Understanding Photographic Styles Using Dense Convolutional Networks

Anonymous ICCV submission

Paper ID 7

Abstract

Photographs are characterized by different compositional attributes like the Rule of Thirds, depth of field, leading lines etc. The presence or absence of one or more of these attributes contribute to the overall artistic value of an image. Historically, the problems researched in Computer Vision have been focused around understanding non-aesthetic physical properties (objects, their relative positions, semantics, extents etc.*) of generic images. However, with several solutions successfully tackling these problems, in recent years, there has been an increased interest towards subtler understanding of images (in terms of aesthetic attributes). In this work, we analyze different photographic style attributes. Based on the recent developments of deep neural networks, we implement a system for automatically predicting the styles that characterize a photograph. In our experiments, we observe that our implementation performs better than the state of the art in photographic style categorization. Additionally, we analyze each of the attributes and observe a correlation between prediction accuracy and the pre-processing transformations used in Convolutional Neural Networks (CNN). We conclude that although traditional CNNs improve the overall performance in style categorization, the global properties of photographs (Rule of Thirds, complementary colours, vanishing lines* etc.*) are less preserved than the local properties (depth of field, macro, image grain* etc.*).*

1. Introduction

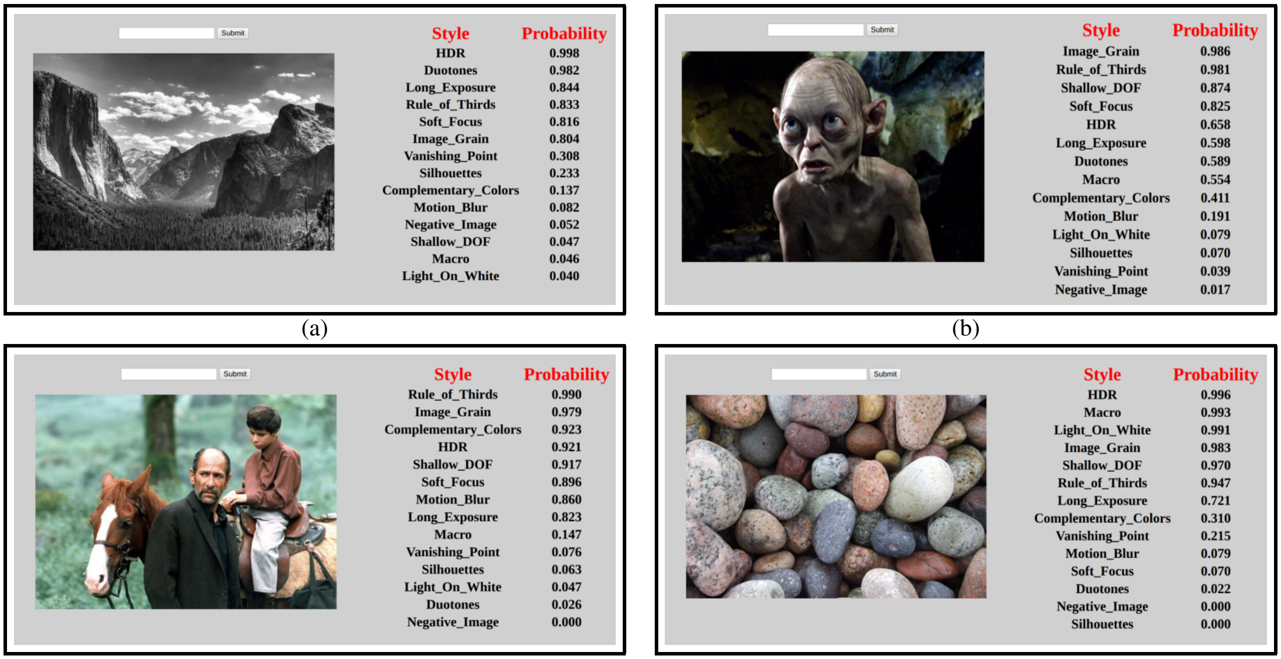
Amid millions of images and video clips uploaded over the internet, there are a few which stand out and leave their mark. For example, Ansel Adam's beautiful captures of the Yosemite National Park are admired widely because of the amazing exposure and geometry that they possess. On the other hand, one cannot help but appreciate the portrayal of the co-existence of the ordinary and the sublime, in renowned Iranian film director Majid Majidi's landscape shots. Browsing through the works of the great photographers or cinematographers, one gets a taste of their originality or style, which make them stand out consistently over the years. Thus, analysing styles is crucial for understanding visual arts.

Traditionally, the primary focus of the Computer Vision community has been around modelling the physical properties of generic images for object detection, localization, segmentation, tracking *etc*. Popular datasets like Caltech [9], Pascal [7] and ImageNet [5] were created for that purpose. The maturing of recognition and scene understanding has resulted in greater interest in the analysis of the subtler, aesthetic based aspects of image understanding. What's more, curated datasets [14, 25] are now available and it is observed that learning from the matured areas of recognition/classification/detection transfers effectively to aesthetics analysis too.

There are two conventions to define the aesthetic properties of an artefact [3]. One is the *content-oriented* approach which defines the properties in terms of objective content like lines, symmetry, colour *etc*. Another is the *affect-oriented* approach which defines the properties in terms of subjective aspects like disinterested and sympathetic experiences of the viewers. In this work, we propose a system which tries to model the content-oriented aesthetic attributes of a photograph and predict its style. Motivated by the recent developments in the field of Deep Convolutional Neural Networks (CNN), our system takes a photograph as an input and predict its styles (ordered, based on probabilities), as listed in Figure 1.

We adopt the DenseNet [11] architecture, trained on the ImageNet [5] dataset and fine-tune the model on the 14 style classes in the AVA dataset[25]. AVA is an image database containing 250,000 images with a quality rating on a scale of 10. A subset of approximately 14000 images contain style annotations corresponding to 14 different photographic attributes as illustrated in Figure 2. We describe AVA in more detail in Section 4. In our experiments, we observe that the DenseNet161 architecture perform better than the state of the art in style-classification [21, 20].

Our second contribution in this work is an analysis of the effect of different data-augmentation strategies on style prediction. Traditionally, CNNs take an input of a fixed size



Style	Probability
HDR	0.998
Duotones	0.982
Long_Exposure	0.844
Rule_of_Thirds	0.833
Soft_Focus	0.816
Image_Grain	0.804
Vanishing_Point	0.308
Silhouettes	0.233
Complementary_Colors	0.137
Motion_Blur	0.082
Negative_Image	0.052
Shallow_DOF	0.047
Macro	0.046
Light_On_White	0.040

(a)

Style	Probability
Image_Grain	0.986
Rule_of_Thirds	0.981
Shallow_DOF	0.874
Soft_Focus	0.825
HDR	0.658
Long_Exposure	0.598
Duotones	0.589
Macro	0.554
Complementary_Colors	0.411
Motion_Blur	0.191
Light_On_White	0.079
Silhouettes	0.070
Vanishing_Point	0.039
Negative_Image	0.017

(b)

Style	Probability
Rule_of_Thirds	0.990
Image_Grain	0.979
Complementary_Colors	0.923
HDR	0.921
Shallow_DOF	0.917
Soft_Focus	0.896
Motion_Blur	0.860
Long_Exposure	0.823
Macro	0.147
Vanishing_Point	0.076
Silhouettes	0.063
Light_On_White	0.047
Duotones	0.026
Negative_Image	0.000

(c)

Style	Probability
HDR	0.996
Macro	0.993
Light_On_White	0.991
Image_Grain	0.983
Shallow_DOF	0.970
Rule_of_Thirds	0.947
Long_Exposure	0.721
Complementary_Colors	0.310
Vanishing_Point	0.215
Motion_Blur	0.079
Soft_Focus	0.070
Duotones	0.022
Negative_Image	0.000
Silhouettes	0.000

(d)

*Figure 1: **Applications** : Screen-shots from our web-based application. For each picture, the left column shows the photograph used and the right column shows the probability values for each attribute. They are ordered in descending order i.e the ones at the top ($> .900$) are the most probable styles. **(a)** This is one of the pictures of Ansel Adam's captures of the Yosemite National Park. It can be observed that the top two attributes are HDR and duotones, which are correct predictions. **(b)** This is a shot from The Lord of the Rings movie. The rule of thirds (the line of the eyes match with the first horizontal line) and shallow depth of field are correct predictions. The image grain prediction is because of a high ISO value of the image. **(c)** This is a shot from Majid Majidi's film The Colours of Paradise. We see that rule of thirds (child's position), shallow depth of field, complementary colours (green background and reddish foreground), image grain (because of the video quality) are all correct predictions. **(d)** An image from Flickr showing repetitive patterns and soft colours. This is one of the scenarios where the system fails. The styles corresponding to this photograph are none among HDR, Macro or Light on White. One of the drawbacks of our system is that the attribute space is limited to 14, which is a small number when it comes to judging photographs.*

(usually a $224 * 224$ patch). The source image is either cropped or re-scaled (warped) to the desired size and passed through the network, in order to pass data in a fixed parametrization while still allowing the processing of a variety of scene content. Due to cropping, information is lost but it works well in scenarios where the object of interest is localized at a particular position (say at the centre). Moreover, as pointed out by [21], it can improve performance by reducing overfitting. On the other hand, warping captures global information better but it loses much of the geometric properties due to distortion and comes with the curse of overfitting.

Intuitively, it appears that the compositional style of a photograph (as in Figure 2) should depend on both local and global factors. For example, to find whether a photograph conforms to the Rule of Thirds principle, the relative positions of the main subjects are required. As an example, in Figure 2, at $(2, 3)$, the position of the bird in the image decides whether the image conforms to the Rule of Thirds or not. This information is completely lost when a fixed-size crop is extracted (though the appearance based properties like Image Grain and HDR are preserved). However, contrary to this intuition, it is shown in [20] that cropping performs much better than warping in style classification. We experiment with several data-augmentation strategies and try to understand the rationale behind such counter-intuitive results. In Section 5.1, we empirically conclude that although cropping improves the overall accuracy, it performs worse in cases of particular classes like Rule of Thirds or Vanishing Lines. We also observe that a hybrid data augmentation strategy, which is a combination of random sized cropping and warping, improves the overall accuracy better than either warping or cropping.

There are several applications of automatic photographic style classification. For example, post-processing images and videos, tagging, organizing and mining large collections of photos for artistic, cultural and historical purposes, scene understanding, building assistive technologies, content creation, cinematography *etc*.

The rest of the paper is organized as follows. In Section 2, we summarize the relevant literature in image quality prediction. In Section 3, we describe the CNN model we adopted and different augmentation strategies we used. In Section 4, we provide a detailed description of the AVA dataset. In Section 5, we provide details of the experiments conducted and analyze the results. In Section 6, we mention some of the possible applications of our system.

2. Related Work

Although image and video classification has always been a fundamental problem in Computer Vision, a considerable part of existing research is focused on identifying and classifying objects or the scene. In other words, the intriguing part for the community has been understanding the quantifiable visual semantics like the class [19, 30], position [34] and number of objects in an image. However, understanding the qualitative aspects from a creative perspective, due to its challenging nature, has only recently started being attacked.

The early works in photograph quality prediction relied on modelling popular attributes like the Rule of the Thirds, colour harmony, exposure etc [4, 16, 22]. Some recent works address the problem in a similar way but with improved features[27, 6, 13, 29, 14]. In [2], Aydin *et al*. point out the *generalist* and *particularist* approach in evaluating a photograph. They propose a system which predicts the contribution of some photographic attributes towards the overall quality of a picture. Using a novel calibration technique they aggregate the scores for different attributes to predict the overall aesthetic score of a photograph. In [25], Marchesotti *et al*. published the AVA dataset which is a collection of about 250,000 images, ratings and style annotations from www.dpchallenge.com. It was the release of AVA and parallel advances in Deep Learning that resulted in a surge of research in this area in the last few years.

In recent years, Deep Learning has performed remarkably well in many Computer Vision tasks like classification [18, 31], detection [8, 28], segmentation [26] and scene understanding [15, 33, 1]. Recent works like [11, 10] have performed well in multi-tasking frameworks for detection and classification. In [10], the authors use a residual framework for tackling training error upon addition of new layers. In [11] the authors use dense connections by connecting outputs from all the previous layers as input to the next layer.

As in many computer vision problems, deep learning has begun to be explored in this domain too. In [20], Lu *et al*. propose two Convolutional Neural Network architectures which capture the local and global properties of a photograph. The same authors use a multiple instance learning-based strategy in [21] and achieve a much better performance for style-classification. In [17], Kong *et al*. learn styles and ratings jointly on a new dataset. Mai *et al*. [23] propose a network that uses a composition-preserving input mechanism. In [24], the authors propose a network that predicts the overall aesthetic score and eight style attributes, jointly.

In principle, our pipeline is similar to [21, 20, 14] in the sense that we also perform a neural style prediction on the AVA dataset. However, our work differs in two important aspects. First, we use an architecture that performs better than generic features [14], the double column [20] or multi-patch aggregation [21] strategies, using a simpler and straightforward training procedure. Second, we make similar observations in per-class precision scores as [14]. But, delving deeper, we attempt to understand the effects of different data-augmentation strategies on the photographic attributes. In [20, 21] the authors suggest that a random crop from the input handles overfitting better than warping and hence perform better. We argue that although a random crop preserves the appearance-based attributes, it fails to preserve the geometric attributes of the photograph like the Rule of Thirds. By examining the effects of different data-augmentation strategies on the different geometric and appearance-based styles, we argue that a hybrid strategy acts as a trade-off between cropping and warping and performs better.

3. Problem Statement

In a CNN, a function ϕ is learnt which maps am image X of $(M \times N)$ size to an attribute space Y of C dimensions. ϕ is a composite function of the form in Equation 1.

$$\phi = F(T(x)) \tag{1}$$

Here, T is a transformation applied on the data as pre-processing. Apart from a data-augmentation strategy as mentioned in 3.3, the input is mean subtracted. Then the processed input is passed through the network. Mathematically,

$$T : X^{M\times N} \rightarrow T(x)^{P\times P} \tag{2}$$

$$F : T(x)^{P\times P} \rightarrow F(T(x))^{C\times 1} \tag{3}$$

where $P \times P$ is the size of the pre-processed image. Essentially, during training a Cross Entropy loss is computed from the last layer of the CNN as shown in Equation 4 and the weights are updated using a Stochastic Gradient Descent

(SGD) procedure . Mathematically,

$$L(\mathbf{w}) = -\frac{1}{N}\sum_{n=1}^{N}\sum_{c=1}^{C}\left[y_n^c \log \hat{y}_n^c + (1-y_n^c)\log(1-\hat{y}_n^c)\right] \tag{4}$$

where, y_n^C and $\hat{y}_n^C$ are the ground truth labels for an image and the logistic regression on the output of the last layer of the network, respectively. A logistic or a sigmoid function is of the form $g(z) = 1/(1+e^{-z})$ which maps the output from the last layer of the CNN to $[0,1]$. N and C are the number of samples used for training and the number of classes, respectively.

3.1. DenseNet

We choose the DenseNet161 [11] network for its superior performance in the ImageNet challenge. Very deep networks suffer from the *vanishing-input* problem *i.e* gradual loss of information as the input passes through several intermediate layers. Recent works like [12, 10, 32] address this problem by explicitly passing information between layers or by dropping random layers while training.
The DenseNet is different from the traditional CNNs in the manner in which each layer receives input from the previous layers. The l^{th} layer in DenseNet receives as an input the concatenated output from all previous $l-1$ layers. We provide more details about the architecture used in Table 1

3.2. Transfer Learning

For an architecture as deep as DenseNet, there is a need for a large amount of training data. However, we have only $11,270$ images for training 14 style attributes of AVA. We address this problem by using a form of transfer learning[35]. We use a DenseNet161 classification model pre-trained on ImageNet, by changing the final fully-connected layer to output 14 instead of 1000 classes and train it on the AVA dataset. This lets us train a very deep network with considerably less amount of data.

3.3. Data Augmentation

As previously mentioned in Section 1, we try five different data-augmentation techniques, to find the correlation of the input transformations and the different style attributes.

- **Centre Crop** (T_{cc}) : Here, a patch of size $(s * s * 3)$ is extracted from the center of the image. This popular strategy is common in scenarios where the object of interest is centrally-localized.
- **Random crop of fixed size** (T_{rc}) : Patches of size $(s * s * 3)$ are extracted from the image at random locations. Although it suffers from data loss, it is effective in scenarios where the object of interest is not centrally localized.

*Table 1: **DenseNet161 Architecture :** There are 161 layers in total. The input is received by the first layer (a 7×7 convolution with a stride size of 2 and pad size of 3), followed by a max pool (3×3, stride size of 2). Then the network has a dense block which is n repetitions of the cell $(2, 2)$. The value of n is 6 for the first dense block. The first dense block is followed by the first transition layer, which is specified in cell $(3, 2)$. The dense and transition blocks are repeated 4 times (with dense block units repeated with different values of n) and finally the architecture ends with a Classification layer.*
The novelty in DenseNet is in introducing the Dense Blocks. Each unit within the dense block is densely connected i.e every l^{th} unit receives as an input which is the concatenated output from all other $l-1$ units within the same block. This setting helps to prevent input information getting lost as the network grows deeper. Additionally, as mentioned in [11], the network has the advantage of fewer parameters and better regularization.

Layer	Unit Specification	# Units
First Layer	7×7 conv, stride 2 3×3 max pool, stride 2	1
Dense Block 1	Batch Normalization RELU 1×1 conv, stride 1 Batch Normalization RELU 3×3 conv, stride 1	6
Transition Layer 1	Batch Normalization RELU 1×1 conv, stride 1 Avergae Pooling	1
Dense Block 2	—	12
Transition Layer 2	—	1
Dense Block 3	—	36
Transition Layer 3	—	1
Dense Block 4	—	24
Transition Layer 4	—	1
Classification Layer	Batch Normalization Fully Connected	1

- **Warping** (T_w) : The input is an-isotropically resized to a fixed dimension of $(s * s * 3)$. This transformation fails to preserve the geometric properties of the input. However unlike a crop, it preserves global information.
- **Isotropic centre crop** (T_{icc}) : In this case, the input is isotropically resized by warping its lower dimension to s. Then, a centre crop of $(s * s * 3)$ is applied. This preserves the geometric properties of the image at the cost of some global information.
- **Random crops of random sizes** (T_{rnc}) : A crop of random size and random aspect ratio is made from the input. Then, it is resized to $(s * s * 3)$. We demonstrate in Section 5, that this strategy addresses the limitations of the previous strategies and performs best.

We train and test our network with all these strategies. In [20], the authors use T_{rc} for a single-column network and both T_{rc} and T_w for a double column network. In [21], the authors use T_{rc}.

It is to be noted here that we did not try appearance-based augmentation techniques like changing contrasts or colours since that would alter the style of the photograph.

4. Dataset

AVA [25] is a dataset containing $250,000$ photographs, selected from www.dpchallenge.com. Dpchallenge is forum for photographers to post their works, based on challenges hosted on the website. Users rate each photograph during the challenge on a scale of 10 and post feedback during and after the challenge. The pictures, with the corresponding scores and comments, provide a rich source of data for style and quality assessment.

Of these, $250,000$ photographs, the authors manually select 72 challenges, corresponding to 14 different photographic styles as illustrated in Table 2. An example from each category is provided in Figure 2. Please note, while style-training images in the dataset are annotated with a single label, the test images have multiple labels associated with them. There are $11,270$, and $2,573$ images available for training and testing the style attributes, respectively.

5. Experiments

5.1. Style Classification

We train style classifiers on 14 AVA style attributes with different data-augmentation strategies described in Section 3.3. We choose one of the transformations and train a model for 25 epochs. We use 11270 images for training and validation and the 2573 images for testing. The experiments are carried out in a NVidia 1080 GTX GPU with 2560 cuda cores and 8 Gigabytes of memory. Training a model takes about 120 minutes with a batch-size of 5. During test phase, we follow the approach suggested by [20, 21]. 50 patches are extracted from the test-image and each patch is passed through the network. The results are averaged to achieve the final scores. The scores are reported in terms of Average Precision (AP) and Mean Average Precision (MAP).

*Table 2: **Style Attributes for AVA Dataset***

Index	Style	# Samples
1.	Complementary Colours	760
2.	Duotones	1041
3.	HDR	317
4.	Image Grain	672
5.	Light on White	960
6.	Long Exposure	676
7.	Macro	1359
8.	Motion Blur	488
9.	Negative Image	768
10.	Rule of Thirds	1112
11.	Shallow Depth of Field	1184
12.	Silhouettes	825
13.	Soft Focus	568
14.	Vanishing Point	540

5.1.1 Best Data-Augmentation

Using the DenseNet [11] architecture, we perform experiments with 25 possible combinations of the 5 different data-augmentation strategies. In Table 3, we report AP and MAP for 25 data-augmentation strategies used with the DenseNet161 architecture. From the results in Table 3, we observe that T_{rnc} performs best when used both during train and test phases. In Table 4, the first four rows list DenseNet161 [11], Resnet152 [10], VGG-19 and VGG-16 [31], trained and tested with T_{rnc}. The last three rows are the current state of the art results reported from the corresponding papers. From Table 4, it is observed that all networks trained with T_{rnc} outperforms the results reported in [20, 21, 14] by a large margin.

One might argue that this improvement can be attributed to the use of a better architecture and we accept this argument. However, we observe from Table 3 that for the same architecture, T_{rnc} performs much better than all other strategies. Therefore, it can be safely concluded that for improving the overall accuracy of style classification, T_{rnc} is the best strategy. We argue that this improvement was due to the fact that T_{rnc} provides both global and local information during training, when the network is trained long enough. Our argument is similar to [20], where they use two parallel columns in the network, one for processing local and

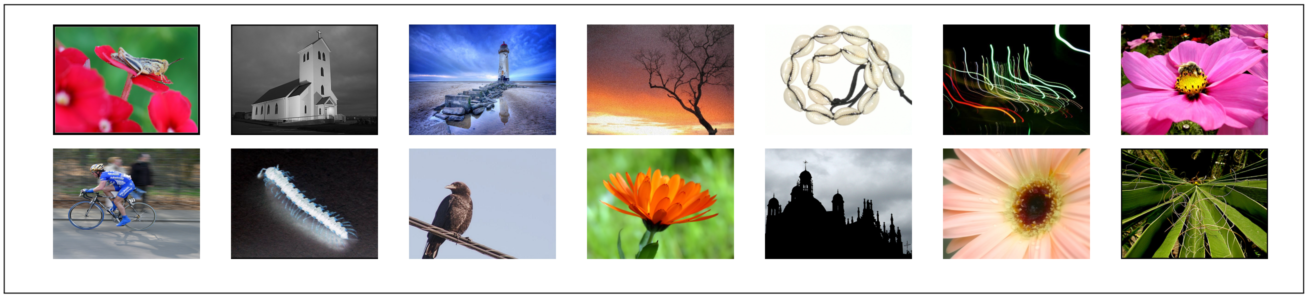


Figure 2: Example images from the AVA dataset corresponding to 14 different styles. (L-R) ***Row 1 :*** *Complementary Colors, Duotones, HDR, Image Grain, Light On White, Long Exposure, Macro.* ***Row 2 :*** *Motion Blur, Negative Image, Rule of Thirds, Shallow DOF, Silhouettes, Soft Focus, Vanishing Point*

another for processing global information. However, with a single column with a deeper architecture and a different data-augmentation strategy we outperform them.

But as pointed out in Section 1, we are also interested in finding out how these strategies perform for different attributes, individually. In the next section, we analyse each style attribute.

5.1.2 Class-wise Precision Scores

In Table 5, we report per-class precision scores for all the 25 combinations. From the table we observe that the precision scores for HDR, Motion Blur, Rule of Thirds and Soft Focus are lower in comparison to other attributes. However, the poor performance of HDR, Motion Blur and Soft Focus could be a result of less samples available for training (see Table 2). However, even with a larger number of samples available, the Rule of Thirds perform poorly. We argue that this is due to the the loss of global information during data augmentation, since this was crucial for this attribute. In Figure 3, we plot the precision for Rule of Thirds for all augmentations. We observe that the per class precision is highest in $T_w - T_w$ augmentation strategy *i.e* warping, which supports the intuitive claim that warping preserves global information better than cropping.

5.1.3 Overfitting

We observe that although warping performs better in classifying Rule of Thirds, its over all performance across 14 classes is poor. We also observe that the network overfits on the training data while training with warping. Thus we argue that during training, warping results in poor generalization performance and hence the drop in test precision. We plot the precision scores for all the classes augmented with T_w, T_{rc} and T_{rnc} in Figure 4. It can be observed from the comparison that T_{rnc} outperforms the other two strategies.

Table 3: ***Style Classification : Comparison of DenseNet161 for different augmentation strategies***

#	Aug-Training	Aug-Testing	AP	MAP
1	T_{cc}	T_{cc}	69.65	63.85
2	T_{cc}	T_w	62.78	56.54
3	T_{cc}	T_{rc}	72.83	65.15
4	T_{cc}	T_{icc}	64.68	58.08
5	T_{cc}	T_{rnc}	70.15	64.37
11	T_w	T_{cc}	60.08	52.81
13	T_w	T_w	69.68	63.75
12	T_w	T_{rc}	66.81	59.62
14	T_w	T_{icc}	69.27	63.35
15	T_w	T_{rnc}	71.35	65.96
6	T_{rc}	T_{cc}	70.59	62.24
8	T_{rc}	T_w	67.93	61.87
7	T_{rc}	T_{rc}	**75.88**	**67.90**
9	T_{rc}	T_{icc}	69.53	63.16
10	T_{rc}	T_{rnc}	73.74	67.20
16	T_{icc}	T_{cc}	62.07	53.76
17	T_{icc}	T_w	67.65	61.03
18	T_{icc}	T_{rc}	70.01	64.52
19	T_{icc}	T_{icc}	70.42	64.31
20	T_{icc}	T_{rnc}	72.42	66.35
21	T_{rnc}	T_{cc}	69.01	60.68
23	T_{rnc}	T_w	71.47	65.42
22	T_{rnc}	T_{rc}	74.17	66.40
24	T_{rnc}	T_{icc}	72.52	66.13
25	T_{rnc}	T_{rnc}	**75.26**	**68.60**

Table 5: ***Per Class Precision Score for 25 augmentation strategies using DenseNet161 :*** *The first 2 rows correspond to the style attributes and the number of samples available for training the CNN. The following 25 rows are divided into 5 blocks corresponding to the five augmentation strategies in the same order as in Table 3. We observe that the precision scores for HDR, Motion Blur, Rule of Thirds, Soft Focus are lower in comparison to other attributes. We argue that the poor performance of HDR, Motion Blur and Soft Focus could be a result of less number of samples available for training.*

.

	ClassLabels													
Augmentation Strategy	Comp Colors	Duotones	HDR	Image Grain	Light On White	Long Exposure	Macro	Motion Blur	Negative Image	Rule of Thirds	Shallow DOF	Silhouettes	Soft Focus	Vanishing Point
	Number of samples available per class for training													
	760	1041	317	672	960	676	1359	488	768	1112	1184	825	568	540
	Precision Scores													
T_{cc} - T_{cc}	55.58	65.66	51.87	75.25	71.34	55.34	56.45	46.01	72.85	29.85	66.95	85.74	39.91	56.03
T_{cc} - T_{w}	55.46	69.67	56.65	37.10	73.37	51.54	51.05	44.96	74.72	24.74	67.23	87.54	31.05	66.50
T_{cc} - T_{rc}	59.12	75.32	62.41	82.16	78.72	58.28	61.66	58.32	81.24	23.32	70.29	91.68	45.57	63.96
T_{cc} - T_{icc}	55.34	69.57	56.61	44.95	73.29	51.69	52.36	48.02	75.71	26.24	68.82	89.17	35.76	65.70
T_{cc} - T_{rnc}	59.45	74.67	67.83	67.10	76.43	59.46	59.23	54.29	81.53	26.64	71.44	91.65	43.76	67.78
T_{w} - T_{cc}	46.90	65.19	54.93	44.36	68.64	46.90	46.65	44.25	73.77	24.75	63.09	80.13	26.68	53.20
T_{w} - T_{w}	56.10	72.86	69.93	65.88	75.70	57.84	56.48	55.56	83.18	32.60	69.43	88.61	36.27	72.14
T_{w} - T_{rc}	53.91	74.82	60.37	50.97	80.65	54.31	53.07	52.53	81.94	23.70	69.72	88.99	30.00	59.83
T_{w} - T_{icc}	57.18	73.52	68.54	66.99	74.64	56.43	54.49	53.81	82.44	30.21	68.51	89.18	40.02	70.96
T_{w} - T_{rnc}	58.95	77.64	72.51	76.20	79.87	58.32	58.15	56.72	85.62	27.76	71.60	90.83	39.18	70.19
T_{rc} - T_{cc}	55.22	74.94	57.43	80.99	72.45	55.17	58.81	52.94	76.66	25.15	75.03	87.51	41.50	57.56
T_{rc} - T_{w}	54.73	80.46	69.21	47.32	79.23	52.03	55.81	48.48	83.80	27.28	77.48	89.47	34.96	65.93
T_{rc} - T_{rc}	59.71	81.12	69.81	84.91	80.40	58.01	63.75	61.24	83.80	27.92	78.87	91.31	46.42	63.39
T_{rc} - T_{icc}	56.13	79.33	72.33	53.96	78.69	51.49	57.41	51.08	82.46	26.38	78.14	90.23	39.42	67.23
T_{rc} - T_{rnc}	59.34	81.78	74.16	71.57	80.55	56.94	61.64	57.04	83.82	27.46	79.72	92.05	45.00	69.77
T_{icc} - T_{cc}	53.15	67.70	46.46	46.82	69.59	51.44	52.66	45.26	67.79	25.23	63.21	80.99	32.04	50.34
T_{icc} - T_{w}	59.26	76.53	64.06	68.22	80.95	56.35	62.04	54.12	77.21	30.51	75.80	89.48	39.98	68.89
T_{icc} - T_{rc}	60.06	77.39	53.97	50.83	81.32	55.78	58.73	56.06	79.52	23.30	68.61	89.09	37.89	61.94
T_{icc} - T_{icc}	60.05	74.83	63.55	70.82	78.60	57.16	62.16	52.92	76.70	31.22	75.60	89.92	40.29	66.63
T_{icc} - T_{rnc}	61.93	79.75	62.91	79.60	80.52	60.06	63.50	56.93	79.96	27.78	75.32	91.38	42.02	67.28
T_{rnc} - T_{cc}	55.72	69.79	53.06	70.74	73.22	54.82	59.98	50.52	75.88	27.06	73.79	86.87	41.67	56.39
T_{rnc} - T_{w}	59.96	75.70	66.99	67.01	82.55	57.83	60.88	49.95	83.50	30.74	79.80	91.16	41.89	68.04
T_{rnc} - T_{rc}	60.89	76.69	61.84	77.13	82.03	58.24	64.14	56.91	83.32	28.00	79.08	91.78	44.83	64.76
T_{rnc} - T_{icc}	61.29	75.13	68.79	72.54	81.17	56.42	61.82	52.09	82.59	29.67	80.36	92.17	42.30	69.61
T_{rnc} - T_{rnc}	62.36	79.62	69.50	81.96	81.74	59.32	65.16	55.71	84.37	30.50	81.20	93.42	46.36	69.29

6. Applications

There are many potential applications of an automatic style and quality estimator in the domain of digital photography. Figure 1(a) shows our system's correct responses on a famous picture of the Yosemite National Park by Ansel Adams.

Interactive and more intelligent cameras, automated photo correction, intelligent photo-organizers are some of the many prospective applications. Such systems could also be useful for developing assistive-technologies like audio-guide .

Our system can be directly extended to video-processing for predicting shot-styles. For example, Figures 1(b)&(c) illustrate two shots taken from Peter Jackson's The Lord of the Rings and Majid Majidi's Colors of Paradise.

However, our attribute space is limited to only 14 styles, which is one major drawback. The set of attributes that make a good photograph is non-exhaustive. For example, in Figure 1(d) the photograph is pleasing because of nice repetitive patterns and soft tones. But our system is unable to describe this due to a limited number of attributes. Moreover, as pointed out in Section 5.1.2, the system needs to be improved for certain geometric styles like the rule of thirds and vanishing lines.

Table 4: ***Style Classification : Different Networks with the best augmentation :*** *The first four rows correspond to our implementations of the top four networks in ImageNet classification. They use a T_{rnc} augmentation, both during training and testing. The last three rows are the state of the art style classifiers, results reported from paper.*

Network	Augmentation	AP	MAP
DenseNet161 [11]	T_{rnc}	**75.26**	**68.60**
ResNet152 [10]	T_{rnc}	74.08	67.25
VGG-19 [31]	T_{rnc}	75.34	67.81
VGG-16 [31]	T_{rnc}	75.29	67.39
RAPID [20]	T_{rc}, T_w	56.93	56.81
Multi-Patch [21]	T_{rc}	69.78	64.07
Fusion [14]	T_{cc}	-	58.10

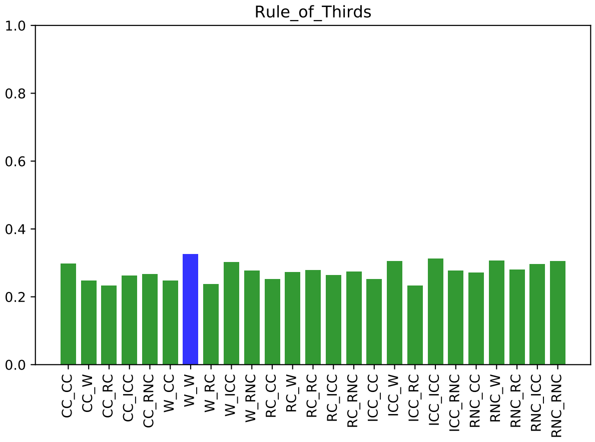


Figure 3: ***Precision scores for Rule of Thirds for 25 augmentation combinations using DenseNet161 :*** *Shown are 25 bins along x-axis, each one corresponding to an augmentation strategy from Table 3. The y axis corresponds to the precision scores. The blue coloured bin is the one with maximum score. In this figure, we observe that the $T_w - T_w$ combination works best for Rule of Thirds, which supports the intuition that global properties are better preserved in warping.*

7. Conclusion

In this work, we utilize the power of very deep neural networks and advance the state of the art in photographic style-classification. We explore different data-augmentation strategies and analyze each of their performance on the 14 style categories of the AVA dataset. We observe, that by using a deeper architecture and a hybrid data augmentation strategy, which incorporates both local and global information of the photographs, it was possible to outperform the state of the art in style classification. However, we also observe that although our strategy improves the overall MAP by a large margin, it under-performs for global attributes like Rule of Thirds or vanishing lines.

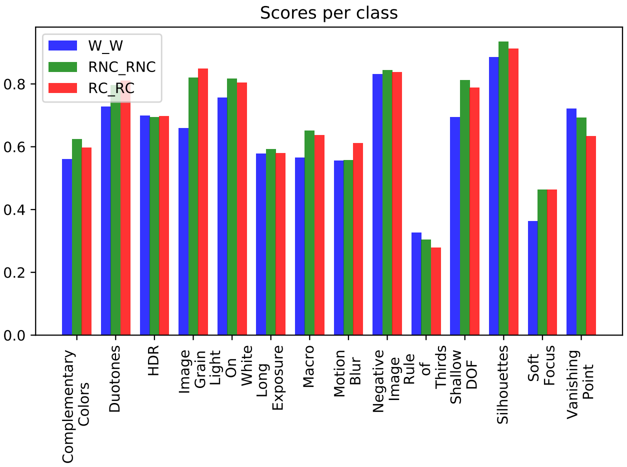


Figure 4: ***Class-Wise Precision scores for T_w, T_{rc} and T_{rnc} using DenseNet161:*** *The x-axis correspond to the 14 attributes. Each attribute has three bins corresponding to the precision scores for the three augmentations. The comparison shows that T_w performs worse and T_{rnc} performs better than T_{rc} in most cases. Only in the two global attributes, Rule of Thirds and Vanishing Lines, T_w performs better, which is consistent with our intuitive assumption.*

There are several possible directions from here. There is a need for a more robust approach for dealing with the global attributes. Building a more generalized model, not limited to a small set of attributes could be another direction. Extending the system to the domain of videos and 360 images could also be possible. We hope that this area will become more active in near future with its challenging and interesting set of problems.

References

[1] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Lawrence Zitnick, and D. Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2425–2433, 2015. 3

[2] T. O. Aydın, A. Smolic, and M. Gross. Automated aesthetic analysis of photographic images. *IEEE transactions on visualization and computer graphics*, 21(1):31–42, 2015. 3

[3] M. C. Beardsley. *Aesthetics, problems in the philosophy of criticism*. Hackett Publishing, 1981. 1

[4] R. Datta, D. Joshi, J. Li, and J. Wang. Studying aesthetics in photographic images using a computational approach. *Computer Vision–ECCV 2006*, pages 288–301, 2006. 3

[5] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database.

In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. IEEE, 2009. 1

[6] S. Dhar, V. Ordonez, and T. L. Berg. High level describable attributes for predicting aesthetics and interestingness. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1657–1664. IEEE, 2011. 3

[7] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010. 1

[8] R. Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015. 3

[9] G. Griffin, A. Holub, and P. Perona. Caltech-256 object category dataset. 2007. 1

[10] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3, 4, 5, 8

[11] G. Huang, Z. Liu, K. Q. Weinberger, and L. van der Maaten. Densely connected convolutional networks. *arXiv preprint arXiv:1608.06993*, 2016. 1, 3, 4, 5, 8

[12] G. Huang, Y. Sun, Z. Liu, D. Sedra, and K. Q. Weinberger. Deep networks with stochastic depth. In *European Conference on Computer Vision*, pages 646–661. Springer, 2016. 4

[13] D. Joshi, R. Datta, E. Fedorovskaya, Q.-T. Luong, J. Z. Wang, J. Li, and J. Luo. Aesthetics and emotions in images. *IEEE Signal Processing Magazine*, 28(5):94–115, 2011. 3

[14] S. Karayev, A. Hertzmann, H. Winnemoeller, A. Agarwala, and T. Darrell. Recognizing image style. In *BMVC 2014*, 2014. 1, 3, 5, 8

[15] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3128–3137, 2015. 3

[16] Y. Ke, X. Tang, and F. Jing. The design of high-level features for photo quality assessment. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 1, pages 419–426. IEEE, 2006. 3

[17] S. Kong, X. Shen, Z. Lin, R. Mech, and C. Fowlkes. Photo aesthetics ranking network with attributes and content adaptation. In *European Conference on Computer Vision*, pages 662–679. Springer, 2016. 3

[18] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. 3

[19] D. Lu and Q. Weng. A survey of image classification methods and techniques for improving classification performance. *International journal of Remote sensing*, 28(5):823–870, 2007. 3

[20] X. Lu, Z. Lin, H. Jin, J. Yang, and J. Z. Wang. Rapid: Rating pictorial aesthetics using deep learning. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 457–466. ACM, 2014. 1, 2, 3, 5, 8

[21] X. Lu, Z. Lin, X. Shen, R. Mech, and J. Z. Wang. Deep multi-patch aggregation network for image style, aesthetics, and quality estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 990–998, 2015. 1, 2, 3, 5, 8

[22] Y. Luo and X. Tang. Photo and video quality evaluation: Focusing on the subject. *Computer Vision–ECCV 2008*, pages 386–399, 2008. 3

[23] L. Mai, H. Jin, and F. Liu. Composition-preserving deep photo aesthetics assessment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 497–506, 2016. 3

[24] G. Malu, R. S. Bapi, and B. Indurkhya. Learning photography aesthetics with deep cnns, 2017. 3

[25] N. Murray, L. Marchesotti, and F. Perronnin. Ava: A large-scale database for aesthetic visual analysis. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2408–2415. IEEE, 2012. 1, 3, 5

[26] H. Noh, S. Hong, and B. Han. Learning deconvolution network for semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1520–1528, 2015. 3

[27] P. Obrador, M. A. Saad, P. Suryanarayan, and N. Oliver. Towards category-based aesthetic models of photographs. In *International Conference on Multimedia Modeling*, pages 63–76. Springer, 2012. 3

[28] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015. 3

[29] J. San Pedro, T. Yeh, and N. Oliver. Leveraging user comments for aesthetic aware image search reranking. In *Proceedings of the 21st international conference on World Wide Web*, pages 439–448. ACM, 2012. 3

[30] J. Schmidhuber. Deep learning in neural networks: An overview. *Neural networks*, 61:85–117, 2015. 3

[31] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 3, 5, 8

[32] R. K. Srivastava, K. Greff, and J. Schmidhuber. Training very deep networks. In *Advances in neural information processing systems*, pages 2377–2385, 2015. 4

[33] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning*, pages 2048–2057, 2015. 3

[34] A. Yilmaz, O. Javed, and M. Shah. Object tracking: A survey. *Acm computing surveys (CSUR)*, 38(4):13, 2006. 3

[35] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson. How transferable are features in deep neural networks? In *Advances in neural information processing systems*, pages 3320–3328, 2014. 4