that the model can better evaluate the picture. We believe that our model does not need to accurately predict the Pawpularity Score of a picture, but should be able to compare the rank of the Pawpularity Score of different pictures. So we used the spearman correlation coefficient to compare the performances of different models in predicting the rank of Pawpularity Score of the picture(Tab.2). The results show that although the models using RMSE as loss function performs worse on quantitive result (RMSE), they can better predict the rank of the Pawpularity Score of the picture. Therefore, using RMSE as the loss function is a more reasonable choice.

Model	RMSE	BCE
ResNet	0.425	0.402
AFDC-Net	0.437	0.418
Swin Transformer	0.458	0.425

Table 2: Spearman correlation coefficient of the models, the bigger the better

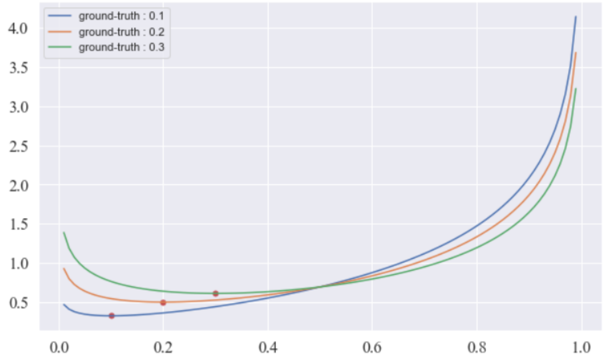


Figure 6: Loss function curve when ground-truth takes different values

Conclusion

In the project, we introduced the most popular model in computer vision filed, ResNet, to animals cuteness rating task as an initial attempt. Then we reproduced the adaptive fractional dilated convolution strategy, further improve its performance by fine-tuning, and confirmed the its effeteness at the same time. Transformer's application to the task is rough in our attempt, but still provided evidence for its effectiveness for the task. We also proposed intuitive assumption of transformer's prospect effectiveness in aesthetic evaluation tasks based on our attempt. Interpretation logic is the most crucial part. Further confirmation of the assumption requires more delicate ablation experiments for rigorous comparison. We should also cast this attempts on a dataset more directly related to aesthetic tasks, such as AVA dataset. Further interpretations of the task should rely on more advanced semantic visualization strategies being proposed in the future.

References

Beal, J.; Kim, E.; Tzeng, E.; Park, D. H.; Zhai, A.; and Kislyuk, D. 2020. Toward transformer-based object detection. *arXiv preprint arXiv:2012.09958*.

Chen, Q.; Zhang, W.; Zhou, N.; Lei, P.; Xu, Y.; Zheng, Y.; and Fan, J. 2020. Adaptive fractional dilated convolution network for image aesthetics assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14114–14123.

Cheon, M.; Yoon, S.-J.; Kang, B.; and Lee, J. 2021. Perceptual image quality assessment with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 433–442.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dhar, S.; Ordonez, V.; and Berg, T. L. 2011. High level describable attributes for predicting aesthetics and interestingness. In *CVPR 2011*, 1657–1664. IEEE.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2015. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 37(9): 1904–1916.

Kaplan, H. C.; Provost, L. P.; Froehle, C. M.; and Margolis, P. A. 2012. The Model for Understanding Success in Quality (MUSIQ): building a theory of context in healthcare quality improvement. *BMJ quality & safety*, 21(1): 13–20.

Ke, Y.; Tang, X.; and Jing, F. 2006. The design of high-level features for photo quality assessment. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 1, 419–426. IEEE.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25: 1097–1105.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. *arXiv preprint arXiv:2103.14030*.

Lu, X.; Lin, Z.; Jin, H.; Yang, J.; and Wang, J. Z. 2014. Rapid: Rating pictorial aesthetics using deep learning. In *Proceedings of the 22nd ACM international conference on Multimedia*, 457–466.

Lu, X.; Lin, Z.; Shen, X.; Mech, R.; and Wang, J. Z. 2015. Deep multi-patch aggregation network for image style, aesthetics, and quality estimation. In *Proceedings of the IEEE international conference on computer vision*, 990–998.

Ma, S.; Liu, J.; and Wen Chen, C. 2017. A-lamp: Adaptive layout-aware multi-patch deep convolutional neural network for photo aesthetic assessment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4535–4544.

Mai, L.; Jin, H.; and Liu, F. 2016. Composition-preserving deep photo aesthetics assessment. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 497–506.

Murray, N.; Marchesotti, L.; and Perronnin, F. 2012. AVA: A large-scale database for aesthetic visual analysis. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 2408–2415. IEEE.

Nishiyama, M.; Okabe, T.; Sato, I.; and Sato, Y. 2011. Aesthetic quality classification of photographs based on color harmony. In *CVPR 2011*, 33–40. IEEE.

Parmar, N.; Vaswani, A.; Uszkoreit, J.; Kaiser, L.; Shazeer, N.; Ku, A.; and Tran, D. 2018. Image transformer. In *International Conference on Machine Learning*, 4055–4064. PMLR.

Rahman, S.; Sermuga Pandian, V. P.; and Jarke, M. 2021. RUITE: Refining UI Layout Aesthetics Using Transformer Encoder. In *26th International Conference on Intelligent User Interfaces*, 81–83.

Sun, X.; Yao, H.; Ji, R.; and Liu, S. 2009. Photo assessment based on computational visual attention model. In *Proceedings of the 17th ACM international conference on Multimedia*, 541–544.

Tong, H.; Li, M.; Zhang, H.-J.; He, J.; and Zhang, C. 2004. Classification of digital photos taken by photographers or home users. In *Pacific-Rim Conference on Multimedia*, 198–205. Springer.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in neural information processing systems*, 5998–6008.

Wang, H.; Zhu, Y.; Adam, H.; Yuille, A.; and Chen, L.-C. 2021a. Max-deeplab: End-to-end panoptic segmentation with mask transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5463–5474.

Wang, Y.; Xu, Z.; Wang, X.; Shen, C.; Cheng, B.; Shen, H.; and Xia, H. 2021b. End-to-end video instance segmentation with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8741–8750.

Yang, F.; Yang, H.; Fu, J.; Lu, H.; and Guo, B. 2020. Learning texture transformer network for image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5791–5800.

Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; and Dai, J. 2020. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*.