StyleBabel: Artistic Style Tagging and Captioning

Anonymous ECCV submission

Paper ID 4976

Abstract. We present StyleBabel, a unique open access dataset of natural language captions and free-form tags describing the artistic style of over 135K digital artworks, collected via a novel participatory method from experts studying at specialist art and design schools. StyleBabel was collected via an iterative method, inspired by 'Grounded Theory': a qualitative approach that enables annotation while co-evolving a shared language for fine-grained artistic style attribute description. We demonstrate several downstream tasks for StyleBabel, adapting the recent ALADIN architecture for fine-grained style similarity, to train cross-modal embeddings for: 1) free-form tag generation; 2) natural language description of artistic style; 3) fine-grained text search of style. To do so, we extend ALADIN with recent advances in Visual Transformer (ViT) and cross-modal representation learning, achieving a state of the art accuracy in fine-grained style retrieval.

Keywords: Datasets and evaluation, Image and video retrieval, Vision + language, Vision applications and systems

1 Introduction

Artistic style is the distinctive appearance of an artwork; i.e. how an artist has depicted their subject matter [14]. Describing the artistic style of digital artwork is an open challenge for computer vision, which has focused on stylization, classification, and search in the style domain. However, automated style description has potential applications in summarization, analytics, and accessibility. For the first time, this paper shows that a set of descriptive tags, or even complete caption sentences, may be automatically generated to describe the fine-grained artistic style of an image – distinct from its content [28,10,55], or the emotions it evokes [2]. Our core contribution enabling this is **StyleBabel**, a novel dataset of fine-grained [54,51,32] annotations describing the artistic style of ∼135K digital artworks[1], collected from expert participants via a novel participatory method that forms a further contribution of this paper. Specifically, our novel contributions are:

1. StyleBabel Dataset. We present a new dataset of over ∼135K {image, tags, natural language (NL) caption} pairs for digital artwork images annotated by a combination of crowd-sourcing and 48 domain experts drawn from design and art schools. StyleBabel is the first dataset containing all three data types for every data sample. Furthermore, the images contained have vastly greater style variation than existing datasets [42,16] which predominantly focus on small subsets of fine art, sometimes further limited to only European or Asian [3,27,53]. Not only is StyleBabel's domain more diverse, but our annotations also differ. StyleBabel focuses on the visual appearance of images (which can include stroke/colouring type, lighting, shading, patterns, shapes, composition, medium,

[1] The dataset will be released for open access (CC-BY 4.0).

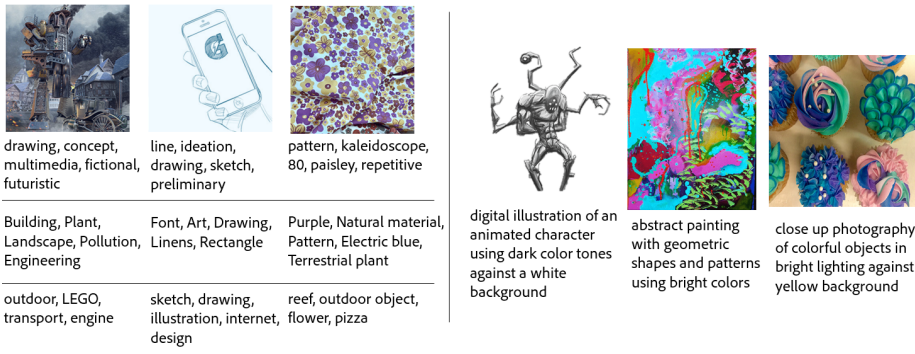


Fig. 1. Results *generated* from models trained on StyleBabel to describe style using free-form tags (left) and captions (right). The top row of tags contains tags generated by our models trained with StyleBabel. The middle row contains tags from a commercial image tagging Google product, and the bottom from a similar commercial product from Microsoft. We show the importance of the style domain-specific information within StyleBabel, for generating relevant style captions.

layout, theme etc.), we avoid external, high level information such as artists, time periods, surrounding meaning, emotions evoked, provenance facts, context or content. This enables new avenues for research not possible before, some of which we explore in this paper.

We do not seek to align our work to a formal ontology or definition of style (something heavily debated even by academics [14,35]). Instead, we explicitly build, evolve, and rely on the emergent structure of style information from the collective, harmonized experience of expert collaborators from art and design universities during our data collection process. In previous work [17], the working definition of style has been the similarity of gram matrices at specific layers in CNN backbones such as VGG [45]. During our data pre-processing/initialization, ahead of style annotation, we similarly use a working definition as similarity in the ALADIN [41] style model's embedding space, previously shown to accurately represent a variety of artistic styles in a metric space.

During annotation, images are grouped into ∼6K moodboards (grids of style-consistent images). Each moodboard is annotated by a group of experts, both with tags and free-form captions, to yield a description of visual style. The vocabulary used for both annotation forms is unconstrained. The moodboard annotations are cross-validated as part of the collection process and refined further via the crowd to obtain individual, image-level fine-grained annotations.

2. Grounded Annotation Methodology. We present a new data annotation methodology inspired by *Grounded Theory* (GT) [44,6], a qualitative research method used in the humanities and social sciences. In GT, participant groups engage in an unconstrained data clustering exercise while simultaneously evolving a shared vocabulary for describing those clusters. Working with these disciplines, we adapted this process to evolve a shared vocabulary for annotation while annotating groups ('moodboards') of images with similar artistic styles. The moodboards are obtained by clustering artworks within the ALADIN fine-grained embedding for style similarity [41]. Our iterative methodology comprises individual and participatory group stages and a validation stage.

3. Artistic Retrieval and Description. We incorporate a Visual Transformer [46] into a fine-grained visual style representation architecture [41], (ALADIN-ViT). ALADIN-

ViT provides state of the art performance at fine-grained style similarity search. We train models for several cross-modal tasks using ALADIN-ViT and StyleBabel annotations. Using CLIP [36], we train with StyleBabel to generate free-form tags describing the artistic style, generalizing to unseen styles. We show that we may apply these tags for text based style search. We similarly demonstrate the synthesis of descriptive natural language captions for digital art.

2 Related Work

Representation learning for visual style has focused primarily on neural style transfer (NST) and style classification.

Style Transfer. Classical approaches learned patch-based representation of style by analogy from paired data. Gatys et al. [17] enabled NST by extracting disentangled representations separating content and style from unpaired data, using a Grammian computed across layers of a pre-trained VGG-19 [45] model. Similarly, feed-forward networks used the Grammian to train fast encoder-decoder models for NST specialized to pre-trained styles [47,23]. Extensions to multi-scale [50] and video [40] NSTS were later presented. To relax the constraint to pre-trained styles, feature based NST was proposed using Adaptive Instance Normalization (AdaIN) [21,18]. This approach was further generalized by the whitening and coloring transform (WCT), which matched feature covariances [29]. Recently unsupervised style transfer was enabled via MUNIT [22] and swapping autoencoder [34]. The latent style codes of encoder-decoder networks for NST were recently adapted for style-based visual representation learning, using weak supervision to learn a metric embedding for fine-grained similarity using the ALADIN model [41].

Style analytics via classification has been explored for digital artwork [52], smaller fine-art collections [57,24], product designs [4], and to identify both painters [7] and genres [43]. Style-based visual search has also been proposed for coarse-grained style using triplet [9] and constrastive training for fine-grained style via ALADIN [41].

Style datasets. No prior dataset of fine-grained artistic style description exists. However, several annotated datasets of artwork have been produced. **Behance Artistic Media – BAM** [52] comprises 2M diverse digital artworks from the `Behance.net` platform, with 7 coarse-grained style and 4 emotion tags and no descriptions. **Omniart** (432K images) and Wikiart (81K) are datasets of fine art with associated metadata but no descriptions or tags. The **SemArt** dataset [16] focuses on very high level contextual semantic information, rarely containing style (visual appearance) information, and narrowly focuses solely on 8th-19th century European fine art. The **BAM-FG** (BAM Fine-grained) dataset comprises 2.62M images grouped into 310K style-consistent groupings but no descriptive text or labels. Our work also uses `Behance.net`, to provide expert style tags and natural language descriptions over 135K images. The **AVA** dataset [5] studies aesthetics information in *photographic* images only, and the follow-up **AVA-captions** dataset [19] adds captions for these, sourced from noisy internet comments sections. Recently, ArtEmis [2] released non-expert annotations capturing the emotions felt by viewers of fine art in WikiArt. Our proposed StyleBabel dataset is aligned to this contemporary work in that it also seeks to ascribe text to visual art. However, our focus is also on digital, not just fine artwork. We also differ in that our annotation is led by expert students in specialized art and design schools, not exclusively by non-expert crowdworkers. Notably, our annotations focus on the style alone, deliberately *avoiding* the

description of the subject matter or the emotions that matter evokes. **ArtEmis** instead describes *exclusively the emotions evoked* by both the style and content of the artwork, both of which feature in the descriptions. To recap, StyleBabel is unique in providing tags and textual descriptions of the artistic style, doing so at a large scale and for a wider variety of styles than existing datasets, with labels sourced from a large, diverse group of experts across multiple areas of art.

Image captioning and visual question answering methods [49] initially learned LSTM language models, leveraging semantic image embeddings e.g. via ResNet/ImageNet. Later image captioning work [28,10,55,56] made use of object detection; regions of interest (ROI). Their relationships yielded improved semantics captioning models, though often due to the bias of co-present context that hinted at the image narrative. This is incompatible with our domain of artistic style, where this localization bias is not something we can use. Recently, there has been an influx of research combining the visual and text domains [36,37,38]. Established methods [13,25], have primarily functioned by generating captions from templates. Though this approach benefits from generating grammatically correct captions, its inflexibility has led to diminished interest in its application, despite recent implementation by OpenAI's CLIP [36]. Here, captioning capabilities generate tags and insert them into templates, using two encoders – for text and image modalities. This model then performs cross-modal training via contrastive loss. More recently, Attention-on-attention [20] can generate state-of-the-art quality captions from an image embedding by filtering out irrelevant attention results. VirTex [11] recently demonstrates that caption annotations are more efficient for representation learning of images, with better or comparable representation quality on ImageNet despite much less training data.

Grounded Theory (GT) [44] is a qualitative method used in the humanities and social sciences to codify data – i.e. to identify and name apparent or emergent patterns across different data sources and types. Through interpretation, participants collaboratively agree on an initial set of summative 'open' codes that are then iteratively combined into larger groups, until eventually, participants arrive, by consensus, at the common themes in the data [6]. We adopt this approach to collecting expert annotation to describe the artistic style of clusters of digital artwork. Free-form textual input from various participants can vary in writing style, creating a very noisy dataset. GT mitigates this by guiding participants to align their responses to a consistent format.

3 StyleBabel Dataset

StyleBabel is a new dataset for cross-modal representation learning. It comprises 135k digital artwork images from the public creative portfolio website `Behance.net` (in turn, available via the BAM dataset). Each image is annotated with a set of keyword tags and natural language descriptions *'captions'* describing its *fine-grained artistic style* – the distinctive appearance of the image – in the English language. We focus mainly on attributes that we can visually depict, rather than more high level and abstract concepts such as emotions [2]. StyleBabel enables the training of models for style retrieval and generates a textual description of fine-grained style within an image: automated natural language style description and tagging (e.g. style2text). We train state of the art proof of concept models for these tasks using our dataset in Sec. 5.

3.1 Study Context

We determined via early initial trials with crowd annotation platforms (AMT) that the quality, coarseness, and diversity of data generated by non-experts is inadequate for style description tasks. We collaborated with graduate schools specializing in digital art and design to address this. Together with academic experts at these schools, we designed a novel multi-staged participatory method to enable novel style vocabulary gathering, tagging, and caption generation, recruiting 48 expert staff and student participants. We particularly sought (but did not make a prerequisite) participants familiar with Behance. The final cohort comprised a representative balance of gender and ethnic background of both graduate and undergraduate cohorts - anecdotally, the gender split was equal. Expertise was at that of a final year undergraduate student, with more senior faculty members participating in the discussions. Their program's specialisms were primarily communication design (graphics, illustration), industrial design, fashion, and animation.

We executed the group exercise online using a *collaborative whiteboard* interactive platform[2], that provides capabilities for multiple users to move and annotate components freely in real time. We built 1300 unique pages of moodboards (Fig.2) to allow participants to move sticky notes with related tags naturally in a collaborative process. The process simulated in-person group collaboration, enabling the formation of tag clusters and aligning to processes of GT codification. The combined use of Miro and Zoom supported real-time spatial organization of information and associated discussion. Workers were paid significantly higher than the national minimum, with a total dataset cost of approximately $160k, which we freely contribute as CC-BY 4.0.

3.2 StyleBabel Grounded Annotation

StyleBabel *does not* aim to develop an ontology to categorize style, with agreement on a diverse ontology eluding art practitioners [14]. Yet, consistency of language is essential for learning of effective representations. Therefore we propose an annotation methodology that enables annotations at scale (multiple participants) and encourages co-evolution of a harmonized natural language to describe the style.

Our annotation process instead is inspired by Grounded Theory (GT) [44,6]; a qualitative method often used for data analysis in the humanities and social sciences. A systematic research process to 'codify' empirical data, identify themes from the data, and associate data with those themes. This process is distinct from fitting (or 'annotating') data to pre-existing categories. GT is an iterative process in which participants co-evolve a language to describe the data as they work on clustering and labeling it with that shared language. Concretely, GT often begins with a discussion around a subset of the data during which clusters are formed. Data is moved freely between clusters during the debate, from which a shared understanding and, ultimately, a shared terminology evolves for describing those clusters. With further data, this language identifies and names patterns apparent or emergent in the data.

Grounded Annotation Process We propose a multi-stage process for compiling the StyleBabel dataset comprised of initial individual and subsequent group sessions and a final individual stage. Each batch of these sessions was estimated to take around 10 hours, run over a week, and run four sets over four weeks.

[2] https://miro.com/

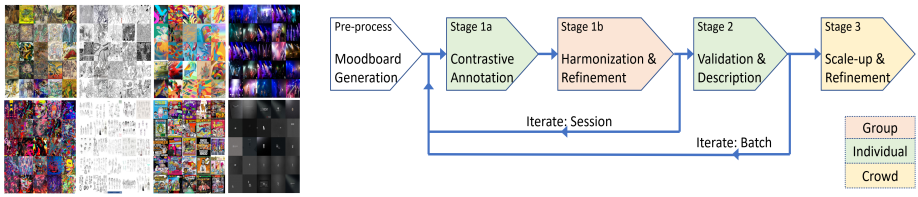


Fig. 2. The StyleBabel Grounded Annotation process. Moodboards (examples, left) are annotated via an iterative process (right) that encourages groups of participants to evolve and converge to a shared language for describing style. See subsec. 3.2 for stage descriptions.

Experts annotate images in small clusters (referred to as image 'moodboards'). Moodboards are obtained by automatically clustering artworks within a fine-grained style embedding [41]. Our annotation process thus pre-determines the clusters for expert annotation. Still, it encourages expert groups to evolve a harmonized language during the iterative annotation process (as in GT) to improve data consistency. We refer to this process as 'grounded annotation'.

Pre-processing - *Moodboard generation (Automated)*

We downloaded an initial dataset of 150 million digital artwork (static image assets) from Behance.net. We encode all images into a metric search embedding [41], that we then clustered into 6.5K clusters, with L_2 distance to identify the 25 nearest image neighbors to each cluster center. The images from each cluster were arranged in a 5×5 grid for presentation as a moodboard. Thus, we start the annotation process using 6,500 moodboards (162.5K images) of 6,500 different fine-grained styles.[3]. The extremely high data density from this internet-scale data corpus ensures that the small clusters formed are very stylistically consistent. Fig. 2 (left) shows examples of moodboards.

Stage 1a - *Contrastive Annotation (Individual)*

Participants were individually presented with a pair of 5x5 moodboards and asked to generate a list of textual tags ('*style attributes*') that occur in one moodboard, but not another, and a list of style attributes which are shared by both. The moodboards were sampled such that they were close neighbors within the ALADIN style embedding. In the annotation, comparative language (e.g. 'X is brighter than Y') was not permitted to encourage standalone descriptions (e.g. 'bright' was allowed). This paired approach encouraged the suggestion of fine-grained style attributes, supporting participants to suggest characteristics that may otherwise not have been considered when looking at individual styles. Multiple participants annotated each moodboard in this way to produce a rich initial set of attributes.

Stage 1b - *Harmonization and Refinement (Group)*

A collaborative workspace was synthesized within Miro, in which 5 moodboards and their associated style tags from Stage 1a are displayed (as 'sticky notes' below each moodboard).

After an initial briefing and group discussion, each group considered moodboards collectively, one moodboard at a time. All participants were asked to add new tags to

[3] We redacted a minimal number of adult-themed images due to ethical considerations

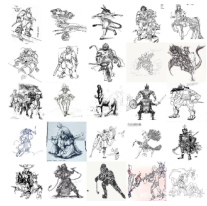
		
Changed (Removed or added tags in Stage 1b)	line, both have the same white backgrounds with very little tonal variations/ just block colours, no difference in tones, bright white background, b+w, constructed by lines, no colour, hand drawing linear, pen and ink	b+w black and white, pen and ink, fantasy, intricate, white space, central composition, linear, illustration
Final tags after Stage 1b cleaning	ink work, sketches, bright, black and white, white background, clean, monotonal, drawing, simple, illustration, linear, pen and ink	graphic, drawing, black and white, pen and ink, fantasy, intricate, white space, central composition, linear, illustration

Fig. 3. Before and after harmonization (Stage 1b), showing the benefits of the GT approach. Two moodboards receive linguistically different tag sets in Stage 1a (individual) but are conformed to the shared vocabulary evolved by the participant group in Stage 1b (group). The moodboard styles and the tag sets differ but draw upon a shared vocabulary (underscored tags) despite the board not being shown in the same session to workers. Tags removed are in red, and additions are in green.

the pre-populated list of tags that we had already gathered from Stage 1a (the individual task), modify the language used, or remove any tags they agreed were not appropriate. Each moodboard was considered 'finished' when no more changes to the tags list could be readily determined (generally within 1 minute). Workers spent at most 2h 15m per session, including all breaks. At least one facilitator was always present throughout to ensure high engagement. Fig. 4 displays an example of moodboards presented during this part of the study via the Miro platform. Fig. 3 provides an illustrative example of language harmonization. Two Stage 1a moodboards with initially very different language but similar style resulted in similar attributes post-Stage 1b.

Stage 2 - *Validation and Description (Individual)*

The participants completed the final stage individually. Each participant was presented with a random collection of 5 moodboards and a set of tags generated during the previous sessions for just one of these moodboards. Participants engaged in ESP-like game [1] to identify which moodboard the tags had been generated to describe, to verify accuracy. Further, we then asked them to create natural language captions, using as many presented tags as possible. We stressed to include as many tags as possible as by now, they had all been thoroughly cleaned and refined.

Stage 3 - *Scale-up and Refinement (Individual)*

Following these sessions with our subject experts, we ran a different task with non-experts (workers on Amazon Mechanical Turk (AMT)). We asked at an *image level* (rather than at the 5x5 cluster/moodboard level as previously) to discard any tags they considered not appropriate. Though the moodboards presented to these non-expert participants are style-coherent, there was still variation in the images, meaning that certain tags apply to most but not all of the images depicted. This step helped us to refine the

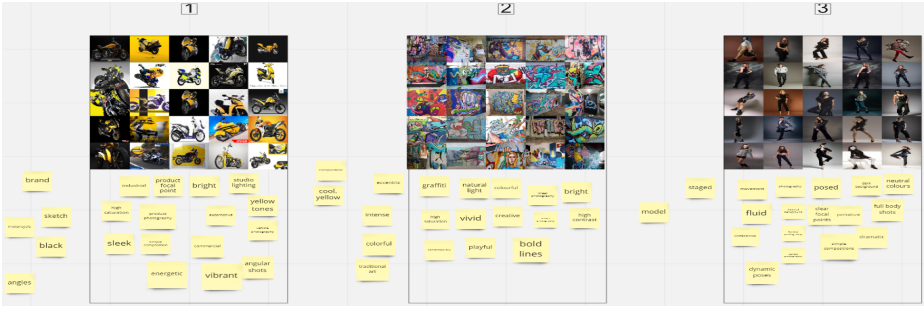


Fig. 4. Virtual environment for collaborative Stage 1b. Style attributes (as 'sticky notes') are added, modified and removed from each moodboard to harmonize and filter language.

final tags to individual images further. Unlike novel vocabulary generation, this verification step can be readily performed by non-experts, as detecting invalid tags is a much easier task than novel tag generation. We collected 3 responses per image and performed majority consensus to determine the final tags to use at the per-image level.

We additionally performed a large-scale crowd annotation exercise to individualize cluster-level captions to individual images. Trained workers were presented with individual images, its tags, and the moodboard caption and were asked to compose (potentially many) natural language captions using the tags and caption, ensuring the full set of tags were incorporated across those sentences. A constant set of workers were trained with feedback, for several months, together with a Quality Control (QC) process to ensure high quality annotation. The QC process included grammatical correctness checking, and rejection of any description of content or emotion the descriptive captions.

Language processing Aside from the crowd data filtering, we cleaned the tags emerging from Stage 1b through several steps, including removing duplicates, filtering out invalid data or tags with more than 3 words, singularization, lemmatization, and manual spell checking for every tag. The spell checking step was carried out in 3 passes. The accuracy of the validation step in Stage 2 was found to be 90%.

The final StyleBabel dataset contains 135k images with an average of 12.8 tags per image, over 6k style groups (of the 6,500 initially sampled, with 6k completed by workers in the available time). The tags dictionary contains 3,151 unique tags, and the captions contain 5,475 unique words. Prepositions, determiners, and conjunctions were filtered out. Cardinals (eg. *80s*) are kept. 9 cardinals, 2098 nouns, 500 verbs, 55 adverbs, and 482 adjectives are captured. Fig. 5 visualizes a word cloud from the 250 most common style attributes in StyleBabel, and Tbl. 1 shows the richness of the tags and captions.

4 Visual Embedding (ALADIN-ViT)

ALADIN is a two branch encoder-decoder network that seeks to disentangle image content and style. It works by pooling Adaptive Instance Normalization (AdaIN) statistics across multiple layers in the style encoder to produce an embedding for fine-grained style similarity using a VGG-19 convolutional backbone for the style encoder. In our

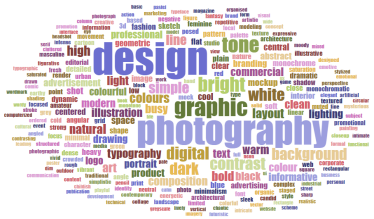


Fig. 5. Visualizing the top 250 tags captured within the StyleBabel annotation over 135K images.

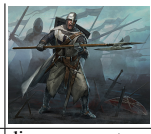
Image				
Tags	dim, concept, action, fantasy, powerful, digital, photography, animated, prototype, masculine, detailed, professional, lighting	abstract, moody, portrait, oil, painting, drawing, artistic, melancholic, pleasing	cold, digital, book, bright, colors, drawing, child, stroke, busy, clear, illustration, festive, blue	experimental, analog, line, development, black, drawing, sketch, figure, commercial, white, scamp, stroke, product, pencil, rough, thin, isometric
Caption	Fantasy themed digital illustration featuring an animated male character, dim highlighting and a hazy, dark and cluttered background. The illustration highlights the powerful masculine character with sharp objects around.	Portrait oil painting of a female character featuring abstract shapes and psychedelic patterns against a dark background. The artistic artwork is melancholic and using thin repetitive strokes and shades.	Digital bright fantasy anthropomorphism cartoon illustration created with soft diffused blended hues, brush strokes, lines, and geometric forms in neutral and cool tones.	Analog experimental sketches with thin pencil strokes and lines. The isometric drawing expresses commercial product development.

Table 1. Excerpt of StyleBabel dataset. Four images and the corresponding tags and captions collected via our Grounded Annotation.

later experiments, we require to use a Visual Transformer [12] (ViT) model for the vision domain. To achieve this, we adapt ALADIN to use ViT as the style encoder backbone; we refer to this as ALADIN-ViT (Fig. 6). We retrain ALADIN-ViT on BAM-FG following the same training method [41], i.e. using both the reconstruction loss and the weakly supervised contrastive loss and using the implicit style grouping in BAM-FG. Having swapped the style encoder for a transformer, it is no longer possible to sample AdaIN statistics from feature maps in the encoder. However, we retain the use of the style code as pairs of values to be split and used in the decoder stage, again keeping the size of the style code at double the number of feature maps in the decoder.

We achieve the state of the art fine grained style retrieval accuracy on the BAM-FG test partition (Tbl. 2) at **64.48** Top-1, beating not only ALADIN (58.98) but also their fused variant (62.18), which incorporates ResNet embeddings into a concatenated embedding. We, in part, attribute the gains in accuracy to the larger receptive input size (in the pixel space) of earlier layers in the Transformer model, compared to early layers in CNNs. Given that style is a global attribute of an image, this greatly benefits our domain as more weights are trained on more global information.

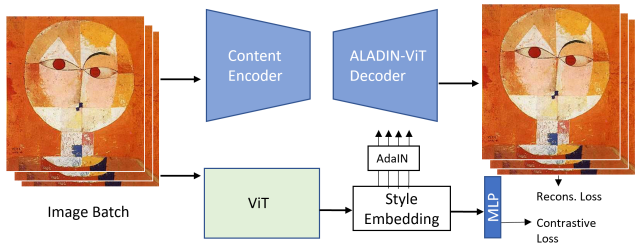


Fig. 6. ViT-ALADIN architecture used for StyleBabel experiments; as per ALADIN [41] but swapping the style encoder for ViT [12] (change in green) and retrained end-to-end on BAM-FG.

Data	Model	IR-1
BAM-FG-C_5	ALADIN	58.98
BAM-FG-C_5	ResNet	45.22
BAM-FG-C_5	ALADIN (Fused)	62.18
BAM-FG-C_5	ViT	57.91
BAM-FG-C_5	ALADIN-ViT	**64.48**

Table 2. Fine grained style retrieval on the BAM-FG dataset [41] of proposed method ALADIN-ViT, compared to previous methods

5 StyleBabel Experimental Setup

Data Partitions. We define train/validation/test partitions within StyleBabel for our experiments as follows. Splits were separated on a moodboard basis, to avoid overlap. The training split has 133k images in 5,974 groups with 3,167 unique tags at an average of 13.05 tags per image. The validation and test splits contain 1k unique images for each validation and test, with 1,256/1,570/10.86 and 1,263/1,636/10.96 unique tags/groups/average tags per image. Captions have an average of 2.4 sentences, with an average of 19.3 words, from 6119 unique words. 1000 samples were extracted for each of the test/validation splits.

Implementation. The models were trained with a maximum batch size of 11k on a 12GB GTX Titan, with a learning rate of 0.003, Adam optimizer, and weight decay of 1e-6. Logit accumulation [41] was employed to reach the maximum batch size possible in the GPU VRAM capacity. We trained all models to convergence. The learning rate was decayed using cosine annealing, as per SimCLR [8]. For the VirTex captioning experiments, a batch size of 105 was used, on a V100, with a learning rate of 2e-4.

6 Experiments and Discussion

We illustrate the potential of our StyleBabel dataset for three cross-modal learning tasks:
1. Style Auto-tagging: (style2text) Using StyleBabel tags, we train a CLIP [36] model to learn a cross-modal embedding between image and text embeddings. We generate several tags for unseen StyleBabel images and explore the generated tags' accuracy and the model's ability to generalize.
2. Style Description: (style2text) We similarly make use of the natural language captions collected in StyleBabel, and showcase the generation of natural language captions describing the style of images.

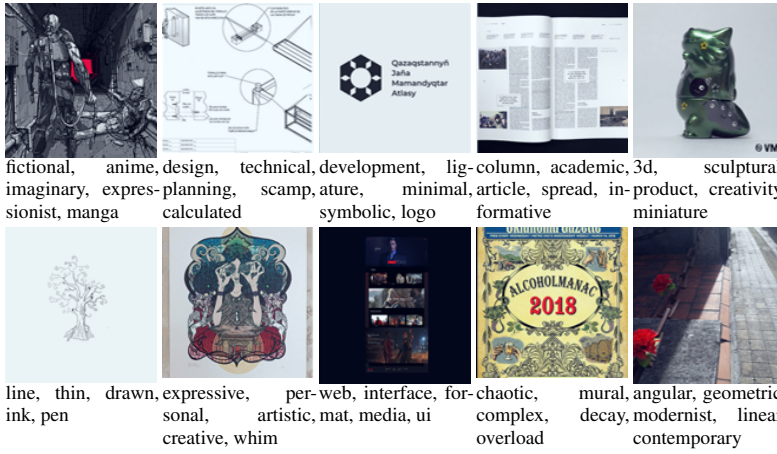


Fig. 7. Style2Text tag generation experiment, using CLIP trained with ALADIN-ViT encoder. Top 5 tags shown for each image.

3. Text Based Style Retrieval: (text2style) We explore the efficacy of our tag generation for text based style search.

6.1 Style Auto-tagging (style2text)

Recent literature in image captioning has transitioned to making use of object detectors in their model pipelines. This makes sense in semantics, as such features are most often localized to a subset of the image. Style, however, is typically a global attribute of an image, and object detectors are not compatible. Style is more abstract and seldom localized to any specific region of an image. We use the CLIP [36] training methodology to learn a joint embedding space between the publicly available CLIP text encoder and our new vision transformer (ALADIN-ViT). CLIP is traditionally formed of two transformers, the first for text encoding and the second for image encoding. Two MLP heads are trained together through contrastive loss to learn a joint text/image embedding, adding invariance to the modality. We freeze both pre-trained transformers and train the two MLP layers (ReLU separated fully connected layers) to project their embeddings to the shared space. We follow CLIP, employing contrastive loss to drive learning, with the same training set-up.

When using the model for inference, we pass the entire dictionary of available tags through the text encoder and multi-modal MLP head to generate text embeddings. Next, we infer the image embedding using the image encoder and multi-modal MLP head, and calculate similarity logits/scores between the image and each of the text embeddings. We evaluate accuracy via WordNet similarity [15] using the *nltk*[4] library to first compute synsets for the N ground truth tags for an image. Next, we sort the tags in the dictionary by the logit scores following the embedding inference similarity. We select the top N "retrieved" tags, and for each, and calculate the WordNet similarity to each ground truth tag. The similarity ranges from 0 to 1, where 1 represents identical tags. We save the top

[4] https://www.nltk.org/

Data	Model	WordNet score
CLIP Webscale	CLIP [36] baseline	0.168
StyleBabel-mturk	ALADIN-ViT	0.164
StyleBabel (coarse)	CLIP [36]	0.187
StyleBabel (coarse)	ALADIN-ViT	0.225
StyleBabel (FG)	CLIP	0.215
StyleBabel (FG)	ALADIN-ViT	**0.352**

Table 3. Ablation experiments for tag generation under the CLIP training setting. We show the benefit of annotating StyleBabel via the proposed GT annotation versus non-expert crowd-sourcing (StyleBabel-mturk). We further demonstrate the benefits of tag annotations individualised to the image-level (FG), compared to cluster-level (coarse).

value (the similarity score for the closest/most relevant ground truth tag) and repeat it for each test image. These values are averaged to form our *WordNet score*. We use this instead of precision/recall, as multiple tags in the dataset can represent very similar (but not identical) concepts. A soft score better encapsulates the perceived accuracy of the tags.

We experiment with training CLIP on variants of StyleBabel, presenting results in Tbl.3. In particular, we quantify the difference in quality between collecting annotation via non-expert crowd annotation (StyleBabel-mturk) and gathering expert annotations using our GA process (StyleBabel/ALADIN-ViT). We also show the value of the final stage, where we refine tags to the image-level (FG) rather than moodboard-level (coarse). We train the MLP heads atop the CLIP image encoder embeddings (the 'CLIP' model) and atop embeddings from our ALADIN-ViT model (the 'ALADIN-ViT' model). The former is not based on the ALADIN-ViT style embedding and underperforms by 40%. The best performing model is the proposed ALADIN-ViT trained via StyleBabel data collected using GA on FG labels, with a WordNet score of **0.352**, double the CLIP [36] baseline. Fig. 7 shows some examples of tags generated for various images, using the ALADIN-ViT based model trained under the CLIP method with StyleBabel (FG).

MS-COCO	a woman with a painting of a face on it	a close up of a remote with a remote	a kite that is hanging in the air	a room with a lot of windows and a clock	a person riding a surfboard in the water
StyleBabel CL+IL	digital illustration of a fictional character using dark color tones against a black background	product photography of business cards with text and logo in soft lighting against a white background	digital illustration featuring characters and typography using bright colors	architectural photography of a modern interior design in bright lighting using a neutral color palette	abstract watercolor painting using paint brush strokes and shading effects

Fig. 8. Examples of natural language captions generated for various art styles; Generated captions are compared from VirTex models trained on MS-COCO and Stylebabel CL+IL, showing the benefit of the style information present in the dataset

We run a user study on AMT to verify the correctness of the tags generated, presenting 1000 randomly selected test split images alongside the top tags generated for each.

For each image/tags pair, 3 workers are asked to indicate tags that don't fit the image. We score tags as correct if all 3 workers agree they belong. The absolute accuracy of this study is 89.86%, indicating high tag generation accuracy.

Finally, we explore the model's generalization to new styles by evaluating the average WordNet score of images from the test split. In Fig.9, we group the data samples into 10 *bins* of distances from their respective style cluster centroid, in the style embedding space. As before, we compute the WordNet score of tags generated using our model and compare it to the baseline CLIP model. Though the quality of the CLIP model is constant as samples get further from the training data, the quality of our model is significantly higher for the majority of the data split. At worst, our model performs similar to CLIP and slightly worse for the 5 most extreme samples in the test split. But for the majority of the test data, our model considerably outperforms CLIP.

Data	Model	Bleu-1	Bleu-2	Bleu-3	Bleu-4	METEOR	Rouge-L	CIDEr
MS-COCO baseline	VirTex	0.162	0.053	0.016	0.005	0.037	0.145	0.022
StyleBabel (CL)	VirTex	0.127	0.049	0.022	0.010	0.054	0.135	0.076
StyleBabel (IL)	VirTex	0.331	0.187	0.113	0.071	0.129	0.288	0.350
StyleBabel (CL+IL)	VirTex	**0.335**	**0.189**	**0.118**	**0.078**	**0.131**	**0.288**	**0.372**
StyleBabel (CL+IL)	ResNet LSTM	0.087	0.021	0.008	0.002	0.033	0.080	0.017
StyleBabel (CL+IL)	ALADIN-ViT LSTM	0.094	0.030	0.013	0.006	0.042	0.089	0.034
VirTex	Artemis	0.185	0.083	0.041	0.023	0.081	0.182	0.146
VirTex	Artemis (SB)	0.120	0.031	0.013	0.005	0.034	0.108	0.029

Table 4. (top) VirTex caption generation metrics on 1k holdout StyleBabel test data. CL represents cluster-level, and IL represents image-level. We also run CL+IL, where we fine-tune a cluster-level model with image-level data labels. (middle) Results with a baseline LSTM model trained over either ResNet or ALADIN image embeddings (bottom) Additional experiments showing performance of a VirTex model trained on an existing dataset (Artemis), and also evaluated on the StyleBabel (SB) test set.

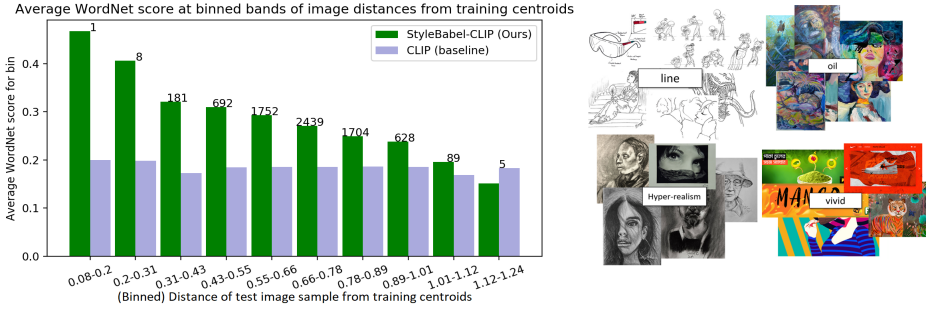


Fig. 9. (left) Generalization experiment for tag generation using baseline CLIP [36] and our StyleBabel trained model using ALADIN-ViT. The test set was sorted by distance in the style embedding space to closest training cluster. Their WordNet scores were binned into 10 quantized distance bands. The numbers atop the bars indicate the number of samples in the corresponding bin. (right) Top 5 style retrieval using textual tags. Using tags generated by ALADIN-ViT/CLIP over the StyleBabel test partition, we perform a keyword based search for artistic style.

6.2 Style Description (style2text)

We explore the feasibility of using the StyleBabel dataset for generating natural language captions. We conduct our experiments using a fusion of the VirTex [11] backbone for the visual representation learning of image/caption pairs, and Attention on Attention (AoA) [20] for the caption decoding. VirTex encodes images without using scene graphs, therefore avoiding issues related to style not being localized in an image. VirTex replaces the Faster-RCNN [39] component in AoA to generate feature maps. We use the pre-trained VirTex on COCO [31], and fine-tune the entire setup, end-to-end on final StyleBabel captions. As per standard practice, during data pre-processing, we remove words with only a single occurrence in the dataset. Removing 45.07% of unique words from the total vocabulary, or 0.22% of all the words in the dataset. We test the caption generation on the StyleBabel test data across Bleu [33], METEOR [26], Rouge [30], and CIDEr [48], as shown in Tab 4. See the supplementary material for further analysis.

We would like to stress that these metrics are not comparable with values from these metrics on standard literature, as we are solving a new task. In literature, these metrics are used for semantic, localized features in images, whereas our task is to generate captions for global, style features of an image. We include an MS-COCO baseline, to show comparative accuracy versus a dataset with no style information. Figure 8 displays captions generated using this method.

6.3 Text based style retrieval (text2style)

We explore the potential of the ALADIN-ViT+CLIP model trained in subsec. 6.1 to perform image retrieval, using textual tag queries. By first indexing the score assigned to each tag in the dictionary at the image level, we can then use a tag query to retrieve images based on the sorted scores for that tag. We use nearest-neighbour search using the image embeddings, reversing the tags generation experiment. Fig 9 shows some example image retrievals using text queries.

To measure the quality of the results, we run all text tags as queries. For each, we compute the WordNet similarity of the query text tag to the kth top tag associated with the image, following a tag retrieval using a given image. We vary k and collect the average scores at values of 1, 5, 10, and 25. The scores at these values are 0.72, 0.467, 0.392, and 0.332, respectively .

7 Conclusion

We proposed StyleBabel, a novel unique dataset of digital artworks and associated text describing their fine-grained artistic style. Our annotation approach was inspired by Grounded Theory (GT) [44], to support the 'emergence' of themes from the corpus of digital artwork – as opposed to fitting images to pre-existing categories [6]. These sessions generated discussion while simultaneously evolving and arriving through consensus at a shared vocabulary for describing image clusters of similar style.

We extended ALADIN [41] to incorporate a visual transformer [12] (ALADIN-ViT) encoder, obtaining state of the art style similarity discrimination, leveraging StyleBabel for the automated description of artwork images using keyword tags and captions. We also showed text-based image retrieval of images based on the generated style tags. Further work could explore use of tags as priors in generating captions, and exploring more downstream tasks using StyleBabel.

References

1. Luis A. and Laura D. Esp: Labeling images with a computer game. pages 91–98, 01 2005. 7
2. P. Achlioptas, M. Ovsjanikov, K. Haydarov, M. Elhoseiny, and L. Guibas. Artemis: Affective language for visual art. In *Proc. CVPR*, 2021. 1, 3, 4
3. Zechen Bai, Yuta Nakashima, and Noa Garcia. Explain me the painting: Multi-topic knowledgeable art description generation. *CoRR*, abs/2109.05743, 2021. 1
4. S. Bell and K. Bala. Learning visual similarity for product design with convolutional neural networks. In *Proc. ACM SIGGRAPH*, 2015. 3
5. Simone Bianco, Luigi Celona, Paolo Napoletano, and Raimondo Schettini. Predicting image aesthetics with deep learning. volume 10016, pages 117–125, 10 2016. 3
6. Kathy C. *Constructing Grounded Theory : A Practical Guide through Qualitative Analysis*. Sage, London, 2006. 2, 4, 5, 14
7. E. Cetinic and S. Grgic. Automated painter recognition based on image feature extraction. In *Proc. ELMAR*, 2013. 3
8. T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020. 10
9. J. Collomosse, T. Bui, M. Wilber, C. Fang, and H. Jin. Sketching with style: Visual search with sketches and aesthetic context. In *Proc. ICCV*, 2017. 3
10. M. Cornia, M. Stefanini, L. Baraldi, and R. Cucchiara. Meshed-memory transformer for image captioning. *arXiv preprint arXiv:1912.08226*, 2020. 1, 4
11. Karan Desai and Justin Johnson. Virtex: Learning visual representations from textual annotations. *CoRR*, abs/2006.06666, 2020. 4, 14
12. A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 9, 10, 14
13. Ali F., Mohsen H., M. Amin S., Peter Y., Cyrus R., Julia H., and David F. Every picture tells a story: Generating sentences from images. In Kostas D., Petros M., and Nikos P., editors, *Computer Vision – ECCV 2010*, pages 15–29, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg. 4
14. Eric F. *Art History and its Methods: A critical anthology*. Phaidon, London, 1995. 1, 2, 5
15. C. Fellbaum. *WordNet: An Electronic Lexical Database*. Bradford Books, 1998. 11
16. Noa Garcia and George Vogiatzis. How to read paintings: Semantic art understanding with multi-modal retrieval. *CoRR*, abs/1810.09617, 2018. 1, 3
17. L. A. Gatys, A. S. Ecker, and M. Bethge. A neural algorithm of artistic style. *arXiv preprint arXiv:1508.06576*, 2015. 2, 3
18. G. Ghiasi, H. Lee, M. Kudlur, V. Dumoulin, and J. Shlens. Exploring the structure of a real-time, arbitrary neural artistic stylization network. *arXiv preprint arXiv:1705.06830*, 2017. 3
19. Koustav Ghosal, Aakanksha Rana, and Aljosa Smolic. Aesthetic image captioning from weakly-labelled photographs. *CoRR*, abs/1908.11310, 2019. 3
20. Lun Huang, Wenmin Wang, Jie Chen, and Xiaoyong Wei. Attention on attention for image captioning. *CoRR*, abs/1908.06954, 2019. 4, 14
21. X. Huang and S. Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proc. ICCV*, 2017. 3
22. X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz. Multimodal unsupervised image-to-image translation. In *ECCV*, 2018. 3
23. J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Proc. ECCV*, 2016. 3
24. S. Karayev, M. Trentacoste, H. Han, A. Agarwala, T. Darrell, A. Hertzmann, and H. Winnemoller. Recognizing image style. In *Proc. BMVC*, 2014. 3
25. G. Kulkarni, V. Premraj, V. Ordonez, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg. Babytalk: Understanding and generating simple image descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(12):2891–2903, 2013. 4

26. Alon Lavie and Abhaya Agarwal. Meteor: An automatic metric for mt evaluation with high levels of correlation with human judgments. pages 228–231, 07 2007. 14
27. Xu Lei, Albert Meroño-Peñuela, Zhisheng Huang, and F. V. Harmelen. An ontology model for narrative image annotation in the field of cultural heritage. In *WHiSe@ISWC*, 2017. 1
28. X. Li, X. Yin, C. Li, P. Zhang, X. Hu, L. Zhang, L. Wang, H. Hu, L. Dong, F. Wei, Y. Choi, and J. Gao. Oscar: Object-semantics aligned pre-training for vision-language tasks. *arXiv preprint arXiv:2004.06165*, 2020. 1, 4
29. Y. Li, C. Fang, J. Yang, Z. Wang, X. Lu, and M. Yang. Universal style transfer via feature transforms. In *Proc. NIPS*, 2017. 3
30. Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. page 10, 01 2004. 14
31. Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312, 2014. 14
32. K. Pang, Y. Yang, T. M. Hospedales, T. Xiang, and Y. Song. Solving mixed-modal jigsaw puzzle for fine-grained sketch-based image retrieval. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10344–10352, Los Alamitos, CA, USA, jun 2020. IEEE Computer Society. 1
33. Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA, 2002. Association for Computational Linguistics. 14
34. T. Park, J.-Y. Zhu, O. Wang, J. Lu, E. Shechtman, A. Efros, and R. Zhang. Swapping autoencoder for deep image manipulation. In *Proc. ECCV*, 2020. 3
35. Andrea Pinotti. *Formalism and the History of Style*, pages 75 – 90. Brill, Leiden, The Netherlands, 2012. 2
36. A. Radford, J. Wook Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021. 3, 4, 10, 11, 12, 13
37. A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever. Zero-shot text-to-image generatio. *arXiv preprint arXiv:2102.12092*, 2021. 4
38. A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever. Zero-shot text-to-image generation. *arXiv preprint arXiv:2102.12092*, 2021. 4
39. Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. *CoRR*, abs/1506.01497, 2015. 14
40. M. Ruder, A. Dosovitskiy, and T. Brox. Artistic style transfer for videos. In *Proc. GCPR*, 2016. 3
41. D. Ruta, S. Motiian, B. Faieta, Z. Lin, H. Jin, A. Filipkowski, A. Gilbert, and J. Collomosse. Aladin: All layer adaptive instance normalization for fine-grained style similarity. *arXiv preprint arXiv:2103.09776*, 2021. 2, 3, 6, 9, 10, 14
42. Babak Saleh and Ahmed Elgammal. Large-scale classification of fine-art paintings: Learning the right metric on the right feature, 2015. 1
43. L. Shamir, T. Macura, N. Orlov, and D. Eckley. Impressionism, expressionism, surrealism: automated recognition of painters and schools of art. *IEEE Trans. Applied Perception*, 2010. 3
44. J. Simondsen and T. Roberton. *Routledge International Handbook of Participatory Design*. Routledge, 2013. 2, 4, 5, 14
45. K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 2, 3
46. A. Srinivas, T.-Yi Lin, N. Parmar, J. Shlens, P. Abbeel, and A. Vaswani. Bottleneck transformers for visual recognition. *arXiv preprint arXiv:2101.11605*, 2021. 2
47. D. Ulyanov, V. Lebedev, A. Vedaldi, and V. Lempitsky. Texture networks: Feed-forward synthesis of textures and stylized images. In *Proc. ICML*, 2016. 3
48. Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. *CoRR*, abs/1411.5726, 2014. 14

49. O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. *arXiv preprint arXiv:1411.4555*, 2015. 4
50. X. Wang, G. Oxholm, D. Zhang, and Y.-F. Wang. Multimodal transfer: A hierarchical deep convolutional neural network for fast artistic style transfer. In *Proc. CVPR*, 2017. 3
51. X.-S. Wei, J.-H. Luo, J. Wu, and Z.-H. Zhou. Selective convolutional descriptor aggregation for fine-grained image retrieval. *arXiv preprint arXiv:1604.04994*, 2017. 1
52. M. J. Wilber, C. Fang, H. Jin, A. Hertzmann, J. Collomosse, and S. Belongie. Bam! the behance artistic media dataset for recognition beyond photography. *arXiv preprint arXiv:1704.08614*, 2017. 3
53. Lei Xu and Xiaoguang Wang. Semantic description of cultural digital images: Using a hierarchical model and controlled vocabulary. *D Lib Mag.*, 21(5/6), 2015. 1
54. B. Yao, A. Khosla, and L. Fei-Fei. Combining randomization and discrimination for fine-grained image categorization. In *CVPR 2011*, pages 1577–1584, 2011. 1
55. P. Zhang, X. Li, X. Hu, J. Yang, L. Zhang, L. Wang, Y. Choi, and J. Gao. Vinvl: Revisiting visual representations in vision-language models. *arXiv preprint arXiv:2101.00529*, 2021. 1, 4
56. L. Zhou, H. Palangi, L. Zhang, H. Hu, J. J. Corso, and J. Gao. Unified vision-language pre-training for image captioning and vqa. *arXiv preprint arXiv:1909.11059*, 2019. 4
57. J. Zujovic, L. Gandy, S. Friedman, B. Pardo, and Pappas T. Classifying paintings by artistic genre: an analysis of features and classifiers. In *Proc. IEEE Workshop on Multimedia Sig. Proc. (MMSP)*, 2009. 3