Unsupervised Image Style Embeddings for Retrieval and Recognition Tasks

[1]Siddhartha Gairola, [1,2]Rajvi Shah* and [1]P.J. Narayanan
[1]CVIT, KCIS, IIIT Hyderabad, India; [2]Facebook Reality Labs, Redmond, WA, USA
{siddhartha.gairola@research.,rajvi.shah@research.,pjn@}iiit.ac.in

Abstract

We propose an unsupervised protocol for learning a neural embedding of visual style of images. Style similarity is an important measure for many applications such as style transfer, fashion search, art exploration, etc. However, computational modeling of style is a difficult task owing to its vague and subjective nature. Most methods for style based retrieval use supervised training with pre-defined categorization of images according to style. While this paradigm is suitable for applications where style categories are well-defined and curating large datasets according to such a categorization is feasible, in several other cases such a categorization is either ill-defined or does not exist. Our protocol for learning style based representations does not leverage categorical labels but a proxy measure for forming triplets of anchor, similar, and dissimilar images. Using these triplets, we learn a compact style embedding that is useful for style-based search and retrieval. The learned embeddings outperform other unsupervised representations for style-based image retrieval task on six datasets that capture different meanings of style. We also show that by fine-tuning the learned features with dataset-specific style labels, we obtain best results for image style recognition task on five of the six datasets.

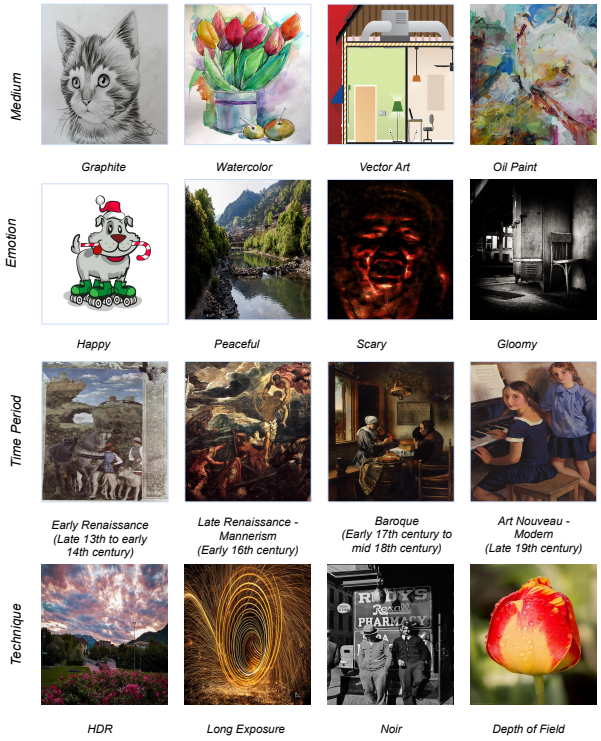

Figure 1. Examples of image style categorization with different meanings of style. Each row corresponds to a category based on a particular understanding of style.

1. Introduction

In visual arts, style is used as a primary apparatus to relate, organize and describe artworks. However, understanding of style is highly contextual and vague. Depending on the context, sense of style is attributed to time period, location, culture, artist, technique, school of design, modality, etc. depicted in Figure 1. A highly subjective construct like style is hence, difficult to model computationally. In the context of computer vision, Karayev et al. [11] presented one of the early works for image style recognition with multiple datasets of photographic and painting images with different types of visual style categorizations such as photographic techniques (Macro, HDR), moods (Serene, Melancholy), themes (Vintage, Romantic, Horror), artistic movements (Renaissance, Post-modern). Later, Wilber et al. [21] presented a large dataset of contemporary artworks – the 'Behance Artistic Media Dataset' (BAM) with crowdsourced labels for media, emotions, and objects. Convolutional Neural Networks (CNN) are found to be very useful for gaining an implicit understanding of images from vast amounts of data for many computer vision tasks. With availability of these datasets and advances in neural learning, developing methods for computational understanding

*This work was done while the author was at IIIT Hyderabad. Project page : https://sidgairo18.github.io/style

of style is becoming an interesting possibility.

Present methods related to style based representations can be divided into two categories - implicit and explicit. Unsupervised style transfer methods [4, 5] model style implicitly as intermediate feature representations learned from an unrelated supervised learning task such as object recognition. Style, in this context typically describes the visual 'look and feel' (texture, tone, and colors) of an image. These methods leverage Gram matrix features which capture the correlation among feature maps extracted from the many layers of a deep CNN (like VGG-19 [18]), typically pre-trained for object classification on a very large dataset like ImageNet [3]. On the other hand, the popular paradigm in computer vision community for explicit style understanding is to treat it as a supervised classification problem. Such methods generally use large datasets with a fixed set of style labels to train a neural network for the style classification task and use the learned feature maps for style representation [1, 2, 8, 11, 21]. The representations learned under this paradigm are effective and efficient for task-specific retrieval but have practical limitations in terms of generalization and scalability, the biggest one being the need for manual curation of large training data. This entire process is not only expensive and inefficient, but also ill-suited for a subjective attribute like artistic style where expert annotations are limited to a few significant works of art, like famous paintings or gallery displays. In contrast, Gram Matrix features are readily computable for any new dataset and provide a specific measure of style disentangled from content to some degree, but it is an inefficient representation for search and retrieval due to high correlation and very high dimensionality.

One of the key motivations of this paper is to investigate the quality of understanding of style that can be achieved by an unsupervised approach which does not rely on categorical labels of style. To this effect, we evaluate state-of-the-art representations and their variants for style-based retrieval. We further propose a protocol for unsupervised learning of style representation by leveraging a proxy measure that provides a loose grouping of images. Our proxy measure is based on Gram matrix features popularized by style transfer methods. These features capture the 'look and feel' of an image by measuring the correlation among feature maps produced by different convolutional layers of a CNN and hence are a good choice for discerning different visual styles. We train a Siamese CNN [20] for learning a style embedding that is relevant for style based search and retrieval. However, instead of leveraging the style class labels specified for a dataset, we do this in an unsupervised fashion for many datasets. We first divide a dataset into k clusters using Gram matrix features and then use the cluster labels for learning the embedding by (i) directly minimizing a cross-entropy loss for cluster label classification, and (ii) minimizing a triplet loss for maximizing the distances between stylistically (look and feel wise) similar and dissimilar samples. The training with a triplet loss further reinforces the stylistic similarity which is depicted in Figure 2. This is of large interest as the unsupervised protocol can be used on unlabelled (no supervision) data for learning stylistically useful representations and help understand a highly subjective concept like style (look and feel) better.

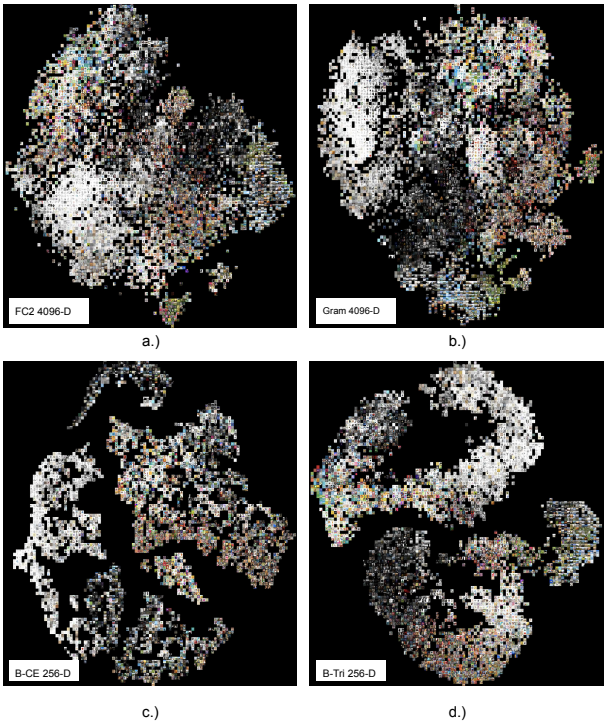

Figure 2. t-SNE [19] visualizations of BAM dataset images based on following feature representations: (top row) FC2 features and PCA-reduced Gram features computed from pre-trained VGG19, (bottom row) embeddings learned using our protocol. It can be observed that using triplet loss (B-Tri) further reinforces the stylistic similarity in comparison to other features (refer to Table 1 and Figure 3 for more details on the representations).

We evaluate the performance of style representations learnt using these unsupervised training protocols across 6 datasets with distinct categories of styles and compare against other known style representations. The triplet loss based unsupervised protocol outperforms other representations for most datasets and our experiments reveal many interesting insights. We also introduce 2 new datasets consisting of curated 'Wall Art Sets' and 'Contemporary Drawings and Paintings'.

To summarize, **our contributions** are two-fold:

- First, we propose an unsupervised protocol for learning a deep neural embedding of visual style of images

by leveraging a proxy measure that provides a loose grouping of stylistically similar images.

- Second, we present a comprehensive comparison with other unsupervised frameworks for image style representation and evaluate the effectiveness of the learned embedding for retrieval and recognition tasks on a variety of datasets, including 2 new datasets. We show that our the proposed approach achieves best overall results across datasets for the retrieval task and best overall results on 5 out of 6 datasets for the recognition task, when compared with several baselines.

To the best of our knowledge, ours is the first work that provides a comprehensive review and evaluation of style representations in an unsupervised setting.

2. Related Work

In recent years, style understanding has become an active field of research in computer vision. In this section, we summarize some of the key works in this area and place our work in context of the state of the art.

Supervised style classification Karayev et al. [11] use many hand-crafted features and features extracted from deep CNNs pre-trained for object recognition task to train linear classifiers in a supervised manner and evaluate recognition performance on three datasets, each with a different meaning of style categories. Aesthetic classification and rating of photographic images has also been explored in [14, 15] using attributes such as depth of field and exposure. Recent methods on style-aware image retrieval and image inpainting [2, 8] use Siamese Networks [20] with a triplet loss for learning style representations and to distentangle style from content. Our choice of triplet loss and some design choices are inspired by success of [2], however the focus of their work is on supervised style retrieval. Recently, Chu and Wu [1] investigated the effectiveness of learned deep correlation features for style classification of paintings and photographs. They use correlation within and across different feature maps (outputs of different convolutional layers) of a pre-trained CNN and train another shallow network on top of these features for dataset-specific style classification.

Representations for automatic style transfer Use of deep correlation features for style representation in [1] is inspired by the seminal work of Gatys et al. [4, 5] for texture synthesis and style transfer. Texture of an image as characterized by deep correlation representation like Gram matrix of feature maps and is shown to disentangle content and style by capturing details like brush strokes, angular geometric shapes, patterns and transition between colours [7]. Lin and Maji [13] also evaluate the efficacy of deep texture representations on texture and scene recognition benchmarks. While style transfer is still an active field of research, in our method we leverage Gram Matrix features as a proxy measure for style similarity.

Automatic discovery of styles Wynen et al. [22] propose an unsupervised learning method to automatically discover, summarize, and manipulate artistic styles from large collections of paintings. They use archetypal analysis on deep image representations (Gram Matrix features [4]) from a collection of artworks, to learn a dictionary of archetypal styles, which are used to characterize a new image by local statistics of deep features. While similar in spirit of unsupervised learning, our work focuses on learning style representation/embedding for retrieval and evaluates it across datasets with different meanings of style.

3. Training Protocol and Data Construction

Instead of leveraging the style class labels specified for a dataset, we learn style representations in an unsupervised manner using data clusters formed using Gram matrix [4, 5]. The details of the training procedures are given later in this section. We first explain the clustering and data construction.

3.1. Training Data Construction

We describe the feature based clustering and triplet formulation which is used later for training a Triplet Network for learning the style representations.

3.1.1 Gram Matrix features based clustering

Feature Extraction As mentioned previously, we wish to learn a style representation without label supervision and use similarity in Gram Matrix as a proxy for loose grouping of dataset images. We use VGG-19 CNN architecture [18] pre-trained for object recognition and localization [16] tasks and extract Gram matrix features as described in [4, 5]. An image is first passed through the CNN and the activations for each layer in the network are computed (shown as $Conv_1$ through $Conv_5$ in Figure 3). As explained in [5] each convolutional layer in the network acts as a non-linear filter bank, and their activations in response to an input image form a set of filtered images referred to as *feature maps*.

A convolutional layer l with N_l distinct filters has N_l feature maps each of size M_l ($M_l = H_l \times W_l$; where H_l and W_l are the height and width of the feature maps in layer l respectively). The responses in layer l can be stored as a matrix $F^l \in R^{N_l \times M_l}$, where $F^l_{i,j}$ is the activation of the i^{th} filter at position *j* in layer *l*. Gram matrix features for layer l are computed as $G^l_{i,j} = \sum_k F^l_{i,k} F^l_{k,j}$. Gram

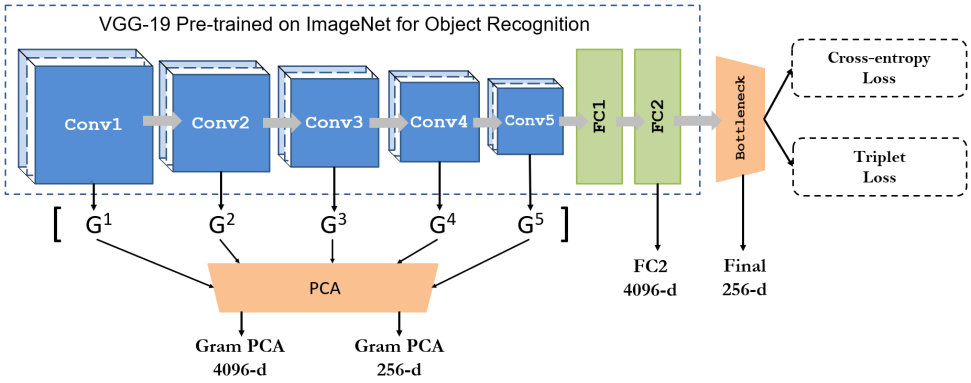

Figure 3. Different feature layers of VGG-19 based CNN used for our experiments.

Matrix features $G^l \in R^{N_l \times N_l}$ are extracted for five layers ($Conv_1$ through $Conv_5$) of the VGG-19 network (shown as G^1 through G^5 in Figure 3).

The resulting Gram Matrix feature vector captures information critical for style texture [6], but has a very high dimensionality (typically of size $\sim 200k$). To make the feature space more compact and computation more efficient, we apply Principal Component Analysis (PCA) to the Gram Matrix style representations and reduce the number of dimensions to 4096 while preserving more than 99% of the variance as shown in [5, 22].

Clustering PCA reduced Gram Matrix features are computed for each image in the training set, followed by soft K-means clustering. The optimal number of clusters for each dataset are determined using *elbow method* as explained in [9]. Clustering on the reduced dimensional Gram Matrix features creates clusters with stylistically similar images coming together. We leverage this style-aware grouping to construct triplets.

3.1.2 Triplet Formulation

Triplet loss tries to enforce a margin between anchor-positive distance and anchor-negative distance in the learned embedding space. Before the training of the Siamese network begins, for every sample in the training data as anchor, K positive and K negative candidates are chosen in an offline pre-processing step as explained below. While training, for each anchor image in a mini-batch, a triplet is formed by randomly selecting a positive and a negative sample for every iteration, from K candidates chosen in the offline process. This procedure is illustrated in Figure 4. This strategy shows a notable improvement in performance than simply pre-selecting the triplets in an offline process.

For selecting positive candidates for an anchor, we pick K nearest neighbors (K-NN) in PCA reduced Gram matrix space (with $K = 40$). Similarly, negative candidates can be selected by picking K furthest neighbors (K-FN). However, due to presence of outliers, this naïve selection strategy results in negative samples with little or no variation irrespective of the anchor image (see last row of Figure 5). For successful learning, we need to mine diverse and informative triplets. Hard negative mining can bring more diversity and relevance to this process [10]. We implement the following two strategies for selecting a diverse pool of negative candidates, but empirically observe the cluster distance based sampling to yield more diverse candidates across queries and datasets.

Random sampling across clusters Given N clusters of training data, for each anchor : (i) randomly sample K images from each cluster except its own, (ii) from the initial set of $(N-1)K$ samples, randomly select K samples as negative candidates.

Cluster distance based sampling Given N clusters of training data, compute a distance between every pair of cluster centers, with D^i_{min} being the nearest cluster distance and D^i_{max} being the furthest cluster distance for cluster i. Let γ denote a value between $(0, 1)$. For an anchor belonging to cluster i, we sample negative candidates as per Gaussian probability distribution with mean (μ) at $\gamma \times \frac{D^i_{min}}{D^i_{max}}$ and standard deviation (σ) as 2% of $(D^i_{max} - D^i_{min})$.

3.2. Training Protocol

We now explain the two training protocols used for style representation learning. The cluster labels are used for learning the embedding by (i) minimizing a cross-entropy loss for cluster label classification, and (ii) minimizing a triplet loss for maximizing the distances between stylistically similar and dissimilar samples.

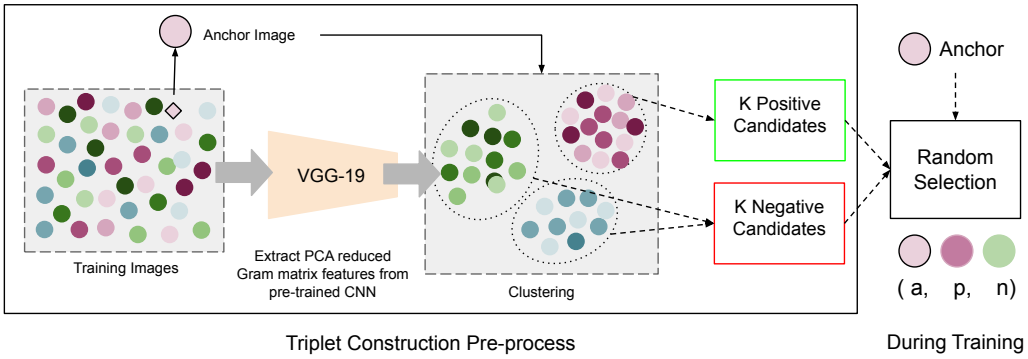

Figure 4. Triplet construction and selection process.

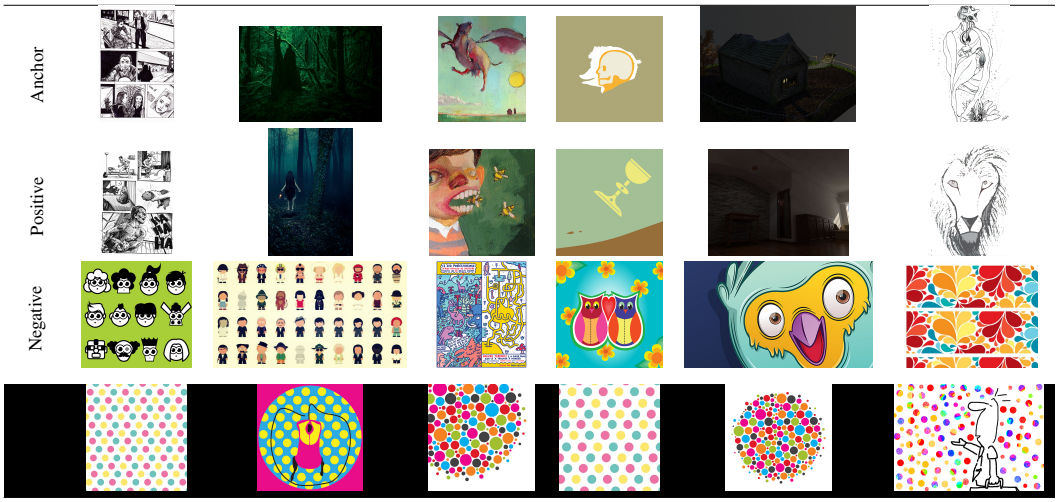

Figure 5. Example triplets sampled with explained procedure in Section 3.1.2 (Cluster distance based sampling). Notice poor diversity for K-FN based negative selection (last row).

Training with cross-entropy loss We train a CNN with VGG-19 architecture [18] augmented by a 256-dimensional bottleneck layer (shown in Figure 3) for 30 epochs and minimize cross-entropy loss for multi-class classification. The use of bottleneck layer results in an improvement in performance for style recognition and retrieval as shown in [2]. During this stage, we simply use the cluster ID for each image as its class label.

Training with triplet loss We train a three branch Siamese network similar to [20] with the same network architecture as above for each branch and minimize a triplet loss similar to [17]. We initialize the network branches with weights from the above protocol and further train the network by minimizing the triplet loss for 50 more epochs. For training a Siamese Network with triplet loss we need triplets (a, p, n) of anchor image a, positive image p (stylistically similar to anchor) and negative image n (stylistically dissimilar) which are sampled as explained in section 3.1.2. The triplet loss is defined as $\mathcal{L}(a, p, n) = max(0, [m + |f(a) - f(p)|^2 - |f(a) - f(n)|^2])$, where m is a margin promoting convergence. The network describes a function $f(.)$ by minimizing the triplet loss defined in equation above. Adam [12] optimization algorithm is used during training of both stages. We will release the network models and training codes along with the paper for ease of reproduction.

4. Datasets

To evaluate our learning protocol and representations across varied style definitions, we use various datasets with diverse media and style categories. We introduce these datasets briefly here and additional details are given in supplementary material.

Behance Artistic Media Dataset (BAM) This dataset by [21] consists of images from *Behance*[1] - a portfolio website

[1] https://www.behance.net/

for professional and commercial artists. The dataset is annotated in a semi-supervised (human-in-the-loop) manner for 7 artistic medium categories (3D renderings, comics, pencil/graphite sketches, pen ink, oil paintings, vector art, watercolor), and 4 emotion categories (happy, gloomy, peaceful, scary). We use a subset of BAM dataset with 121K images (sampled similar to Behance-Net-TT 110K set in [2]) balanced across media and emotional styles, and with a Train, Validation and Test split as 80:5:15.

AVA Style Dataset Introduced in [15, 11], AVA dataset comprises of 14 photographic style labels on 14K images such as Complementary Colors, Duotones, HDR, Image Grain, Light On White, Long Exposure, Macro, Motion Blur, Negative Image, Rule of Thirds, Shallow DOF, Silhouettes Soft Focus, Vanishing Point. Train:Val:Test split is 85:5:10

Flickr This dataset, introduced in [11] captures several different aspects of visual style in photographic images, including photographic techniques (Macro, HDR), composition styles (Minimal, Geometric), moods (Serene, Melancholy), genres (Vintage, Romantic, Horror), and types of scenes (Hazy, Sunny). There are 20 visual styles available on 80,000 images. The Train:Val:Test split is 60:20:20, similar to [11].

Wikipaintings A dataset [11] of paintings annotated with historical art style labels, ranging from Renaissance to Modern Art. We select 25 different styles, and harvest a subset of 25,000 images balanced in style labels. Train:Val:Test split is 85:5:10.

DeviantArt Dataset DeviantArt[2] is a website similar to Behance for amateur artists, with different art style labels. We harvest a dataset of 6500 images from this website for Traditional Art and Digital Art categories. These are further divided into Paintings, Drawings and Mixed Media leading to 5 style classes. Train:Val:Test split is 85:5:10.

WallArt Dataset Wall Art dataset is scraped by us from a home accessories marketplace website *Junique*[3]. The site features handpicked wall art sets each of 2 or 3 artworks that go well together, selected by their in-house curators. Each set is also categorized into one of 13 broader style/theme labels by the curators such as, Country Living, Fashionista, Minimal Monochrome, Fine Art Photography, New Romantic, Shades of Summer, Abstract & Colourful, etc. We mainly use this dataset for qualitative evaluation of retrieval due to the interesting 2-level hierarchy of style relevance (within each set and within each theme).

[2] https://www.deviantart.com/ [3] https://www.juniqe.com/wall-art/inspiration

5. Experiments and Results

Abbreviation	Feature	Dimension	Loss/Training
GM-L	Gram PCA 1	4096	Pretrained
GM-S	Gram PCA 2	256	Pretrained
F×C	Fusion×Content[11]	4000	Pretrained
FC2	Fully Connected[18]	4096	Pretrained
B-Tri	Bottleneck	256	Triplet
B-CE	Bottleneck	256	Cross-entropy

Table 1. Details of feature representations used for performance evaluation and comparison. Refer to Figure 3 for depiction of these representations.

In this section, we evaluate performance of the style representations learned using our proposed approach against other known representations such as PCA-reduced Gram Matrix features and features of [11] on datasets discussed in the previous section. Table 1 provides a summary list of these features with abbreviations for brevity. We use these representations in two ways to establish their effectiveness, (i) for retrieval tasks, to retrieve stylistically similar images in the nearest neighbor sense (ii) for recognition tasks, where we train a softmax classifier on top of the learned representations for image style recognition.

5.1. Retrieval Task

We use the learnt representation to perform retrieval of stylistically similar images on 6 datasets. To evaluate the retrieval performance, we form query sets for each dataset by randomly sampling 10% of the images from the test partition of each dataset (denoted by #Q in Table 2). For every query, we sort the test split samples based on L_2 distance in individual representation space and calculate Average Precision (AP) using dataset specific class labels. The mean Average Precision (mAP) for each dataset and feature representation is provided in Table 2. A Combined Dataset Score (CDS) is computed for each feature, which is the weighted average (in terms of number of queries) of the mAP across datasets. These results demonstrate that the proposed unsupervised learning protocol improves retrieval performance across all but one dataset over pre-trained features. The triplet loss based representation B-Tri does better than cross-entropy based representation B-CE over all datasets as expected, with B-CE being the 3rd best overall. For Wall Art dataset, training was done using a subset of BAM samples due to small size.

Since we do not use class labels for training but use 4096 dimensional PCA reduced Gram features (GM-L) as proxy measure for clustering images, we were initially expecting the 256-dimensional learned representation to at best do as well as GM-L representation. However, B-Tri shows notable improvement in mAP over GM-L. This improvement is the result of the max-margin nature of triplet loss and diverse negative sampling, thus showing the effectiveness of

Dataset	#Q	Random	Feat. Dim: $\sim$ 4096			Feat. Dim : 256		
			F×C	FC2	GM-L	GM-S	B-CE (Ours)	B-Tri (Ours)
AVA Style	200	8.70	19.39	18.98	20.63	20.30	19.87	**21.34**
Flickr	2000	5.63	16.42	15.10	16.21	15.44	16.58	**17.72**
WikiPainting	250	4.56	15.72	15.64	16.99	15.20	17.10	**19.22**
BAM	1000	10.40	27.03	26.57	**34.5**	33.07	28.32	30.54
Deviant Art	100	21.33	35.51	32.82	36.00	35.12	38.80	**40.17**
WallArt	100	8.12	24.96	22.43	27.00	21.15	27.31	**27.53**
CDS (non-weighted)		9.78	23.17	21.92	25.22	23.38	24.66	**26.09**
CDS (weighted)		7.53	26.80	23.77	27.42	25.79	27.06	**28.53**

Table 2. mAPs computed for retrieval on different datasets and features. The learning procedure (Section 3) produces a compact representation B-Tri (256-D) which achieves best performance on 5 out of 6 datasets and best overall CDS. #Q indicate number of query images and CDS indicate Combined Dataset Score (both weighted and non-weighted).

Dataset	Feat. Dim : $\sim$ 4096				Feat. Dim : 256		
	GM-L (All Conv)	GM-L (Conv 5)	F×C [11]	FC2	B-Tri (Ours)	B-CE (Ours)	GM-S (All conv)
AVA Style	48.32	46.96	**58.10**	57.90	53.86	40.74	38.19
Flickr	40.47	39.25	38.80	33.60	**42.15**	36.58	35.80
WikiPainting	51.02	50.92	47.30	35.60	**52.36**	44.37	36.47
BAM	87.81	86.20	82.40	80.10	**89.30**	84.21	80.76
Deviant Art	56.77	55.39	53.20	51.78	**59.74**	52.06	49.03

Table 3. mAPs computed for recognition task on different datasets by training a softmax classifier on top of the features. B-Tri (Ours) performs best on all but the AVA Style dataset, improving the recognition mAP by at least 1.3.

the triplet training.

5.2. Recognition Task

Starting with different unsupervised representations shown in Figure 1, we train a softmax max classifier on the training splits of all datasets and evaluate style classification performance on test splits. The mean Average Precision calculated across all style labels for all datasets is given in Table 3. It can be seen that the triplet loss based unsupervised representation (B-Tri) outperforms pre-trained feature representations for all but the AVA Style dataset. This experiment shows effectiveness of the learned representation for task-specific fine tuning when labels are available.

For AVA Style dataset the Fusion×Content features of [11] performs better. These features combine activations of independently trained content classifier with Fusion features in outer product sense. Karayev et al. [11] suggest that some style categories are inherently content-dependent, hence combining content-classifier activations improves performance. Since labelled data training is not the main focus of this work, we did not pursue this reasoning with our representations.

Also, the combined Gram Matrix features (*All Conv*) perform better than standalone layers($Conv_1$ to $Conv_5$). For detailed information see the supplementary material.

5.3. Qualitative Results for Style based Search

Figure 6 shows the top 4 results for query images from different datasets. As discussed before, style labels are often contextual and convey a limited meaning of style. This indicates that a low precision score does not necessarily imply poor quality of visual similarity. The retrieved results that are highlighted by a black box don't have the same style label as the query, despite obvious visual similarity. For example, the first query (row1, left) belongs to style class 'comic' and retrieved results belong to the classes 'Pen Ink', 'Graphite', 'Pen Ink', 'Gloomy'. We also observe that some style classes are visually more similar as compared to other classes. Figures 2 shows the t-SNE [19] visualisations of the learned representations (B-CE and B-Tri) as compared with pre-trained Gram Matrix features and FC2 features. This further strengthens the fact that triplet based learning improves the stylistic similarity (look and feel wise) after training.

We provide more results and statistics such as confusion matrix per dataset for retrieval task, feature visualizations, clustering performance, and additional qualitative results in the supplementary material.

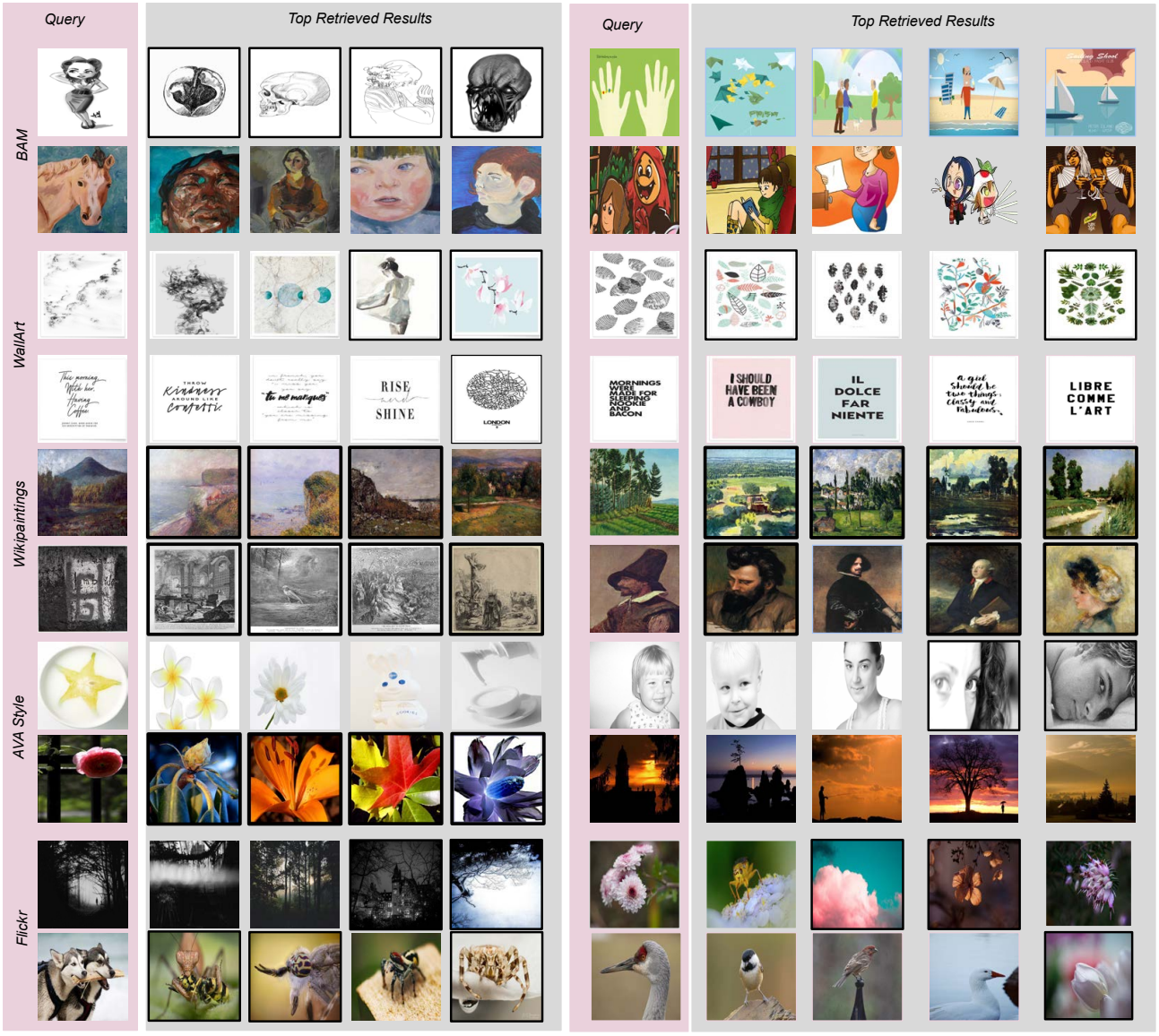

Figure 6. Retrieval results using the best performing representation B-Tri for example queries from different datasets. Images highlighted by black border have style labels different from query style labels although they are visually similar.

6. Conclusion and Future Work

In this work, we proposed a protocol for unsupervised learning of image style representation using Gram Matrix (deep feature correlation map) as a proxy measure of stylistic similarity. Since style is a context-dependent notion, we evaluated performance of the learned representation on a number of datasets with very different definitions of style categorization. We showed that triplet loss based training indeed learns an effective representation that outperforms traditional representations despite being more compact. The sampling scheme introduced for diverse negative sample mining proves useful for improved training. We observed that visual stylistic similarity or 'look and feel' notion of style is not always correlated with style categorization and showed this both qualitatively and quantitatively.

In future, we wish to explore the applications of our protocol with other proxy measures for style-aware grouping, e.g. semantic descriptions for fashion image search. We also wish to expand our unsupervised learning framework such that hierarchies of styles or multiple notions of style can be captured and represented simultaneously.

References

[1] W. Chu and Y. Wu. Image style classification based on learnt deep correlation features. *IEEE Transactions on Multimedia*, 20(9), 2018.

[2] J. Collomosse, T. Bui, M. Wilber, C. Fang, and H. Jin. Sketching with style: Visual search with sketches and aesthetic context. In *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017.

[3] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *Proceedings CVPR*, 2009.

[4] L. A. Gatys, A. S. Ecker, and M. Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[5] Leon Gatys, Alexander S Ecker, and Matthias Bethge. Texture synthesis using convolutional neural networks. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*. 2015.

[6] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. A neural algorithm of artistic style. *CoRR*, abs/1508.06576, 2015.

[7] Leon A. Gatys, Alexander S. Ecker, Matthias Bethge, Aaron Hertzmann, and Eli Shechtman. Controlling perceptual factors in neural style transfer. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[8] Andrew Gilbert, John Collomosse, Hailin Jin, and Brian Price. Disentangling structure and aesthetics for style-aware image completion. In *2018 Conference on Computer Vision and Pattern Recognition (CVPR'18)*, 2018.

[9] Cyril Goutte, Peter Toft, Egill Rostrup, Finn Årup Nielsen, and L. K. Hansen. On clustering fmri time series. *NeuroImage*, 9:298–310, 1999.

[10] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *CoRR*, 2017.

[11] Sergey Karayev, Matthew Trentacoste, Helen Han, Aseem Agarwala, Trevor Darrell, Aaron Hertzmann, and Holger Winnemoeller. Recognizing image style. In *Proceedings of the British Machine Vision Conference*. BMVA Press, 2014.

[12] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2014. URL `http://arxiv.org/abs/1412.6980`. cite arxiv:1412.6980Comment: Published as a conference paper at the 3rd International Conference for Learning Representations, San Diego, 2015.

[13] Tsung-Yu Lin and Subhransu Maji. Visualizing and understanding deep texture representations. In *Proceedings of IEEE CVPR*, 2016.

[14] Luca Marchesotti, Naila Murray, and Florent Perronnin. Discovering beautiful attributes for aesthetic image analysis. *Int. J. Comput. Vision*, 113(3):246–266, July 2015.

[15] Florent Perronnin. Ava: A large-scale database for aesthetic visual analysis. In *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[16] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115 (3), 2015.

[17] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[18] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[19] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

[20] 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 *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[21] Michael J. Wilber, Chen Fang, Hailin Jin, Aaron Hertzmann, John Collomosse, and Serge Belongie. Bam! the behance artistic media dataset for recognition beyond photography. In *The IEEE International Conference on Computer Vision (ICCV)*, 2017.

[22] Daan Wynen, Cordelia Schmid, and Julien Mairal. Unsupervised learning of artistic styles with archetypal style analysis. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 6584–6593. 2018.