This ICCV Workshop paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

PFAGAN: An aesthetics-conditional GAN for generating photographic fine art

Naila Murray
NAVER LABS Europe
naila.murray@naverlabs.com

Abstract

In this work we consider the problem of generating aesthetically pleasing photography, sometimes termed photographic fine art (PFA). We cast this problem as a generative modeling task and use a conditional GAN framework. Recent works have shown that conditioning based on semantic information is beneficial for improving photo-realism. In this work we propose a novel GAN architecture which is able to generate photo-realistic images with a specified aesthetic quality by conditioning on both semantic and aesthetic information. To condition the generator, we propose a modified conditional batch normalization layer. To condition the discriminator, we use a joint probabilistic model of semantics and aesthetics to estimate the compatibility between an image (either real or generated) and the conditioning variable. We show quantitatively and qualitatively that our model, called PFAGAN, is able to generate images conditioned on semantic categories and aesthetic scores.

1. Introduction

Most works in the computer vision community related to image aesthetics seek to assess [19, 20, 23, 15, 27] or enhance [36, 22, 15, 5] the aesthetic quality of a given image. In this work we aim to *generate aesthetically pleasing images* using a generative model. Recent years have seen an explosion in works aiming to model natural image distributions. In particular, models based on generative adversarial networks (GANs) [8] and variational auto-encoders [14, 31] have been used to learn generative models of images. Initially, only fairly simplistic image distributions such as digit images from MNIST were successfully modeled. More recently, visually complex images from ImageNet or the CelebA faces dataset [12, 26, 2] can be modeled with a high degree of realism.

GANs in particular have been shown to produce highly photo-realistic images but are notoriously difficult to train. This is due to several factors, including the sensitivity of the minimax objective function to minor changes in the model architecture and hyper-parameters [8, 33, 30]. One method for improving stability and photo-realism is to condition the generative model using semantic information, *e.g.* class labels, which encourages the model to generate an image depicting the corresponding class [24, 29]. This also has the benefit of serving to regularize the model, which is of particular importance for small-scale training datasets. Class-conditional generation can also enable applications where some control over the synthesized images is desired.

Figure 1: Synthetic images generated using PFAGAN: columns show images conditioned to have high aesthetic quality and to depict the "lake", "meadow", "sunset", and "barn" categories.

In this work, we aim to generate aesthetically pleasing images and propose a novel GAN architecture which can be conditioned using both semantic and aesthetic information. We use the AVA dataset [28] to train and evaluate our model. AVA is, to the best of our knowledge, the largest image dataset with both semantic and aesthetic annotations. The aesthetic annotations of AVA are histograms of scores; each image in AVA is associated with a histogram of scores given by different observers. Most works that use this data don't use these score histograms directly, but convert the histogram into a binary label by thresholding the mean score of each image. However, the chosen threshold is arbitrary and can introduce noise when training [27]. In fact, [28] found that, when training aesthetic classification models using thresholded scores as labels, removing training images with scores close to the threshold resulted in faster model convergence and similar test-time performance. An addi-

tional issue with removing low-quality images is that it removes useful training data for learning to condition on semantics. Given these two issues, we elect to directly use the (normalized) score histograms to condition our generative model.

In doing so we make the following main contributions:

- We propose a novel discriminator which conditions on both semantics and aesthetics. To aesthetically-condition the discriminator, we propose a projection-based compatibility function between score histograms and image representations in the discriminator.
- We propose a novel mixed-conditional batch normalization method to condition the generator on both variables. To do this we map score histograms to parameters that condition batch normalization layers in the generator.

Our quantitative and qualitative results show that PFAGAN is able to generate images that conform to semantic and aesthetic conditioning (*cf.* Figure 1).

2. Related work

A broad literature exists covering deep generative image models, and there are a plethora of works in recent years which are able to generate highly realistic images using methods based on variational auto-encoders [14, 31] and GANs [8]. We focus on works that propose conditional GANs and GANs for aesthetics, as these are most closely related to our own.

Conditional GANs: GANs were originally formulated to train a generator to mimic some target distribution, and are often used to model datasets such as MNIST [17], CIFAR [16] and ImageNet [32]. These datasets all contain images belonging to one of a distinct set of categories. Therefore, generating realistic images conforming to the statistics of these datasets necessarily requires the generator to implicitly learn categorical information. In order to more explicitly encode this categorical information, Chen *et al*. [3] introduce structure in a subset of the input random vector to the generator using information-theoretic regularization. Using this regularization they were able to encode variations in MNIST, such as digit rotation and digit category, in this subset of the input random variables.

Rather than learning to disentangle in this manner, works on conditional GANs have sought to explicitly encode variations of interest within the random vector [24, 29]. Categorical disentanglement is of particular interest, where one would like an explicit mapping between one or more variables in the random vector and a given object category in the target dataset. Typically, the categorical information is encoded with a one-hot vector that is appended to a random noise vector. To improve conditional generation, several works have augmented the objective function with loss functions targeting the categorization task [29].

Alternatives to concatenating the category embedding with the random vector have been proposed. For generators for example, works such as [26, 2] use conditional batch normalization [4]. In particular, category-specific scale and shift parameters are used in the batch normalization layers to generate category-specific images. For discriminators, [26] propose a projection-based alternative to concatenation which introduces categorical information using a compatibility function between the category embedding and the image representation. In this work, we propose a modified version of conditional batch normalization for continuous rather than discrete conditional information. We also extend the projection discriminator to allow for conditioning on both discrete and continuous conditioning variables.

GANs for aesthetics: Deng *et al*. [5] proposed a model for aesthetics-driven image enhancement which takes as input an image and generates a modified image with enhanced aesthetic properties. There have been may works addressing the related problem of style transfer, where the goal is to transfer the style from a source image to an existing target image [7, 6]. One work which aims to transfer aesthetically-pleasing styles is [34], where the authors perform style transfer on headshot photos using styles preferred by professional photographers when taking portraits. There has been very little work exploring aesthetics-aware training of GANs. The only work of which we are aware is that of [39], which includes two additional objectives in the loss function for training the GAN. One is a content-aware loss which captures the distance between feature maps given by the generator and those given by VGGNet [35]. The assumption is that VGGNet encodes semantic properties and so minimizing this loss will enhance the semantic properties of the generated image. In contrast to this, we aim to condition the generator using a specific semantic category. The second loss is an aesthetics-aware loss which aims to maximize the aesthetic score of generated images (using an auxiliary network to score the images according to their aesthetic quality). While this encourages the generator to synthesize aesthetically pleasing images, it does not allow the generator to be conditioned on an aesthetic random variable as in our case.

3. Method

We propose to use a GAN to learn a generative image model conditioned on semantic and aesthetic properties. The model consists of a generator $G(z, y_s, y_a; \theta_g)$ and a discriminator $D(x, y_s, y_a; \theta_d)$, where z is the input noise vector, y_s is the conditional semantic information, y_a is the

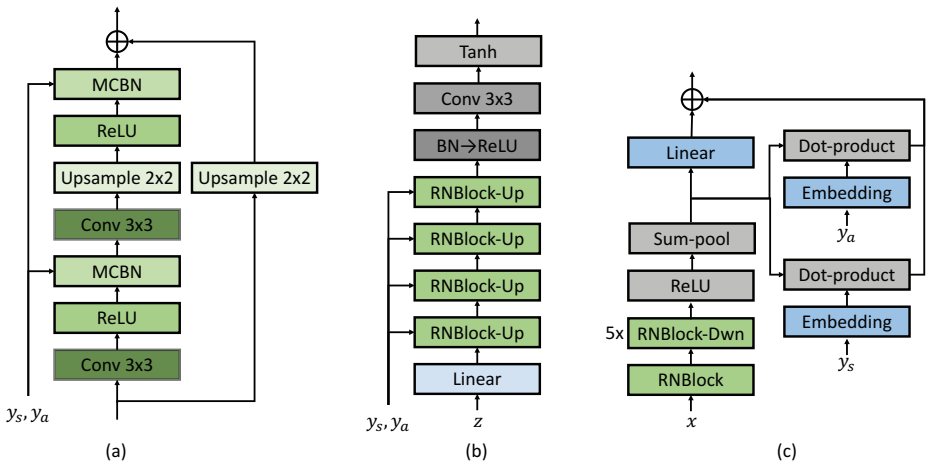


Figure 2: Schema of the components of PFAGAN: (a) the ResNet block with upsampling (RNBlock-Up) for the generator with MCBN layers; (b) the full generator network; (c) the full discriminator network. The RNet-Dwn layer differs from the RNBlock-Up layer in that the input is downsampled rather than upsampled and regular batch normalization is used.

conditional aesthetic information, x is the generated image, and θ_g and θ_d are the parameters of G and D respectively. We train the model using the standard two-player adversarial game described in [8], using a hinge loss [38, 18]:

$$\begin{aligned}\min_D\ & E_{q(y_s,y_a)}\left[E_{q(x|y_s,y_a)}\left[\max(0, 1 - D(x, y_s, y_a))\right]\right] + \\ & E_{q(y_s,y_a)}\left[E_{p(z)}\left[\max(0, 1 + D(G(x, y_s, y_a), y_s, y_a))\right]\right] \\ \min_G\ & - E_{q(y_s,y_a)}\left[E_{p(z)}\left[D(G(z, y_s, y_a), y_s, y_a)\right]\right],\end{aligned} \tag{1}$$

where q and p represent the true distribution and that of the generator G respectively.

3.1. Conditioning the generator

To condition the generator we propose a normalization procedure for convolutional layers that is related to both conditional batch normalization [4] and conditional instance normalization [6]. We first review these two normalization procedures before describing our mixed-conditional batch normalization approach.

Conditional batch normalization was proposed to condition visual representations on continuous language embeddings, and is formulated as follows:

$$o_{i,c} = \hat{\lambda}_c \left(\frac{h_{i,c} - \mu_c}{\sigma_c} \right) + \hat{\beta}_c, \tag{2}$$

where $h_{i,c}$ is element i of channel c, and μ_c and σ_c are the computed batch statistics. The scaling parameters, $\hat{\lambda}_c$ and $\hat{\beta}_c$, are computed by first applying two affine transformations to the language embedding to compute the vectors $\Delta\lambda \in \mathbb{R}^{|\mathcal{C}|}$ and $\Delta\beta \in \mathbb{R}^{|\mathcal{C}|}$, where $|\mathcal{C}|$ is the number of channels. $\hat{\lambda}$ and $\hat{\beta}$ are computed as:

$$\hat{\lambda} = \lambda + \Delta\lambda; \qquad \hat{\beta} = \beta + \Delta\beta \tag{3}$$

The affine transformation was learned after fixing all other parameters in the network.

Conditional instance normalization was originally designed to condition visual representations on different visual styles. It is formulated as:

$$o_{i,c} = \lambda_{s,c} \left(\frac{h_{i,c} - \mu_c}{\sigma_c} \right) + \beta_{s,c}. \tag{4}$$

Here, $s \in \mathcal{S}$, where $\mathcal{S}$ is the set of (style or semantic) categories, and $\lambda_{s,c}$ and $\beta_{s,c}$ are the category- and channel-specific scaling and shifting parameters. [6] showed that this simple mechanism is sufficient to learn a generator conditioned on categorical information. The $\lambda_{s,c}$ and $\beta_{s,c}$ parameters are stored in a look-up table and trained via back-propagation with the rest of the network.

Mixed-conditional batch normalization: To condition on both categorical information related to semantics and continuous information in the form of score distributions, we propose a three-step approach that draws inspiration from the two previous normalization methods. First, we select a set of parameters for a semantic category s from a look-up table. This set is $\mathcal{A}_s = \{W_{\lambda,s}, b_{\lambda,s}, W_{\beta,s}, b_{\beta,s}\}$, where $W_{\lambda,s}, W_{\beta,s} \in \mathbb{R}^{R \times |\mathcal{C}|}$, $b_{\lambda,s}, b_{\beta,s} \in \mathbb{R}^{|\mathcal{C}|}$, and R is the dimension of the score histogram. The set parametrises two affine transformations of the aesthetic information y_a, which is encoded as a normalized score histogram. In the

second step, these two affine transforms produce $\lambda_{s,a} \in \mathbb{R}^{|\mathcal{C}|}$ and $\beta_{s,a} \in \mathbb{R}^{|\mathcal{C}|}$ as follows:

$$\begin{aligned} \lambda_{s,a} &= y_a^\top W_{\lambda,s} + b_{\lambda,s}; \\ \beta_{s,a} &= y_a^\top W_{\beta,s} + b_{\beta,s}. \end{aligned} \tag{5}$$

Lastly, elements of $\lambda_{s,a}$ and $\beta_{s,a}$ are used, as in eqn 4, to condition on semantics and aesthetics. In Figure 2 we illustrate how we incorporate our mixed-conditional batch normalization layer, which we call (MCBN), into the generator. We learn the affine transformation parameters end-to-end with the rest of GAN parameters. MCBN is similar to conditional batch normalization in that affine transformations are used to compute the scaling and shifting parameters. It is also related to conditional instance normalization in that the affine parameters are selected via a look-up table and trained end-to-end.

3.2. Conditioning the discriminator

We extend the projection discriminator of [26], which maps a categorical or continuous conditioning variable into a shared space with an image representation. The dot product between the projection of the conditioning variable and the image representation serves as a compatibility function. This function is then maximized when training the generator, and maximized (resp. minimized) when training the discriminator with real (resp. generated) images.

To arrive at the projection formulation, [26] showed that the optimal solution for the hinge discriminator loss can be decomposed into the sum of two likelihood ratios. In our case in which there are two conditioning variables, the optimal solution D^* can be decomposed similarly:

$$\begin{aligned} D^*(x, y_s, y_a) &= \log \frac{q(x|y_s, y_a)q(y_s, y_a)}{p(x|y_s, y_a)p(y_s, y_a)} \\ &= \log \frac{q(y_s, y_a|x)}{p(y_s, y_a|x)} + \log \frac{q(x)}{p(x)}, \end{aligned} \tag{6}$$

where the semantic information is encoded as a one-hot vector y_s. To model $q(y_s, y_a|x)$ and $p(y_s, y_a|x)$ we assume that y_s and y_a are conditionally independent. While works have shown that aesthetic properties are content-dependent this simplifying assumption worked well in practice. We then formulate the optimal solution as follows:

$$D^*(x, y_s, y_a) = \log \frac{q(y_s|x)}{p(y_s|x)} + \log \frac{q(y_a|x)}{p(y_a|x)} + \log \frac{q(x)}{p(x)}. \tag{7}$$

Assuming a log-linear model for $q(y_s|x)$ gives [26]:

$$\log\, q(y_s|x) = v_s^{q\top} \phi(x) - \log Z_s^q(\phi(x)), \tag{8}$$

where s is the semantic category, $\phi(x)$ is the image representation and Z_s^q is the partition function of $q(y_s|x)$. Modeling $p(y_s|x)$ analogously gives:

$$\log \frac{q(y_s|x)}{p(y_s|x)} = (v_s^q - v_s^p)^\top \phi(x) - \log \frac{Z_s^q(\phi(x))}{Z_s^p(\phi(x))}. \tag{9}$$

If we model $q(y_a|x)$ and $p(y_a|x)$ as Gaussian distributions then, as shown in [26], we can obtain the following similar form:

$$\log \frac{q(y_a|x)}{p(y_a|x)} = \kappa + y_a^\top U \phi(x) - \log \frac{Z_a^q(\phi(x))}{Z_a^p(\phi(x))}, \tag{10}$$

where κ is a constant dependent only on y_a that can be ignored in the optimization, and U is a projection from y_a to the image representation space.

We estimate $\log \frac{q(x)}{p(x)} - \log \frac{Z_s^q(\phi(x))}{Z_s^p(\phi(x))} - \log \frac{Z_a^q(\phi(x))}{Z_a^p(\phi(x))}$ as $\psi(\phi(x))$, where ψ is a fully-connected layer in our case. We can then parametrize D as:

$$D(x, y_s, y_a) = y_s^\top V \phi(x) + y_a^\top U \phi(x) + \psi(\phi(x)), \tag{11}$$

where $y_s^\top V = v_s = v_s^q - v_s^p$. The semantic and aesthetic embedding functions V and U are trained end-to-end along with the other GAN parameters.

3.3. Implementation and network architecture

We use a ResNet[10]-like architecture for both D and G, similar to those used in [26]. G uses two MCBN layers within each ResNet convolutional block. We applied spectral normalization [25] to all of the weight tensors in D. We generate images of resolution 128×128. The architectures for D and G are shown in Figure 2. We trained both networks with a learning rate of 0.0002 and updated the discriminator 5 times for every update of the generator. We used the Adam optimization algorithm [13] for both networks, with $\beta_1 = 0$ and $\beta_2 = 0.9$. We used a batch size of 256 and used early stopping to terminate training. Our model, implemented in PyTorch, took 40 hours to train using 2 Nvidia V100 GPUs.

4. Dataset creation

There is no publicly available large-scale dataset with both reliable semantic annotations and reliable aesthetic annotations. Aesthetics datasets sourced from `dpchallenge.com`, such as AVA [28], contain reliable aesthetic annotations, in the form of a histogram of scores ranging from 1 to 10. Images in AVA have on average 210 scores. However, images from `dpchallenge.com` only contain weak and incomplete semantic annotations in the form of tags given to images by photographers. Photographers are limited to a maximum of 2 tags, chosen from a pre-defined list, so additional tags that might be relevant (*e.g.* an image can be relevant to "nature", "landscape", "black and white" and "rural" tags) cannot be added. In addition, different taggers have different conceptions of the semantics of different tags and no guidance is given in using them. As a result, images with a given tag tend to have a high variety of visual content and such tags are too noisy to be reliably used to train our model. Datasets sourced from

Flickr or photo.net have similar limitations to collecting semantic annotations. Their aesthetic annotations are also less interpretable and more sparse.

To obtain semantic annotations we adopt a semi-supervised approach. We use the AVA dataset [28], which contains 255K images from dpchallenge.com. A subset of 20K images from the AVA dataset was weakly annotated with 8 semantic categories using tags obtained from dpchallenge.com [28], with roughly 5K images per category. For each of these 20K images, we queried the entire AVA database to retrieve visually similar images. For this image retrieval procedure, we extracted representations for each database image using the model of Gordo *et al*. [9] and ranked the databases images in decreasing order of their dot-product similarity to each query image. While this representation was trained on images of landmarks, we found that the representations worked well for our task. Figure 3 shows representative query results for several images.

Among the top 5000 retrieved images for each query, we retain all images with a similarity score higher than 0.65. This gives 8 sets of retrieved images, one per category. For each set of images, we clustered their associated image representations using spectral clustering, with the number of clusters set to 100, resulting in 800 image clusters. We manually inspected and grouped similar clusters and discarded clusters that were incoherent or had fewer than 500 members. After this procedure, we obtained a dataset of 38506 images with 11 pseudo-labels corresponding to: "barn", "beach", "bird", "cat", "flower", "lake", "meadow", "mountain", "portrait", "sunset", "trees". We call this dataset AVA-Sem. Image samples from each category are shown in Figure 4, along with the number of images per category. One can observe that the clusters are semantically coherent, although there is still a high variance of style and composition, and some false positives.

5. Experiments

We now describe our evaluation of PFAGAN using quantitative and qualitative experiments. For testing, we use y_a and y_s pairs from the training set to condition random variable z. As an alternative to using conditioning information from the training set, one could model the distribution $m(y_a, y_s)$ of pairs and sample from it. However if the model is not sufficiently accurate this would run the risk of using unrealistic vectors to condition our model.

Metrics: Several metrics have been proposed to evaluate generative image models [33, 21]. The two most widely-used ones are the inception score (IS) [33] and the Fréchet inception distance (FID) [11]. User studies have also been proposed but are difficult to replicate consistently [33].

The inception score (IS) is derived from the softmax predictions of a deep classification model, InceptionV3 [37], trained on ImageNet. It was conceived to evaluate two desirable properties of a generative image model. The first is that it should generate images for which the inception model gives high confidence for one class, *i.e.*, there should be a clear and recognizable object that belongs to one of the classes in the dataset. The second is that it should generate images for which the inception model gives diverse predictions, *i.e.* the model should generate diverse content and avoid mode collapse. However there are serious drawbacks to the IS [1, 21] and for this work it is less appropriate because our model wasn't trained with ImageNet images and was trained to generate scenic images in addition to object-centered images. However we observed that it correlated to some extent with more realism and report it for completeness, using code provided by the authors of [33].

The FID is derived from the image representations extracted from InceptionV3. It is a measure of the distance between the distribution of representations extracted from real images and that of generated images. Both distributions are modeled as Normal distributions. IntraFID is a variant of FID introduced by [26] to evaluate category-conditional GANs, and is the average of the FID scores calculated between generated images and real images for each category. In our case, we would like to evaluate both the semantics-conditional generation and the aesthetics-conditional generation. We report results on semantics by computing the IntraFID. That is, for generated and real images with semantic category s, we calculate the statistics of their distributions. We then average across the categories. Because the aesthetics conditioning uses a continuous random vector y_a we can not directly use FID or IntraFID. We create two aesthetics-related categories, HiQ and LoQ by retaining real and generated images with mean scores (as computed using their normalized score histograms) higher than 6.5 and lower than 4.5, respectively. We then calculate the FID separately for these two categories.

Using a pre-trained model: To learn to generate realistic images, one needs large-scale datasets, ideally with many samples per category. As our dataset is fairly small-scale, we experimented with initializing our model using a model pre-trained on AVA, which has 255K images, compared to the 38506 images in AVA-Sem. Because the full AVA dataset does not have semantic annotations, we trained a version of PFAGAN in which the parameters for semantic conditioning were removed. Specifically, for each batch normalization layer in the generator, a single set of affine transformation parameters $\mathcal{A}$ was used to map aesthetic annotations to scaling and shifting parameters. For the discriminator, the $y_s^\top V\phi(x)$ term was removed from equation 11. We then initialize each $\mathcal{A}_s$ in PFAGAN with the corresponding pre-trained $\mathcal{A}$ for that layer. We initialize the V embedding with random Gaussian noise.

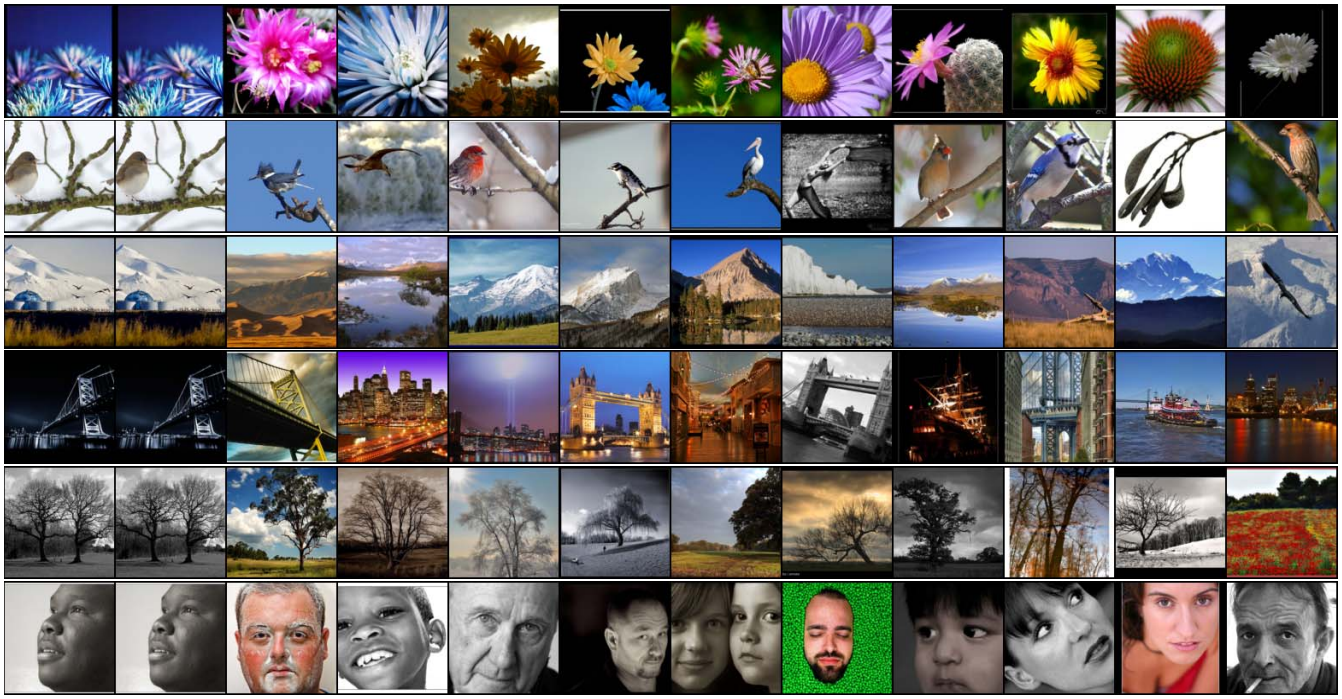
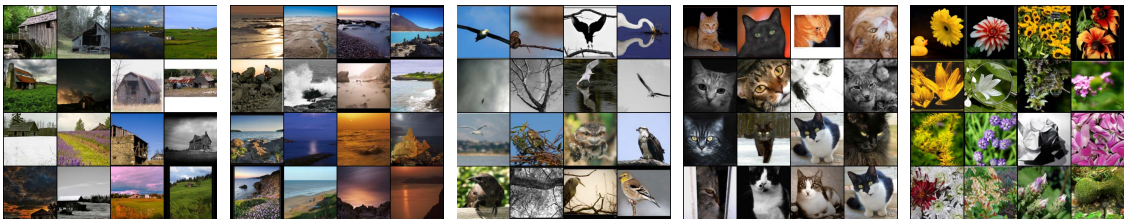
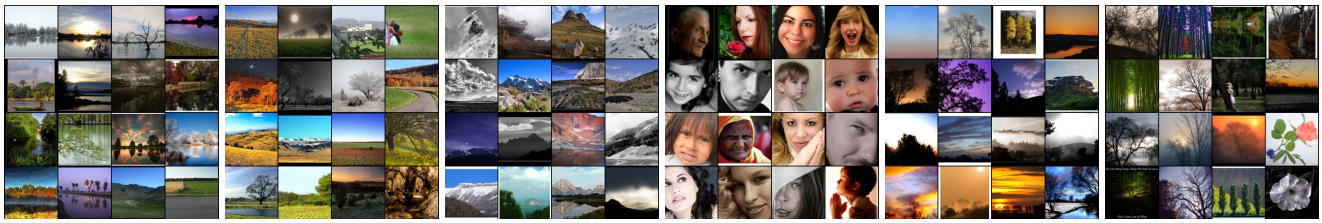


Figure 3: Query results for a sample of AVA images. The first column contains the query image while the remaining columns show the retrieved images in decreasing order of similarity. Note that the first retrieved image is always the query image itself. See section 4 for details.



(a) barn (1808) (b) beach (1622) (c) bird (1519) (d) cat (2051) (e) flower (10868)



(f) lake (3280) (g) meadow (2092) (h) mountain (720) (i) portrait (7871) (j) sunset (3319) (k) trees (3356)

Figure 4: Images from each of the 11 pseudo-categories of AVAc. In parentheses we show the number of images for each category. The categories are highly semantically coherent, despite a few false positives such as the images of flowers in the "trees" category. See section 4 for details.

5.1. Results

In Table 1 we report results on the AVA-Sem dataset with (PFAGAN-pt) and without (PFAGAN) pre-training. We report the inception score (IS) calculated over all images. We also report the Intra-FID score across semantic categories (Intra-FID-Sem), the FID for high-quality images (FID-HiQ), and the FID for low-quality images (FID-LoQ). Unsurprisingly, pre-training the model significantly improves performance for all FID-based metrics. In addition, FID-

Method	IS	Intra-FID-Sem	FID-HiQ	FID-LoQ
PFAGAN	3.81	117.50	113.68	120.71
PFAGAN-pt	**3.95**	**75.25**	80.62	**66.80**
PDGAN [26]	3.59	86.89	**76.96**	74.54

Table 1: Results on AVA-Sem dataset with (PFAGAN-pt) and without (PFAGAN) pre-training, as well as the projection discriminator GAN [26]. We report the inception score (IS) and different FID metrics (*cf.* section 5.1).

HiQ is a fair bit higher than FID-LoQ. This may be due in part to increased complexity in modeling high quality images. We discuss this further in section 5.2.

Comparison to baseline: We compare to the projection discriminator GAN (PDGAN) [26], which can be considered the baseline version of PFAGAN in which the aesthetics projection term in equation 11 and the MCBN are removed. We explicitly condition PDGAN using only semantic information y_s, and condition on aesthetics implicitly by training using only the subset of AVA-Sem (17948 images) with a mean score greater than 5.5 (the midpoint of the rating scale). In Table 1 we see that the FID-LoQ is significantly worse than for PFAGAN-pt, which is to be expected as PDGAN was trained using only HiQ images. The FID-HiQ is better, which can be explained by the same fact. However the Intra-FID-Sem performs much worse. This is unsurprising as the dataset now has fewer images with which to train the model to condition semantically. Note that pre-training PDGAN with AVA did not improve performance, indicating that AVA is not effective for pre-training if the final GAN will be conditioned only on semantics.

5.2. Qualitative results:

In Figure 6 we show randomly generated images for all 11 categories, ordered by decreasing FID score. HiQ images are shown in the top 5 rows while LoQ images are shown in the bottom ones. We generated the images using PFAGAN-pt. For several categories, the model is able to effectively replicate key aesthetic characteristics. For example, reflection on water is visible for images in the "lake" and "sunset" categories, and rule-of-thirds composition is present in the "sunset" and "barn" categories. Landscape categories such as "lake", "beach" and "sunset" tend to show more realism than categories corresponding to animals, plants and people, such as "portrait" and "cat". This is likely because images in the former category are dominated by low-frequency and/or repetitive textures that are relatively easy to model. Additional causes of low realism for some categories include too few training examples, and a high degree of intra-class variability. A degree of mode collapse is evident for several semantic-aesthetic configurations.

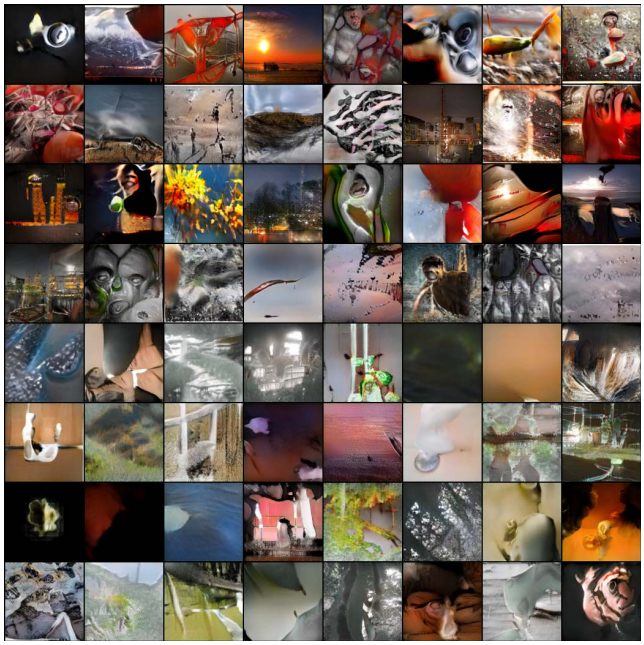


Figure 5: Generated images with a model conditioned only on aesthetic information. The first 4 rows contain images conditioned on HiQ aesthetic score distributions, while the last 4 rows contain images conditioned on LoQ aesthetic score distributions. While general aesthetic properties of high- and low-quality images have been learned, the images exhibit limited semantic structure.

The HiQ generated images clearly have different characteristics when compared to their LoQ counterparts. For example, HiQ images tend to exhibit high colour contrast and saturation, particularly for landscape categories where a dramatic sky is desirable. In other categories such as "flower" and "portrait", they exhibit less clutter than LoQ images. We validated this quantitatively by computing the FID between generated HiQ images and real LoQ images (denoted HiQG-vs-LoQR) and vice versa (denoted LoQG-vs-HiQR). We obtained FID=86.13 for HiQG-vs-LoQR and FID=88.95 for LoQG-vs-HiQR, both of which are higher than our FID-HiQ and FID-LoQ scores. This indicates that our generated HiQ images (resp. generated LoQ images) are indeed closer to real HiQ images (resp. real LoQ images) than our generated LoQ images (resp. generated HiQ images), and that PFAGAN is able to effectively use the aesthetics-conditioning information to modulate image generation.

Conditioning on semantics was key to generating realistic images. To illustrate this, Figure 5 shows images generated using the version of PFAGAN which is conditioned using only aesthetic information (and was used for pre-training). The top 4 rows are generated with HiQ annota-

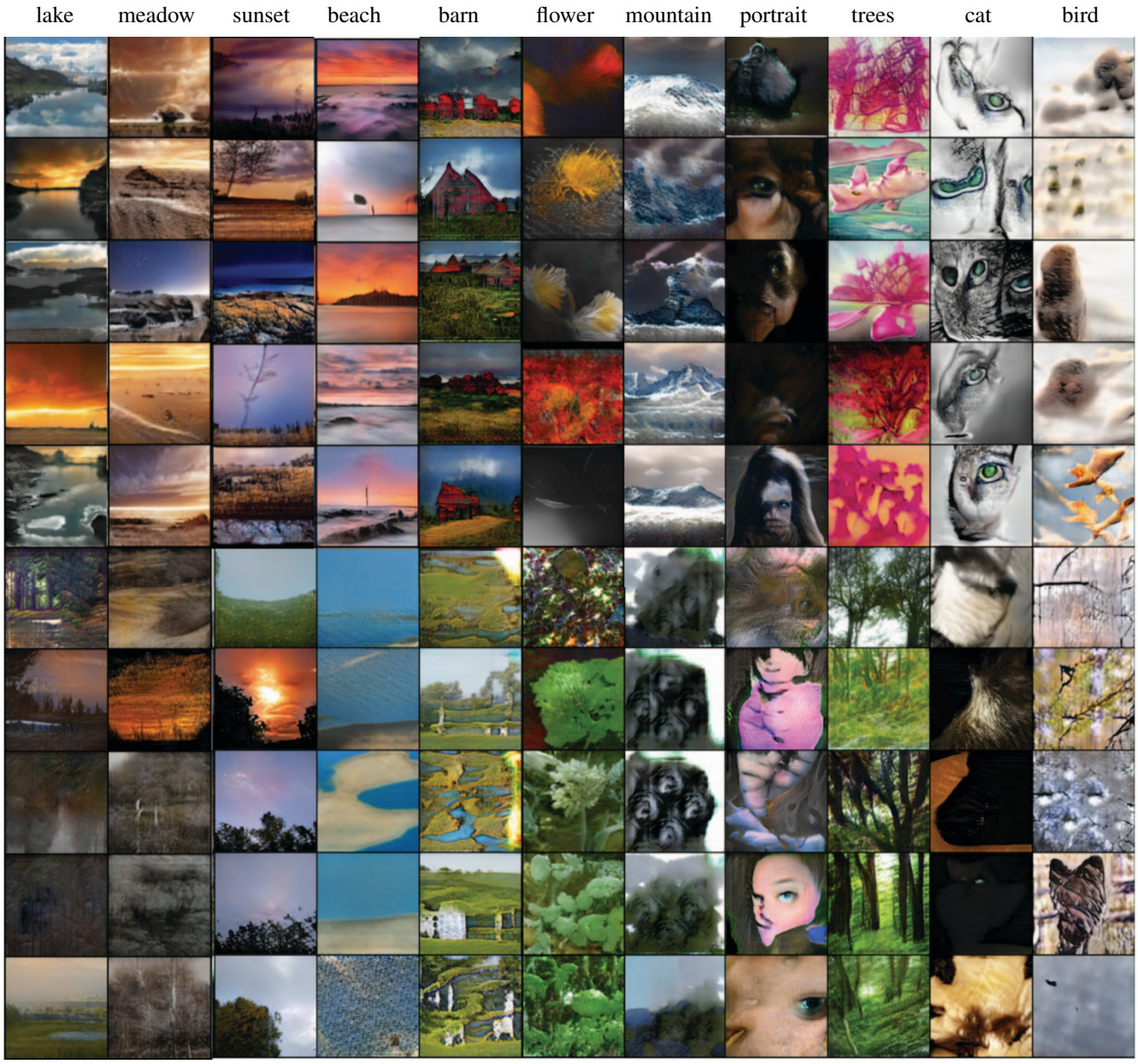


Figure 6: Generated images for the 11 categories in AVAc. The columns are ordered by decreasing FID score. The first 5 rows contain images conditioned on HiQ aesthetic score distributions, while the last 5 rows contain images conditioned on LoQ aesthetic score distributions.

tions and the bottom 4 with LoQ ones. While some differences in color saturation can be observed between the two sets of images, there is little recognizable structure.

6. Conclusion

We introduce the problem of aesthetics-conditional image generation and propose a conditional GAN model, PFAGAN, that generates photographic fine art by conditioning the generator on both semantic and aesthetic criteria. We propose a mixed-conditional batch normalization layer to condition the generator. We use a projection-based conditioning method for the discriminator that assumes conditional independence of the image aesthetics and semantics given the image. We show that PFAGAN is able to capture both aspects, aesthetic and semantic, in the generative model.

References

[1] S. Barratt and R. Sharma. A note on the inception score. 2018. 5

[2] A. Brock, J. Donahue, and K. Simonyan. Large scale gan training for high fidelity natural image synthesis. *ICLR*, 2019. 1, 2

[3] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *NIPS*, 2016. 2

[4] H. De Vries, F. Strub, J. Mary, H. Larochelle, O. Pietquin, and A. C. Courville. Modulating early visual processing by language. In *NIPS*, 2017. 2, 3

[5] Y. Deng, C. C. Loy, and X. Tang. Aesthetic-driven image enhancement by adversarial learning. In *ACMMM*, 2018. 1, 2

[6] V. Dumoulin, J. Shlens, and M. Kudlur. A learned representation for artistic style. In *ICLR*, 2017. 2, 3

[7] L. A. Gatys, A. S. Ecker, and M. Bethge. Image style transfer using convolutional neural networks. In *CVPR*, 2016. 2

[8] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *NIPS*, 2014. 1, 2, 3

[9] A. Gordo, J. Almazan, J. Revaud, and D. Larlus. Deep image retrieval: Learning global representations for image search. In *ECCV*, 2016. 5

[10] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 4

[11] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NIPS*, 2017. 5

[12] T. Karras, T. Aila, S. Laine, and J. Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *ICLR*, 2018. 1

[13] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv*, 2014. 4

[14] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *ICLR*, 2014. 1, 2

[15] S. Kong, X. Shen, Z. Lin, R. Mech, and C. Fowlkes. Photo aesthetics ranking network with attributes and content adaptation. In *ECCV*, 2016. 1

[16] A. Krizhevsky. Learning multiple layers of features from tiny images. Technical report, 2009. 2

[17] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, et al. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 1998. 2

[18] J. H. Lim and J. C. Ye. Geometric gan. *arXiv*, 2017. 3

[19] X. Lu, Z. Lin, H. Jin, J. Yang, and J. Z. Wang. Rapid: Rating pictorial aesthetics using deep learning. In *ACM MM*, 2014. 1

[20] X. Lu, Z. Lin, H. Jin, J. Yang, and J. Z. Wang. Rating image aesthetics using deep learning. *IEEE Transactions on Multimedia*, 2015. 1

[21] M. Lucic, K. Kurach, M. Michalski, S. Gelly, and O. Bousquet. Are gans created equal? a large-scale study. In *NIPS*. 2018. 5

[22] R. S. M. S. Subhabrata Bhattacharya. framework for photo-quality assessment and enhancement based on visual aesthetics. *ACM MM*, 2011. 1

[23] L. Marchesotti, N. Murray, and F. Perronnin. Discovering beautiful attributes for aesthetic image analysis. *IJCV*, 2015. 1

[24] M. Mirza and S. Osindero. Conditional generative adversarial nets. *arXiv*, 2014. 1, 2

[25] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida. Spectral normalization for generative adversarial networks. 2018. 4

[26] T. Miyato and M. Koyama. cgans with projection discriminator. In *ICLR*, 2018. 1, 2, 4, 5, 7

[27] N. Murray and A. Gordo. A deep architecture for unified aesthetic prediction. *arXiv*, 2017. 1

[28] N. Murray, L. Marchesotti, and F. Perronnin. Ava: A large-scale database for aesthetic visual analysis. In *CVPR*, 2012. 1, 4, 5

[29] A. Odena, C. Olah, and J. Shlens. Conditional image synthesis with auxiliary classifier gans. In *ICML*, 2017. 1, 2

[30] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *ICLR*, 2016. 1

[31] D. J. Rezende, S. Mohamed, and D. Wierstra. Stochastic backpropagation and approximate inference in deep generative models. *ICML*, 2014. 1, 2

[32] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *IJCV*, 2015. 2

[33] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved techniques for training gans. In *NIPS*, 2016. 1, 5

[34] Y. Shih, S. Paris, C. Barnes, W. T. Freeman, and F. Durand. Style transfer for headshot portraits. *ACM TOG*, 2014. 2

[35] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015. 2

[36] M. S. Subhabrata Bhattacharya, Rahul Sukthankar. A coherent framework for photo-quality assessment and enhancement based on visual aesthetics. In *ACM MM*, 2010. 1

[37] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016. 5

[38] D. Tran, R. Ranganath, and D. M. Blei. Hierarchical implicit models and likelihood-free variational inference. *NIPS*, 2017. 3

[39] R. Zhang, X. Liu, Y. Guo, and S. Hao. Image synthesis with aesthetics-aware generative adversarial network. In *Pacific Rim Conference on Multimedia*, 2018. 2