Self-Supervised Multi-Task Pretraining Improves Image Aesthetic Assessment

Jan Pfister, Konstantin Kobs, Andreas Hotho
University of Würzburg
Germany
{pfister,kobs,hotho}@informatik.uni-wuerzburg.de

Abstract

Neural networks for Image Aesthetic Assessment are usually initialized with weights of pretrained ImageNet models and then trained using a labeled image aesthetics dataset. We argue that the ImageNet classification task is not well-suited for pretraining, since content based classification is designed to make the model invariant to features that strongly influence the image's aesthetics, e.g. *style-based features such as brightness or contrast.*

We propose to use self-supervised aesthetic-aware pretext tasks that let the network learn aesthetically relevant features, based on the observation that distorting aesthetic images with image filters usually reduces their appeal. To ensure that images are not accidentally improved when filters are applied, we introduce a large dataset comprised of highly aesthetic images as the starting point for the distortions. The network is then trained to rank less distorted images higher than their more distorted counterparts. To exploit effects of multiple different objectives, we also embed this task into a multi-task setting by adding either a self-supervised classification or regression task. In our experiments, we show that our pretraining improves performance over the ImageNet initialization and reduces the number of epochs until convergence by up to 47 %. *Additionally, we can match the performance of an ImageNet-initialized model while reducing the labeled training data by* 20 %. *We make our code, data, and pretrained models available.*

1. Introduction

Assessing the aesthetics of an image automatically can be used to choose the most aesthetic images [23], sort image collections [9], or optimize image editing filter parameters [5]. Modern Image Aesthetic Assessment (IAA) methods are based on Convolutional Neural Networks (CNNs) that receive an image as input and output a score that is higher for more aesthetic images. Such models are usually initialized with weights trained on the ImageNet classification task [4] to build on the already learned features by fine-tuning the network on a labeled dataset such as AVA [29]. We argue that the ImageNet classification task is not well-suited for IAA models, since it is not designed to teach the network aesthetically relevant features. Due to the classification objective, it even discourages features important for aesthetic assessment. For example, a classification network should be invariant to the image's lighting conditions and thus discourages features taking the image's brightness into account. We therefore propose to pretrain the model on pretext tasks that are specifically designed to let the network learn relevant features to assess the aesthetics of an image.

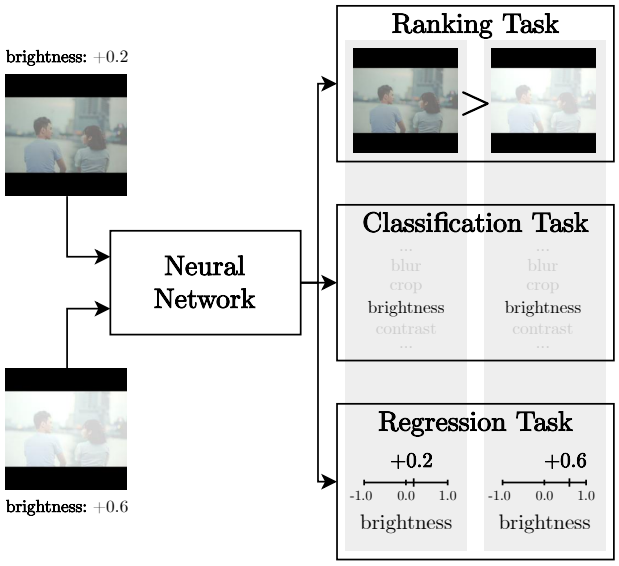

Figure 1. A schematic overview of our novel pretext tasks. Images are singly distorted using image filters in different intensities. The neural network then learns to output higher scores for less distorted images. In a multi-task setting, the network additionally classifies the distortion or estimates the distortion's intensity.

In the related task of *technical* Image Quality Assessment (IQA), the method RankIQA [26] pretrains a CNN to rank images based on the intensity of an applied *technical* distortion. Images of high technical quality are distorted by applying artificial technical distortions such as noise addi-

tion, blurring, or JPEG compression. Since technical distortions clearly degrade an image, the network can be trained to assess images with higher distortion intensity to be of lower quality in a self-supervised fashion. We propose to adapt and extend this self-supervised pretraining for the task of Image Aesthetic Assessment. Instead of just technical distortions, we apply image filters that usually change an image's *technical quality* (*e.g.* noise), *style* (*e.g.* contrast), or *composition* (*e.g.* cropping). However, applying style or composition filters to ordinary images can lead to undesired improvements in the image aesthetics, violating the assumption of the self-supervised task. Thus, we introduce a large dataset of highly *aesthetic* images, consisting of the most popular images from the stock photo website pexels.com. Showing that applying filters to those images results in less appealing images, we can adopt the same ranking task for IAA.

According to related IQA literature [19, 27, 39, 8, 40], embedding similar approaches into a multi-task training setting can improve the prediction performance of the resulting image assessor by making use of additionally available information. Therefore we propose to combine this ranking task with a classification or regression task adapted for IAA: While the ranking task itself focuses on predicting relative changes in image aesthetics, the classification learns to predict the applied distortion and the regression task estimates the intensity of the currently applied distortion. Figure 1 shows a high-level overview of our proposed pretext tasks.

In our experiments, we use the IAA method NIMA [36] as a reference and replace its pretrained weights with the weights from our other common self-supervised pretext tasks before fine-tuning on the labeled AVA dataset [29]. We find that we can improve the performance over the baseline methods while reducing the number of epochs until convergence by up to $47\,\%$ over the ImageNet-initialized NIMA. In an analysis, we find that we match the fully supervised model's correlation and Mean Absolute Error metrics while requiring approximately $20\,\%$ less training data.

Our main contributions are: 1. We propose and evaluate self-supervised *aesthetic-aware* pretext tasks for Image Aesthetics Assessment. 2. We introduce a new dataset containing aesthetically pleasing images from a popular stock photo website. 3. We make code, data, and models available to encourage the trend of *aesthetic-aware* pretraining methods[1].

The paper is structured as follows: Section 2 gives an overview on related work. Section 3 explains the method we propose to learn aesthetic-aware features. In Section 4, we introduce our dataset and explain the experimental setup. Results are given in Section 5. Section 6 and Section 7 discuss the results and conclude the work, respectively.

[1] https://github.com/janpf/self-supervised-multi-task-aesthetic-pretraining

2. Related Work

General Image Aesthetic Assessment (IAA) aims to automatically assess the aesthetics of images. Most recent top-performing approaches [35, 36, 28, 14] predicting image aesthetics on the AVA dataset [29] train exclusively on human-sourced aesthetic scores like the Mean Opinion Score (MOS). Notably Talebi *et al.* [36] introduce NIMA, a CNN that is trained to minimize the Earth Mover's Distance loss (EMD) [15] on the human-sourced voting *distribution* from the AVA dataset [29]. Due to its solid performance despite its comparably simple architecture we select NIMA for our extensive studies of our different pretext tasks. We use the lightweight and efficient MobileNet [33] architecture as proposed in [36] to enable us to test several different setups and configurations in this paper.

No-Reference Image Quality Assessment A task related to IAA is *No Reference Image Quality Assessment (NR-IQA)*, which assesses the *technical* quality of images. As for IAA, many existing NR-IQA approaches make heavy use of human-sourced quality scores [1, 18, 36, 8]. Under the assumption that distorting images degrades their technical quality, RankIQA [26] distorts technical aspects of high-quality images and learns to rank the resulting images against each other using an efficient ranking loss, which we adopt in this work. In addition, we embed this task into a multi-task setting.

While in the literature most multi-task NR-IQA approaches apply a classification task [19, 27, 39, 8] predicting the type of distortion, we also include a regression task [40] estimating the distortion intensity.

Self-Supervised Feature Learning Our approach aims to learn useful representations without manual annotations. Many self-supervised pretext tasks for CNNs [17, 7, 10, 21] teach universally applicable image features using tasks like: rotation classification [22], image part shuffling, or super-resolution [17]. Sheng *et al.* [34] transfer distortion-based self-supervised NR-IQA pretext tasks to the IAA task. They are able to improve the prediction performance for the downstream IAA task on AVA over the de-facto default ImageNet [4] pretext task [35, 36, 28, 20] by pretraining on mainly *technical* quality focused distortions in a self-supervised multi-task pretext setup. We expand on this approach by explicitly introducing style and composition aware distortions and evaluating different pretext task combinations against each other.

3. Methodology

Our method is based on the observation that singly distorting a high-quality image results in a lower-quality image, which is exploited in the *technical Image Quality As-*

sessment method RankIQA [26]. We transfer their approach to the Image Aesthetic Assessment setting by using image filters that degrade not only technical quality, but different aspects of image aesthetics. To overcome the challenge that applying image filters such as contrast adjustments might enhance images, we define requirements for the dataset used in the pretext tasks. By training a neural network to output higher scores for less distorted versions of an image, the network has to learn what filter intensity is more appropriate between two images. We propose to also optionally add one of two other pretext tasks in a multitask setting, making use of other objectives besides the relative ranking: 1. Distortion identification guides the network to differentiate between distortions and 2. distortion intensity estimation lets the model learn about the absolute intensity. Our three proposed pretext tasks are depicted in Figure 1.

Following the common transfer learning setting, the resulting model acts as initialization for fine-tuning on a labeled image aesthetic dataset in a supervised fashion. We later show that our pretrained models learn more suitable features for the IAA task than the common ImageNet [4] initialization [35, 36, 28, 20].

3.1. Aesthetic-Aware Image Distortions

From related works [36, 34] we identify five essential aspects of general image aesthetics: technical quality, style, composition, content, and semantics. For our pretext tasks we aim to select image filters that distort aesthetic base images regarding these aspects. Since it is non-trivial to find image filters that distort image content and semantics in a self-supervised way [34], we focus on the remaining aspects *technical quality*, *style*, and *composition*, even though it is possible to include content or semantic based distortions if such image filters are found. In the following, we discuss these aspects in general, while in Section 4.1, we select specific image distortions.

Technical Quality Technical distortions can appear on an image under real world use cases, when taking, saving or transmitting an image, *e.g.* compression, blurring or the addition of noise. It is well documented that applying such technical distortions to an image lowers its perceived quality [19, 27, 39, 40, 8, 31]. For our method, a set of distortions D_{tech} influencing this aspect has to be chosen.

Image Style Image style is mostly described through image properties such as contrast, brightness, or saturation. Thus, style based distortions consist of image filters that are often used for color correction and color grading. Applying a filter from the chosen filter set D_{style} to an aesthetic image results in a degradation in quality, especially when using high filter intensities.

Image Composition Image composition concerns the location of subjects and objects in the image, thus it is highly correlated with the chosen image crop. We assume that for highly aesthetic images, the original framing is selected in an aesthetically pleasing way, *e.g.* by following common guides such as the rule of thirds. Operations like crops or rotations then destroy such alignments. Applying these distortions D_{comp} in pretraining has the additional benefit of making the network learn to recognize structures in images in general, which has been shown to be useful in similar image pretext tasks [17, 7, 10, 21].

3.2. Highly Aesthetic Dataset

The dataset chosen for pretraining is important for our proposed method, as the purpose of applying any distortion described in Section 3.1 to an image from the dataset is to degrade its aesthetics regarding the corresponding aspect. In RankIQA, high-quality images with regard to the technical aspect are required to make sure that applying a technical distortion does in fact degrade an image's quality [26]. The application of style and composition filters thus has to consistently degrade the associated aspect of aesthetic quality of an image from the dataset. We therefore derive two requirements for the pretraining dataset used in our method:

Highly Aesthetic The dataset needs to contain only highly aesthetic images with regard to their *technical quality*, *style*, and *composition*. This minimizes the risk that applying a filter accidentally improves the image's aesthetics, since the undistorted image presumably already has optimal filter parameters.

Diverse in Style and Content To prevent the network from overfitting on a specific editing style or image content, the dataset needs to contain a wide variety of different images. High content diversity ensures that the model learns to generally correlate content with style features, *e.g.* sunsets with orange tints or portraits with natural skin tones. Consequently the dataset needs to be of sufficient size to meet the requirement regarding its diversity.

3.3. Self-Supervised Aesthetic-Aware Pretext Tasks

Applying the selected aesthetic-aware image distortions (Section 3.1) to the high-quality images (Section 3.2) results in a dataset containing the original, unedited images and some automatically generated lower quality image variants (regarding technical quality, style, and composition). While there is no absolute aesthetic score for neither the original image nor any of the generated images, we can access the intensity of the applied distortion and the fact that a higher intensity of an applied distortion makes the image look less aesthetically pleasing. In the following we introduce our main ranking-based self-supervised pretext task as well as

two additional tasks based on classification and regression. We combine the ranking task with each of the other tasks in a multi-task setting to guide the network to learn features related to image aesthetics. In our experiments, we then assess the effects of the pretext tasks on the downstream IAA performance.

Given an image I and a set of distortions that unites all image filters defined for technical quality, style, and composition $D_{all} = D_{tech} \cup D_{style} \cup D_{comp}$. Each distortion $d \in D_{all}$ has a set of possible intensity values $V^{(d)} = \{v | v \in \mathbb{R}\}$ that can be positive or negative. Applying the filter with these distortion intensities to the image I, we obtain a set of distorted images $I^{(d)} = \{I_v^{(d)} | v \in V^{(d)}\}$. An intensity value of zero equals the original image $I_0^{(d)} = I$.

In general, a neural network pretrained on ImageNet is used as the base model. For each task, the last layer is replaced in order to conform to the desired output.

Ranking Aesthetic Value Our main pretext task is a ranking task based on the RankIQA method for technical image assessment [26]. Given all distorted versions of an image, the original image $I_0^{(d)}$ is assumed to be the most aesthetically pleasing and should therefore be rated with the highest score. Images $I_i^{(d)}$ with a higher absolute distortion intensity value $|i| > 0$ should decrease the image aesthetics, *e.g.* increasing the brightness results in a less appealing image and a lower score compared to an image with a weaker increase in brightness. Given these assumptions, we utilize the ranking-based pretext task from [26] that ranks all images with respect to their intensity value. With our dataset and aesthetic-aware image distortions, we are able to apply this method to Image Aesthetic Assessment. The neural network is supposed to output scalars that are larger for higher-quality images and predict smaller values for images with larger distortion intensities. We let the last layer of the base neural network output one scalar and apply a sigmoid activation function.

For one image, the ranking loss is defined as

$$L_{\text{rank}} = \sum_{\substack{I_i^{(d)}, I_j^{(d)} \in I^{(d)}, \\ d \in D_{all}, \\ |i| < |j|, \\ \text{sign}(i) = \text{sign}(j)}} \max\left(0, (f_{rank}(I_i^{(d)}) - f_{rank}(I_j^{(d)})) - m\right), \tag{1}$$

where m is the margin denoting the desired minimal output difference between two differently distorted images when fed through the ranking network f_{rank} which returns a score between 0 and 1.

Classifying Applied Distortions In addition to the ranking task, we optionally train a distortion identification task in a self-supervised manner as done in some NR-IQA methods [19, 27, 39, 8]. We replace the network's last layer with three separate layers, one for each quality aspect $a \in \{tech, style, comp\}$. Each layer has $|D_a|$ outputs followed by a softmax activation function to output probabilities $f_{class}^a(I)$ for a given input image I.

We consider each quality aspect separately to allow cross synergies, *e.g.* adding noise to an image often results in lower saturation. A single probability distribution across all distortions could not consider such synergies and would increase the loss since multiple changes were correct.

The loss for this pretext task is thus defined as

$$L_{\text{class}} = \sum_{\substack{I_i^{(d)} \in I^{(d)}, \\ d \in D_a, \\ a \in \{tech, style, comp\}}} L_{CE}\left(f_{class}^a\left(I_i^{(d)}\right), d\right), \tag{2}$$

where L_{CE} is the Categorical Cross-Entropy loss function taking the output probability distribution and the index of the applied distortion.

Estimating Intensity of Applied Distortions While the distortion identification task only classifies the distortion, another task we can add to our multi-task setting is to explicitly predict the distortion intensity. The last network layer is replaced to output $|D_{all}|$ values, one for each distortion. We calculate the loss by applying the squared error to the output for the applied distortion. Only calculating the error at the output index of the applied distortion makes it possible to model cross relations between distortions.

The loss function for the regression task is

$$L_{regr} = \sum_{\substack{I_i^{(d)} \in I^{(d)}, \\ d \in D_{all}}} \left(f_{regr}\left(I_i^{(d)}\right)_d - \text{norm}(i)\right)^2, \tag{3}$$

where $\text{norm}(i)$ is the normalized intensity value to be predicted by the regression network f_{regr} such that all non-negative intensities are normalized to the range $[0, 1]$ and all non-positive intensity values are normalized to the range $[-1, 0]$. This scales all intensity values to similar ranges, lowering the influence of single distortions on the loss.

3.4. Multi-Task Pretraining and Fine-Tuning

Related IQA approaches [19, 27, 39, 8, 40] have shown that a multi-task training setup improves the resulting image assessor due to the additional information being available during training. For our proposed multi-task setting, we combine the ranking task with either the classification or regression task by adding the losses for the given pretext tasks using a loss balancing scheme [24]. The resulting overall loss is

$$L = h\left(L_{rank}\right) + h\left(L_{class}\right) \tag{4}$$

Aspect	Distortion
Technical	JPEG compression[a], Defocus blur[a], Motion blur[a], Pixelate[a], Gaussian noise[a], Impulse noise[a]
Style	Brightness[b], Contrast[b], Saturation[b], Exposure[b], Shadows[c], Highlights[d], Temperature[e], Tint[b], Vibrance[f]
Composition	Rotation[b], Horizontal crop[b], Vertical crop[b], Left Diagonal crop[b], Right Diagonal crop[b], Image Ratio[b]

[a][0,...,5] [b][-5,...,5] [c][-5,...,3] [d][-3,...,5] [e][-4,...,5] [f][-2,...,4]

Table 1. The distortions applied in our pretext tasks and their intensity value ranges, grouped by their corresponding aesthetics aspect. Each distortion also contains the original image (intensity 0). Parameters passed to the libraries can be found in Appendix A.

when adding classification to the ranking task and

$$L = h\left(L_{rank}\right) + h\left(L_{regr}\right) \tag{5}$$

when adding the regression task. Here, $h(L_{task}) = \exp(\frac{L_{task}}{T})$ evens the individual loss values. We set $T = 50$ according to the author's recommendation [24].

After pretraining the model with the proposed self-supervised tasks, the network should have learned important visual features to identify and judge image aesthetics. We can fine-tune the pretrained model on a labeled image aesthetics dataset, which should achieve better performance on the labeled dataset.

4. Experiments

4.1. Aesthetic-Aware Image Distortions

During our self-supervised pretraining we apply *technical*, *style*, and *composition* filters to the high-quality images. Table 1 shows our selected distortions with corresponding intensity values.

To change the technical quality of an image, we use the library "imagenet-c" [12], introduced to benchmark the robustness of image classification models against distortions on the ImageNet [4] dataset. For changes in style we use the graphics suite darktable [37] that provides common color correction and color grading filters. All compositional distortions such as cropping, rotation, or adjusting image ratios are implemented in Python using Pillow [3]. We resize and pad all images to 224×224 pixels. This ensures that we do not accidentally destroy the image's composition.

4.2. Highly Aesthetic Dataset

For our dataset, we download the 130 000 most popular images of all time[2] from the free stock photo website pexels.com. As detailed in Section 3.2, the dataset needs to contain *highly aesthetic* images of *diverse style and content* to be usable in our self-supervised pretext tasks.

[2] pexels.com/popular-photos/all-time/ now offline

Figure 2. Random examples from our unlabeled dataset from pexels.com used for our self-supervised pretraining.

Highly Aesthetic Stock photos are inherently made to be aesthetically pleasing, which makes a stock photo website an ideal source for our dataset. It can safely be assumed that stock photos in general are quite well lit and color coordinated. By taking only the most popular photos, we are certain that these images are liked by the general public. Figure 2 shows randomly selected images from the dataset. To spot-check the aesthetics, we conduct a survey: A random image from the dataset and the same image with an image filter applied are shown. The subject clicks on the image that they find to be more aesthetically pleasing. 73 annotators have rated 6182 image pairs. For 74 % of these pairs, the unedited image is preferred, when compared to *any* style filter in *any* intensity. Besides the survey, we apply a trained version of NIMA[3] [36] to the original images, which classifies approximately 88 % of them as high-quality.

Diverse in Style and Content According to pexels.com, the stock photos are from a wide variety of art styles and topics. Available categories range from dominant colors over common photo tags like *food*, *fashion*, or *people* to emotional aspects like *moody*, *wellbeing*, or *happy*. This induces a high diversity in image contents and styles. Assuming that the most popular images are a somewhat representative sample of all images, our dataset covers a wide range of different styles while each individual image remains well edited. Additionally, we apply a pretrained DenseNet121 [16] for image classification and RetinaNet [25] for object detection to our dataset to make sure that the images are diverse in content. We find that the images of our dataset spread across many different classes and contain a wide variety of objects and subjects. Detailed results can be found in Appendix B.

[3] https://github.com/kentsyx/Neural-IMage-Assessment

4.3. Self-Supervised Aesthetic-Aware Pretraining

As the main pretext task, we train the network to rank an image with differently intense filter settings for one filter. In addition, we add one of two tasks, as described in Section 3.4, resulting in three pretext task combinations: *ranking*, *ranking+classification*, and *ranking+regression*. We set the hyperparameter m, which describes the desired minimal margin between two images with different distortion intensities, to 0.2 in our experiments.

Prior to training, we randomly split our dataset into $100\,000$ training images, $15\,000$ validation, and $15\,000$ test images. Every image is then distorted using each of the filters described above with all respective intensity values, resulting in 173 different variants per image, including the original image. Overall, this makes $100\,000{\times}173{=}17\,300\,000$ training images and $15\,000{\times}172{=}2\,595\,000$ images for validation and testing.

As the neural network architecture, we use MobileNetV2 [33], since it speeds up training times due to fewer parameters while still performing well [36]. This allows us to train comparatively quickly even on the approximately 17 million training images.

For each pretext single-task or multi-task setting, we train the model for 20 epochs and then select the epoch with the lowest validation loss as the initialization for fine-tuning. We select an initial learning rate of 10^{-4} for all model configurations based on a hyperparameter search. All other training settings are taken directly from the original MobileNetV2 paper [33], *i.e*. optimizing using RMSProp [13] with decay and momentum set to 0.9, weight decay regularization of 4×10^{-5}, and an exponential learning rate decay of 0.98 per epoch. As batch size, we take eight images and all their distorted variants as a batch, thus resulting in 1384 images in each batch.

4.4. Baselines

We compare our method to three other models. On the one hand, we train a model that is initialized with weights trained on the ImageNet classification task [4], a common initialization for IAA methods [35, 36, 28, 20]. On the other hand, we employ two common self-supervised pretext tasks. **RotNet** [22] classifies the rotation $\in \{0°, 90°, 180°, 270°\}$ applied to an input image. **SimCLR** [2] aims to output similar representations for different augmentations of the same image. For a fair comparison, both tasks are applied to the ImageNet-initialized MobileNet architecture [33] using our collected dataset. See Appendix C for details.

4.5. Fine-tuning Pretrained Models

After pretraining, each model is then fine-tuned on the AVA dataset [29]. For our experiments, we follow the training method from NIMA [36], since it is simple and elegant: We replace the last layer of the network with a fully connected layer with ten outputs, indicating the possible AVA scores one to ten. Outputs are normalized to a distribution using the softmax function. The training procedure then uses the Earth Mover's Distance Loss [15] on the rating distribution of the annotator's votes for each image.

Training parameters are kept close to the original values proposed by Talebi *et al*. [36], only incorporating small practical improvements by Lennan *et al*. [23] and changes to adapt for the different pretext tasks: We train using the Adam [30] optimizer with separate learning rates for the pretrained layers (10^{-4}) and for the new fully connected layer (10^{-3}) and a weight decay regularization of 4×10^{-5}. Additionally, we use a learning rate schedule that monitors the validation loss and halves the learning rate if the loss does not significantly improve for five consecutive epochs. This schedule is a compromise between the originally proposed schedule in NIMA [36] and the more aggressive decline in learning rate used by Lennan *et al*. [23]. As suggested by Lennan *et al*. we keep the weights of all layers but the very last frozen until the validation loss plateaus for the first time. We train on the AVA dataset and employ an early stopping strategy that stops training once the validation loss does no longer improve for ten consecutive epochs.

4.6. Evaluation Setup

For evaluating the trained models, we mostly follow the setup by Talebi *et al*. [36]. Given the ten-dimensional output denoting a distribution across the possible rating scores, a mean score can be computed by calculating the expected score given the output distribution. We let all models predict the mean aesthetic scores for the test split of the AVA dataset and label images with scores above five as aesthetic and below five as not aesthetic. We can then calculate the accuracy [32] based on this binarization and the ground truth mean opinion scores by AVA.

In most applications for IAA, however, it is necessary to rank different images based on their aesthetic score. Thus, metrics measuring the performance of the predicted numerical score are of higher interest. We calculate the Spearman (SRCC) [32] and Pearson (LCC) [6] correlation coefficients as well as the Mean Absolute Error (MAE) [32] between the predicted and ground truth mean scores.

In addition to the image aesthetic assessment dataset AVA, we — like Talebi *et al*. [36] — also test our models on the TID2013 [31] dataset that is designed to measure the performance of technical image quality assessment models. Since our pretext tasks incorporate distortions leading to technically degraded images, we suspect that the performance of pretrained models improves on this dataset as well. We calculate the SRCC, LCC, and MAE between the predicted and ground truth scores from the dataset.

		AVA				TID2013		
Pretext Task	**Epochs**	LCC ↑	SRCC ↑	ACC ↑	MAE ↓	LCC ↑	SRCC ↑	MAE ↓
ImageNet [4]	147	0.5491	0.5372	0.7548	0.4716	0.4760	0.3897	0.9747
RotNet [22]	79	0.3272	0.3139	0.7112	0.5376	0.1282	0.1075	1.1334
SimCLR [2]	**65**	0.5483	0.5360	0.7540	0.4718	0.4824	0.3944	0.9688
ranking	96	0.5536	0.5414	0.7560	0.4698†	0.4842	0.3946	0.9679†
ranking + classification	89	0.5535	0.5409	0.7550	0.4702	**0.5009**	**0.4111**	**0.9620**†
ranking + regression	77	**0.5541**	**0.5420**	**0.7582**	**0.4697**†	0.4855	0.3971	0.9672†

Table 2. Performance results: Shown are the Accuracy (ACC) on the binary task, Pearson (LCC) and Spearman (SRCC) correlation, as well as the Mean Absolute Error (MAE) between the ground truth and the scores returned by our model. The best value for each metric is printed in bold. For our pretext tasks, † indicates significantly better results compared to the ImageNet pretext task.

5. Results

In Table 2, the fine-tuned models are identified on the basis of their respective pretext task they were initialized with. In addition to the evaluation metrics specified in Section 4.6, we also provide the number of training epochs before halted by early stopping.

Aesthetic Pretraining Improves Performance Models pretrained on any of our proposed pretext tasks show better evaluation metrics than all baseline models. Compared to the ImageNet baseline, any additional pretraining reduces the number of epochs during fine-tuning by $29\,\%$ to $55\,\%$, while only our pretext tasks consistently improve the performance of the resulting model at the same time.

Adding classification to the ranking task in a multitask setting does not improve the performance of the ranking task on AVA, however, results in the best model evaluated on the TID2013 dataset. This supports the findings of previous work that used classification pretext tasks for technical image quality assessment [19, 27, 39, 8]. We suspect that this is due to the difficulty of distinguishing some categories of distortions. In the classification task, we categorize all image filters into the categories *technical*, *style*, and *composition* and train the network to identify the distortion in the corresponding category. We observe that the classification accuracy per category of the pretrained model on the testset of our collected stock photo dataset is higher for technical distortions than for style or composition changes (*c.f.* Appendix D). The pretrained model has learned to extract features that are especially useful for technical image distortions, making the model more suitable to be used as an initialization for IQA.

In contrast to classification, adding the regression task yields the best performance metrics on AVA in our experiments. Our intuition is that this is due to two effects. First, the loss only takes exactly one distortion into account, thus making all distortions independent of each other. For the classification task, we have to combine multiple distortions based on their type in order to be able to calculate the Cross Entropy Loss, making their outputs dependent on each other. Second, and more importantly, letting the network predict an intensity per distortion encourages the extraction of features that already have a notion of order. These features are then effectively used during fine-tuning and result in better performance. In fact, Bonferroni-corrected [11] one-sided Wilcoxon signed rank tests [38] on the Mean Absolute Errors (MAE) show that using our pretext tasks over ImageNet produces *significantly better* results at an α level of $1\,\%$. This means, on average, the returned score is significantly closer to the mean human annotated score than when using the ImageNet initialization.

The self-supervised baselines RotNet [22] and SimCLR [2] are explicitly designed to learn content based features in order to improve image classification performance. Thus, similar to the ImageNet pretraining, learned features are mostly invariant to aesthetically relevant changes such as lighting. Our method explicitly embraces these features, leading to useful features for the downstream task.

Aesthetic Pretraining Learns Cross-Distortion Relations In the following we analyze the outputs of the pretext regression network f_{regr}. Note that we study the pretrained model that was *not* yet fine-tuned on the AVA dataset. During pretraining, the regression loss is only computed on the corresponding output dimension denoting the currently applied distortion, as described in Section 3.3. All other outputs are not used for loss calculation, thus can independently predict intensities for their respective distortions. This allows the network to learn synergies between different distortions, *e.g.* an increase in brightness usually reduces the contrast (see Figure 1) and vice versa.

To check if the model has learned these relations, we feed all test images of our collected dataset with all brightness changes through the network. Then, we plot the model's predicted brightness and contrast changes as a violin plot in Figure 4. Given higher or lower brightness intensities, the predicted outputs are also increasing or decreasing, respectively. An increase in brightness also results in a decrease in contrast, which shows the model learned cross-

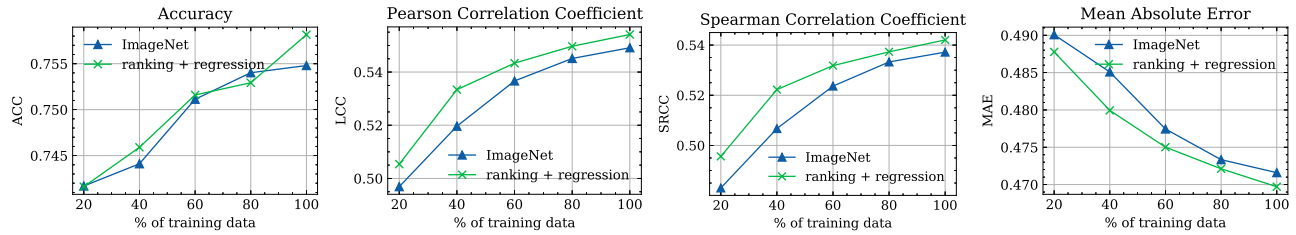

Figure 3. Accuracy, Pearson/Spearman correlation coefficient and Mean Absolute Error for different sizes of the labeled dataset.

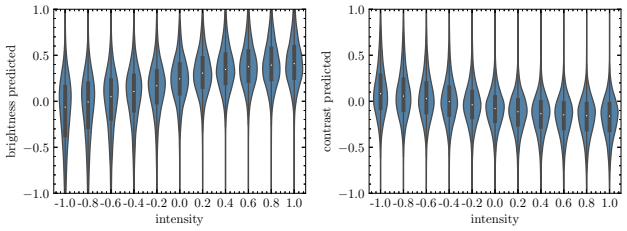

Figure 4. Outputs of the brightness (left) and contrast (right) neuron for different intensities (x-axis) of the brightness distortion.

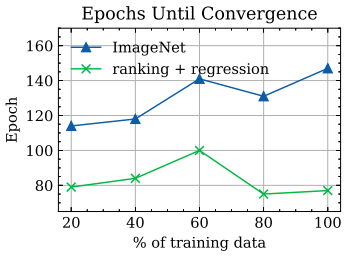

Figure 5. Epoch in which early stopping occurred by dataset size.

distortion relations, which it was never explicitly trained on.

Aesthetic-Aware Pretrained Models Need Less Labeled Data According to our results, models pretrained on our proposed tasks were able to learn useful image features for aesthetic image assessment, resulting in significantly improved prediction performance and training time. We hypothesize that this also reduces the need for labeled training images in order to achieve comparable performance to the ImageNet baseline. To verify this, we fine-tune the *ranking+regression* pretrained model and the ImageNet baseline on subsets of the AVA training data: 20 %, 40 %, 60 %, 80 %, 100 %. Figure 3 shows the Accuracy, Pearson and Spearman Correlation Coefficients, as well as the Mean Absolute Error on the AVA testset for both models. While our model seems to match the baseline's accuracy on the binary task, it consistently outperforms the baseline on both correlation coefficients and the Mean Absolute Error. Due to the higher correlations, we can assume that our model matches the relative aesthetic order of images better than the baseline. Furthermore, we find that we match the ImageNet model in terms of correlation and Mean Absolute Error while needing approximately 20 % less training data. Our pretrained model also converges faster than the ImageNet baseline across all training data sizes (see Figure 5). Overall, pretraining the model on our proposed tasks thus provides a better initialization for fine-tuning than the ImageNet classification task.

6. Discussion

In our self-supervised pretraining, our models do not learn to rank different images against each other but only to differentiate between distortions and intensities applied to one image. Our models are therefore not able to learn some potentially relevant aesthetic features and relations between different images during pretraining. To furthermore enable our models to *e.g.* rank different base images during the pretext task, we suggest exploring additional losses: A loss minimizing the distance between scores of two images of the same distortion and same intensity could be introduced. Additionally, we can extend the ranking loss and not only rank distortions of the same image against each other, but against other base images as well. Under the assumption that all images in the dataset are roughly equally aesthetic, this allows our network to learn relationships between aesthetic features of different base images.

7. Conclusion

In this paper, we have proposed self-supervised aesthetic-aware pretext tasks optimized for fine-tuning Image Aesthetic Assessment models. For this, we have introduced a large dataset of highly aesthetic images that were systematically degraded in quality using distortions in three aspects of image aesthetics: technical quality, image style, and composition. We have applied the pretext tasks in a multi-task setting and have shown that ranking as well as estimating distortion intensities improves performance over the employed baselines and converges faster than starting with ImageNet initialization. An analysis has shown that our pretext tasks are able to teach the neural network meaningful and relevant features about image aesthetics, without access to an explicit human opinion as reference. We thus encourage researchers to use our pretrained models as initialization for their CNN-based IAA methods.

References

[1] Sebastian Bosse, Dominique Maniry, Klaus-Robert Muller, Thomas Wiegand, and Wojciech Samek. Deep neural networks for no-reference and full-reference image quality assessment. *IEEE Transactions on Image Processing*, 27(1):206–219, jan 2018.

[2] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[3] Alex Clark. Pillow (pil fork) documentation.

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. Ieee, IEEE, jun 2009.

[5] Michael Fischer, Konstantin Kobs, and Andreas Hotho. NICER: Aesthetic image enhancement with humans in the loop. *arXiv preprint arXiv:2012.01778*, 2020.

[6] David Freedman, Robert Pisani, and Roger Purves. Statistics (international student edition). *Pisani, R. Purves, 4th edn. WW Norton & Company, New York*, 2007.

[7] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations.

[8] S. Alireza Golestaneh and Kris Kitani. No-reference image quality assessment via feature fusion and multi-task learning. 2020.

[9] Google. Google photos. photos.google.com.

[10] Priya Goyal, Dhruv Mahajan, Abhinav Gupta, and Ishan Misra. Scaling and benchmarking self-supervised visual representation learning. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, oct 2019.

[11] Winston Haynes. *Bonferroni Correction*, pages 154–154. Springer New York, New York, NY, 2013.

[12] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *Proceedings of the International Conference on Learning Representations*, 4, 2019.

[13] Geoffrey Hinton, Nitish Srivastava, and Kevin Swersky. Overview of mini-batch gradient descent.

[14] Vlad Hosu, Bastian Goldlucke, and Dietmar Saupe. Effective aesthetics prediction with multi-level spatially pooled features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9375–9383, 2019.

[15] Le Hou, Chen-Ping Yu, and Dimitris Samaras. Squared earth mover's distance-based loss for training deep neural networks.

[16] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4700–4708. IEEE, jul 2017.

[17] Longlong Jing and Yingli Tian. Self-supervised visual feature learning with deep neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2020.

[18] Le Kang, Peng Ye, Yi Li, and David Doermann. Convolutional neural networks for no-reference image quality assessment. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, jun 2014.

[19] Le Kang, Peng Ye, Yi Li, and David Doermann. Simultaneous estimation of image quality and distortion via multi-task convolutional neural networks. In *2015 IEEE International Conference on Image Processing (ICIP)*. IEEE, sep 2015.

[20] Keunsoo Ko, Jun-Tae Lee, and Chang-Su Kim. PAC-net: Pairwise aesthetic comparison network for image aesthetic assessment. In *2018 25th IEEE International Conference on Image Processing (ICIP)*. IEEE, oct 2018.

[21] Alexander Kolesnikov, Xiaohua Zhai, and Lucas Beyer. Revisiting self-supervised visual representation learning. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jun 2019.

[22] Nikos Komodakis and Spyros Gidaris. Unsupervised representation learning by predicting image rotations. In *International Conference on Learning Representations (ICLR)*, 2018.

[23] Christopher Lennan, Hao Nguyen, and Dat Tran. Image quality assessment. https://github.com/idealo/image-quality-assessment, 2018.

[24] Sicong Liang and Yu Zhang. A simple general approach to balance task difficulty in multi-task learning. 14:arXiv–2002, 2020.

[25] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. Focal loss for dense object detection. In *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, oct 2017.

[26] Xialei Liu, Joost Van De Weijer, and Andrew D. Bagdanov. RankIQA: Learning from rankings for no-reference image quality assessment. In *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, oct 2017.

[27] Kede Ma, Wentao Liu, Kai Zhang, Zhengfang Duanmu, Zhou Wang, and Wangmeng Zuo. End-to-end blind image quality assessment using deep neural networks. *IEEE Transactions on Image Processing*, 27(3):1202–1213, mar 2018.

[28] Shuang Ma, Jing Liu, and Chang Wen Chen. A-lamp: Adaptive layout-aware multi-patch deep convolutional neural network for photo aesthetic assessment. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4535–4544. IEEE, jul 2017.

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

[30] Jimmy Omony. Constrained stochastic space search method for parameter estimation in biological networks. *British Journal of Mathematics & Computer Science*, 4(7):952–968, jan 2014.

[31] Nikolay Ponomarenko, Oleg Ieremeiev, Vladimir Lukin, Karen Egiazarian, Lina Jin, Jaakko Astola, Benoit Vozel, Kacem Chehdi, Marco Carli, Federica Battisti, and C.-C. Jay Kuo. Color image database tid2013: Peculiarities and preliminary results. In *European Workshop on Visual Information Processing (EUVIP)*, volume 30, pages 106–111. IEEE, 2013.

[32] Claude Sammut and Geoffrey I. Webb, editors. Springer US, Boston, MA, 2010.

[33] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. MobileNetV2: Inverted residuals and linear bottlenecks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, jun 2018.

[34] Kekai Sheng, Weiming Dong, Menglei Chai, Guohui Wang, Peng Zhou, Feiyue Huang, Bao-Gang Hu, Rongrong Ji, and Chongyang Ma. Revisiting image aesthetic assessment via self-supervised feature learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(04):5709–5716, apr 2020.

[35] Kekai Sheng, Weiming Dong, Chongyang Ma, Xing Mei, Feiyue Huang, and Bao-Gang Hu. Attention-based multi-patch aggregation for image aesthetic assessment. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 879–886, 2018.

[36] Hossein Talebi and Peyman Milanfar. NIMA: Neural image assessment. *IEEE Transactions on Image Processing*, 27(8):3998–4011, aug 2018.

[37] The Darktable Development Team. Darktable. https://www.darktable.org.

[38] Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80, dec 1945.

[39] Qingbo Wu, Hongliang Li, King N. Ngan, Bing Zeng, and Moncef Gabbouj. No reference image quality metric via distortion identification and multi-channel label transfer. In *2014 IEEE International Symposium on Circuits and Systems (ISCAS)*. IEEE, jun 2014.

[40] Yi Zhang and Damon M. Chandler. Opinion-unaware blind quality assessment of multiply and singly distorted images via distortion parameter estimation. *IEEE Transactions on Image Processing*, 27(11):5433–5448, nov 2018.