Sketching with Style: Visual Search with Sketches and Aesthetic Context

John Collomosse[1,2] Tu Bui[1] Michael Wilber[3] Chen Fang[2] Hailin Jin[2]

[1]CVSSP, University of Surrey [2]Adobe Research [3]Cornell Tech

Abstract

We propose a novel measure of visual similarity for image retrieval that incorporates both structural and aesthetic (style) constraints. Our algorithm accepts a query as sketched shape, and a set of one or more contextual images specifying the desired visual aesthetic. A triplet network is used to learn a feature embedding capable of measuring style similarity independent of structure, delivering significant gains over previous networks for style discrimination. We incorporate this model within a hierarchical triplet network to unify and learn a joint space from two discriminatively trained streams for style and structure. We demonstrate that this space enables, for the first time, style-constrained sketch search over a diverse domain of digital artwork comprising graphics, paintings and drawings. We also briefly explore alternative query modalities.

1. Introduction

Determining user intent from a visual search query remains an open challenge, particularly within Sketch based Image Retrieval (SBIR) where free-hand sketched queries often present ambiguous or incomplete descriptions of the desired visual content [8]. Traditionally, SBIR literature considers images to be similar if they contain objects of similar shape (*e.g.* fine-grained [35, 39, 29]) or semantics (*e.g.* category retrieval [9, 13, 25]). However, such definitions do not scale well to larger datasets, where a sketched shape can closely match a diverse range of content. Additional visual modalities have been explored within the sketch, such as color [31, 3], explicit labeling of sketched parts [6, 14], and motion (for video) [7, 1, 15] to better constrain the query and so improve the relevance of results.

This paper proposes a novel definition of visual similarity for SBIR, in which the sketched query is constrained using one or more secondary (contextual) images to specify the desired aesthetic of the results. For example, a sketch of a dog accompanied by a contextual set of watercolor paintings, or scary images, would yield watercolor paintings of dogs, or images of scary dogs, respectively. Importantly, we do not require the contextual images to contain the desired subject matter (*e.g.* dogs). Rather, we seek to disentangle the notions of content (structure) and aesthetics (style) enabling independent specification of both within the query. Visual style remains a highly challenging and understudied area of Computer Vision. Our exploration of style as a modality for visual search complements recent advances *e.g.* in style transfer [10] enabled by deep learning.

Constraining visual search to match a user's intended 'look and feel' is a promising novel direction for enhancing search relevance, particularly over aesthetically diverse imagery. Our work leverages a recent large-scale dataset of contemporary artwork covering a breadth of styles, media and emotions (BAM [37]) from which we learn a model of aesthetic style. Concretely, we propose a hierarchical triplet convolutional neural network (convnet) architecture to learn a low-dimensional joint embedding for structure and style. Each branch of this network unifies complementary information on scene structure and aesthetics derived from two discriminatively trained convnet streams, which are themselves of triplet architecture.

Our technical contributions are three-fold. First, we propose a triplet convnet to learn an embedding for aesthetic style, showing this novel model to outperform by a large margin, previous attempts to use deep convnets for measuring visual style similarity. Second, we build upon our model, incorporating a state of the art convnet for sketch-photo similarity [4] to develop a hierarchical triplet convnet for learning a joint space for structural and style similarity over a diverse domain of digital artwork (comprising not only photos, but also paintings, 3D renderings, hand-drawn and vector-art drawings, in a variety of media). Third, we demonstrate and evaluate the performance of our model

within a novel SBIR framework that uniquely accepts a set of contextual images alongside the sketched query shape, enabling stylistic constraint of the visual search. Although our study is scoped primarily to leverage sketches as means for specifying structure in queries, we also experiment with alternative modalities such as artwork and text.

2. Related Work

Visual style remains a sparsely researched topic that has received greatest attention from the synthesis perspective *e.g.* style transfer through learning visual analogies [12] or more recently deep representations [10]. The Gram matrix computed across layers of a pre-trained model (*e.g.* VGG-16 on ImageNet) has been shown to abstract content from style in diverse imagery, and has been exploited for texture description [20]. Deep convnets have also been shown effective at classifying artwork style [16]. Although targetting photographs rather than general artwork, aesthetic classification and rating of images has also been explored via attributes such as depth of field and exposure [23, 21]. More generally, attributes [17, 27] (including relative attributes [18]) have been used to assist visual search. However attribute models often require explicit definition of a fixed set of attribute categories, or pre-training for their detection.

This paper explores effective representations for leveraging style as a constraint in a visual search. Rather than attempting to classify aesthetic attributes, we develop a definition of visual similarity that disentangles image structure and style, enabling independent specification of each in a visual search query. We leverage SBIR as a platform for this study. Sketches primarily describe structure, and we propose such descriptions be augmented with one or more images exemplifying the desired visual style of results. Effective methods for augmenting shape description in SBIR are becoming urgently needed to resolve query ambiguity as SBIR matures to larger datasets. Our use of convnets to unify structure and style complements recent deep approaches for shape matching in SBIR [39, 29, 2, 4, 5] which have been shown to outperform shallow-learning models [33, 9, 13, 28]. Contrastive loss networks have been used to map sketches to photo edge-maps or rendered 2D views of 3D models [25, 34]. Triplet networks have also been leveraged for both fine-grained [39, 29] and category-level SBIR [4]. Convnets were fused with a complex pipeline incorporating object proposals, query expansion and re-ranking for retrieval [2]. All these methods address SBIR via cross-domain learning; the gap between sketch and photos is bridged via regression explicitly seeking to discard appearance properties. Our goal is not proposing another shape matching technique. Rather, we incorporate a leading model [4] to explore aesthetic constraint of SBIR.

Relevance feedback (RF) is often used disambiguate user intent particularly when multiple modalities exist in a query [14]. RF requires user interaction to iteratively refine search results, learning a per-query re-weighting of the feature space. Classically this is learned via linear classifier [30], or recently, online learning of shallow convnets using positive and negative images [36]. Our work also explores re-weighting using convnets, but pre-learns a corpus-wide model for combining structural and style modalities. Unlike RF, we do not iteratively request feedback from the user.

3. Methodology

We describe the proposed network architecture and training methodology for learning a feature embedding for visual search with aesthetic context.

3.1. Behance Artistic Media (BAM) Dataset

Our work leverages *BAM*; a dataset of $\sim$65 million contemporary artworks from `https://behance.net` [37] annotated using a large-scale active learning pipeline [38]. The annotations in BAM include semantic categories (bicycle, bird, cars, cat, dog, flower, people, tree); seven labels for a wide breadth of different artistic media (3D renderings, comics, pencil/graphite sketches, pen ink, oil paintings, vector art, watercolor); four emotion labels of images likely to induce certain emotions in the viewer (happy, gloomy, peaceful, scary); and short textual captions for a small subset of images. The dataset's semi-automated label annotations come in the form of likelihood scores which may be thresholded at desired quality targets to control the trade-off between dataset size and label precision. In our work, we use images labeled at a precision of 90%. We adopt BAM due to the high diversity of artistic content spanning drawings, paintings, graphics and vector art in contemporary and classic styles. In contrast, AVA [23] comprises largely photographic content proposed for aesthetic attribute mining.

3.2. Hierarchical Network Architecture

Triplet networks are commonly applied to learn low-dimensional feature embeddings from data distributions, and have recently been applied to photo-based object instance retrieval [11, 26]. Our proposed architecture is also of triplet design, each branch unifying two discriminatively trained network streams that independently learn a feature embedding for image *structure* and *style* respectively (Fig. 2). Furthermore, the network stream for each modality has, itself, a triplet sub-structure. The architecture and training of the style stream is given in Fig. 1 and described in detail within Sec. 3.2.1. The *structure branch is not a contribution of this paper*, and reproduces a state of the art model [4] for shape retrieval 3.2.2. We describe how the overall triplet model integrates these streams in Sec. 3.2.3, to learn a joint feature embedding within which we measure visual similarity for search (Sec. 3.3).

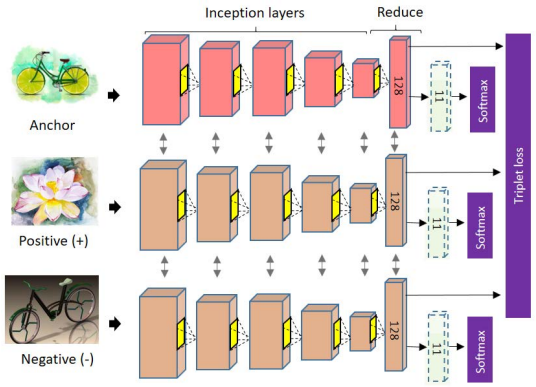

Figure 1. The triplet convnet for learning the style model comprises three fully shared (siamese) branches formed from GoogleNet with a bottleneck layer appended to pool5. Training proceeds via classification loss, followed by triplet loss guided by hard negatives exhibiting shared semantics and differing style.

3.2.1 Style Network

We first describe the model forming the style stream within each branch of our hierarchical triplet architecture. The network comprises three branches, each augmenting GoogLeNet [32] through addition of a 128-D inner-product layer to serve as a bottleneck after *pool5* layer and prior to drop-out. The bottleneck is later shown to be performance critical both for style classification and for search (c.f. Sec. 4) due to the increased sparsity of *pool5* features when training on diverse artwork rather than photos. Fig. 1 illustrates the fully-shared (siamese) configuration of the branches.

The model is trained from scratch, independent of the wider network, using an 88k artwork training set (Behance-Net-TT, Sec. 4.1) evenly partitioned into 11 style categories (S), each balanced across the 8 semantic categories (Z). Training proceeded initially via classification loss (soft-max loss, 30 epochs) and then by refinement under triplet loss (50 epochs). Triplets were formed using a randomly selected anchor image $a = (s \in S, z \in Z)$, a randomly selected hard positive image $p = (s, z' \in Z \setminus z)$ and a hard negative image $n = (s' \in S \setminus s, z)$. The network describes a function $f(.)$ minimising:

$$\mathcal{L}(a,p,n) = \left[m + |f(a) - f(p)|^2 - |f(a) - f(n)|^2\right]_+ \quad (1)$$

where $m = 0.2$ is a margin promoting convergence, and $[x]_+$ indicates the non-negative part of x. Triplet refinement improves decorrelation between semantics and style (Fig. 3), discouraging learned correlations with objects (*e.g.* trees $\rightarrow$ peaceful, skulls $\rightarrow$ scary scenes). This refinement is later shown to yield significant performance gain (Sec. 4.2).

3.2.2 Structure Network

The triplet model of Bui *et al.* [4], fine-tuned over BAM, comprises the structure stream. The network incorporates an anchor branch accepting a sketch query and positive/negative branches that accept a photographic image as input. The image branch closely resembles AlexNet [19], and the sketch branch Sketch-A-Net (a short-form AlexNet optimal for sketch recognition [40]). The network learns a joint embedding from exemplar triplets comprising query sketches, positive photos that match those sketches, and negative photos that do not. We fix the output layer of the network to 128-D and inhibit sharing of network weights across branches *i.e.* training yields separate functions for embedding the sketch $g_s(.)$ and for the image $g_i(.)$ content. These functions are embedded within our larger network (Fig. 2).

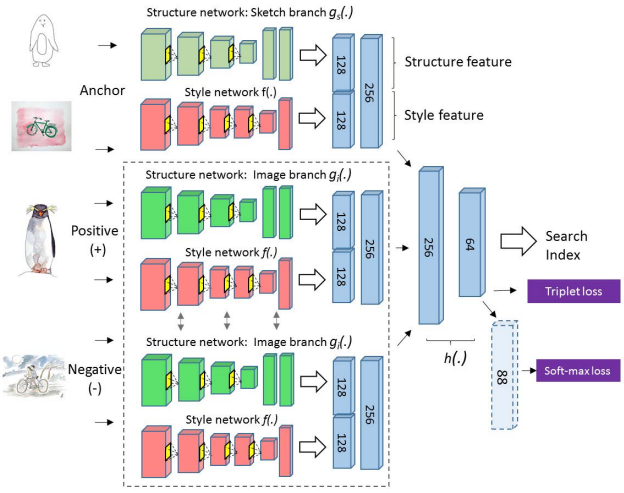

Figure 2. Hierarchical triplet convnet combining vectors from the style (Fig. 1) and structure [4] streams. Joint embedding of the two modalities is learned (Sec. 3.2.3) from the concatenated features initially via classification loss, followed by hard negative mining resulting in a 64-D descriptor for indexing.

Training follows the four-stage training process outlined in [4]. The process utilises the TU-Berlin sketch dataset (for the anchor) augmented with social media sourced photographs (for the positive/negative pair). The final step recommends fine-tuning the network using triplets sampled from representative imagery; we use random artwork images sampled from BAM with 480 TU-Berlin sketches having category overlap.

3.2.3 Hierarchical (Multi-modal) network

The outputs of the structure and style models are normalised and concatenated to form a 256-D input vector, forming the structure of each branch of the larger triplet network (Fig. 2). Note that the anchor branch integrates $g_s(.)$ and the positive/negative branches $g_i(.)$. The branches feed forward to two final inner product layers of 256-D and 64-D separated by ReLU activation, which learn projection $h(.)$ over the two 128-D subspaces for visual search.

A careful training protocol is required to ensure conver-

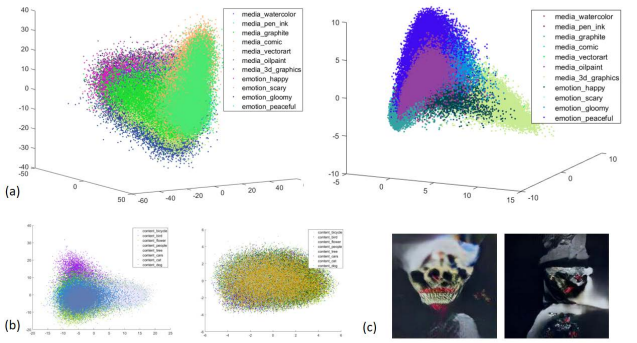

Figure 3. PCA visualizations showing discrimination and decorrelation in the style (a) and semantic (b) spaces, before (left) and after (right) triplet refinement. (c) Activation maximization for 'Scary' category using method of [24] yields interesting qualitative insight into learned visual attributes for style; objects (teeth and skull) disappear in the decorrelated space but strong color cues remain.

gence in these final layers. Since triplet loss is a loose regularization (loss is computed on the relative distance between the anchor-positive and anchor-negative pairs) we have found it more effective to initialise training with a stricter regularization (soft-max loss). We initially train the network with an additional classification layer that recognises each of the 8×11 semantic-style combinations in the dataset. Training proceeds minimising the hybrid loss:

$$\mathcal{L}'(a,p,n) = \sum_{i \in \{a,p,n\}} \phi_s \mathcal{S}(i) + \phi_t \mathcal{L}(a,p,n), \qquad (2)$$

where weight type (ϕ_s, ϕ_t) is set $(1.0, 0.0)$ initially, relaxed to $(0.2, 0.8)$ after the initial 2000 epochs.

Training proceeds by passing a sketched query and a homogeneous style set of artwork (randomly varying between 1-10 images) to the structure $g_s(.)$ and style $f(.)$ arms of the anchor branch, yielding a 256-D query vector. The output of the style stream is averaged over all images in the style set. The positive and negative vectors are formed via $g_i(.)$ and $f(.)$ each using a single artwork image selected at random. During initial training, triplets are first formed randomly: the positive exemplar an artwork image selected at random from those images sharing the same semantics and style as the anchor, while the negative exemplar differs in either semantic or style or both. In the later training phase (from epoch 1000) , we narrow down the negative list further by choosing the negative sample from top k returned images using the current network weights as a visual search system (Sec. 3.3). Query sketches are subject to random affine perturbation (rotation, scale, translation) during training. Although our network takes raster data as input the TU-Berlin dataset is provided in vector form, enabling the random omission of sketched strokes for further data augmentation. Note that all training is carried out on the Behance-Net-TT and sketch datasets described in Sec. 4.1.

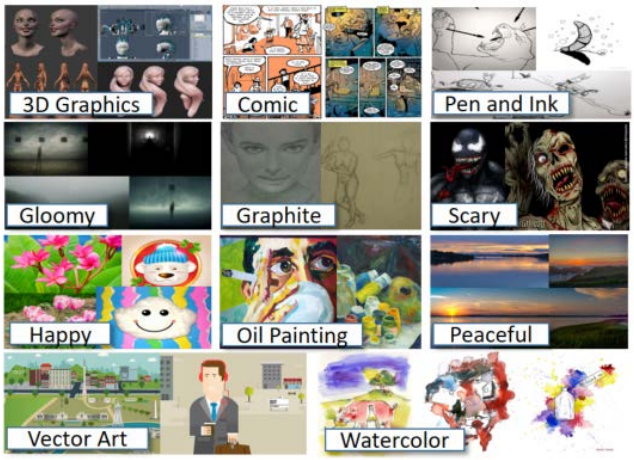

Figure 4. Examples sampled from image sets used to specify style within the SBIR queries evaluated in Sec. 4.3.

3.3. Visual Search with Aesthetic Context

We index a 879k dataset (Behance-VS, Sec. 4.1) of artworks $D = \{d_1, ..., d_n\}$ forward-passing each image d_i via through the hierarchical network to form a 64-D image descriptor d'_i.

$$d'_i = h(\begin{bmatrix} g_i(d_i) & f(d_i) \end{bmatrix}). \qquad (3)$$

Descriptors are stored within a distributed (map-reduce) index hosting multiple kd-trees across several machines. Given query sketch q and set of contextual (style) images $C = \{c_1, ..., c_m\}$ the search descriptor q' is:

$$q' = h(\begin{bmatrix} g_s(q) & \sum_{l=1}^{m} \omega_m f(c_l) \end{bmatrix}). \qquad (4)$$

We perform a k nearest-neighbor look-up ranking results by $|q' - d'_i|$ distance for relevance. Typically the context set describes a common style and so c_i has weight $\omega_i = \frac{1}{m}$, however it is possible to blend styles (Sec. 4.4).

4. Experiments and Discussion

We evaluate the retrieval performance of our style model (Sec. 4.2), and the SBIR+style search that incorporates it (Sec. 4.3). We experiment with style interpolation (Sec. 4.3.2) and alternative modalities for queries (Sec. 4.4).

4.1. Dataset Descriptions

Our experiments make use of BAM [37], and the TU-Berlin [9] dataset of free-hand sketches:

Network Training and Test (Behance-Net-TT, 110k) Taking the outer product of style and a subset of 8 semantic attributes in BAM we form 11×8 sets of artwork images. Attribute scores are thresholded using per-attribute thresholds distributed with the dataset for $p = 0.9$ positive confidence. Images showing positive for both attribute pairs only, are sorted by the sum of their scores. The top 1.25k images are taken in each case yielding a balanced dataset of

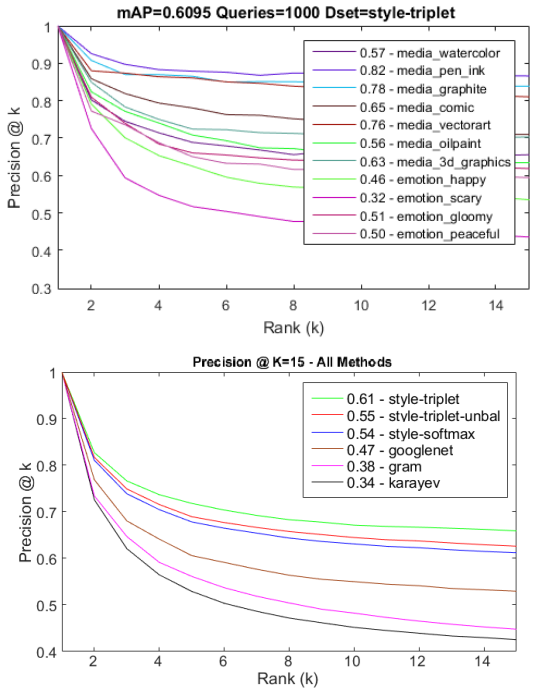

Figure 5. Style retrieval performance over 1000 random queries (Precision @ K=15). Top: best performing style network (style-triplet loss, mAP 61.0%). Bottom: performance comparison for all baselines.

110k images. The dataset is partitioned 80:5:15 into training, validation and test sets, ensuring balanced semantics.

Visual Search Evaluation (Behance-VS, 879k) Subsets of BAM are created for each of the 11 style categories, by thresholding attribute scores at $p = 0.9$. Within each set, a random sample of $90k$ images is selected ensuring balanced semantics *i.e*. ensuring the same count of positive flags are present for each semantic category. Images are excluded if they appear in Behance-Net-TT. Images collated in multiple sets are discarded. This yields a 879k test corpus for visual search. A small subset (10 images $\times 11$ styles) is held out to serve as style sets for the test queries (Fig. 4).

Sketch set (TU-Berlin, 0.5k) 480 sketches from the TU-Berlin dataset [9] (60×8 object categories overlapping with BAM) are used to fine-tune the structural branches of the network (Sec. 3.2.2), and to train the hierarchical network. A queryset of 24 sketches are used to drive the visual search evaluation (Sec. 4.3).

4.2. Evaluating Visual Search: Style only

We first evaluate the accuracy of the style model (Sec. 3.2.1) at retrieving artworks of similar style from the test partition of Behance-Net-TT. A query set of 1000 images are selected at random, with even semantic coverage, from the test partition. Mean Average Precision (mAP) is computed for the test set (Table 1).

Comparing the models at the initial classification (style-softmax) and triplet refinement (style-triplet) training

Method	mAP (%)
style-triplet*	**61.0**
style-triplet-unbal*	55.0
style-softmax*	53.6
GoogLeNet* [32]	46.8
Gram / Gatys *et al*. [10]	38.4
Karayev *et al*. [16]	34.1

Table 1. Style-only retrieval performance (mAP, 1000 queries; * indicates Behance-Net-TT trained methods.

stages, the latter shows a performance gain of 7.31% mAP. A qualitative improvement in style discrimination and semantics decorrelation is also observable when visualizing the spaces before and after triplet refinement (Fig. 3).

We include an early experiment performed using a semantically unbalanced version of Behance-Net-TT (style-triplet-unbal) which yielded poorer results suggesting overfitting to semantic bias (certain objects became expected for certain styles) and further motivating decorrelation of style and semantics. Precision @ K curves are presented for each style for the leading (triplet) model, and for all models in Fig. 5. Emotion retrieval is evidently more challenging than media type, and strong performance is shown in visually distinctive media *e.g*. vector art.

Performance is base-lined against three techniques: Karayev *et al*. [16] who use pre-trained CaffeNet features (DECAF_6) with no fine-tuning; GoogLeNet [32] features trained from scratch over the 11 style categories (pool5); Gatys *et al*. [10] where Gram matrices computed across multiple convolutional layers ($\text{conv1\_1} - \text{conv5\_1}$) of a pre-trained VGG-19 network model, shown in [10] to decorrelate style from content for image stylization. The retrieval mAP (Table 1) and precision @ K=15 curves (Fig. 5) indicate significant performance gains in the learned models, even versus contemporary texture descriptors (Gram [10]). A surprising result was the degree to which addition of the bottleneck in GoogLeNet enhanced retrieval performance (6.8% mAP). The analysis supports incorporation of the triplet refined model within the wider network.

4.3. Evaluating Visual Search: Sketch+Style

Mechanical Turk (MTurk) was used to evaluate retrieval accuracy over the top 15 ranked results returned from 264 search queries (24 sketches $\times 11$ style sets), since no annotation for structure and style relevance was available. Each query required $\sim$ 4k annotations over the top 15 results with three repetitions ensuring that no participant re-annotated the same result. Retrieval was performed using six experimental configurations to search the Behance-VS dataset of 879k artwork images. The configurations explored were: *ss-triplet-64* the proposed combination of sketch+style features re-weighted and projected to 64-D by the learned layers of the multimodal network; *ss-triplet-128* similar network but with 128-D output layer; *ss-concat* a degenerate case in which the 256-D concatenated sketch+style vector

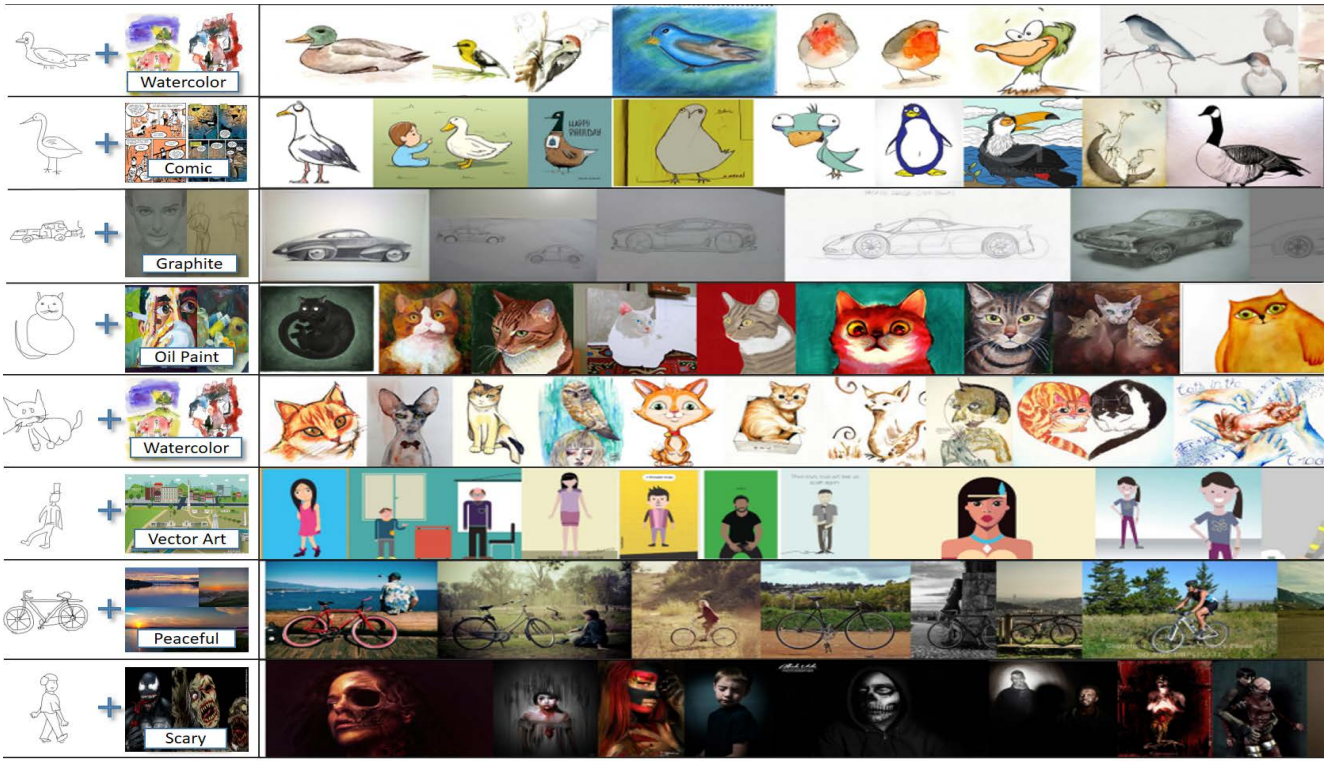

Figure 6. Representative structure+style visual search results. Query comprising sketch and a sample of the 10 image style set (inset left).

Method	Wcol.	Pen	Graphite	Comic	Vector	Oil	3D	Happy	Scary	Gloomy	Peace	mAP
ss-triplet-64	51.2	**48.1**	53.4	**48.1**	**60.1**	55.2	**63.0**	38.9	42.8	**49.7**	**68.6**	**52.7**
ss-triplet-128	**57.7**	46.5	**56.1**	38.6	47.3	**60.5**	56.2	**40.7**	38.7	39.8	52.8	48.6
ss-concat	55.2	40.7	37.6	37.3	57.0	50.6	49.9	39.0	**43.3**	46.5	62.8	47.2
GoogLeNet *et al*. [32]	43.2	31.2	38.2	20.2	36.9	39.1	46.3	32.6	37.2	38.2	64.0	38.8
Gatys *et al*. [10]	24.9	32.3	39.8	29.1	40.7	36.5	8.40	29.9	27.9	39.8	43.7	32.1
Karayev *et al*. [16]	32.3	23.6	34.6	30.6	46.1	23.3	10.3	37.7	27.3	37.4	41.8	31.4

Table 2. Retrieval performance (mAP %)over 24 sketches $\times$ 11 styles; mAP computed to rank 15 using MTurk annotation of search results.

is used directly for search *i.e*. no joint learning; *three baseline style model*s of Sec. 4.2 substituted for our proposed model in the full network. Participants were provided with a tutorial containing examples of each style, and for each query/result asked to indicate both whether the visual style matched and if any of the main objects in the image matched the sketched shape.

4.3.1 Retrieval performance

We compute mAP (Table 2) and precision @ K=15 (Fig.8) of the proposed method to each of the five baselines, breaking down performance per style category. A result is considered relevant if both structure and style match.

The best performing output layer configuration for the hierarchical network is 64-D, showing a performance lead of 4.1% over use of a 128-D layer (offering also a compactness advantage for indexing). Fig. 7 summarises the relative performance of the techniques, and a t-test was run for all configuration pairs. When considering performance across all styles, all proposed methods (*ss-**) outperform existing baselines by a very significant margin ($p<0.01$) and the use of 64-D triplet learning in the hierarchical network outperforms direct concatenation of the structure and style vectors (256-D) by significant margin ($p<0.15$).

All techniques that were explicitly trained on artwork (BAM) outperformed those that were not, and the trend of Table 1 is mirrored in the structure+style retrieval results (Table 2). It is an interesting result that forcing learning to explicitly de-correlate semantics and style during training of the style stream of the network (triplet refinement, Sec 3.2), following by explicit learning of the projection of structure and style together in the hierarchical network, yields a performance boost, so justifying our network design. Assessing structure relevance only (*i.e*. considering only shape responses from MTurk) confirmed little variation (62.5 ± 2.3) across methods for the 24×11 query set. Since this is simply evaluating prior work [4] on Behance-VS we do not

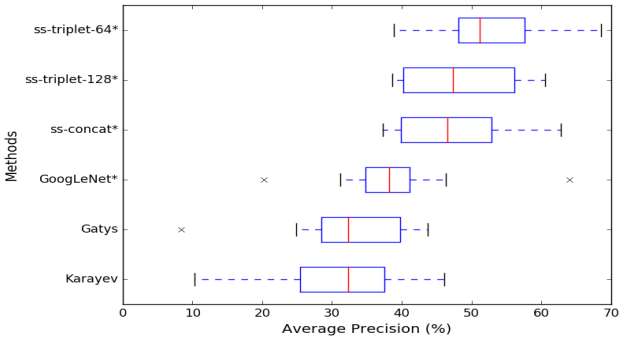

Figure 7. Relative performance of proposed method (*ss-triplet-64*) versus baselines; (*) indicates learned over BAM.

consider this further, beyond controlling for style. Considering style relevance only reveals performances of 64.5%, 63.3% and 58.9% for *ss-triplet-64, ss-triplet-128,* and *ss-concat* respectively against 49.0%, 47.8% and 46.5% for baselines [32], [10], [16] reflecting Table 2. To benchmark average query time over $897k$ images we use a single machine (Xeon E5-2637 3.5Ghz) and achieve $\sim$ 700ms using (*ss-triplet-64*) features.

Figs. 6,11 contain representative results from queries run under the *ss-triplet-64* model. Fig. 12 highlights some incorrect top-15 results. A vector-artwork and a photo are returned for a 3D Graphics query, likely due to their flat backgrounds typical in 3D renderings (in the latter example, content matches shape but not semantics). The action sports shot of the cyclist is arguably not 'peaceful' but the countryside setting is a strong prior. Two photos are misidentified as oil paintings due to flaking paint visible on the wall, and an incomplete oil painting is returned as the final example.

4.3.2 Blending visual styles

We qualitatively explore interpolation of style in the query set (Fig. 9). Three pairs of styles (watercolor-graphite, oilpaint-penink, happy-scary) were combined by varying the weights ω_i (eq.4) on the averaging function using to combine style vectors from images in the context set. For example, a cat was queried using 25% watercolor and 75% graphite as the style context. The results in that case showed a sketched cat with grey watercolor wash in the background, whilst a 50:50 blend showed a half-finished watercolor (many artists upload in-progress work to Behance). We also explore interpolation within a style, here two watercolor context images are used for style: one coarse washes, the other using harsh black outlines depicting fine detail with a background wash. We observe a trend toward the more heavily weighted style (media or emotion) suggesting linear interpolation in this space exert useful control.

4.4. Alternative query modalities

Significant advantages of visual search over text are that no annotation of the search corpus is required, nor any adherence to a tag ontology. Here we contrast visual search against the use of free-text keywords for specifying aesthetic context or structure. A small portion ($\sim 2\%$) of the BAM data is accompanied by free-text crowd annotation, describing content and style in $\sim 3-10$ words. Approximately $4.2k$ images within Behance-VS are accompanied by this annotation, enabling text-based experiments.

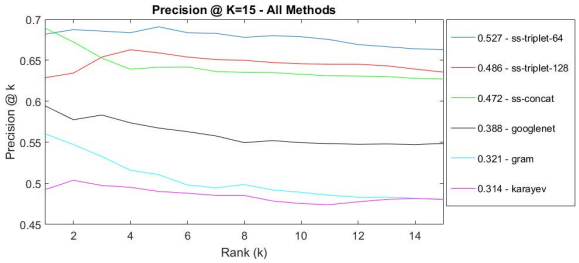

Figure 8. Relative performance of proposed method (*ss-triplet-64*) versus baselines (precision @ K=15).

We substituted the style stream of our network for a popular skip-gram network (Word2Vec [22]) with 128-D output trained over the text corpus in BAM (holding out Behance-VS). The resulting features were directly input to our framework in *ss-concat* configuration. Instead of querying with a sketch + set of style images, we queried the network using a sketch + free-text phrase: *watercolor painting*, *oil painting*, *pen and ink*, *graphite*, *3d graphics*, *vector art*, *comic*, *scary*, *gloomy*, *happy*, *peaceful*. In all, 88 queries (a random subset of 8 sketched queries across 11 styles) were run on this reduced dataset. The results were evaluated through MTurk under identical protocol to Sec. 4.3. Interestingly, the results of visually specifying the aesthetic context were significantly higher ($p<0.05$) than use of text (40.1% versus 18.6% mAP). We similarly explored substitution of keywords for sketches as the structural component of the query, which indicated more strongly (but without significance) in favour of the visual query (40.1% versus 36.4% mAP). We did not pursue this further, as evaluation of [4] versus text

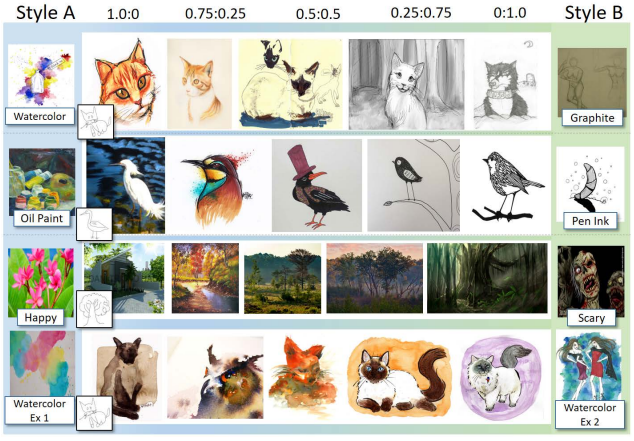

Figure 9. Qualitative results combining pairs of style sets A and B using different weights for a given sketched query (inset). Examples of inter- and intra- style blending (view at 400% zoom).

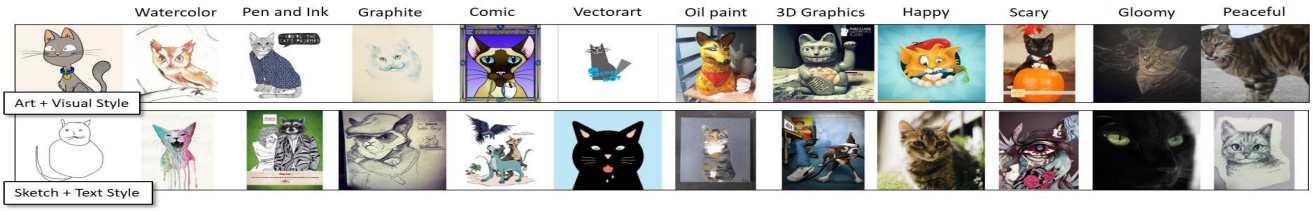

Figure 10. Alternative query modalities. Top-1 result using (top) Behance artwork as structure; (bottom) text instead of image set as style.

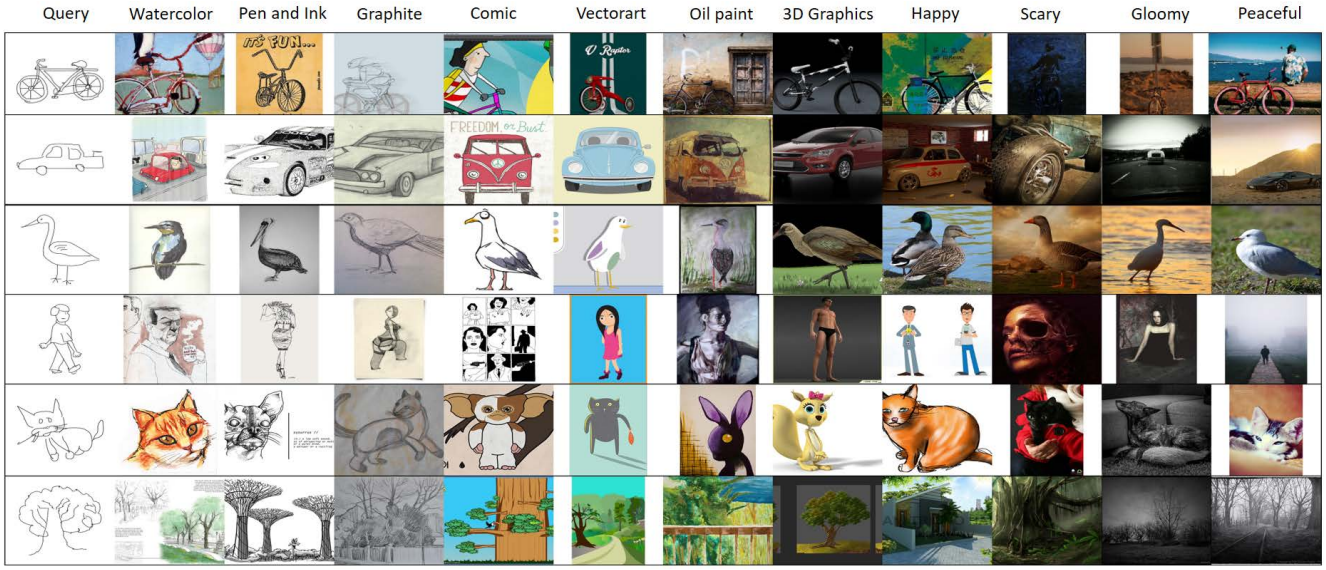

Figure 11. Matrix visualising top result returned for outer-product of all style visual contexts over seven representative query sketches.

search is outside the scope of our contribution. Rather, we explored substitution of the sketch for an artwork image. We swap the structure stream $g_s(.)$ for $g_i(.)$ in the anchor branch of the hierarchical network. As with sketch, we observed results to correctly disentangle content and style. For example, a comic of a cat accompanied by watercolor imagery returned watercolor cats (Fig. 10).

5. Conclusion

We demonstrated a novel visual search framework using a structural query (sketch) augmented with a set of contextual images that specify the desired visual style of results. The framework is underpinned by a learned representation for measuring visual similarity for SBIR that disentangles content and aesthetics. We first proposed a triplet convnet comprising siamese branches adapted from GoogLeNet, showing significant performance increase over the state of art for style retrieval alone. We then combined this network with a shape-matching network [4] to create a multi-modal network of triplet design. We demonstrated that learning a projection of structure and style features through this network further enhances retrieval accuracy, evaluating performance against baselines in a large-scale MTurk experiment.

Interesting directions for search could further explore

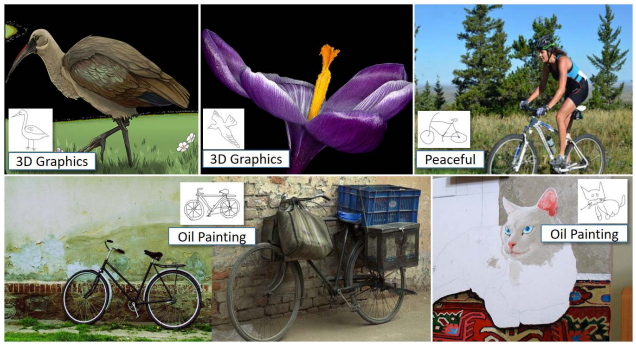

Figure 12. Failure cases from top-15, discussed in Sec. 4.3.1.

blending multiple styles in queries or even style extrapolation. Content recommendation systems, or relevance feedback interfaces, could harness style-space to enhance browsing and discovery. Exploring alternative modalties for structure or for style specification (Sec. 4.4) also appears an interesting avenue, although we do not believe such extensions necessary to demonstrate the novelty and promise of visual search constrained by aesthetic context.

Acknowledgements

We thank Aaron Hertzmann for feedback and discussions. An InnovateUK SMART grant supported the first author.

References

[1] Motion-sketch based video retrieval using a trellis levenshtein distance. pages 121–124, 2010. 1

[2] S. D. Bhattacharjee, J. Yuan, W. Hong, and X. Ruan. Query adaptive instance search using object sketches. In *Proc. ACM Multimedia*, pages 1306–1315, 2016. 2

[3] T. Bui and J. Collomosse. Scalable sketch-based image retrieval using color gradient features. In *Proc. ICCV Workshops*, pages 1–8, 2015. 1

[4] T. Bui, L. Ribeiro, M. Ponti, and J. Collomosse. Generalisation and sharing in triplet convnets for sketch based visual search. *CoRR Abs*, arXiv:1611.05301, 2016. 1, 2, 3, 6, 7, 8

[5] T. Bui, L. Ribeiro, M. Ponti, and J. Collomosse. Compact descriptors for sketch-based image retrieval using a triplet loss convolutional neural network. *Computer Vision and Image Understanding (CVIU)*, 2017. 2

[6] T. Chen, M.-M. Cheng, P. Tan, A. Shamir, and S.-M. Hu. Sketch2photo: Internet image montage. In *Proc. ACM SIGGRAPH Asia*, pages 124:1–124:10, 2009. 1

[7] J. P. Collomosse, G. McNeill, and Y. Qian. Storyboard sketches for content based video retrieval. In *Proc. ICCV*, pages 245–252, 2009. 1

[8] J. P. Collomosse, G. McNeill, and L. Watts. Free-hand sketch grouping for video retrieval. In *Proc. ICPR*, 2008. 1

[9] M. Eitz, J. Hays, and M. Alexa. How do humans sketch objects? In *Proc. ACM SIGGRAPH*, volume 31, pages 44:1–44:10, 2012. 1, 2, 4, 5

[10] L. A. Gatys, A. S. Ecker, and M. Bethge. Texture synthesis using convolutional neural networks. In *Proc. NIPS*, pages 262–270, 2015. 1, 2, 5, 6, 7

[11] A. Gordo, J. Almazán, J. Revaud, and D. Larlus. Deep image retrieval: Learning global representations for image search. In *Proc. ECCV*, pages 241–257, 2016. 2

[12] A. Hertzmann, C. E. Jacobs, N. Oliver, B. Curless, and D. H. Salesin. Image analogies. In *Proc. ACM SIGGRAPH*, pages 327–340, 2001. 2

[13] R. Hu and J. Collomosse. A performance evaluation of gradient field HOG descriptor for sketch based image retrieval. *Computer Vision and Image Understanding (CVIU)*, 117(7):790–806, 2013. 1, 2

[14] S. James and J. Collomosse. Interactive video asset retrieval using sketched queries. In *Proc. CVMP*, 2014. 1, 2

[15] S. James, M. Fonseca, and J. Collomosse. Reenact: Sketch based choreographic design from archival dance footage. In *Proc. ICMR*, 2014. 1

[16] S. Karayev, M. Trentacoste, H. Han, A. Agarwala, T. Darrell, and a. H. W. Aaron Hertzmann. Recognising image style. In *Proc. BMVC*, 2014. 2, 5, 6, 7

[17] A. Kovashka and K. Grauman. Attribute adaption for personalised image search. In *Proc. ICCV*, 2013. 2

[18] A. Kovashka and K. Grauman. Attribute pivots for guiding relevance feedback in image search. In *Proc. ICCV*, 2013. 2

[19] A. Krizhevsky, I. Sutskever, and G. Hinton. Imagenet classification with deep convnets. In *Proc. NIPS*, 2012. 3

[20] T.-Y. Lin and S. Majo. Visualizing and understanding deep texture representations. In *Proc. CVPR*, 2016. 2

[21] L. Marchesotti, F. Perronnin, and F. Meylan. Discovering beautiful attributes for aesthetic image analysis. *Intl. Journal Computer Vision (IJCV)*, 113(3):246–266, 2015. 2

[22] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv preprint*, arXiv:1301.3781, 2013. 7

[23] N. Murray, L. Marchesotti, and F. Perronnin. AVA: A large-scale database for aesthetic visual analysis. In *Proc. CVPR*, 2012. 2

[24] 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. 4

[25] Y. Qi, Y.-Z. Song, H. Zhang, and J. Liu. Sketch-based image retrieval via siamese convolutional neural network. In *Proc. ICIP*, pages 2460–2464. IEEE, 2016. 1, 2

[26] F. Radenović, G. Tolias, and O. Chum. CNN image retrieval learns from BoW: Unsupervised fine-tuning with hard examples. In *Proc. ECCV*, pages 3–20, 2016. 2

[27] M. Rastegari, A. Diba, D. Parikh, and A. Farhadi. Multi-attribute queries: To merge or not to merge? In *Proc. CVPR*, 2013. 2

[28] J. M. Saavedra and B. Bustos. Sketch-based image retrieval using keyshapes. *Multimedia Tools and Applications*, 73(3):2033–2062, 2014. 2

[29] P. Sangkloy, N. Burnell, C. Ham, and J. Hays. The sketchy database: Learning to retrieve badly drawn bunnies. In *Proc. ACM SIGGRAPH*, 2016. 1, 2

[30] L. Setia, J. Ick, H. Burkhardt, and A. I. Features. Svm-based relevance feedback in image retrieval using invariant feature histograms. In *Proc. ACM Multimedia*, 2005. 2

[31] X. Sun, C. Wang, A. Sud, C. Xu, and L. Zhang. Magicbrush: Image search by color sketch. In *Proc. ACM Multimedia*. IEEE, 2013. 1

[32] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Proc. CVPR*, 2015. 3, 5, 6, 7

[33] C. Wang, Z. Li, and L. Zhang. Mindfinder: image search by interactive sketching and tagging. In *Proc. WWW*, pages 1309–1312, 2010. 2

[34] F. Wang, L. Kang, and Y. Li. Sketch-based 3d shape retrieval using convolutional neural networks. In *Proc. CVPR*, pages 1875–1883, 2015. 2

[35] J. Wang, Y. Song, T. Leung, C. Rosenberg, J. Wang, J. Philbin, B. Chen, and Y. Wu. Learning fine-grained image similarity with deep ranking. In *Proc. CVPR*, pages 1386–1393, 2014. 1

[36] X. Wang, K. M. Kitani, and M. Hebert. Contextual visual similarity. *arXiv preprint*, arXiv:1612.02534, 2016. 2

[37] M. J. Wilber, C. Fang, H. Jin, A. Hertzmann, J. Collomosse, and S. Belongie. BAM! the behance artistic media dataset for recognition beyond photography. In *Proc. ICCV*, 2017. 1, 2, 4

[38] 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. 2

[39] Q. Yu, F. Liu, Y.-Z. Song, T. Xiang, T. M. Hospedales, and C.-C. Loy. Sketch me that shoe. In *Proc. CVPR*, pages 799–807, 2016. 1, 2

[40] Q. Yu, Y. Yang, Y.-Z. Song, T. Xiang, and T. Hospedales. Sketch-a-net that beats humans. In *Proc. BMVC*, 2015. 3