

# BAM! The *Behance Artistic Media* Dataset for Recognition Beyond Photography

Michael J. Wilber<sup>1,2</sup> Chen Fang<sup>1</sup>  
 John Collomosse<sup>1</sup>  
<sup>1</sup> Adobe Research

Hailin Jin<sup>1</sup> Aaron Hertzmann<sup>1</sup>  
 Serge Belongie<sup>2</sup>  
<sup>2</sup> Cornell Tech

## Abstract

Computer vision systems are designed to work well within the context of everyday photography. However, artists often render the world around them in ways that do not resemble photographs. Artwork produced by people is not constrained to mimic the physical world, making it more challenging for machines to recognize.

This work is a step toward teaching machines how to categorize images in ways that are valuable to humans. First, we collect a large-scale dataset of contemporary artwork from Behance, a website containing millions of portfolios from professional and commercial artists. We annotate Behance imagery with rich attribute labels for content, emotions, and artistic media. Furthermore, we carry out baseline experiments to show the value of this dataset for artistic style prediction, for improving the generality of existing object classifiers, and for the study of visual domain adaptation. We believe our Behance Artistic Media dataset will be a good starting point for researchers wishing to study artistic imagery and relevant problems. **This dataset can be found at <https://bam-dataset.org/>**

## 1. Introduction

“Art is an effort to create, beside the real world, a more humane world.” – André Maurois

Recent advances in Computer Vision have yielded accuracy rivaling that of humans on a variety of object recognition tasks. However, most work in this space is focused on understanding *photographic imagery* of everyday scenes. For example, the widely-used COCO dataset [17] was created by “gathering images of complex everyday scenes containing common objects in their natural context.” Outside of everyday photography, there exists a diverse, relatively unexplored space of *artistic imagery*, offering depictions of the world as reinterpreted through artwork. Besides being culturally valuable, artwork spans broad styles that are not found in everyday photography and thus are not available to current machine vision systems. For example, current object clas-



Figure 1. State of the art object detectors such as SSD trained on Pascal VOC can reliably detect objects in everyday photographs (top row), but do not generalize to other kinds of artistic media (see scores under each image). In this work, **we create a large-scale artistic dataset** spanning a breadth of styles, media, and emotions. We can use this dataset to improve the generality of object classifiers—our object classifier’s scores are above 0.95 for all these images.

sifiers trained on ImageNet and Pascal VOC are frequently unable to recognize objects when they are depicted in artistic media (Fig. 1). Modeling artistic imagery can increase the

generality of computer vision models by pushing beyond the limitations of photographic datasets.

In this work, we create a large-scale artistic style dataset from Behance, a website containing millions of portfolios from professional and commercial artists. Content on Behance spans several industries and fields, ranging from creative direction to fine art to technical diagrams to graffiti to concept design. Behance does not aim to be a historical archive of classic art; rather, we start from Behance because it represents a broad cross-section of contemporary art and design.

Our overall goal is to create a dataset that researchers can use as a testbed for studying artistic representations across different artistic media. This is important because existing artistic datasets are too small or are focused on classical artwork, ignoring the different styles found in contemporary digital artwork. To solidify the scope of the problem, we choose to explore three different facets of high-level image categorization: object categories, artistic media, and emotions. These artistic facets are attractive for several reasons: they are readily understood by non-experts, they can describe a broad range of contemporary artwork, and they are not apparent from current photographic datasets.

We keep the following goals in mind when deciding which attributes to annotate. For object categories, we wish to annotate objects that may be drawn in many different visual styles, collecting fewer visually distinct categories but increasing the density (instances per category) and breadth of representation. ImageNet and COCO, for example, contain rich fine-grained object annotations, but these datasets are focused on everyday photos and cover a narrow range of artistic representation. For media attributes, we wish to annotate pictures rendered with all kinds of professional media: pencil sketches, computer-aided vector illustration, watercolor, and so on. Finally, emotion is an important categorization facet that is relatively unexplored by current approaches.

There are several challenges, including annotating millions of images in a scalable way, defining a categorization vocabulary that represents the style and content of Behance, and using this resource to study how well object recognition systems generalize to unseen domains. According to our quality tests, the precision of the labels in our dataset is 90%, which is reasonable for such a large dataset without consortium level funding.

**Our contributions** are twofold:

- **A large-scale dataset**, the Behance Artistic Media Dataset, containing almost 65 million images and quality assurance thresholds. We also create an **expert-defined vocabulary** of binary artistic attributes that spans the broad spectrum of artistic styles and content represented in Behance. This dataset can be found at <https://bam-dataset.org/> upon release late

	Size	Scope	Annotations
A-SUN [25]	0.014m	Photos of scenes	Objects, context
Behance-2M (Private) [6]	1.9m	Contemporary artwork	User/View behavior
Recognizing Image Style [13]	0.16m	Photos, paintings	Art genre, photo techniques
AVA [21]	0.25m	Photos	Aesthetics, content, style
Visual sentiment ontology [1]	0.31m	Photos, videos	Adj/Noun pairs
OpenImages [14]	9.2m	Photos	Content labels
<b>Behance Artistic Media</b>	<b>65m</b>	<b>Contemporary artwork</b>	<b>Emotion, Media, Objects</b>

Table 1. A comparison of several related datasets. Our Behance Artistic Media dataset is much larger than the others and includes a broad range of contemporary artwork.

Spring 2017.

- An investigation of the **representation gap** between objects in everyday ImageNet photographs and objects rendered in artistic media on Behance. We also explore how models trained on one medium can transfer that performance to unseen media in a **domain adaptation setting**. To investigate aesthetics and art styles, we compare performance of different kinds of features in predicting emotion and media and show how Behance Artistic Media can be used to improve style classification tasks on other datasets. Finally, we briefly investigate **style-aware image search**, showing how our dataset can be used to search for images based on their content, media, or emotion.

We believe this dataset will provide a starting foundation for researchers who wish to expand the horizon of machine vision to the rich domain of artwork.

## 2. Related Work

Attributes and other mid-level representations [28, 7] have a long and rich history in vision. Attributes have been applied to aesthetics and other artistic qualities, usually with a focus on photography. For instance, Obrador *et al.* [23], Dhar *et al.* [5], and Murray *et al.* [21] collect descriptive attributes such as interestingness, symmetry, light exposure, and depth of field. Work by Peng *et al.* [26], You *et al.* [31], Jou *et al.* [12], and Borth *et al.* [1] study emotional attributes in photographs. Others describe image style not in attributes, but in terms of low-level feature correlations as in work done by Gatys *et al.* [8], Lin *et al.* [18], and others. We are more concerned about high-level image categorization than low-level texture transfer.

Ours is not the only dataset focused on artwork. We compare related artistic datasets in Tab. 1. Most are focused exclusively on everyday photographs [21, 25, 1], but some [13, 2, 9] include classical paintings. Likewise, Ginosar *et al.* [9] discuss person detection in cubist art. The work of Fang *et al.* [6] also studies Behance imagery, but does not collect descriptive attributes. Recently, Google released the “Open Images” dataset [14] containing some media-related labels including “comics”, “watercolor paint”, “graffiti”, etc. However, it is unclear how the quality of the labeling was evaluated. Each of these labels contain less than 400 human-verified images and there are no labels that cap-



Figure 2. Top: Sampling of images within projects with the “Cat” tag. Projects with the “Cat” tag may contain other animals (1), title cards (3,5), or unrelated pictures (5,6). Bottom: Top classifications from a classifier trained to distinguish the “Cat” tag. Images are more related, but this tends to learn many small animals. The precision of cats in the top 100-scoring images is only 36%.

ture emotions. Our work is most similar in spirit to Karayev *et al.* [13], which studies photographic image style. They collect annotations for photographic techniques, composition, genre, and mood on Flickr images, as well as a set of classical painting genres on Wikipaintings. Our focus is on non-photorealistic contemporary art. To our knowledge, our work is the first work seeking to release a large-scale dataset of a broad range of contemporary artwork with emotion, media, and content annotations.

### 3. The Behance Media Dataset

Our dataset is built from <http://behance.net>, a portfolio website for professional and commercial artists. Behance contains over ten million projects and 65 million images. Images on Behance are grouped into Projects, the fundamental unit of categorization. Each Project is associated with metadata, including a title, description, and several noisy user-supplied tags.

Artwork on Behance spans many fields, such as sculpture, painting, photography, graphic design, graffiti, illustration, and advertising. Graphic design and advertising make up roughly one third of Behance. Photography, drawings, and illustrations make up roughly another third. This artwork is posted by professional artists to show off samples of their best work. We encourage the reader to visit <http://behance.net> to get a sense of the diversity and quality of imagery on this site. Example images from Behance are shown in Fig. 3.

**Selecting attribute categories.** In this work, we choose to annotate our own artistic binary attributes. Attribute names are rendered in sans serif font. Our attributes capture three categorization facets:

- **Media attributes:** We label images created in 3D computer graphics, comics, oil painting, pen ink, pencil sketches, vector art, and watercolor.
- **Emotion attributes:** We label images that are likely to make the viewer feel calm/peaceful, happy/cheerful, sad/gloomy, and scary/fearful.

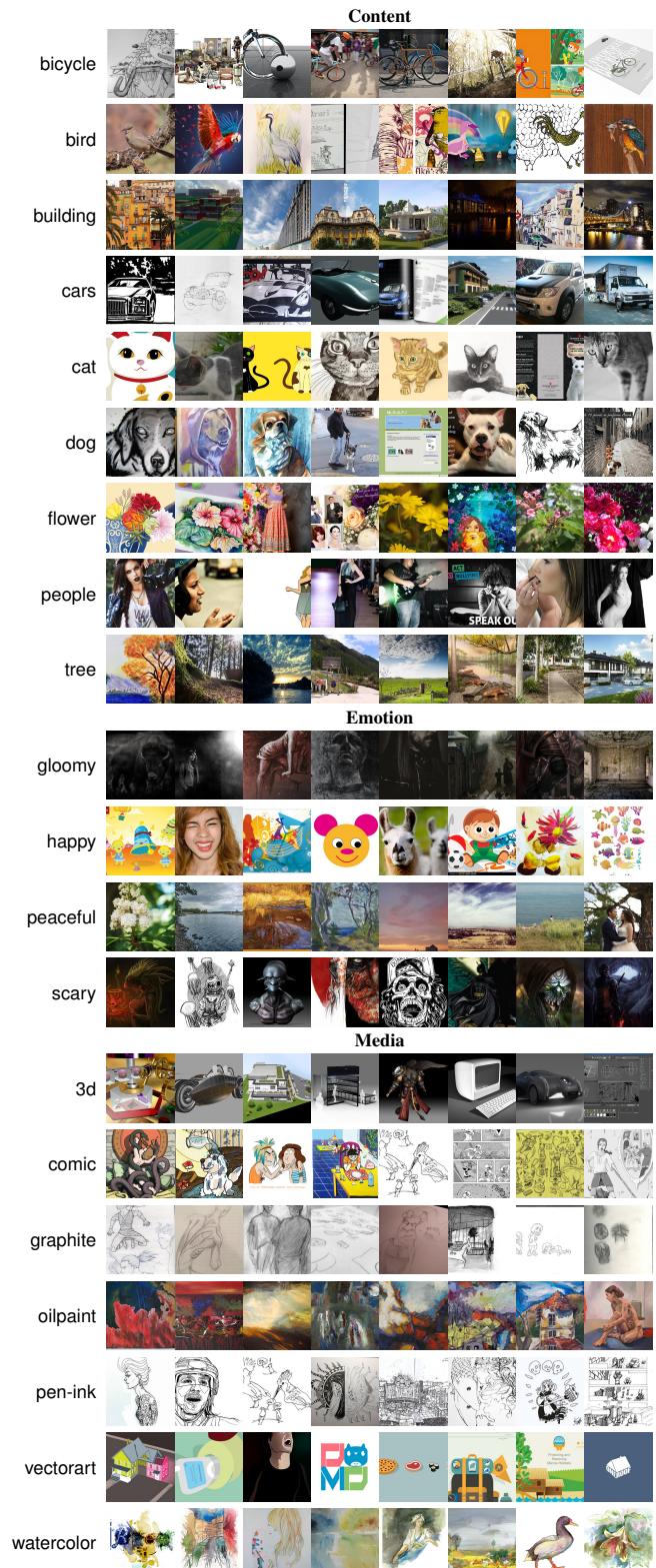


Figure 3. Example images from Behance Artistic Media. We encourage the reader to zoom in for more detail.

- **Entry-level object category attributes:** We label images containing bicycles, birds, buildings, cars, cats, dogs, flowers, people, and trees.

We chose these attributes as follows: The seven media attributes were chosen on the expert advice of a resident artist to roughly correspond with the genres of artwork available in Behance that are easy to visually distinguish. Our goal is to strike a balance between distinctive media while covering the broad range available in Behance. For instance, oil paint and acrylic are considered to be different media by the artistic community, but are very hard for the average crowdworker to distinguish visually. The four emotion attributes are seen on Plutchik’s *Wheel of Emotions* [27], a well-accepted model for emotions that was also used in [12]. From this model, we chose the emotions that are likely to be visually distinctive. The content attributes represent entry-level object categories and were chosen to have some overlap with Pascal VOC while being representative of Behance content. We focus on entry-level categories because these categories are likely to be rendered in a broad range of styles throughout Behance.

Although this work is only concerned with a small set of labels (arguably a proof-of-concept), the dataset we release could itself be the basis for a real PASCAL/COCO-sized labeling effort which requires consortium-level funding.

**Tags are noisy.** Behance contains user-supplied tags, and one may wonder whether it is feasible to train attribute classifiers directly from these noisy tags alone, such as in previous work [11, 20]. However, unlike that work, we cannot create our dataset from tags alone for two reasons. First, not all of our attributes have corresponding tags. Second, tags are applied to each *project*, not each image. For example, even though a project called “Animal sketches 2012” may have the “Dog” tag, we do not know which image that tag should apply to. Training on tags alone is too noisy and reduces the final classifier precision. To demonstrate, we train a binary classifier on the “Cat” tag, but from manual inspection, it only learns to distinguish different small animals and is not fine-grained enough to find cats; see Fig. 2. The precision of cats among the top 100 detections is only about 36%. To increase this accuracy, we must rely on human expertise to collect labels.

### 3.1. Annotation pipeline

Our dataset requires some level of human expertise to label, but it is too costly to collect labels for all images. To address this issue, we use a hybrid human-in-the-loop strategy to incrementally learn a binary classifier for each attribute. Our hybrid annotation strategy is based on the LSUN dataset annotation pipeline described in [32], which itself shares some similarity with other human-in-the-loop collection systems [32, 15, 4]. An overview of this process is shown in Fig. 4. At each step, humans label the most informative samples in the dataset with a single binary attribute

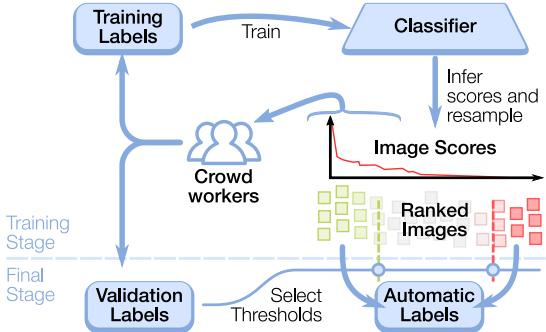


Figure 4. A diagram of our crowdsourcing pipeline. First, we train a set of classifiers on all labels collected so far. We then use this classifier to rank a random sample of images. High-scoring images are sent back to the crowd, and the resulting labels are added to the training and validation set. After four iterations, the validation set is used to select positive and negative thresholds with certain precision and recall targets. Images meeting these thresholds are added to the automatic label set.

label. The resulting labels are added to each classifier’s training set to improve its discrimination. The classifier then ranks more images, and the most informative images are sent to the crowd for the next iteration. After four iterations, the final classifier re-scores the entire dataset and images that surpass a certain score threshold are assumed to be positive. This final threshold is chosen to meet certain precision and recall targets on a held-out validation set. This entire process is repeated for each attribute we wish to collect.

**Crowdsourcing task.** The heart of our human-in-the-loop system is the actual human annotation task. We collect annotations for each attribute independently. To do this, we rely on Amazon Mechanical Turk, a crowdsourced marketplace. Crowdworkers (“Turkers”) complete Human Intelligence Tasks for a small cash payment. In each HIT for a given attribute, we show the Turker 10 handpicked positive/negative example images and collect 50 binary image annotations. Turkers indicate whether each image has the attribute of interest. Each HIT only collects labels for a single attribute at a time to avoid confusion. For quality control, we show each image to two separate Turkers and only use answers where both Turkers agree. We also collect sparse text annotations for a subset of these images. Every 10 images, we present an annotation recently provided by the Turker and ask for a brief 3-word caption to justify their choice. This has the effect of encouraging annotators to carefully consider and justify their choices.

It is always important to balance the trade-off between squeezing high-quality work out of annotators while being respectful of their effort and abilities. The subjectivity of our task makes this trade-off harder to manage. To address this issue, we prevent spam by only accepting work from crowdworkers who previously completed 10,000 MTurk tasks with

a 95% acceptance rate, and we only use labels where both workers agree. Finally, the quality of the entire dataset is ensured by setting appropriate label thresholds on held-out validation data; see Sec. 3.1

**Iterative learning.** Starting from a small handpicked initial label set, the dataset is enlarged by an iterative process that alternates between training a classifier on the current label set, applying it to unlabeled images, and sending unconfident images back to the crowd for more labeling. On each iteration, we train a deep learning classifier using 10/11<sup>th</sup>s of the total collected crowd labels. The last 1/11<sup>th</sup> is always held out for validation. We apply this classifier to the entire dataset. The crowd then labels 5,000 images that score higher than a threshold set at 50% precision measured on validation data. This way, we show our Turkers a balance of likely-positive and likely-negative images each time.

After four iterations, we arrive at a final classifier that has good discrimination performance on this attribute. We score the entire dataset with this classifier and use thresholds to select the final set of positives and negatives. The positive score threshold is chosen on validation data such that the precision of higher-scoring validation images is 90%, and the negative threshold is chosen such that the recall of validation images above this threshold is 95%. In this way, we can ensure that our final labeling meets strict quality guarantees.

It is important to note that the resulting size of the dataset is determined solely by the number of relevant images in Behance, our desired quality guarantees, and the accuracy of the final classifier. A better attribute classifier can add more images to the positive set while maintaining the precision threshold. If we need more positive data for an attribute, we can sacrifice precision for a larger and noisier positive set.

**Classifier.** For content attributes, our classifier is a fine-tuned 50-layer ResNet [10] originally trained on ImageNet. For emotion and media attributes, we found it better to start from *StyleNet* [6]. This model is a GoogLeNet [30], fine-tuned on a style prediction task inferred from user behavior. Each network is modified to use binary class-entropy loss to output a single attribute score. To avoid overfitting, we only fine-tune for three epochs on each iteration. See Fig. 3 for examples of Behance images.

**Resulting dataset statistics** Our final dataset includes positive and negative examples for 20 attributes. The median number of positive images across each attribute is 54,000, and the median number of negative images is 8.7 million. The “People” attribute has the most positive images (1.74 million). Humans are commonly featured as art subjects, so this is not surprising. The attribute with the least positives is “Cat” with 19,244 images. We suspect this is because our final labeling model cannot easily distinguish cats from other cat-like renditions. Cats on Behance are commonly rendered in many different styles with very high intra-class variation. Statistics for all attributes are shown in Fig. 5.

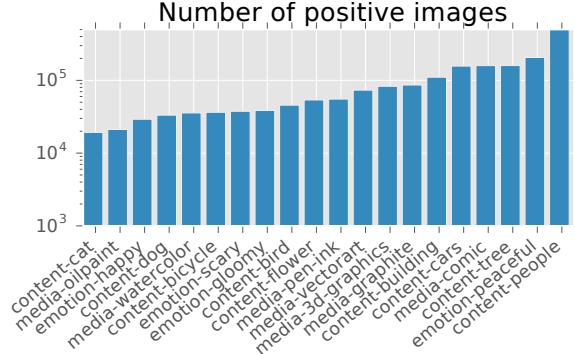


Figure 5. Top: Number of positive images in the final set.

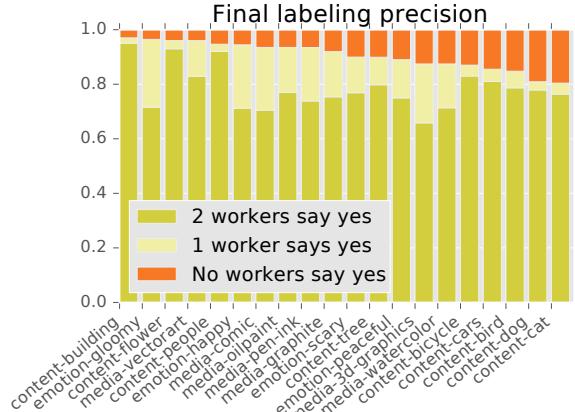


Figure 6. Final quality assurance: Showing worker agreement of automatically-labeled positive images in the final dataset.

Our automatic labeling model can amplify the crowd’s annotation effort. The ratio of automatic positive labels to crowd-annotated positive labels is 17.4. The amplification factor for negative labels is much higher—about 505—because automatic systems can quickly throw away easy negatives to focus the crowd’s attention on potentially relevant images.

**Final quality assurance** As a quality check, we tested whether the final labeling set meets our desired quality target of 90% precision. For each attribute, we show annotators 100 images from the final automatically-labeled positive set and 100 images from the final negative set using the same interface used to collect the dataset. Fig. 6 shows worker agreement on the positive set as a proxy for precision. The mean precision across all attributes is 90.4%, where precision is the number of positive images where at least one annotator indicates the image should be positive. These checks are in addition to our MTurk quality checks: we only use human labels where two workers agree and we only accept work from turkers with a high reputation who have completed 10,000 tasks at 95% acceptance.

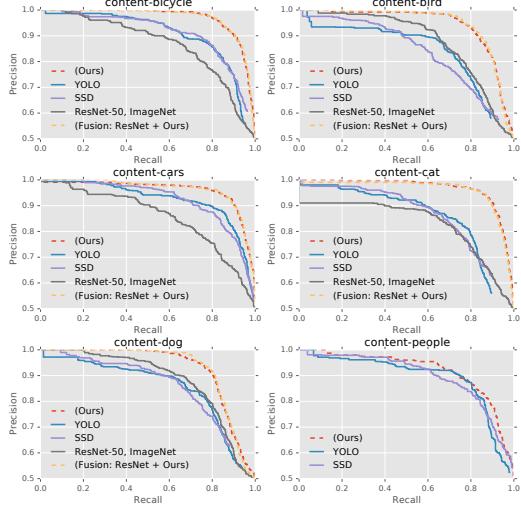


Figure 7. PR curves for object categories comparing our model, YOLO, SSD, ResNet-50, and fusion of ours and ResNet-50.

## 4. Experiments

We can use Behance Artistic Media to study recognition across artistic domains as well as aesthetics and style. First, we investigate the *representation gap* between objects that appear in everyday photographs and objects that appear in artwork. We find that ordinary object detectors do not adequately recognize artistic depictions of objects, showing that there is room for improvement. The existence of this gap leads us to explore the relationship between object representations as rendered across different artistic media. We pose this as a *domain transfer* problem and measure the extent to which knowledge about objects in one medium can apply to objects in an unseen medium. In addition to objects, we briefly consider *style and aesthetics* by comparing different features on emotion/media classification and using our style labels to improve aesthetic prediction tasks on other art datasets. Finally, we conclude with an experiment of *learning feature spaces (feature disentangling)* to build a task-specific search engine that can search for images according to their content, emotion, or media similarity.

### 4.1. Bridging the representation gap

**Detecting objects in artwork.** How different are objects in everyday photographs compared to the stylized objects found in our dataset? We expect that existing pre-trained object detectors might not recognize objects in artwork because existing object detectors trained on ImageNet or VOC are only exposed to a very narrow breadth of object representations. Objects in photographs are constrained by their real-world appearance.

To investigate the representation gap between our dataset and everyday photographs, we consider 6 content attributes that correspond to Pascal VOC categories: **Bicycle**, **Bird**,

	AP	Ours	Yolo	SSD	RN50	Fusion
Bicycle	<b>0.9703</b>	0.9008	0.9116	0.8702	— <sup>1</sup>	0.9704
People	<b>0.9103</b>	0.8863	0.8952	— <sup>1</sup>	—	—
Bird	<b>0.9400</b>	0.8516	0.8387	0.8768	0.9453	—
Cat	<b>0.9660</b>	0.8583	0.8620	0.8026	0.9501	—
Cars	<b>0.9551</b>	0.9140	0.9194	0.8519	0.9628	—
Dog	<b>0.9272</b>	0.8510	0.8582	0.8818	0.9293	—
Average	<b>0.9448</b>	0.8770	0.8801	0.8567	0.9512	—

Table 2. Average precision across different VOC categories using our model, YOLO, SSD, ResNet-50, and fusion of ours and ResNet-50. <sup>1</sup>: We do not report people results because there are relatively few ImageNet people categories.

**Cars, Cat, Dog, People.** We then extract scores for these attributes using two object detectors trained on VOC: YOLO [29] and SSD [19]. For the sake of comparison, we use these detectors as binary object classifiers by using the object of interest’s highest-scoring region from the detector output. We also compare to ResNet-50 classifiers [10] trained on ImageNet, taking the maximum dimension of the ImageNet synsets that correspond with the category of interest. In this way, we can measure how well existing object detectors and classifiers already find objects in art without extra training. We also compare to our final attribute classifier trained in Sec. 3.1, the fine-tuned ResNet-50 that was used to automatically label the final dataset.

We evaluate these methods on 1,000 positives and 1,000 negatives on each attribute’s *human-labeled* validation set to avoid potential bias from the automatic labeler. The results are shown as precision/recall curves in Fig. 7 and AP is shown in Tab. 2. Vision systems trained on photography datasets like VOC (YOLO, SSD) and ImageNet (RN50) perform worse than vision systems that saw objects in artwork during training. From manual inspection, most false negatives of these systems involve objects rendered with unique artistic styles. Specific failure cases are shown in Fig. 1. We can improve performance slightly by fusing ImageNet and Behance scores together with a simple linear combination. The resulting “Fusion” model performs slightly better than our own model and ResNet-50 on all but two attributes. These results show that in terms of object recognition, there is a *representational gap* between photography and artwork.

**Object representation across artistic media.** The existence of this representational gap leads us to question how objects are represented across different artistic media. How well do models trained on one medium generalize to unseen media, and which media are most similar? We can answer these questions within the context of domain adaptation, which has been extensively studied in the vision literature [3, 24]. A good model should know that although cats rendered in drawings are more “cartoony” and abstract than the realistic cats seen in oil paint and ImageNet, they both contain the same “cat” semantic concept, even though the

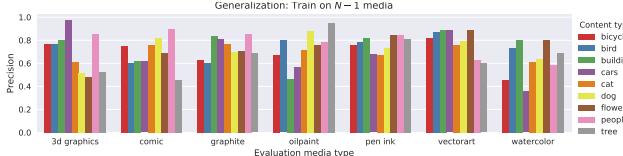


Figure 8. Object recognition performance on an unseen domain. We evaluate an object classifier on the given artistic medium while training it on the 6 other media.

context may vary.

We retrieve the 15,000 images that maximize  $\sigma(x_{i,c})\sigma(x_{i,m})$  for every pair of content and media labels  $(c, m)$ , where  $\sigma$  is the sigmoid function and  $x_i$  is that image’s label confidence scores. We set aside 1/11th of these as the validation set. Note that this validation set is a strict subset of the validation set used to train the automatic labeler. We then fine-tune a pre-trained ResNet for one epoch. The last layer is a 9-way softmax.

In the first set of experiments, we measure an object classifier’s ability to generalize to an unseen domain by learning the representation styles across the other 6 media types and evaluating on only the 7th media type. Results are summarized on the last row of Tab. 3 and broken down by object categories in Fig. 8. Generally, objects that are iconic and easily recognizable within each medium have the highest performance (for example, 3D+cars, watercolor+flowers, graphite+people), but objects that are unlikely to be drawn consistently within each style have the worst generalization performance (watercolor cars/bicycles, 3D flowers). Even though the frequency was controlled by sampling a constant number of images for every  $(object, medium)$  pair, this could be because the artist is less familiar with uncommon objects in their medium and has more individual leeway in their portrayal choices.

These experiments reveal how well classifiers can generalize to unseen domains, but they do not reveal the correlations in object style between different media types. To capture this, our second set of experiments trains an object classifier on only a single media type and evaluates performance on a second media type. As an additional photography medium, we also retrieve 15,000 images for each object from its corresponding ImageNet synset. Average object classification accuracy is shown in Tab. 3. The  $N - 1$  baseline model is trained on all other types. This metric gives a rudimentary comparison of the similarity between artistic media; for instance, comic, graphite, and pen ink are similar to each other, as are oil paint and watercolor. In addition to the gap between ImageNet and Behance (compare last two rows), these results illustrate the gap between each meium’s stylistic depictions. Our dataset can be used to explore these relationships and other similar domain adaptation problems.

Train \ Test	3D	3d	comic	graphite	oil paint	pen ink	vectorart	watercolor	ImageNet
Train	3D	81	37	54	56	45	59	45	66
Comic	66	77	67	60	75	78	54	62	
Graphite	62	46	77	58	65	48	47	62	
Oil paint	48	38	42	82	36	54	60	68	
Pen ink	55	56	67	54	79	58	51	63	
Vector art	65	56	44	48	55	86	41	51	
Watercolor	51	47	60	76	58	56	68	65	
ImageNet	52	32	50	60	43	48	48	82	
<i>N - 1 Baseline</i>	70	69	73	73	77	79	63	73	

Table 3. Domain transfer from one medium to another. This is mean object recognition performance when trained on a single artistic medium (row) and evaluated on a single medium (col). The “ $N - 1$  Baseline” model was trained on the other artistic media types.

## 4.2. Style and aesthetics

Turning away from object categories for a moment, we now consider tasks related to stylistic information using the emotion and media labels in our dataset. We first investigate the effectiveness of different pre-trained features on emotion and media classification, and then show how to improve aesthetic and style classifiers on other artistic datasets.

**Feature comparison.** How well can object recognition models transfer to emotion and media classification? Do models fine-tuned for style tasks forget their object recognition capabilities? To find out, we compare a linear SVM trained on pre-trained ResNet features to two style prediction models: a linear SVM trained on StyleNet features [6] and a StyleNet fine-tuned on Behance Artistic Media. The original StyleNet model was a GoogLeNet that was trained for a style prediction task. We hypothesize that it may outperform ResNet on tasks related to emotion and media classification.

We evaluate these models on held-out human labels for each attribute. Performance for six attributes is shown in Fig. 9. For all four emotion attributes and 4/6 media attributes, the AP of linear classifiers on StyleNet features outperformed ImageNet-derived features. However, ImageNet-derived features have higher AP than StyleNet features on all nine content attributes. Different features are useful for content tasks compared to emotion/media tasks, and our dataset can help uncover these effects.

**Aesthetic classification on other datasets.** Other artistic datasets such as Wikipaintings and AVA contain photographic style annotations. How well do models trained on our dataset perform on these datasets? We show that automatic labels from Behance Artistic Media can slightly improve style classification on existing datasets. We evaluate on the three datasets introduced in [13]: 80,000 images in 20 photographic styles on Flickr, 85,000 images from the top 25 styles on Wikipaintings, and the 14,000 images with 14 photographic styles from the hand-labeled set of AVA [21]. For comparison to previous work [6], we report AVA classification accuracy calculated only on the 12,000

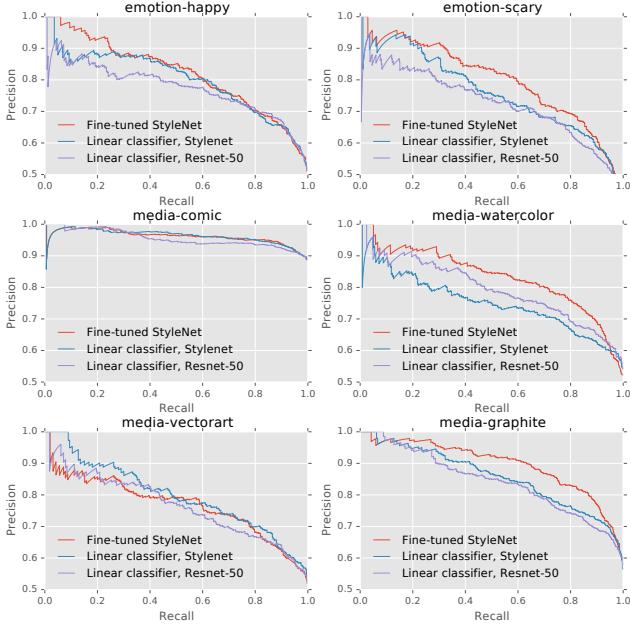


Figure 9. Performance of different features on style attributes.

images that have a single style label.

To solve this task, we train a joint attribute model (*JAM*) that outputs all attribute scores simultaneously. Each training sample  $(x, i, \ell)$  is a tuple of image  $x$ , attribute index  $i$ , and label  $\ell \in \{-1, 1\}$ . It is not suitable to train this model using ordinary cross entropy because each attribute is not mutually exclusive. Thus, we must use a loss function with two properties: each attribute output should be independent of other attributes and unknown attribute values should not induce any gradient. We lift image  $x$  to a 20-dimensional partial attribute vector  $\hat{y} \in \mathcal{R}^{20}$ , where  $\hat{y}_{j \neq i} = 0$  and  $\hat{y}_{j=i} = \ell$ . This allows us to train using a soft-margin criterion,  $loss(x, y) = \frac{1}{20} \sum_i \log(1 + \exp(-\hat{y}_i y_i))$ . Our JAM model is a fine-tuned ResNet-50 model with a linear projection from 1,000 to 20 dimensions. We trained our model for 100 epochs, starting with a learning rate of 0.1 and multiplying it by 0.93 every epoch. The training set includes roughly 2 million images evenly sampled between attributes and evenly distributed between positive and negative images drawn from the automatically-labeled images in Behance Artistic Media.

Results are shown on Table 4. On all three challenges, our model shows improved results compared to both the original ResNet-50 and StyleNet. This shows that Behance imagery is rich and diverse enough to improve style recognition tasks on other datasets. This is particularly interesting because Flickr and AVA are both focused on photographic style. Categories in AVA are chosen to be useful for aesthetic quality prediction tasks. This shows that models can train on our dataset to improve performance on other aesthetic classification datasets.

	JAM	ResNet-50 (ImageNet)	StyleNet [6]
Flickr	<b>0.389</b>	0.376	0.372
Wikipaintings	<b>0.508</b>	0.505	0.414
AVA	<b>0.615</b>	0.603	0.560

Table 4. Performance of our joint model for style detection on other datasets

### 4.3. Visual subspace learning

Finally, we conclude by showing how to learn task-specific subspaces to retrieve images according to content, emotion, or media similarity. One disadvantage of the joint attribute model mentioned above is the lack of separate feature spaces for each task. Consider the case of image retrieval where the goal is to find images that share the content of a query image but not necessarily its artistic medium. We can use Behance Artistic Media to solve this task by treating it as a visual subspace learning problem. Starting from a pre-trained ResNet shared representation, we remove the top layer and add three branches for content, emotion, and media. Each branch contains a linear projection down to a 64-dimensional subspace and a final projection down to label space. The final model is trained similarly to the model in Sec. 4.2. Only the initial ResNet weights are shared; the embedding is separate for each task. We qualitatively show three images close to the query within each task-specific embedding.

The results show that this simple strategy can learn sensible task-specific embeddings. Neighbors in latent-content space generally match the content of the query and neighbors in latent-media space generally match the query’s artistic medium. The effect is qualitatively weaker for emotion space, perhaps because of the limited label set. From a human inspection of 100 random queries, the precision-at-10 for content, media, and emotion is 0.71, 0.91, and 0.84 respectively. Media and emotion precision-at-10 are slightly improved compared to our shared feature baseline of 0.80, 0.87, 0.80, which could be explained if the shared representation focuses almost exclusively on content. One limitation of this approach is that without any conditioning, the three learned subspaces tend to be correlated: objects close in media-space or emotion-space sometimes share content similarity. Our dataset could provide a rich resource for feature disentangling research.

### 4.4. Visualizing the learned model

We qualitatively explore the kind of visual style cues learnable from the proposed dataset in Fig. 11. A dataset of 110k images was formed by sorting all 65m Behance Artistic Media images by likelihood score for each of the 7 media and 4 emotion attributes, and sampling the top 10k images in each case. Duplicate images selected across attributes were discarded. A modified Alexnet[16] (fc6 layer 1024-D, fc7 layer 256-D) was trained from scratch on the



Figure 10. Retrieval results showing a query image, three content neighbors, and three neighbors from another facet

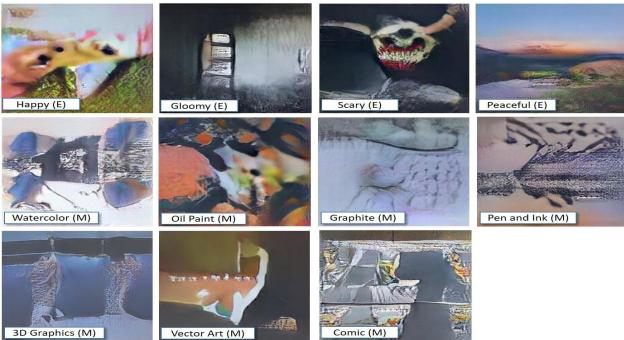


Figure 11. Visualizing learning: maximum activations for each of the media and emotion styles.

11 style (media and emotion) attributes for 40 epochs via SGD with learning rate 0.01. Nguyen *et al.* [22] recently proposed a deep generator network (DGN) based visualization technique for synthesizing stimuli preferred by neurons through combination of a truncated (ImageNet trained) CaffeNet and up-convolutional network initialized via white noise. We run the DGM-AM variant of their process for 200 iterations, using a learning rate of 2.0 and weighting factor 99. The images synthesized for several media types (*e.g.* graphite, oil-paint and watercolor paintings) epitomize textures commonly encountered in these art forms although styles exhibiting structural combination of flatter regions are less recognizable. Fragments of objects commonly recognizable within emotion-based styles (*e.g.* teeth for *scary*, bleak windows in *gloomy* or landscapes in *peaceful*) are readily apparent.

## 5. Conclusion

Computer vision systems need not be constrained to the domain of photography. We propose a new dataset, “Behance Artistic Media” (BAM!), a repository of millions of images posted by professional and commercial artists representing a broad snapshot of contemporary artwork. We

collected a rich vocabulary of emotion, media, and content attributes that are visually distinctive and representative of the diversity found in Behance.

However, though Behance does include tag metadata, we showed that these tags are too noisy to learn directly. Further, the scale of Behance makes brute-force crowdsourcing unattractive. To surmount these issues, we collected labels via a hybrid human-in-the-loop system that uses deep learning to amplify human annotation effort while meeting desired quality guarantees.

The resulting dataset is useful for several computer vision tasks. We use it to highlight the representation gap of current object detection systems trained on photography, showing that Behance captures a wider gamut of representation styles than current sets such as VOC and ImageNet. Different artistic media in Behance have unique aesthetics, providing an interesting test bed for domain transfer tasks, and different features prove useful for content tasks compared to media/emotion classification. We also use Behance to improve the performance of style classification on other datasets, showing that researchers can train on our dataset for a marked improvement in performance. Finally, we conclude with a subspace learning task for retrieving images based on their content or artistic media.

We believe our dataset provides a good foundation for further research into the underexplored realm of large-scale artistic imagery.

## Acknowledgments

This work is partly funded by an NSF Graduate Research Fellowship award (NSF DGE-1144153, Author 1), a Google Focused Research award (Author 6), a Facebook equipment donation to Cornell University, and Adobe Research.

All images in this paper were shared with Creative Commons attribution licenses and we thank the creators for sharing them: -jose-, AARON777, AbrahamCruz, AndreaRenosto, AndrewLili, AngelDecuir, AnnaOfsa, Ari-anne, AstridCarolina, Aubele, ChumaMartin, DannyWu, DubrocDesign, Elsimova, EtienneBas, ExoticMedia, Folken-sio, GinoBruni, GloriaFl, HattyEberlein, Hectorag, ImaginaryFS, IsmaelCG, J\_Alfredo, JaclynSovern, JanayFra-zier, JaneJake, JodiChoo, Joshua\_Ernesto, KimmoGrab-herr, Kiyorox, Marielazuniga, Mirania, NIELL, NikolayDimchev, NourNouralla, OlivierPi, Ooldigital, Pleach-flies, PranatiKhannaDesign, PraveenTrivedi, Radiotec1000, RamshaShakeel, RaoulVega, Roxxo2, Royznx, RozMogani, SaanaHellsten, Sanskruta, Slavina\_InkA, Swann-Smith, TCaetano, THE-NEW, Thiagoakira, TomGrillo, WayneMiller, Yipori, ZaraBuyukliiska, adovillustration, alobuloe61b, amos, animatorcreator, answijnberg, anthonyrquigley, antoniovizzano, apedosmil, artofmy, asariotaki, asjabor, assaadawad, atanuschopski, averysauer, ballan, bartosz, beissola, bilbo3d, bluelaky, brainless-

rider, brandplus, brittany-naundorff, bucz, campovisual, candlerenglish, carlosmessias, carlosoporas, castilloseas, cghostgun, charbelvanille, charlottestrawbridge, chin2off, cknara, coyote-spark, crazydiamondstudio, cristianbw, crivelaro, cube74, customshoot, daconde, dafrankish, danielstarrason, dariomaggiore, deboralstewart, dedos, designrahul, digital-infusion, ememonkeybrain, esanaukkarinen, evanwitek, fabsta, flochau, florencia\_agra, foundationrugs, gatomontes, giuseppchristiano, gumismydrug, gurudesigner, heroesdesign, higoszi, ifyouprefer, ikuchev, illustrationdesign, irf, irwingb, ivancalovic, jabaarte, jakobjames, janaramos, janeschmittbarra, jdana, jota6six, juliencoquentin, juliensps, justindruss, kalalinks, karolienssoete, katesimpressions, katiemott, kpkeane, larawillson, leovelasco, leturk, lindseyrachael, lisa\_smirnova, littleriten, manf\_v, margotsztajman, martabellvehi, matildedigmann, maviemamuse, mechiparra, michaelkennedy, monroyilustrador, mvtacca, nataliamon, nati24k, nazpa, nickelcurry, nicopregelj, nobleone, nspanos, nunopereira, olaolaolao, olivierventura, ollynicholson, omarmongepinal, onthebrinkartstudios, osamaaboezz, pamelapg, paperbenchdesign, parfenov, pengyangjun, pilarcorreia, pilot4ik, polyesterdress, priches, pumaroo, ragamuffin, rainingsuppie, raquelirene, rasgaltwc, rebeccaahan, redric, rikishi, roosariski, rrabbstyjke, rufusmediateam, sairarehman, sakettattoocrew1, samii69, sandhop, santarromana, santocarone, sashaanikusko, scottchandler, sebagresti, sehriyar, shaneferns, shaungmakeupartist, siqness, snowballstudio, softpill, sompson, sophieecheetham, starkdesigns, studiotuesday, tadzik, tatianaAdz, taylorboren, taylored, tensil, theCreativeBarn, tomtom\_ungor, toniitonen, travisany, tweek, ukkrid, visual3sesenta, zarifeozcalik, zmahmoud, zwetkow

## References

- [1] D. Borth, R. Ji, T. Chen, T. Breuel, and S.-F. Chang. Large-scale visual sentiment ontology and detectors using adjective noun pairs. In *Proc. MM*, 2013.
- [2] E. J. Crowley and A. Zisserman. In search of art. In *Workshop on Computer Vision for Art Analysis, ECCV*, 2014.
- [3] G. Csurka. Domain adaptation for visual applications: A comprehensive survey. *CoRR*, abs/1702.05374, 2017.
- [4] Y. Cui, F. Zhou, Y. Lin, and S. J. Belongie. Fine-grained categorization and dataset bootstrapping using deep metric learning with humans in the loop. In *Proc. CVPR*, 2016.
- [5] S. Dhar, V. Ordonez, and T. L. Berg. High level describable attributes for predicting aesthetics and interestingness. In *CVPR*, 2011.
- [6] C. Fang, H. Jin, J. Yang, and Z. L. Lin. Collaborative feature learning from social media. *CoRR*, abs/1502.01423, 2015.
- [7] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing objects by their attributes. In *CVPR Workshops 2009*, pages 1778–1785, 2009.
- [8] L. A. Gatys, A. S. Ecker, and M. Bethge. Image Style Transfer Using Convolutional Neural Networks. In *Proc. CVPR*, 2016.
- [9] S. Ginosar, D. Haas, T. Brown, and J. Malik. Detecting people in cubist art. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 8925, pages 101–116, 2015.
- [10] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proc. CVPR*, 2016.
- [11] H. Izadinia, B. C. Russell, A. Farhadi, M. D. Hoffman, and A. Hertzmann. Deep Classifiers from Image Tags in the Wild. In *Proc. Multimedia COMMONS*, 2015.
- [12] B. Jou, T. Chen, N. Pappas, M. Redi, M. Topkara, and S.-F. Chang. Visual Affect Around the World: A Large-scale Multilingual Visual Sentiment Ontology. In *Proc. MM*, pages 159–168, 2015.
- [13] S. Karayev, M. Trentacoste, H. Han, A. Agarwala, T. Darrell, A. Hertzmann, and H. Winnemöller. Recognizing image style. In *Proc. BMVC*, 2014.
- [14] I. Krasin, T. Duerig, N. Alldrin, A. Veit, S. Abu-El-Haija, S. Belongie, D. Cai, Z. Feng, V. Ferrari, V. Gomes, A. Gupta, D. Narayanan, C. Sun, G. Chechik, and K. Murphy. Open-images: A public dataset for large-scale multi-label and multi-class image classification. *Dataset available from https://github.com/openimages*, 2016.
- [15] J. Krause, B. Sapp, A. Howard, H. Zhou, A. Toshev, T. Duerig, J. Philbin, and L. Fei-Fei. The unreasonable effectiveness of noisy data for fine-grained recognition. In *Proc. ECCV*, 2016.
- [16] A. Krizhevsky, I. Sutskever, and G. Hinton. Imagenet classification with deep convolutional neural networks. In *Proc. NIPS*, 2012.
- [17] T. Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: Common objects in context. In *Proc. ECCV*, pages 740–755, 2014.
- [18] T.-Y. Lin and S. Maji. Visualizing and understanding deep texture representations. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.
- [19] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, and A. C. Berg. SSD: Single Shot MultiBox Detector. In *Proc. ECCV*, 2016.
- [20] I. Misra, C. L. Zitnick, M. Mitchell, and R. Girshick. Seeing through the Human Reporting Bias: Visual Classifiers from Noisy Human-Centric Labels. In *CVPR*, 2016.
- [21] N. Murray, L. Marchesotti, and F. Perronnin. AVA: A large-scale database for aesthetic visual analysis. In *Proc. CVPR*, 2012.
- [22] A. Nguyen, A. Dosovitskiy, J. Yosinski, T. Brow, and J. Clune. Synthesising the preferred inputs for neurons in neural networks via deep generator networks. In *Proc. NIPS*. IEEE, 2016.
- [23] P. Obrador, M. A. Saad, P. Suryanarayan, and N. Oliver. Towards category-based aesthetic models of photographs. In *Proc. MMM*, 2012.
- [24] V. M. Patel, R. Gopalan, R. Li, and R. Chellappa. Visual domain adaptation: A survey of recent advances. *IEEE Signal Processing Magazine*, 32(3):53–69, May 2015.
- [25] G. Patterson and J. Hays. Sun attribute database: Discovering, annotating, and recognizing scene attributes. In *Proc. CVPR*, 2012.

- [26] K.-C. Peng, A. Sadovnik, A. Gallagher, and T. Chen. Where do emotions come from? predicting the emotion stimuli map. In *Proc. ICIP*, 2016.
- [27] R. Plutchik. The nature of emotions: Human emotions have deep evolutionary roots. *American Scientist*, 89(4):344–350, 2001.
- [28] N. Rasiwasia, P. J. Moreno, and N. Vasconcelos. Bridging the gap: Query by semantic example. *IEEE Trans. Multimedia*, 9:923–938, 2007.
- [29] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You Only Look Once: Unified, Real-Time Object Detection. In *CVPR 2016*, pages 779–788, 2016.
- [30] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. E. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Proc. CVPR*, 2016.
- [31] Q. You, J. Luo, H. Jin, and J. Yang. Building a large scale dataset for image emotion recognition: The fine print and the benchmark. *CoRR*, abs/1605.02677, 2016.
- [32] F. Yu, Y. Zhang, S. Song, A. Seff, and J. Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop, 2015. arXiv:1506.03365.