Image Representations for Style Retrieval, Recognition and Background Replacement Tasks

Thesis submitted in partial fulfillment
of the requirements for the degree of

Master of Science
in
Computer Science and Engineering by Research

by

Siddhartha Gairola
201402068
siddhartha.gairola@research.iiit.ac.in

International Institute of Information Technology
Hyderabad - 500 032, INDIA
April 2020

Copyright © Siddhartha Gairola, 2020
All Rights Reserved

International Institute of Information Technology
Hyderabad, India

CERTIFICATE

It is certified that the work contained in this thesis, titled "**Image Representations for Style Retrieval, Recognition and Background Replacement Tasks**" by **Siddhartha Gairola**, has been carried out under my supervision and is not submitted elsewhere for a degree.

_____ _____

Date Adviser: Prof. Narayanan P J

To a beautiful Life, a gift from Almighty

Acknowledgments

The completion of this master's thesis and the related publications was the outcome of a great team behind me, without which this would have been an impossible task and I would like to extend my gratitude to all those who made this possible.

First and foremost, I would like to express my sincere gratitude to the captain of our team and my advisor Prof. P.J. Narayanan for his constant and relentless support throughout the course of my research work over the past three years. Computer Graphics will forever be one of the best courses I have taken as a student at IIIT Hyderabad, and it was only back in the Spring of 2016 when Prof. P.J. Narayanan was our course instructor for Graphics, that I was first drawn towards and became curious about the exciting field of Computer Graphics, Computer Vision and Image Processing. His undying love for the subject and crisp delivery of the lectures will be remembered fondly always. I thank him for taking out time from his busy schedule and always having the patience to listen to my raw ideas and providing valuable insight which were key to finding the right answer. He always encouraged me to see the bigger picture and not take the easy way out, at the same time giving me the freedom to try out new things and take responsibility for my actions.

I am indebted to Rajvi Shah (soon to be Dr. Rajvi Shah), my senior at CVIT, a mentor and a friend who mentored me from the very moment I joined the lab. She has been a pillar of support for me, helping me, guiding me, pushing me and when necessary letting me figure things out on my own. She has been like an angel always watching out for me.

I am thankful to all my friends and colleagues at IIIT who made this journey memorable and were by my side through thick and thin. I would like to thank Saumya, my friend and a co-author in my first ever conference publication, for her diligence, help and fruitful discussions we had about work and life. Our gang which comprised of Anshuman, Allen, Dhruv, Jaipal, Jerin, Mahtab, Pinkesh, Punyaslok, Sarthak, Shaleen, and Shubham, thank you for always having my back and getting me out of trouble.

I would like to thank my seniors at the lab Parikshit, Saurabh and Pritish for the insightful discussions we had and providing the much needed guidance whenever required.

I would also wish to thank my dearest friend and bud, Kritika for always being by my side, possessing the ability to make any day awesome and providing valuable reviews and critique to my work.

Finally, I would like to thank my parents and brother for their unconditional love and support towards all my endeavours and being with me every step of the way. It would not have been possible without their sacrifices and belief in me.

Abstract

Replacing overexposed or dull skies in outdoor photographs is a desirable photo manipulation. It is often necessary to color correct the foreground after replacement to make it consistent with the new sky. Methods have been proposed to automate the process of sky replacement and color correction. However, many times a color correction is unwanted by the artist or may produce unrealistic results.

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.

In this thesis, we primarily study various image representations and their applications in understanding visual style and automatic background replacement. First, we propose a data-driven approach to sky-replacement that avoids color correction by finding a diverse set of skies that are consistent in color and natural illumination with the query image foreground. Our database consists of $\sim$1200 natural images spanning many outdoor categories. Given a query image, we retrieve the most consistent images from the database according to L_2 similarity in feature space and produce candidate composites. The candidates are re-ranked based on realism and diversity. We used pre-trained CNN features and a rich set of hand-crafted features that encode color statistics, structural layout, and natural illumination statistics, but observed color statistics to be the most effective for this task. We share our findings on feature selection and show qualitative results and a user-study based evaluation to show the effectiveness of the proposed method.

Next, we propose an unsupervised protocol for learning a neural embedding of visual style of images. 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.

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. Our findings along with the curated outdoor scene database would be useful to the community for future research in the direction of sky-search and sky-replacement.

Contents

Chapter Page

Abstract vi

Introduction xii

1 Introduction 1
1.1 Motivation 1
1.2 Thesis Overview 5
1.2.1 Automatic Sky Search and Replacement 5
1.2.2 Understanding Visual Style 5
1.3 Contributions 7
1.4 Thesis Workflow 8

2 Background and Related Work 9
2.1 Representations for automatic style transfer 9
2.2 Supervised style classification 11
2.3 Automatic discovery of styles 14
2.4 Automatic sky-search and sky-replacement 15
2.5 Realistic image composition 15
2.6 Image Feature Measures 15
2.6.1 Image Statistical Measures 15
2.6.2 Illumination Features 17
2.6.3 Pre-trained CNN Features: 18

3 Data-driven Sky Search and Replacement 19
3.1 Database Collection 19
3.2 Proposed System 20
3.2.1 Feature Description 21
3.2.1.1 Foreground Features 21
3.2.1.2 Illumination Features 21
3.2.2 Candidate search and composition 21
3.2.3 Feature Selection based on composite realism 22
3.2.4 Re-ranking for realism and diversity 23
3.3 Results and Discussion 24
3.4 Summary 28

4 Unsupervised Image Style Embeddings for Retrieval and Recognition . . . 29
4.1 Dataset Creation . . . 29
4.1.1 Training Data Construction . . . 29
4.1.1.1 Gram Matrix features based clustering . . . 30
4.1.1.2 Triplet Formulation . . . 31
4.1.2 Training Protocol . . . 32
4.2 Datasets . . . 33
4.3 Experiments and Results . . . 34
4.3.1 Retrieval Task . . . 34
4.3.2 Recognition Task . . . 35
4.3.3 Qualitative Results for Style based Search . . . 36
4.4 Summary . . . 38

5 Additional Results . . . 39
5.1 Qualitative Results . . . 39
5.2 Confusion Matrix . . . 46
5.3 t-SNE Visualizations . . . 52
5.4 Samples from clustering . . . 57
5.5 Dataset Details . . . 58
5.6 Additional Plots and Tables . . . 62

6 Conclusions and Future Work . . . 64
6.1 Conclusions . . . 64
6.2 Future Work . . . 65

Related Publications . . . 66

Bibliography . . . 67

List of Figures

Figure Page

1.1 For a query image with a dull sky (left), examples of consistent (middle) and inconsistent (right) sky replacements. . . . 2
1.2 An example of paintings by two different artists. Left: Water Lilies by French artist Monet, Right: Dutch artist Van Gogh's Cafe Terrace at Night . . . 3
1.3 Examples of image style categorization with different meanings of style. Each row corresponds to a category based on a particular understanding of style. . . . 4
1.4 t-SNE [41] 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 4.1 and Figure 4.1 for more details on the representations). . . . 6

2.1 Examples of images that combine the content of a photograph with the style of several well-known artworks. The original photograph **A** is used as the content image in all the examples. The painting that provided the style for the respective generated image is shown in the bottom left corner of each panel. (Gatys et al. [7]) . . . 10
2.2 Content representations and style representations from a deep CNN (VGG-19 [33]) computed across different layers. (Gatys et al. [7]) . . . 12
2.3 An illustration of the VGG-19 model [33] network architecture : conv means convolution, FC means fully connected (Zheng et al. [52]). . . . 13
2.4 Figure 2 from [50]: Archetypes learned from the GanGogh collection and van Gogh's paintings. Each row represents one archetype. The leftmost column shows the texture representations, the following columns the strongest contributions from individual images in order of descending contribution. Each image is labelled with its contribution to the archetype. . . . 14
2.5 Correlated Color Temperature (CCT) chart . . . 16
2.6 Top: HSV color cylinder; Bottom: Illumination features computed as a function of sun zenith angle ($\Delta\phi_s$), sun azimuth angle (θ_s) and a binary variable (v_s) for sun visibility. 17

3.1 For a query image with a dull sky (left), examples of consistent (middle) and inconsistent (right) sky replacements. . . . 19
3.2 An overview of our sky-replacement pipeline. Candidate composites are created using skies from database images with most similar foregrounds. Final composites are re-ranked to maximize realism and diversity of the presented set. . . . 20

3.3 An overview of the sky replacement step. 22
3.4 Ablation study : Running average of realism scores for composites sorted on feature distances 23
3.5 Comparison with existing state of the art method, [38]. We tested our model on the same input image, the results obtained are just as aesthetically appealing using a completely different pipeline. 25
3.6 Example illustrating the efficacy of the re-ranking. 26
3.7 Failure of colour transfer methods in the in-house implementation of [38] as compared with our method which chooses skies that are already compatible with the foreground. 26
3.8 Failure cases, from left, (i) segmentation error, (ii) inconsistent sky reflection in water, (iii) bright (sun) spot, (iv) better composition achieved with use of illumination map. . 27
3.9 Example results of our diverse and compatible sky-replacement system 27

4.1 Different feature layers of VGG-19 based CNN used for our experiments. 30
4.2 Triplet construction and selection process. 30
4.3 Example triplets sampled with explained procedure in Section 4.1.1.2 (Cluster distance based sampling). Notice poor diversity for K-FN based negative selection (last row). . 31
4.4 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. 37

5.1 Nearest Neighbour retrieval results for select queries from BAM subset test split. Notice that for rows 1 and 2, the queries and neighbours are very similar looking but the labels do not match. This indicates the lower mAP scores for retrieval using unsupervised methods. ‘Oil Paint’ and ‘Water Colour’ are hard to differentiate, similarly ‘Gloomy’ and ‘Peaceful’ 40
5.2 Retrieval Results for Query and Top Neighbours Deviantart dataset. 41
5.3 Retrieval Results for Query and Top Neighbours AVA Style dataset. 42
5.4 Retrieval Results for Query and Top Neighbours Wikipaintings Subset dataset.It is interesting to see the retrieved results and their relevance with respect to the query image. Notice row 7 where, ‘Abstract Expressionism’ labelled query retrieves ‘Ukiyo-e’, ‘Cubism’ and ‘Pop Art’ paintings. 43
5.5 Retrieval Results for Query and Top Neighbours Flickr Test Set. 44
5.6 Retrieval Results for Query and Top Neighbours WallArt dataset. The style themes for this dataset have been manually curated by experts, the retrieved samples show similarity both in terms of appearance and style themes. 45
5.7 Confusion Matrix for Top 100 retrievals for 1000 Query images on Behance Subset Test set using learnt representations. Here we see the following pairs confusing with each other - ‘Watercolor’ with ‘Oilpainting’ since both are very colourful, ‘Graphite’ and ‘Pen Ink’ both are hand-drawn and dull, and ‘3D Graphics’ with ‘Vectorart’. 46
5.8 Confusion Matrix for Top 100 retrievals for 1000 Query images on Wikipaintings Subset Test set using learnt representations. 47
5.9 Confusion Matrix for Top 100 retrievals for 1000 Query images on Flickr Test set using learnt representations. 48
5.10 Confusion Matrix for Top 20 retrievals for 100 Query images on WallArt Test set using learnt representations for 13 style themes. 49

5.11 Confusion Matrix for Top 100 retrievals for 200 Query images on AVA Style Test set using learnt representations. 50
5.12 Confusion Matrix for Top 50 retrievals for 100 Query images on Deviant Art Test set using learnt representations for 5 labels. 51
5.13 t-SNE visualization on BAM dataset for FC2 pre-trainined features (4096-D) from VGG19. 52
5.14 t-SNE visualization on BAM dataset for PCA-reduced Gram Matrix (4096-D) pre-trained features from VGG19. 53
5.15 t-SNE visualization on BAM dataset for PCA-reduced Gram Matrix (256-D) pre-trained features from VGG19. 54
5.16 t-SNE visualization on BAM dataset for B-CE (256-D) features learnt when training with cross-entropy loss using cluster cluster id for each image as its class label. 55
5.17 t-SNE visualization on BAM dataset for B-Tri (256-D) features learnt when training with triplet loss. Notice that using triplet loss (B-Tri) further reinforces the stylistic similarity in comparison to other features as can be seen from Figures 5.13, 5.14 and 5.15. 56
5.18 Each row shows examples drawn randomly from seven clusters, for clustering applied to BAM [48] subset. It can be seen that clustering in Gram matrix space groups stylistically similar images together.(Each row only contains samples from a single cluster) 57
5.19 Dataset wide mAP scores for style based classification using different features. Notice that B-Tri features clearly show improvement over other features across most datasets. 62
5.20 Dataset wide mAP scores for retrieval performance using different features. Notice that B-Tri and B-CE features clearly show improvement over other features across most datasets. 62

List of Tables

Table		Page
3.1	Statistics on user preferences	25
4.1	Details of feature representations used for performance evaluation and comparison. Refer to Figure 4.1 for depiction of these representations.	34
4.2	mAPs computed for retrieval on different datasets and features. The learning procedure (Section 4.1) 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).	35
4.3	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.	35
5.1	Ava Style dataset (a subset of AVA dataset [26]) similar to [17] style categories and the number of images in each category.	58
5.2	DeviantArt dataset style categories and the number of images in each category.	58
5.3	Flickr dataset [17] style categories and the number of images in each category.	59
5.4	WallArt dataset style categories and the number of images in each category.	59
5.5	Wikipaintings Subset dataset, which is a subset of Wikipaintings dataset [17] style categories and the number of images in each category as used for our experiments.	60
5.6	Behance Style Subset datset style classes and the number of images in each category as used for our experiments, which is a subset of BAM dataset [48] very similar to the Behance-Net-TT used in [4].	61
5.7	mAPs for gram matrices computed for different layers (conv1-conv5) of VGG19 [33] Network for recognition using a softmax classifier on different datasets and features. Evidently a combination of all convolutional layers performs best.	63

Chapter 1

Introduction

The primary objective of this thesis is to study various image representations and their applications in understanding visual style, image retrieval, image recognition and background replacement. In this chapter, we describe the motivation of our research, use of image representations in understanding visual style and automatic background replacement. We then describe our contributions followed by the organization of our thesis work.

1.1 Motivation

Background replacement: With the ubiquity of smart phone cameras, photography has become a democratized hobby with millions of photos uploaded to social media platforms like Instagram, Flickr, Facebook every day. Along with this comes the need for sharing perfect photographs, however, the captured shots are often unattractive due to undesirable backgrounds, occlusions, poor lighting or exposure, motion blur, lack of smile, presence of eye blinks, etc. In recent years, many methods have been proposed for a number of automatic photo enhancements. This thesis focuses on the problem of automatic sky-replacement.

Sky is often the hardest part of the scene to perfect in outdoor photography. Depending upon the geographic location and weather conditions, sky could persistently be gray and dull, or too bright. Even when the sky is perfect blue with white clouds and looks beautiful to the naked eye, it most often gets washed out in a single exposure shot captured with a standard smart-phone camera. Professional outdoor photographers often prefer the golden hour (when sun is closer to the horizon) or use specifically designed filters and polarizers to overcome this problem. Multi-exposure (HDR) photography can alleviate this problem to some extent, however, not much can be done if at the time of capture sky is just dull.

Professional digital artists perfect the bad-sky photographs by manually replacing the original sky with a desirable one and performing a series of interactive corrections to make the sky and the foreground consistent with each other while keeping the final composite 'plausible'. This is a non-trivial and time consuming edit that is too cumbersome for a naïve user to perform. Recently Tao et al. [37] proposed an automatic method for sky-replacement that performs semantic-aware color transform

Figure 1.1: For a query image with a dull sky (left), examples of consistent (middle) and inconsistent (right) sky replacements.

on the foreground to achieve natural looking composites. However, color-correction is not always desirable. Hence, we propose a different approach to sky-replacement that avoids or minimizes the need for post-replacement color corrections[27].

Visual Style: 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 1.3 [1]. A highly subjective construct like style is hence, difficult to model computationally. In the context of computer vision, Karayev et al. [17] 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. [48] presented a large dataset of contemporary artworks – the 'Behance Artistic Media Dataset' (BAM) with crowd-sourced labels for media, emotions, and objects.

The style of an image, a photograph or art plays a very important role in how it is perceived and felt by the viewer. A beautiful piece of art form can generate happy or cheerful emotions within us, on the other hand a dull image may cause the viewer to be pensive or reflective. One may contrast the different artistic styles, painting techniques of different painters, different art forms or even different artistic time periods. Art experts may easily describe to us how an artistic piece by Monet may be very different from one by Van Gogh based on colors, brush strokes and emotions (see Figure 1.2). In recent times, apps like Prisma[2] (which transforms an image into an artistic effect) have gained huge popularity, which is nothing but a deep convolutional neural network performing style transfer.

Implicit vs Explicit Style: 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 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 [7, 8] model style implicitly as intermediate feature representations learned from an unrelated supervised learning task such as object recognition. Style, in

[1] Note: The images shown in the figure are part of the BAM [48], Wikipaintings [17], Flickr [17] and AVA Style [17, 26] datasets.

[2] *https://prisma-ai.com/*

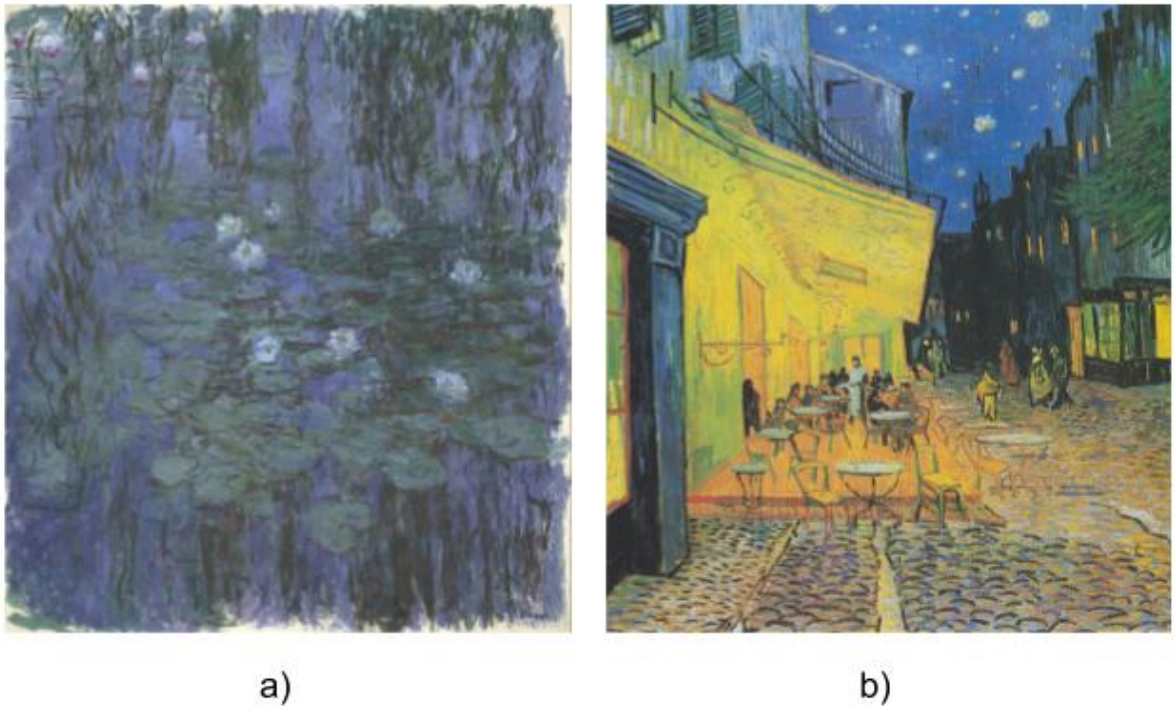


Figure 1.2: An example of paintings by two different artists. Left: Water Lilies by French artist Monet, Right: Dutch artist Van Gogh's Cafe Terrace at Night

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 [33]), typically pre-trained for object classification on a very large dataset like ImageNet [5]. 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 [3, 4, 11, 17, 48]. 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 thesis 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.

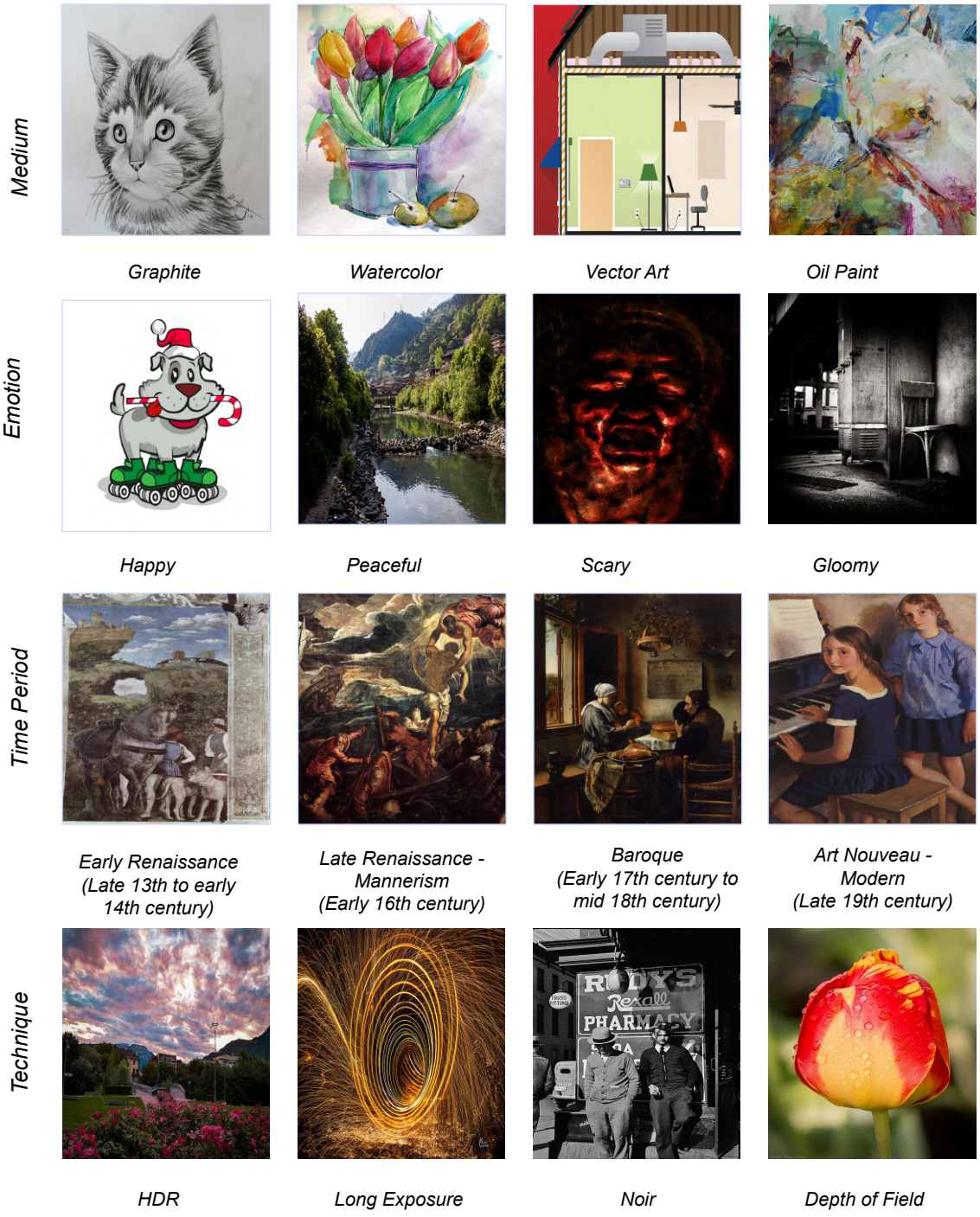


Figure 1.3: Examples of image style categorization with different meanings of style. Each row corresponds to a category based on a particular understanding of style.

To summarize, the key motivations of this thesis are (i) to develop a method for automatic sky-search and sky-replacement without the need for post-replacement color corrections, and (ii) to investigate the quality of understanding of visual style which can be achieved by an unsupervised approach which does not rely on any categorical labels of style.

1.2 Thesis Overview

1.2.1 Automatic Sky Search and Replacement

We develop a method for automatic sky-replacement without the need for post-replacement color corrections. Our approach is data-driven and centered around the idea of 'compatible' sky-search. Given a query image with a problematic sky, our method first finds images with similar foregrounds and natural illumination. It then creates candidate composites by replacing the query image sky with the retrieved image skies and ranks the composites based on realism and diversity. The user is finally presented with the top-k candidate composites as replacement outcomes without color transfer thereby retaining the natural color composition of the foreground in the original image. We demonstrate the effectiveness of our method with qualitative results and a comprehensive user study. Figure 3.2 summarizes the proposed system with a block diagram.

For retrieving compatible yet useful images, we curated a dataset of 1246 outdoor images spanning many outdoor categories with interesting skies from ADE20K dataset [54] and the dataset of [38]. To achieve compatible sky-search, we use an ensemble of hand-crafted features such as Color Statistics (Correlated Color Temperature (CCT), Luminance, and Saturation histograms), GIST [25], Bag of Words[34], and natural illumination statistics [20] (represented as a probability map of sun position in the sky), as well as CNN features (pre-trained). These features encode rich information about color distribution, structural layout, semantics, and natural illumination. We finally select the color statistical features, as we found based on an ablation study that the composites produced using the retrieval results with these features were most realistic. We evaluate the composite images using RealismCNN [55] – a discriminative model trained to predict realism of an image. Section 3.2 explains the data collection, feature selection, and re-ranking based on realism and diversity in detail.

1.2.2 Understanding Visual Style

We wish to examine the quality of understanding of visual style which can be achieved by an unsupervised approach that does not rely on any 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

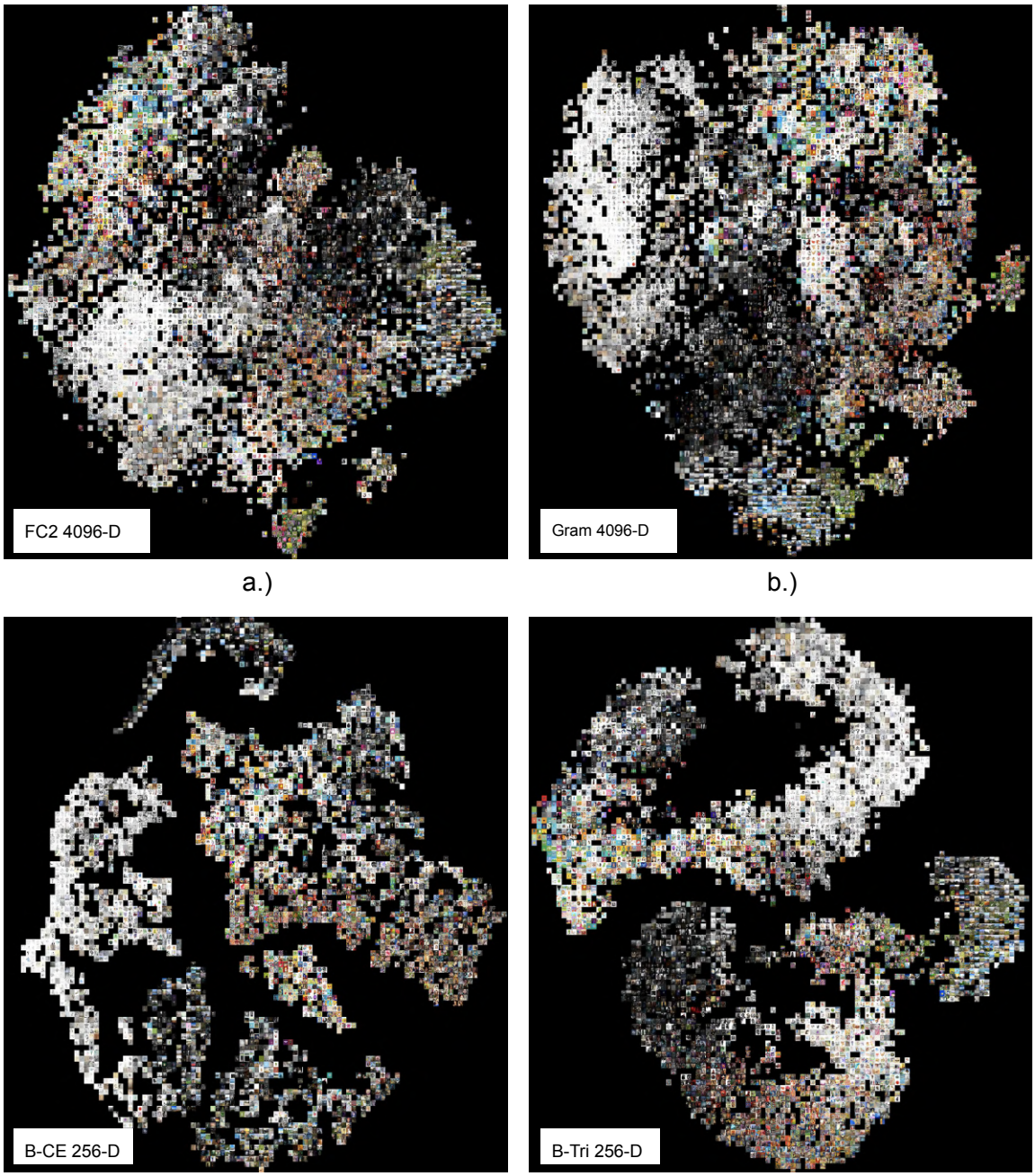


Figure 1.4: t-SNE [41] visualizations of BAM dataset images based on following feature representations: (top row) FC2 features and PCA-reduced Gram features computed from pre-trained VGG19, (bottom row) embeddings learned using our protocol. It can be observed that using triplet loss (B-Tri) further reinforces the stylistic similarity in comparison to other features (refer to Table 4.1 and Figure 4.1 for more details on the representations).

for discerning different visual styles. We train a Siamese CNN [45] 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 by the t-SNE [41] visualizations in Figure 1.4. 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.

1.3 Contributions

In this thesis, we propose a data-driven approach to sky-replacement that avoids color correction by finding a diverse set of skies that are consistent in color and natural illumination with the query image foreground. Given a query image, we retrieve the most consistent images from a database of ~ 1200 natural images spanning many outdoor categories, according to L_2 similarity in feature space and produce candidate composites. The candidates are then re-ranked based on realism and diversity. We used pre-trained CNN features and a rich set of hand-crafted features that encode color statistics, structural layout, and natural illumination statistics, but observed color statistics to be the most effective for this task. We share our findings on feature selection and show qualitative results and a user-study based evaluation to show the effectiveness of the proposed method.

Also we propose an unsupervised protocol for learning a neural embedding of visual style of images. The proposed protocol does not leverage categorical labels but a proxy measure for forming triplets of anchor, similar and dissimilar images. We use these triplets to 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 labes, we obtain best results for image style recognition tasks on five of the six datasets.

To summarize, the following are the **key contributions** of this thesis:

- We present a novel pipeline for compatible-search based sky-replacement that is a useful alternative or prelude to automatic color transfer based methods.
- We curated a large database of outdoor images with interesting skies and evaluated usefulness of a large number of features (both hand-crafted and deep learned) for this task.
- 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.

- 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. Our findings along with the curated outdoor scene database would be useful to the community for future research in this direction.

1.4 Thesis Workflow

This thesis comprises of 6 main chapters. Chapter 1 discusses the motivation of the problem being solved, provides an overview to visual style, implicit vs explicit style and automatic sky search and replacement, and finally summarizes the key contributions of the thesis.

Chapter 2 gives an overview of the prior work on methods for automatic sky-search, sky-replacement and realistic image compositing techniques, representations used for automatic style transfer, supervised style classification methods and protocols for automatic discovery of image styles.

Chapter 3 explains the proposed method for the task of color consistent sky search and replacement.

Chapter 4 discusses the proposed method for learning unsupervised image style embeddings for the tasks of retrieval and recognition.

Chapter 5 provides additional results in the form of qualitative results, dataset details, confusion matrices, feature visualizations and some additional plots and tables.

Finally, chapter 6 summarizes and concludes the main findings of the thesis, discusses the limitations and provides final remarks for future consideration.

Chapter 2

Background and 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. This is followed by discussing the existing methods which are closely related to the task of image rendering and compositing within the context of sky background stylization and replacement. Finally there is a brief explanation of the statistical and deep learned features used for understanding the realism of image composites, which are used in our proposed method for sky-search and sky-replacement.

2.1 Representations for automatic style transfer

Style Transfer: Use of deep correlation features for style representation [3] is inspired by the seminal work of Gatys et al. [7, 8] for texture synthesis and style transfer. Texture of an image as characterized by deep correlation representations like Gram matrix of feature maps is shown to disentangle content and style by capturing details like brush strokes, angular geometric shapes, patterns and transition between colours [10]. This Gram matrix representation from different convolutional layers of a network pre-trained for the task of object classification on the ImageNet [5] dataset is used for the task of *style transfer*, that transfers a photo into an image with a style similar to a given target painting. Given a photographic image P and a reference painting image I with the target style, Gatys et al. [7] propose to adjust the output image $\hat{P}$ in such a way that the Gram matrix (style features) of feature maps are similar to those of I, while the content features (regular feature maps) of $\hat{P}$ are similar to those of the original input image P. To perform style transfer gradient descent is performed on the white noise initialized image $\hat{P}$ to match the style of the source image while preserving the content of the target image. See Figure 2.1 for examples of output image after performing the aforementioned style transfer procedure.

The method described above [7] is computationally very expensive. A few works were proposed to make it more efficient and fast. Johnson et al. [16] proposed perceptual loss functions for constructed transformation networks. These are defined based on high level features from a loss network. The perceptual loss functions are shown to capture image similarities more robustly, and the transformation networks make the process of style transfer much more efficient. Later, Ulyanov et al. [39] proposed a feed-forward convolutional neural network to shift the optimization step of the style transfer process to

Figure 2.1: Examples of images that combine the content of a photograph with the style of several well-known artworks. The original photograph **A** is used as the content image in all the examples. The painting that provided the style for the respective generated image is shown in the bottom left corner of each panel. (Gatys et al. [7])

the learning stage, and this made style transfer much more light weight. They designed a new normalization technique and a learning formulation to further improve the quality and diversity of stylization [40]. The learnt networks described are efficient but are restricted to a single style only. To transfer a specific style onto an image, a separate network has to be trained from scratch. Dumoulin et al. [6] proposed a conditional instance normalization technique, and developed a generic network to capture the properties of different image styles in a flexible manner. Further, Ruder et al. [29] extented the image style transfer procedure to videos. In addition to simply performing style transfer on each video frame, they introduced a temporal consistency loss and a multi-pass algorithm to make the video transformations results smooth. Tanno et al. [36] extended the style transfer network [16] to further learn multiple artistic styles at the same time, and also reduced the computational requirement to develop a real-time style transfer application on mobile devices. The style transfer performed using convolutional neural networks is not semantics aware. To take semantics into account, Champandard [2] consider a semantic map corresponding to the input image, such that a user can sketch a spatial layout associated with semantic meanings, and following this the proposed system can synthesize a fine artwork with the specified style conforming both to the semantics and sketched layout. Despite neural style transfer achieving amazing results, why these correlations between the feature maps from a pre-trained neural network can capture style so well is unclear. Li et al. [21] proposed a novel interpretation to show how matching the Gram matrix from feature maps is equivalent to minimizing the maximum mean discrepancy with the second order polynomial kernel.

Lin and Maji [22] 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.

2.2 Supervised style classification

Additionally to style transfer, a few efforts have been made to use style representations in image style analysis tasks like style classification, style search and style retrieval. Wang et al. [47] construct a style representation and color representation for handbags, based on discovery of discriminative patches and dominant color features respectively. The handbags are first classified into different style classes, and further within each class, handbags of varied colors are discriminated. These style and color representations are used to measure the inter-class style similarity and intra-class color variations respectively. Based on the Gestalt theory (which is a psychological study of how human visions organize the visual perception), Shen and Cheng [32] aim to improve the usability of local features in images of the same content but in different styles, by proposing Gestalt feature points. It was demonstrated that these Gestalt feature points give superior performance over the existing local features.

Karayev et al. [17] use many hand-crafted features and features extracted from deep CNNs pretrained 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. They showed that

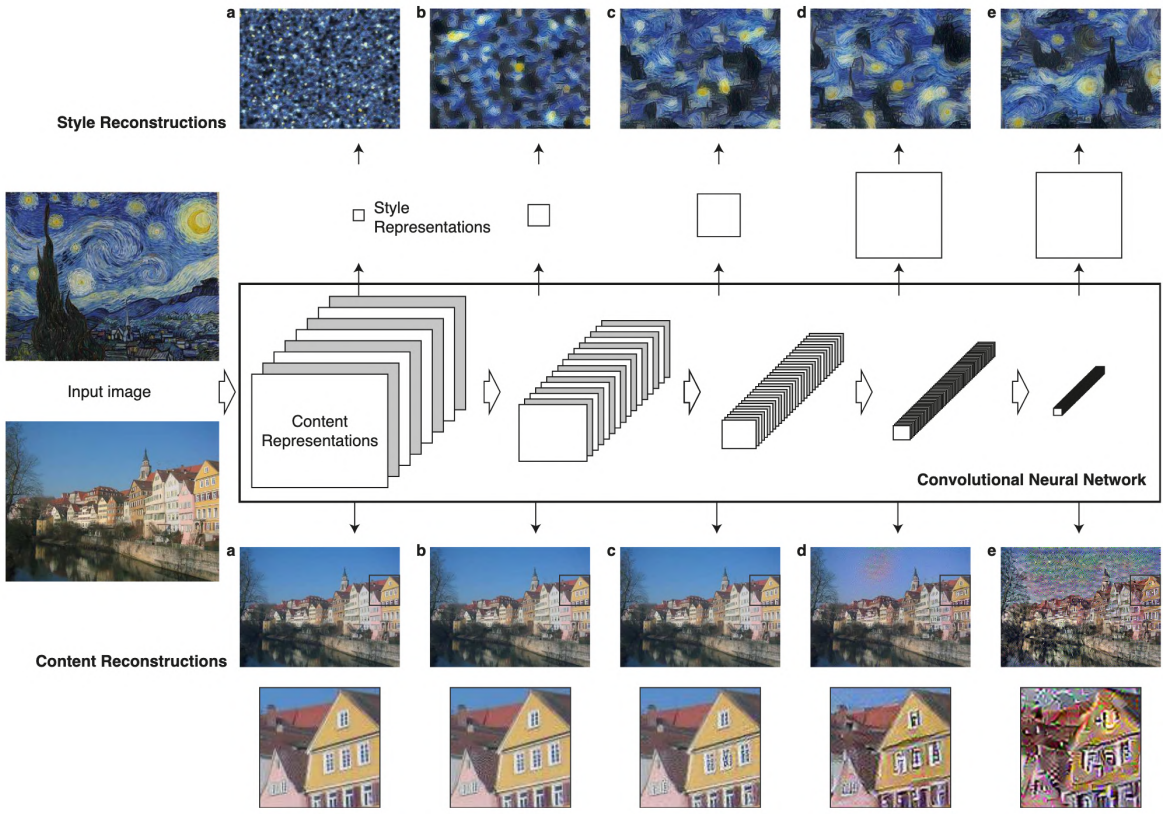


Figure 2.2: Content representations and style representations from a deep CNN (VGG-19 [33]) computed across different layers. (Gatys et al. [7])

the deep features yield performance that is much better than the conventionally used hand-crafted features in recognizing different image styles. Aesthetic classification and rating of photographic images has also been explored in [23, 26] using attributes such as depth of field and exposure. Recent methods on style-aware image retrieval and image inpainting [4, 11] use Siamese Networks [45] 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 [4], however the focus of their work is on supervised style retrieval.

Deep Correlation Features: Recently, Chu and Wu [3] 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. In addition to the Gram matrix based style representations [7], they investigate the performance of various other correlations that are well defined in statistics theory.

VGG-19 [33] network is used which has been pre-trained on ImageNet[5] dataset for the task of image classification. Filter responses are extracted from different layers of VGG-19. This network

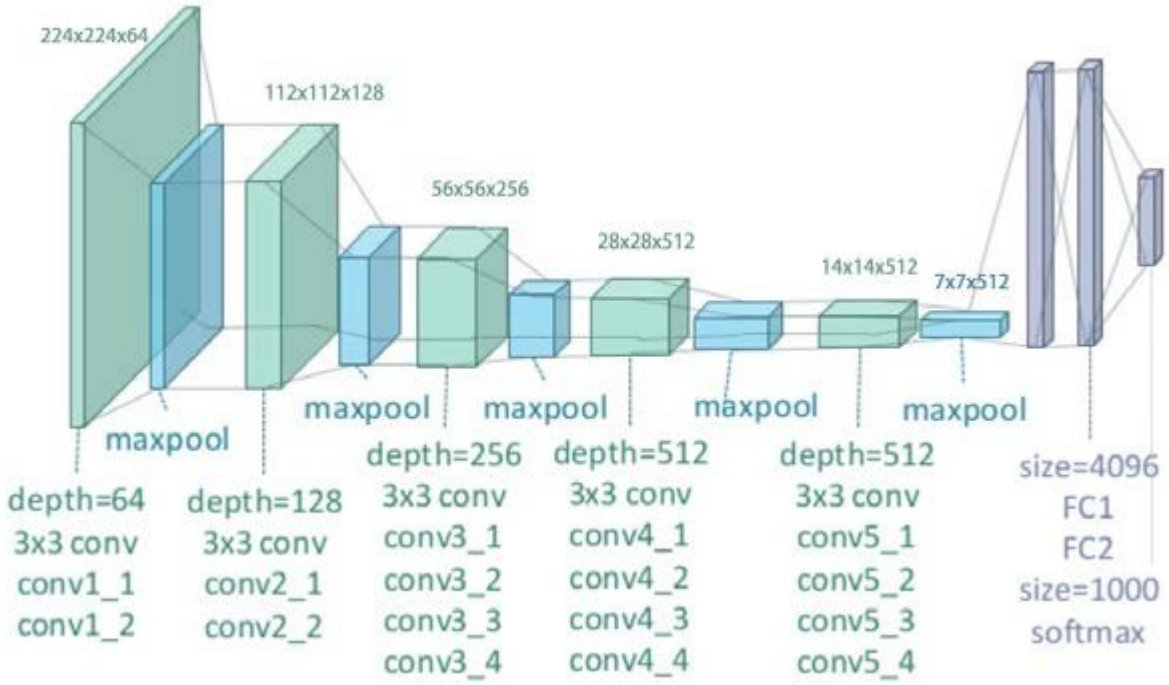


Figure 2.3: An illustration of the VGG-19 model [33] network architecture : conv means convolution, FC means fully connected (Zheng et al. [52]).

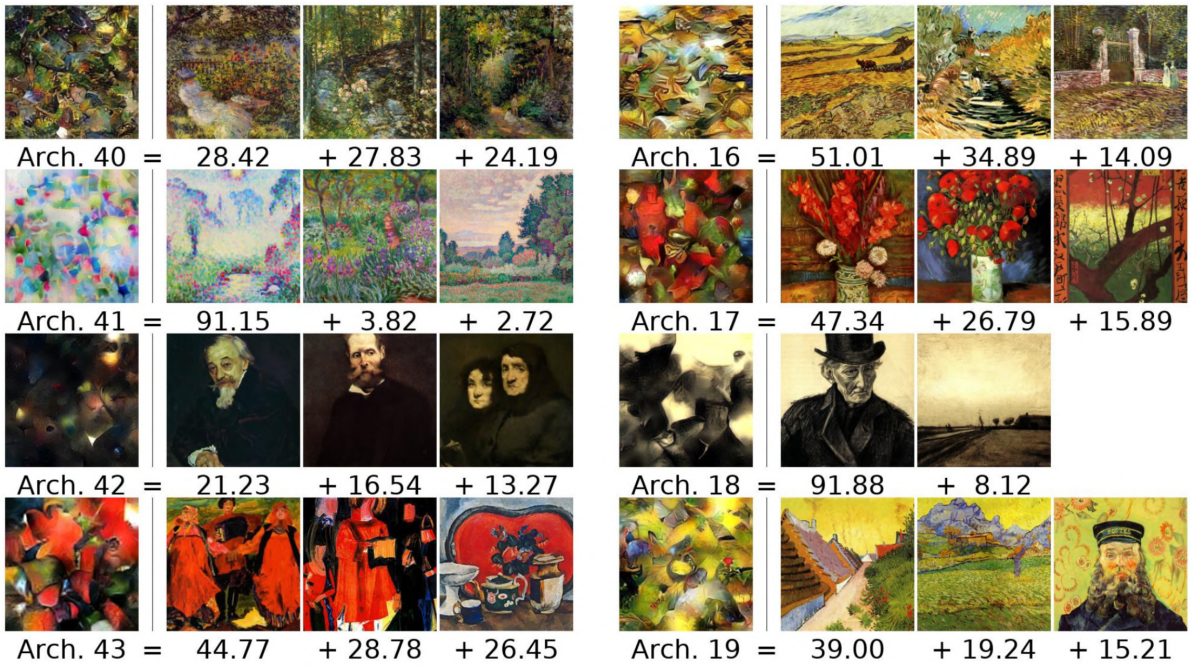
consists of sixteen convolutional layers and three fully-connected layers. The receptive field at each convolutional layer is fixed to 3×3 with a stride of 1 pixel. Spatial pooling is performed out by five max-pooling or average-pooling layers, which follow the 2^{nd}, 4^{th}, 8^{th}, 12^{th} and 16^{th} convolutional layers. Max-pooling (or average-pooling) is performed over a 2×2 window, with a stride of 2. The convolutional layers in the network are divided into five groups due to the presence of five pooling layers. In [7], the convolutional layers described were named as 'conv1_1', 'conv1_2', 'conv2_1', 'conv2_2', 'conv3_1', 'conv3_2', 'conv4_1', 'conv4_2', and so on. For example, the 'conv2_1' layer is the 3rd convolutional layer that just follows the first pooling layer, see Figure 2.2, 2.3 and 4.1.

Gram Matrix Features: As introduced before in section 2.1, Gatys et al. [7] construct a style representation based on correlations between feature maps (filter responses), which is used to transform a photograph into an image with a particular target style. The style representations are correlations represented by the Gram Matrix $G^l \in R^{N^l \times N^l}$, where G^l_{ij} is the inner product between the vectorized feature map i and j in layer l, i.e.

$$G^l_{ij} = \sum_k F^l_{ik} F^l_{kj}$$

, where F^l_{ik} is the activation of the i^{th} filter response at the position k in layer l. For example, the 'conv5_1' layer of the VGG-19 [33] network model, there are 512 feature maps with both width and height of each feature map as 14. Each feature map is then vectorized into $14 \times 14 = 196$ dimensional

vector. These 512 feature maps are stacked together to form F^5 and the resulting Gram Matrix is a symmetric 512×512 dimensional matrix. Chu and Wu [3] use these Gram matrix features as style vector representations which are then classified by an SVM classifer.



(a) Archetypes from GanGogh collection.

(b) Archetypes from van Gogh's paintings.

Figure 2.4: Figure 2 from [50]: Archetypes learned from the GanGogh collection and van Gogh's paintings. Each row represents one archetype. The leftmost column shows the texture representations, the following columns the strongest contributions from individual images in order of descending contribution. Each image is labelled with its contribution to the archetype.

2.3 Automatic discovery of styles

Wynen et al. [50] 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 [7]) 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. To visualise what an archetype 'looks like', the authors in [50] synthesise a texture from an image filled with random noise, using the style representation of the archetype. Figure 2.4 shows some examples, with the synthesised archetypal textures in the left-most column and the three images next to them on each row showing the individual images that made the strongest contribution to the archetype.

While similar in spirit of unsupervised learning, our work focuses on learning style representations/embeddings for retrieval and evaluates it across datasets with different meanings of style.

2.4 Automatic sky-search and sky-replacement

Tao et al. [37] proposed an interactive search system using a set of semantic sky attributes (category, layout, richness, horizon, etc.) and showed how it can be used for controllable sky replacement. However the sky segmentation and consequently horizon detection introduce errors in sky replacement. The quality of the sky replacement procedure is measured by a simple geometric metric to score compatibility, and global color transfer is used to match the output image appearance. Tsai et al. [38] proposed a data-driven sky search scheme based on semantic layout of the input image. To re-compose the stylized sky with the original foreground naturally, an appearance transfer method is developed to match statistics locally and semantically. However, the color transfer algorithm is linked with label matching between the source and the target which adds both complexity and a limitation on the kind of source images that can be used. Also, color transfer may be undesirable and may introduce artefacts in the foreground regions. In contrast, we do not rely on similar sky replacement methods and also do not need to use appearance transfer methods.

2.5 Realistic image composition

Much work has been done for realistic image composition [46] and for evaluating realism of composites[44, 49]. Lalonde and Efros [19] propose an object insertion technique that searches for objects that are consistent with the input photograph in terms of camera orientation, lighting, resolution, etc. and uses feature based assessment of composite realism. Xue et al. [51] determine the key statistical measures that influence the realism of a composite and then adjust these in a given query composite automatically using a data-driven algorithm. In this work, we leverage the implicit correlation between background and foreground regions in natural images for compatible sky-search that lead to more realistic composites.

2.6 Image Feature Measures

In this section, we explain the underlying statistical and deep learned features used for understanding the realism of image composites, which are used in our proposed method for sky-search and sky-replacement.

2.6.1 Image Statistical Measures

The image pixels are first inversely Gamma corrected [35], before we compute the image statistics on a given input image. Then we transform the image statistics in such a way that they are approximately linear to human visual perception. Weber's law is followed to convert luminance and saturation into log domain, and CCT is defined by mired [24].

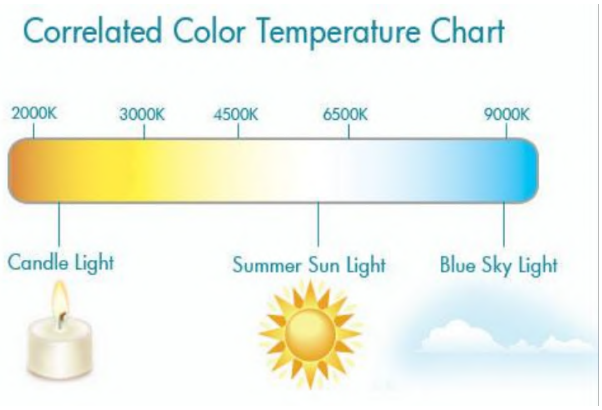


Figure 2.5: Correlated Color Temperature (CCT) chart

Correlated Color Temperature (CCT): These capture the colour characteristics of light. CCT feature vector gives a general indication of the apparent "warmth" or "coolness" of the light emitted by the source. As was studied in [51], we use "mired" as the unit of CCT.

$$1\, mired = 106/K\ ,$$

where K is the $Plankian$ color temperature in $Kelvin$, clipped in the range of $[1500, 20000]$ that is also the normal range of natural lighting. CCT is computed using the package Opt-Prop (using [43]).

Equation 2.1 gives the computation for CCT features for an RGB image

$$CCT = 449 \times n^3 + 3525 \times n^2 + 6823.3 \times n + 5520.33\ , \tag{2.1}$$

where $n = ((0.23881) \times R + (0.25499) \times G - (0.58291) \times B)/((0.11109) \times R - (0.85406) \times G + (0.52289) \times B)$ and R, G, B are the red, green, blue color intensity values at that particular pixel location of the image respectively.

Luminance: Luminance is the amount of energy perceived by an observer from a light source. Similar to [51], $log_2 Y$ is used, where Y (normalized to $[\epsilon, 1.0]$) is the luminance channel of xyY space, where $\epsilon = 3.03 \times 10^{-4}$ (corresponding to intensity 1 in a $0 - 255$ grayscale image before inverse Gamma correction is performed) and is used to avoid undefined log values. The unit of difference in log_2 domain is a $stop$.

Saturation: It is the degree to which a pure color is dilated by white light. It is computed as $log_2 S$, where $S \in [\epsilon, 1.0]$ is the saturation channel of HSV space. HSV color space has a cylindrical geometry Figure 2.6, with hue, their angular dimension, starting at the red primary at $0°$ passing through the green primary at $120°$ and the blue primary at $240°$, and then wrapping back to red at $360°$. The central vertical axis comprises the neutral, achromatic, or gray colors, ranging from black at lightness 0 or value 0, the bottom, to white at lightness 1 or value 1, the top.

H: H is a circular value in $[0.0, 1.0]$ (or $[0, 360]$), the hue channel of HSV space.

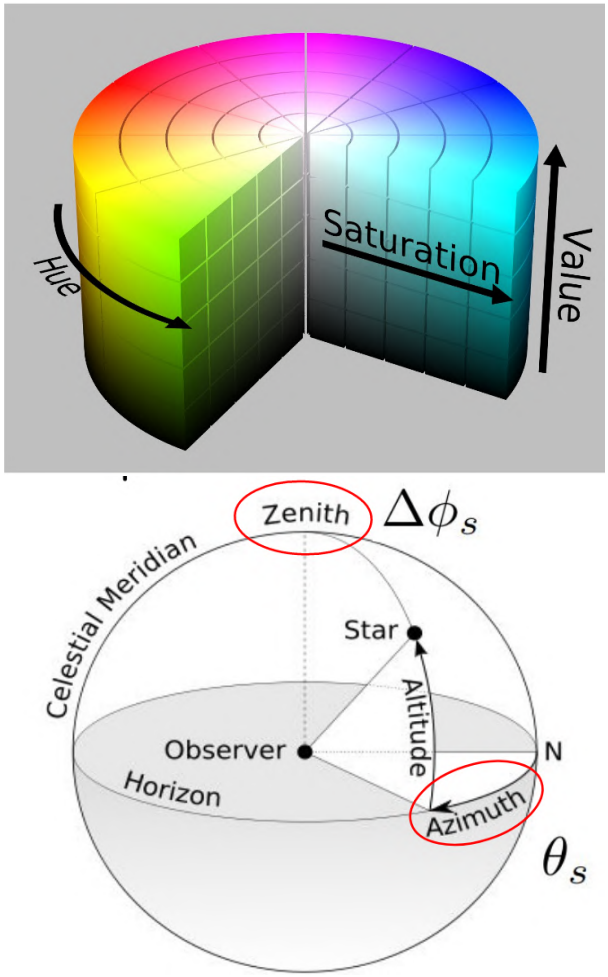


Figure 2.6: Top: HSV color cylinder; Bottom: Illumination features computed as a function of sun zenith angle ($\Delta\phi_s$), sun azimuth angle (θ_s) and a binary variable (v_s) for sun visibility.

2.6.2 Illumination Features

Sun position & visibility: Approach worked on by [20]. Function of sun zenith angle ($\Delta\phi_s$), sun azimuth angle (θ_s) and a binary variable (v_s) for sun visibility. This method estimates a probability distribution over sun position in the sky (azimuth and zenith angles) and visibility using a combination of weak cues (sky pixel intensities, cast shadows on ground, vertical surface shading) and a data-driven prior. Apart from foreground similarity, images with illumination similar to the query would be better candidates for sky replacement. Illumination I is given by:

$$I = \{\theta_s, \Delta\phi_s, v_s\} \tag{2.2}$$

2.6.3 Pre-trained CNN Features:

Pre-trained CNN features: Pre-trained features from convolutional neural networks pre-trained on large image datasets such as ImageNet [5] have been shown to perform well for image tasks in visual understanding. These deep features implicitly learn the spatial layout and object semantics at the deeper layers of the network when trained on large datasets. We use the output of FC7 (fully-connected) layer (4096 dim.) as feature representation from VGG-19 [33] architecture trained on different datasets.

More details on how these features are used with respect to our sky-search and sky-replacement procedure along with an extensive qualitative and quantitative analysis are given in chapter 3.

Chapter 3

Data-driven Sky Search and Replacement

Figure 3.1: For a query image with a dull sky (left), examples of consistent (middle) and inconsistent (right) sky replacements.

In this chapter, we describe in detail our method for automatic sky-search and sky-replacement. Our approach is data-driven and centered around the idea of 'compatible' sky-search. Given a query image with a problematic sky, our method first finds images with similar foregrounds and natural illumination. It then creates candidate composites by replacing the query image sky with the retrieved image skies and ranks the composites based on realism and diversity. The user is finally presented with the top-k candidate composites as replacement outcomes without color transfer thereby retaining the natural color composition of the foreground in the original image. We demonstrate the effectiveness of our method with qualitative results and a comprehensive user study. Figure 3.2 summarizes the proposed system with a block diagram. To summarize, this chapter presents our novel pipeline for compatible-search based sky-replacement that is a useful alternative or prelude to automatic color transfer methods. We describe our curated large database of outdoor images with interesting skies and evaluate the usefulness of a large number of features for this tasks.

3.1 Database Collection

The database of 1246 images used with the proposed system consists of 415 Flickr images with diverse skies (collected by [38]) and 831 outdoor images curated from the ADE20K Dataset[54]. ADE20K dataset consists of $\sim$ 22K images with 150 semantic categories like sky, road, grass. The

Project page : *https://cvit.iiit.ac.in/research/projects/cvit-projects/findmeasky*

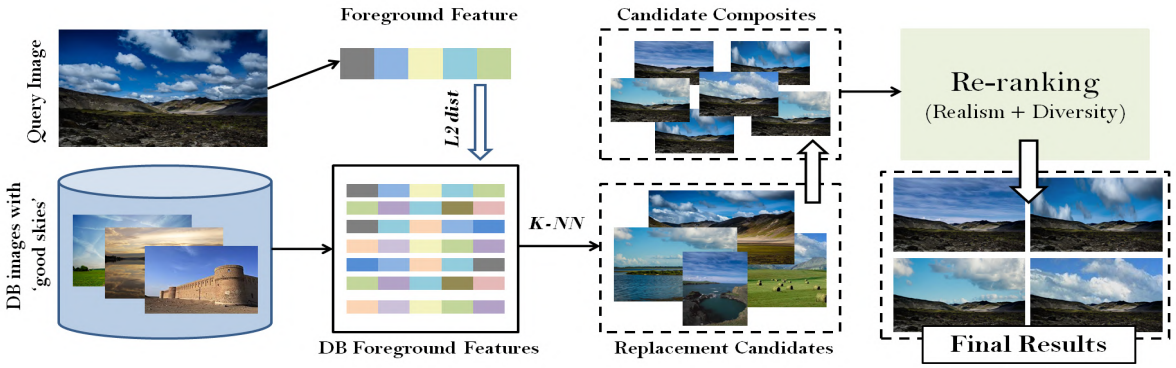


Figure 3.2: An overview of our sky-replacement pipeline. Candidate composites are created using skies from database images with most similar foregrounds. Final composites are re-ranked to maximize realism and diversity of the presented set.

images with sky category were first filtered to a set of $\sim$ 6K useful images for which the sky region made $> 40\%$ of the total image. These images were manually rated between 1 to 5 for interestingness and aesthetic appeal of the skies by two human raters and only the images with average scores higher than 3 were added to the final database.

3.2 Proposed System

The motivation behind our sky replacement method is to find naturally consistent yet interesting skies for a query image. Our system is based on the following hypothesis. Given two images, (i) if their foreground regions are similar (in color, layout, and semantic makeup), and (ii) if the estimated natural illumination (predicted positions of the sun in the sky) is similar, swapping their skies would lead to highly realistic composites that wouldn't need foreground color correction. This hypothesis is validated with experiments (discussed later). We first curate a database of outdoor images with interesting and aesthetically appealing skies along with their foreground masks. We represent each database image with image features corresponding to its foreground region and illumination. Similarly, given a query image (and its foreground mask), we compute its foreground features and natural illumination. For each query, we retrieve the top-K nearest neighbor images from the database based on the L_2 distance in feature space and use the sky regions in these images as viable candidates for replacement. We evaluate all candidate composites for realism and diversity and re-rank the candidates to provide most realistic yet diverse alternatives to the query image. This procedure is outlined in Figure 3.2.

3.2.1 Feature Description

3.2.1.1 Foreground Features

Color statistics: Xue et al. [51] studied the relation between the background and foreground regions for realistic composition using various 2D statistical measures and identified correlated color temperature (CCT), luminance, and saturation to be the most significant measures in determining realism of a composite. We use this finding and represent the image foreground using histograms of these color statistics computed at every pixel (using [43]).

Bag of Visual words and GIST: Hand-crafted features such as Bag of visual words (BoW) [34] and GIST have been popularly used for measuring object-level and scene-level similarities between images. For our task, BoVW features are computed by quantizing densely extracted local descriptors (like SIFT) from foreground region of an image into a large visual vocabulary and building a normalized histogram of these word occurrences. GIST features are designed to capture spatial envelope of the scene and use histogram representation of gabor filter responses applied at multiple scales and orientations. We use VLFeat library [42] to extract BoW and GIST features.

Pre-trained CNN features: Image descriptors computed using convolutional neural networks (CNNs) pre-trained on large data such as ImageNet have proven to be very effective for a number of visual understanding tasks. The success of these features can be attributed to implicit learning of spatial layout and object semantics at later layers of the network from very large datasets. We use two different pre-trained networks, (i) VGG19 architecture [33] trained on ILSVRC-2012 (ImageNet) dataset, and (ii) VGG16 architecture trained on Places205 dataset [53], and extract two variants of CNN features. With both architectures, we use the output of FC7 (fully-connected) layer (4096 dim.) as feature representation. Between these two, ImageNet pre-trained CNN features performs better. We did not fine-tune these networks for our task due to lack of labeled data.

3.2.1.2 Illumination Features

Sun position & visibility: Apart from foreground similarity, images with illumination similar to the query would be better candidates for sky replacement. We compare the sun position in the sky estimated using [20]. This method estimates a probability distribution over sun position in the sky (azimuth and zenith angles) and visibility using a combination of weak cues (sky pixel intensities, cast shadows on ground, vertical surface shading) and a data-driven prior.

3.2.2 Candidate search and composition

Candidate Search: The query and the candidates are compared using a combination of foreground distance (d_{fg}) and the sun position distance (d_{il}) as follows,

Figure 3.3: An overview of the sky replacement step.

$$d(I_q, I_c) = d_{fg}(I_q, I_c) + \alpha\, d_{il}(I_q, I_c) \tag{3.1}$$

Foreground features are compared using L_2 distance. For comparing illumination, instead of comparing two probability distributions, we directly compute the angular distance (zenith and azimuth) between the query and the candidate images. If the highest probability is below 0.5, the parameter α is 0, we do not consider the illumination distance as reliable and discard it otherwise α is 1. Distances are normalized between 0 and 1.

Composition: The database images are stored with an alpha mask corresponding to the sky/foreground segmentation. We assume the availability of alpha mask for query image also. Tsai et al. [38] explain an automatic method to obtain accurate sky segmentation. Alternatively a semi-automatic method [28] can be used to obtain a reliable alpha mask for the query image. Given the query and the candidate images with corresponding segmentation masks, we first crop the tightest rectangle consisting only of the sky pixels from the candidate image and scale it to match the size of the maximum bounding rectangle of the query image. We then replace the query image sky patch by the scaled candidate sky patch as illustrated in Figure 3.3 and perform laplacian pyramid based blending [1] along the seam to reduce composition artifacts.

3.2.3 Feature Selection based on composite realism

Given a query, the ideal feature is the one that yields candidate images with most suitable skies for replacement. Suitability of an image for this task is determined by perceived realism and aesthetic appeal of the final composite. These properties are highly subjective and hence obtaining ground-truth rankings/ratings for a large number of query images requires extensive human annotation effort. Recently Zhu et al. [55] trained a discriminative model to predict realism of an image (RealismCNN). While, this is not an accurate indicator of 'goodness' of a candidate for our task, it is a useful alternative to validate usefulness of the features in absence of any ground-truth/baseline. We created a validation set of 100 query images for this ablation study. For each query, we retrieved the top-100 candidates using L_2 distance of the five foreground features and also using a combination of foreground and sun position distances. This leads to 100 composites per query per feature (100K composites per feature). Using the predictive model explained above [55], we obtain a realism score for each composite.

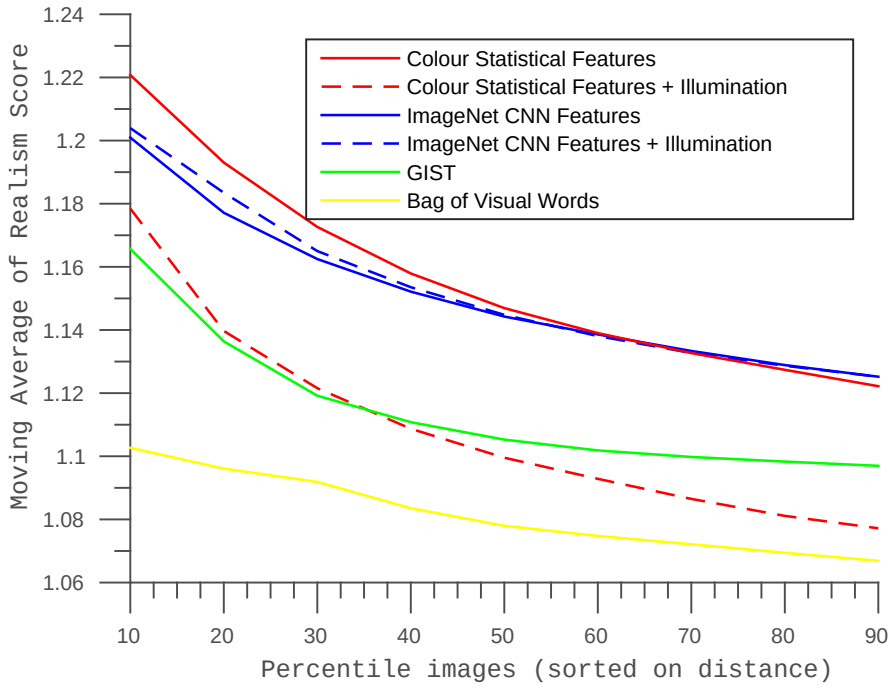


Figure 3.4: Ablation study : Running average of realism scores for composites sorted on feature distances

Figure 3.4 shows the running average of realism scores for incremental subsets of top-K composites (10%, 20%, ..., 100%). As discussed before, our hypothesis is that using skies of images with most similar foregrounds and/or illumination would lead to most realistic composites. If this hypothesis is valid, with increase in value of K, average realism score of the top-K composites should be decreasing. This trend can be observed for all features, validating our hypothesis. Among all foreground features, color statistics feature yields the highest average realism scores, CNN feature is a close second (ImageNet pre-trained). We study the effect of these two features combined with illumination feature (sun positions) as per equation 1. While the average scores drop for combined illumination and color features, these features are helpful to avoid physically implausible composites. But since the performance is significantly lower we finally only use color statistic features for finding the suitable candidate skies.

3.2.4 Re-ranking for realism and diversity

While the candidates obtained using feature based distances are compatible and the resulting composites are realistic, presenting all composites to the user is unnecessary and often undesirable. Many composites can potentially be redundant if the replaced sky is similar to the query and/or to other composites. We propose to select a small and diverse subset of highly realistic composites.

To achieve this, the composites are re-ranked based on the realism score (RealismCNN) and a diversity measure. This is done by casting this problem to a max-sum diversification objective and optimizing this objective using a facility dispersion algorithm as proposed by [12].

For relevant and diverse retrieval, we wish to select a subset that maximizes total relevance ($\sum w$) and total dissimilarity ($\sum d$). Consider U is the set of all candidate composites for a query image I_q and $S \subseteq U$ is the desired subset. The bi-criteria objective ($f(S)$) that achieves this can be given by Equation 3.2 [12] (where $\lambda > 0$ is a trade-off parameter).

$$f(S) = (k-1)\sum_{u \in S} w(u) + 2\lambda \sum_{u,v \in S} d(u,v) \tag{3.2}$$

$$d'(u,v) = w(u) + w(v) + 2\lambda d(u,v) \tag{3.3}$$

To recast the objective as *max-sum dispersion* (that maximizes sum of all pairwise distances in the subset S), [12] introduces a new pairwise distance given in Equation 3.3. For our task, we want the composite to be realistic and the sky regions to have comparable aspect ratios hence, (i) relevance w for each composite is a product of it's min-max normalized realism score and the scale factor (i.e scaling applied to the candidate sky patch), and (ii) the dissimilarity d is the L_2 distance between between two *sky regions* in color feature space.

3.3 Results and Discussion

Our system is implemented in MATLAB with binary bindings for realism evaluation and blending. Currently, the code is not optimized for performance and takes around a minute to produce 100 candidates for a query image, of which, we show the top-4. To evaluate the effectiveness of our method, we show qualitative results for a few query images and discuss findings of the user-study based evaluation conducted for a larger query set.

Qualitative results: Figure 3.9 illustrates the 4 best composites for the query images on the top. The query images shown include a variety of scene types and configurations such as aerial/ground shots, presence/absence of foreground objects (person, tower), dull/interesting skies. It can be seen that for all queries, the composites are diverse, natural looking, and aesthetically appealing. Figure 3.6 shows the usefulness of the re-ranking algorithm. The images before re-ranking have similar backgrounds to the input image. But after re-ranking we get images which are both diverse and relevant. Figure 3.5 compares the results from the given pipeline and the results given by [38]. The comparison clearly shows that our method produce results which are similar in aesthetic appeal. Figure 3.7 depicts the failure of the color transfer techniques used by [38] as the specularity and reflection from the roofs in the houses is clearly visible. There is no need for such correction in our method as it chooses skies that are already compatible with the foreground of the input image.

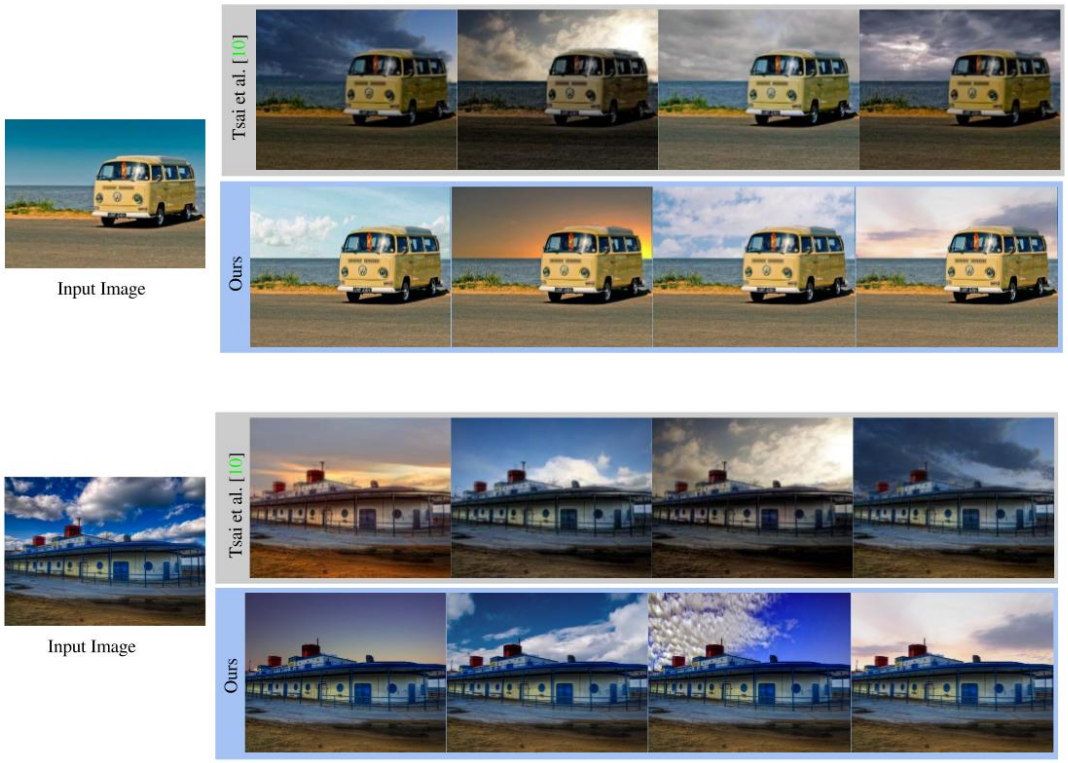


Figure 3.5: Comparison with existing state of the art method, [38]. We tested our model on the same input image, the results obtained are just as aesthetically appealing using a completely different pipeline.

	min	max	mean	median
$\mathcal{R}_q >$ all $\mathcal{R}_c$	0%	52%	12.72%	8.6%
any $\mathcal{R}_c \geq \mathcal{R}_q$	48.48%	100%	87.32%	91.4%
any $\mathcal{R}_c > \mathcal{R}_q$	10.5%	81.67%	43.38%	43.31%

Table 3.1: Statistics on user preferences

User-study evaluation: To assess the performance of our replacement system, we conducted a user study where we asked the users to rate groups of images based on their naturalness and aesthetic appeal. Each group included a query image and top-3 composites in a randomized (and anonymous) fashion. The user study was conducted for a set of 30 query groups and each group was rated by at least 40 participants. The participants belonged to age group 20 to 35 and had varying degrees of photography and composition expertise, with a larger segment self-identifying as amateur or casual photographers. Each image was rated between scores 1 to 5 which correspond to 'very bad', 'bad', 'okay', 'good', and 'very good' descriptions. In absolute terms, the median score (across users and queries) for the original image is 2.82 (below 'okay') while for the composites, it is 3.12 (above 'okay') indicating that the composites were perceived to be equally or more attractive than the original images, Relatively,

83.33% of the times at least one out of three composites received a rating strictly higher than the query image indicating preferable aesthetic appeal of the composites. We also report statistics on the fraction of times a query image I_q is rated $>$, $=$, $<$ any of the composite images I_c in Table 3.1. It shows user agreement for various cases, e.g. for the criteria any $\mathcal{R}_c > \mathcal{R}_q$ (where $\mathcal{R}_i$ is rating of an image i), the worst performing query set (column corresponding to min) 10.5% users agree, the best set has 81.67% users in agreement, and on average over all query sets 43.38% users agree. The survey results clearly indicate merit in our replacement system.

Figure 3.6: Example illustrating the efficacy of the re-ranking.

Figure 3.7: Failure of colour transfer methods in the in-house implementation of [38] as compared with our method which chooses skies that are already compatible with the foreground.

Failure Cases: Figure 3.8 shows a few cases where our pipeline fails. Figure 3.8 (i) illustrates that like any composition system, success of our system also assumes accurate segmentation and incorrect segmentation can lead to inconsistent composites. In case of scenes with specular surfaces like in (ii–iii), inconsistent reflection of the sky can lead to unnatural looking compositions.

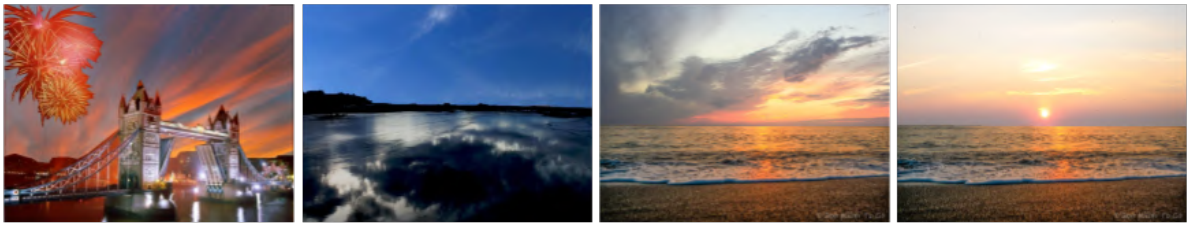


Figure 3.8: Failure cases, from left, (i) segmentation error, (ii) inconsistent sky reflection in water, (iii) bright (sun) spot, (iv) better composition achieved with use of illumination map.

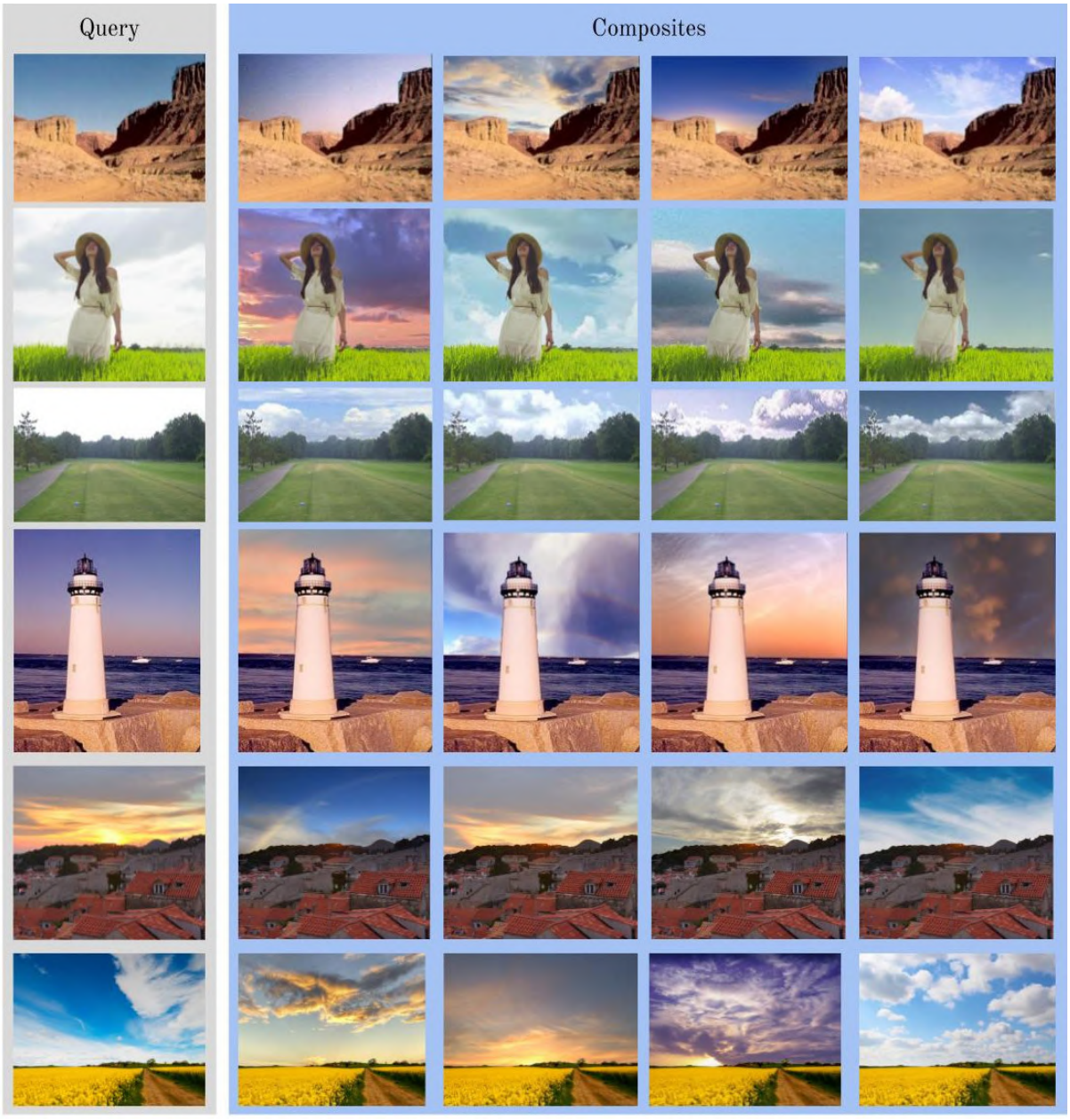


Figure 3.9: Example results of our diverse and compatible sky-replacement system

3.4 Summary

In this chapter, we explained in detail our proposed data-driven method that given a query image produces interesting and realistic composites with different skies without using color transfer as a post-processing step. To achieve interesting replacements, we curated a new dataset of outdoor images with interesting skies. To achieve realism without color transfer, we proposed a foreground similarity hypothesis and validated it using a realism prediction model. We also experimented with a variety of image based features for this task and observed color statistical features to be very effective. We further showed a re-ranking technique to achieve both realism and diversity in the final subset presented to the user. The effectiveness of our method is evaluated by conducting a thorough user study.

In the next chapter, we describe our proposed protocol for unsupervised learning of image style representation using Gram Matrix (deep feature correlation map) as a proxy measure of stylistic similarity.

Chapter 4

Unsupervised Image Style Embeddings for Retrieval and Recognition

In the previous chapter, we introduced and explained our proposed method for the task of color consistent sky search and replacement, and demonstrated the effectiveness of our method with qualitative results and a comprehensive user study. In this chapter, we examine the quality of understanding of visual style which can be achieved by an unsupervised approach that does not rely on any categorical labels of style. As mentioned in the previous chapters, for explicit style understanding previous work tend 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. Such representations are efficient and effective for only a specific task and have limitation in terms of generalization and the need for a manually curated large database for training. Therefore, making the entire process expensive and inefficient as well as ill-suited for a subjective attribute like artistic style where expert annotations are very limited. On the other hand, we 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. The details of which are further explained in this chapter including dataset creating, learning protocol and evaluation results.

4.1 Dataset Creation

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 [7, 8]. The details of the training procedures are given later in this section. We first explain the clustering and data construction.

4.1.1 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.

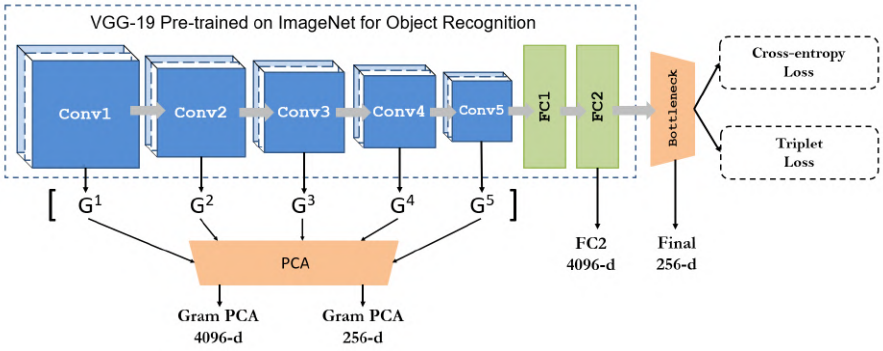


Figure 4.1: Different feature layers of VGG-19 based CNN used for our experiments.

4.1.1.1 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 [33] pre-trained for object recognition and localization [30] tasks and extract Gram matrix features as described in [7, 8]. 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 4.1). As explained in [8] 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 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 4.1).

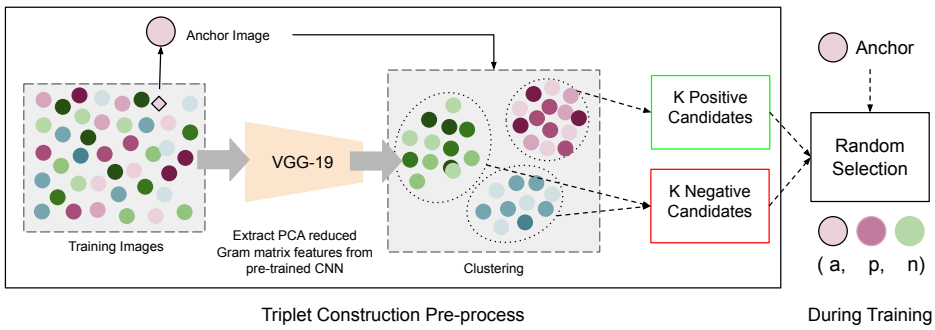


Figure 4.2: Triplet construction and selection process.

The resulting Gram Matrix feature vector captures information critical for style texture [9], but has a very high dimensionality (typically of size $\sim 200k$). To make the feature space more compact and

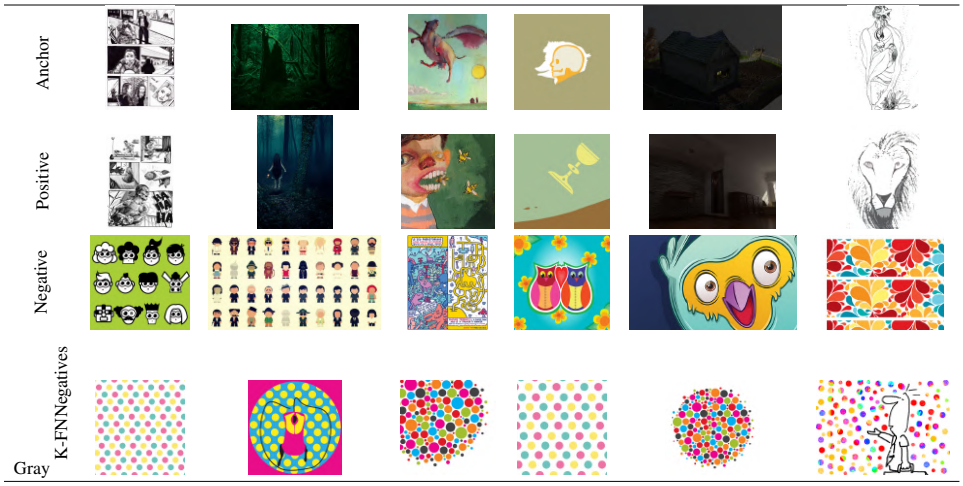


Figure 4.3: Example triplets sampled with explained procedure in Section 4.1.1.2 (Cluster distance based sampling). Notice poor diversity for K-FN based negative selection (last row).

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 [8, 50].

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 [14]. 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.

4.1.1.2 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 4.2. 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 4.3). For

successful learning, we need to mine diverse and informative triplets. Hard negative mining can bring more diversity and relevance to this process [15].

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 $(D^i_{max} - D^i_{min})$.

4.1.2 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.

Training with cross-entropy loss: We train a CNN with VGG-19 architecture [33] augmented by a 256-dimensional bottleneck layer (shown in Figure 4.1) 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 [4]. 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 [45] with the same network architecture as above for each branch and minimize a triplet loss similar to [31]. 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 4.1.1.2. 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 [18] optimization algorithm is used during training of both stages[1].

[1]We will release the network models and training codes along with the related publication for ease of reproduction *https://sidgairo18.github.io/style*

4.2 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 [48] consists of images from *Behance*[2] - a portfolio website 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 [4]) balanced across media and emotional styles, and with a Train, Validation and Test split as 80:5:15.

AVA Style Dataset: Introduced in [26, 17], 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 [17] 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 [17].

Wikipaintings: A dataset [17] 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[3] 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*[4]. 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

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

evaluation of retrieval due to the interesting 2-level hierarchy of style relevance (within each set and within each theme).

4.3 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[17]	4000	Pretrained
FC2	Fully Connected[33]	4096	Pretrained
B-Tri	Bottleneck	256	Triplet
B-CE	Bottleneck	256	Cross-entropy

Table 4.1: Details of feature representations used for performance evaluation and comparison. Refer to Figure 4.1 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 [17] on datasets discussed in the previous section. Table 4.1 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.

4.3.1 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 4.2). 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 4.2. 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

Dataset	#Q	Random	Feat. Dim: $\sim$ 4096			Feat. Dim : 256		
			F$\times$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 4.2: mAPs computed for retrieval on different datasets and features. The learning procedure (Section 4.1) 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$\times$C [17]	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 4.3: 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.

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 the triplet training.

4.3.2 Recognition Task

Starting with different unsupervised representations shown in Figure 4.1, 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 4.3. 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$\times$Content features of [17] performs better. These features combine activations of independently trained content classifier with Fusion features in outer product sense. Karayev et al. [17] suggest that some style categories are inherently content-dependent, hence combin-

ing 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.

4.3.3 Qualitative Results for Style based Search

Figure 5.1 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. Figure 1.4 shows the t-SNE [41] 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 chapter 5.

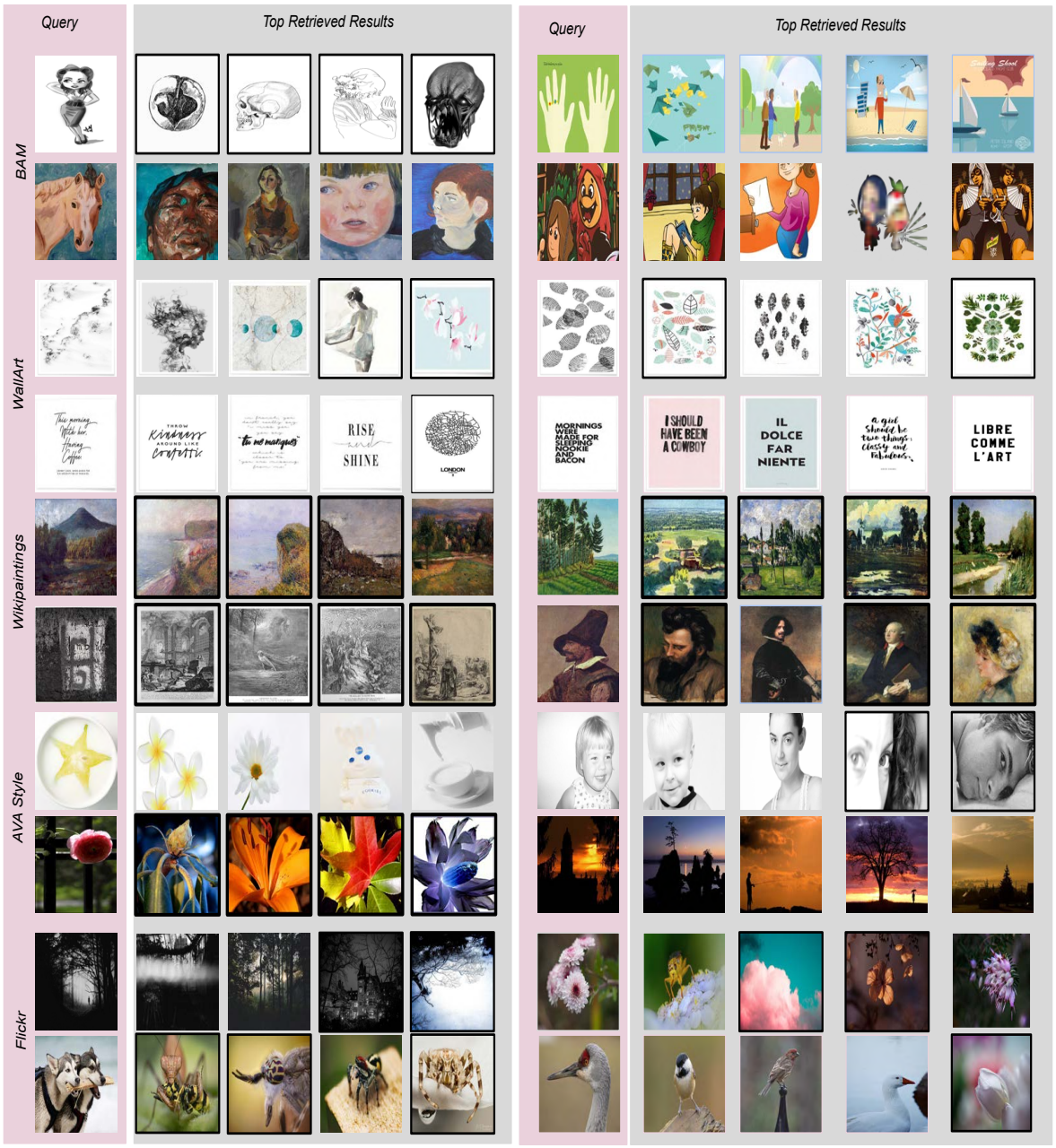


Figure 4.4: Retrieval results using the best performing representation B-Tri for example queries from different datasets. Images highlighted by black border have style labels different from query style labels although they are visually similar.

4.4 Summary

In this chapter, we described our proposed 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.

Chapter 5

Additional Results

This chapter contains additional results for the different experiments performed and detailed statistics for the different datasets used.

5.1 Qualitative Results

In this section we present a qualitative analysis of our learnt style representation for the task of retrieval across six datasets (Refer to section 4, Datasets in chapter 4).

The qualitative results further bring to light that our learnt representations and Gram Matrix features do capture the look and feel of an image. Two images could be very similar looking but could have different style labels in different style context.

Figures 5.1, 5.2, 5.3, 5.4, 5.5, 5.6 show results of nearest neighbor retrieval for example queries from each dataset with triplet loss based representation (B-Tri). Since style labels are often contextual and convey a limited meaning of style, a low precision score does not necessarily imply poor quality of visual similarity. The retrieved results that are highlighted by a black bounding box don't have the same style label as the query, despite obvious visual similarity.

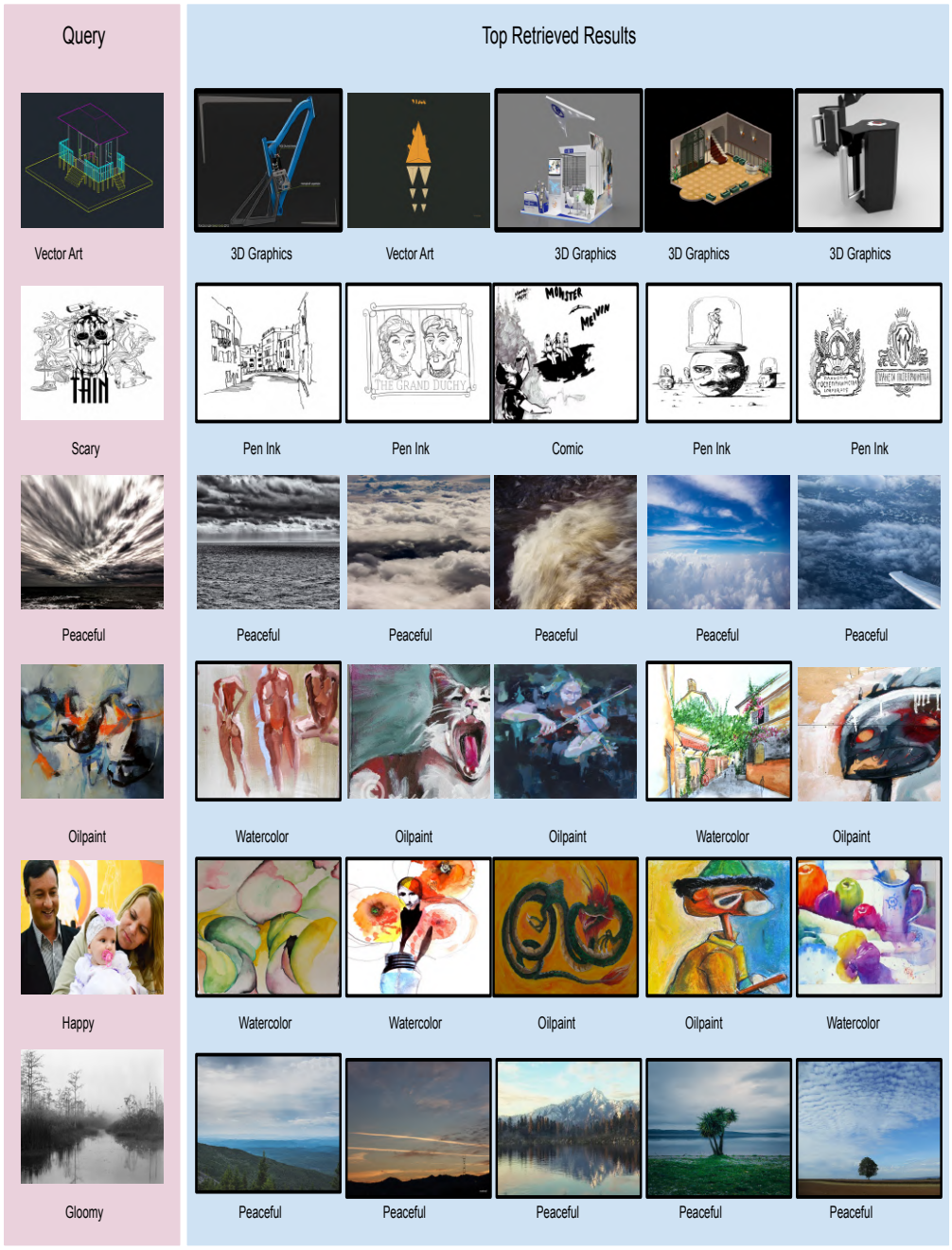


Figure 5.1: Nearest Neighbour retrieval results for select queries from BAM subset test split. Notice that for rows 1 and 2, the queries and neighbours are very similar looking but the labels do not match. This indicates the lower mAP scores for retrieval using unsupervised methods. 'Oil Paint' and 'Water Colour' are hard to differentiate, similarly 'Gloomy' and 'Peaceful'

Figure 5.2: Retrieval Results for Query and Top Neighbours Deviantart dataset.

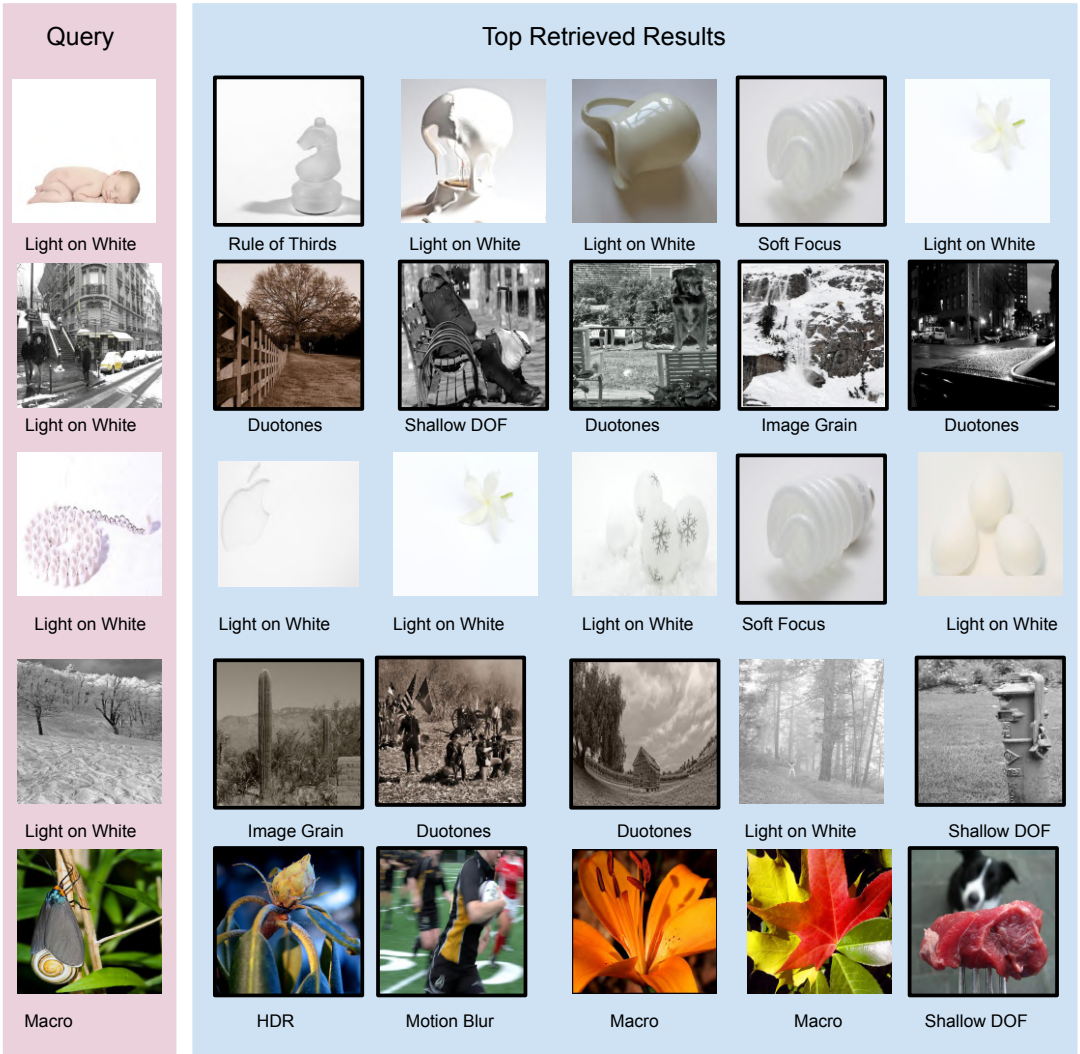


Figure 5.3: Retrieval Results for Query and Top Neighbours AVA Style dataset.

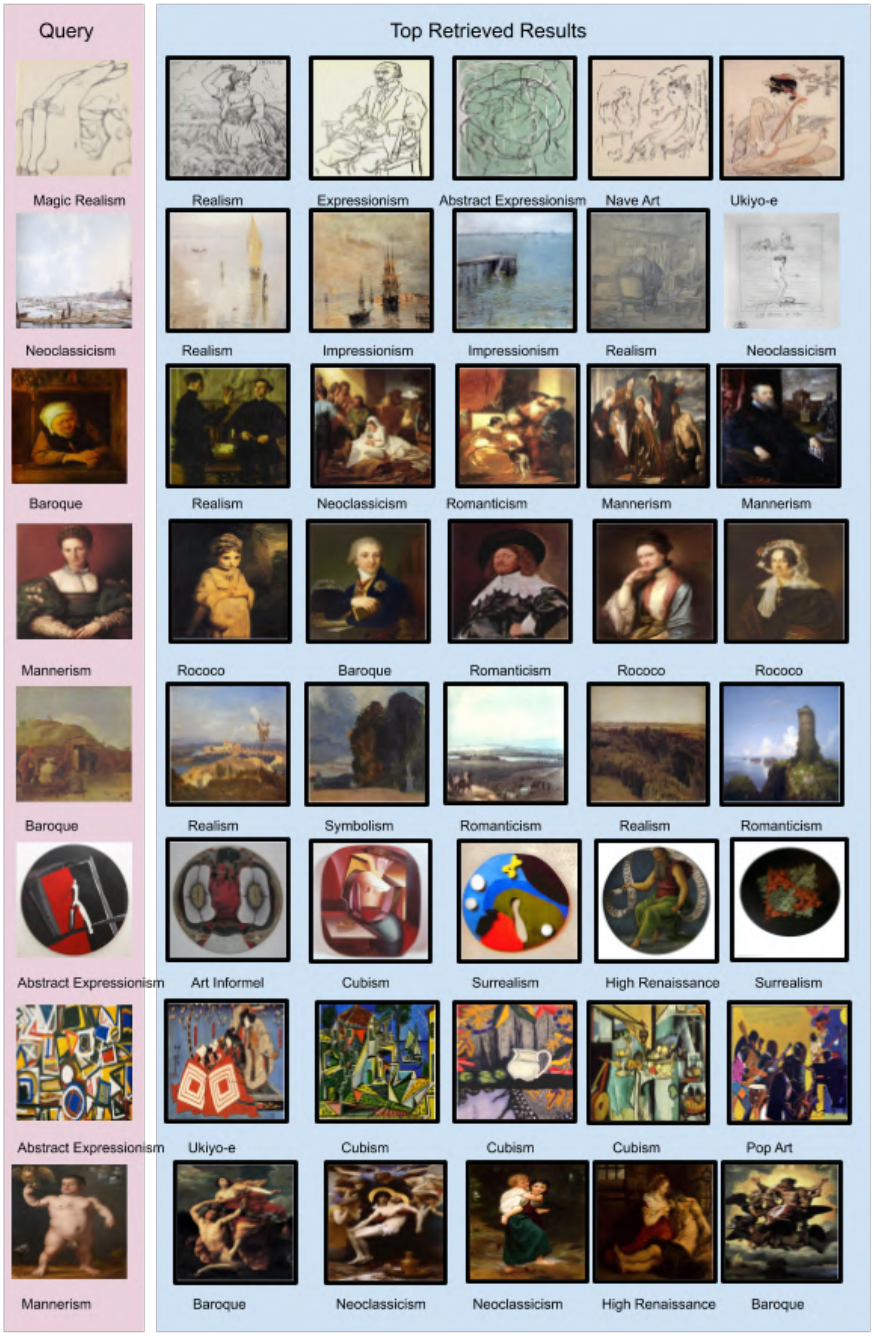


Figure 5.4: Retrieval Results for Query and Top Neighbours Wikipaintings Subset dataset.It is interesting to see the retrieved results and their relevance with respect to the query image. Notice row 7 where, 'Abstract Expressionism' labelled query retrieves 'Ukiyo-e', 'Cubism' and 'Pop Art' paintings.

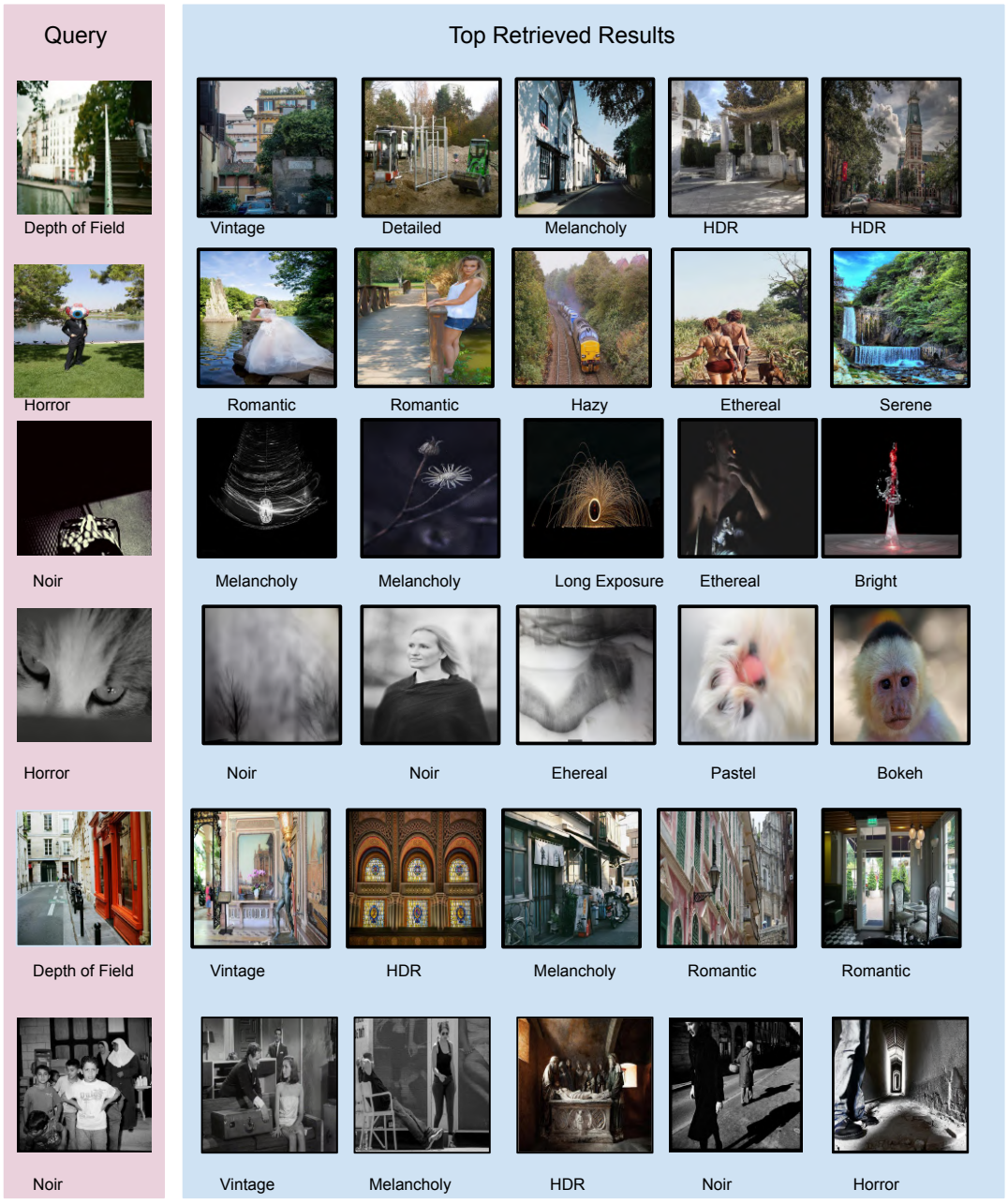


Figure 5.5: Retrieval Results for Query and Top Neighbours Flickr Test Set.

Figure 5.6: Retrieval Results for Query and Top Neighbours WallArt dataset. The style themes for this dataset have been manually curated by experts, the retrieved samples show similarity both in terms of appearance and style themes.

5.2 Confusion Matrix

Figures 5.7, 5.8, 5.9, 5.12, 5.10, 5.11 show class-wise confusion matrix for retrieval for each dataset. It can be observed that style classes that are more visually similar as compared to other classes are confused more.

Figure 5.7: Confusion Matrix for Top 100 retrievals for 1000 Query images on Behance Subset Test set using learnt representations. Here we see the following pairs confusing with each other - 'Watercolor' with 'Oilpainting' since both are very colourful, 'Graphite' and 'Pen Ink' both are hand-drawn and dull, and '3D Graphics' with 'Vectorart'.

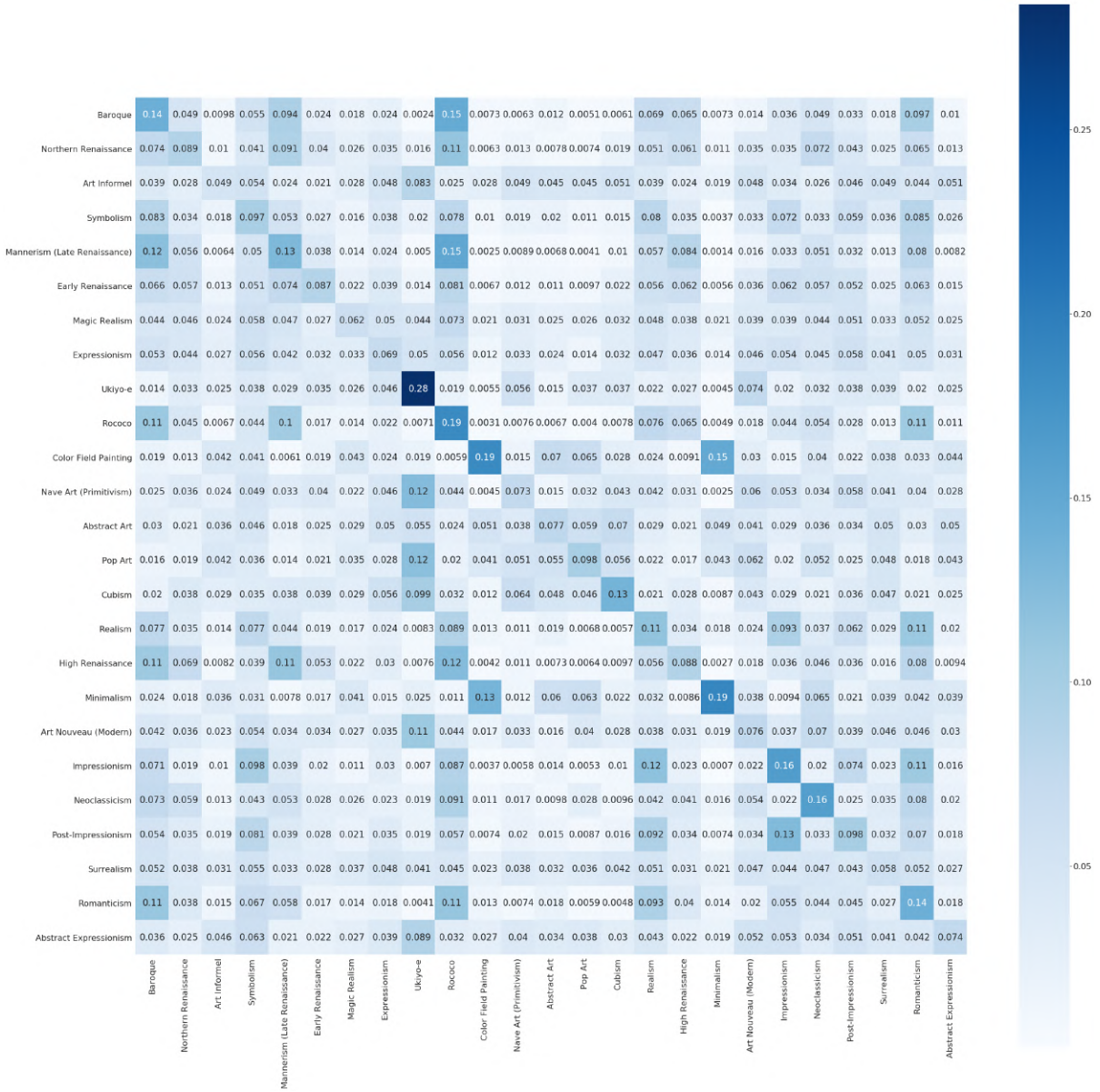


Figure 5.8: Confusion Matrix for Top 100 retrievals for 1000 Query images on Wikipaintings Subset Test set using learnt representations.

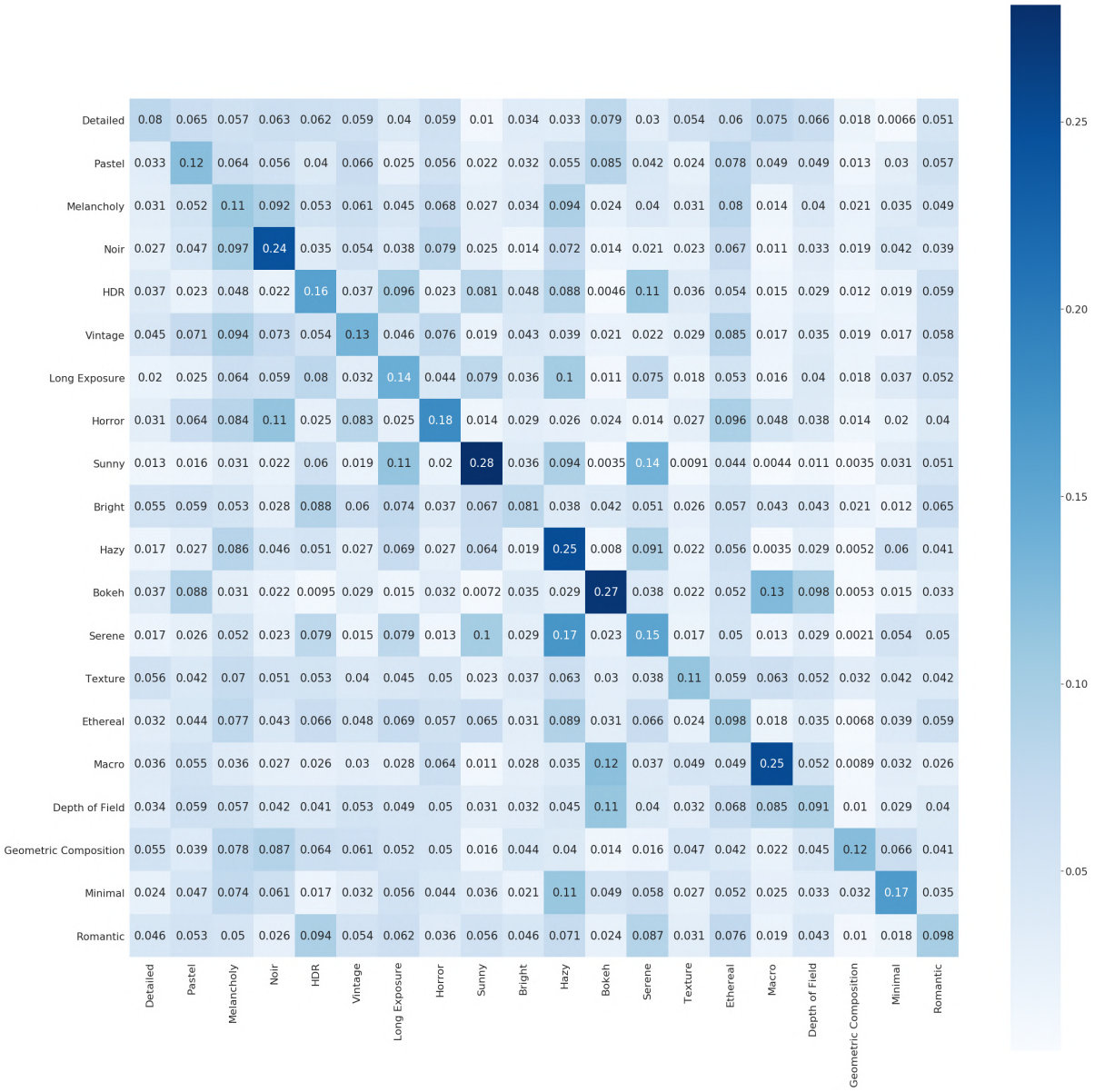


Figure 5.9: Confusion Matrix for Top 100 retrievals for 1000 Query images on Flickr Test set using learnt representations.

5.4 Samples from clustering

Figure 5.18 shows randomly drawn images from different clusters formed using PCA reduced Gram features for BAM dataset.

Figure 5.18: Each row shows examples drawn randomly from seven clusters, for clustering applied to BAM [48] subset. It can be seen that clustering in Gram matrix space groups stylistically similar images together.(Each row only contains samples from a single cluster)

5.5 Dataset Details

Tables 5.1, 5.2, 5.3, 5.4, 5.5, 5.6 provide details of number of images per class for each dataset discussed in chapter 4.

Ava Style	
Style	**Number of Images**
Rule of Thirds	839
Silhouettes	1043
Complementary Colors	388
Shallow DOF	1819
Motion Blur	833
Macro	779
Duotones	1216
Vanishing Point	620
Light On White	1059
Negative Image	1326
HDR	735
Soft Focus	642
Long Exposure	1612
Image Grain	932

Table 5.1: Ava Style dataset (a subset of AVA dataset [26]) similar to [17] style categories and the number of images in each category.

Deviant Art	
Style	**Number of Images**
Digital Art Mixed Media	1521
Digital Art Drawings & Paintings	1122
Traditional Art Drawings	1559
Traditional Art Mixed Media	1322
Traditional Art Paintings	627

Table 5.2: DeviantArt dataset style categories and the number of images in each category.

Flickr	
Style	**Number of Images**
HDR	3994
Noir	3999
Sunny	399
Horror	4000
Long Exposure	3999
Detailed	4000
Vintange	4000
Melancholic	4000
Macro	4000
Minimal	4000
Ethereal	4000
Depth of Field	3998
Geometric Composition	4000
Texture	4000
Serene	4000
Hazy	4000
Romantic	4000
Bright	4000
Pastel	4000
Bokeh	4000

Table 5.3: Flickr dataset [17] style categories and the number of images in each category.

Wall Art	
Style	**Number of Images**
Country Living	20
Scandi Chic	40
Fashionista	107
Coastal Views	84
Young at Heart	124
Minimal Monochrome	9
Fine Art Photography	31
Pastel Dreams	179
New Romantic	111
Modern Dandy	107
Bold and Contemporary	21
Shades of Summer	94
Abstract and Colourful	13

Table 5.4: WallArt dataset style categories and the number of images in each category.

Wikipaintings Subset	
Style	**Number of Images**
Realism	999
Pop Art	999
Post-Impressionism	999
Color Field Painting	1000
Ukiyo-e	998
Art Informel	969
Nave Art (Primitivism)	999
Baroque	997
Neoclassicism	998
Abstract Expressionism	996
Early Renaissance	1000
Abstract Art	998
Minimalism	993
Romanticism	996
Impressionism	1000
High Renaissance	998
Cubism	1000
Northern Renaissance	999
Expressionism	997
Mannerism (Late Renaissance)	999
Rococo	990
Symbolism	997
Art Nouveau (Modern)	999
Surrealism	1000
Magic Realism	991

Table 5.5: Wikipaintings Subset dataset, which is a subset of Wikipaintings dataset [17] style categories and the number of images in each category as used for our experiments.

Behance Style Subset											
Style	other	bicycle	cat	tree	bird	dog	building	flower	cars	people	Total
Watercolor	780	35	221	503	2190	1441	542	555	39	2560	8866
Pen Ink	559	85	152	121	3031	1860	258	59	57	2483	8665
Graphite	936	45	147	123	1540	1344	297	56	95	4259	8842
Comic	178	77	207	20	1534	2181	142	59	53	4361	8812
Vectorart	1936	74	100	29	1680	1243	689	52	106	2883	8792
Oilpaint	1188	15	110	602	977	1332	349	391	28	3757	8749
3d graphics	2697	149	25	165	415	525	1413	88	900	2455	8832
Happy	287	33	630	247	1918	1357	27	1681	2	2718	8900
Scary	779	21	141	266	1722	1579	89	397	7	3763	8764
Gloomy	945	61	51	1558	438	454	1745	27	49	3428	8756
Peaceful	1403	23	70	4100	625	364	695	581	61	900	8822

Table 5.6: Behance Style Subset datset style classes and the number of images in each category as used for our experiments, which is a subset of BAM dataset [48] very similar to the Behance-Net-TT used in [4].

5.6 Additional Plots and Tables

Figures 5.19, 5.20 show bar plots for retrieval and recognition mAPs for different feature representations.

Table 5.7 shows the recognition performance (in terms of mAP) of gram matrices computed across different layers ($Conv_1$ to $Conv_5$) of VGG19 [33] Networks for different datasets. A combination of all the layers performs the best.

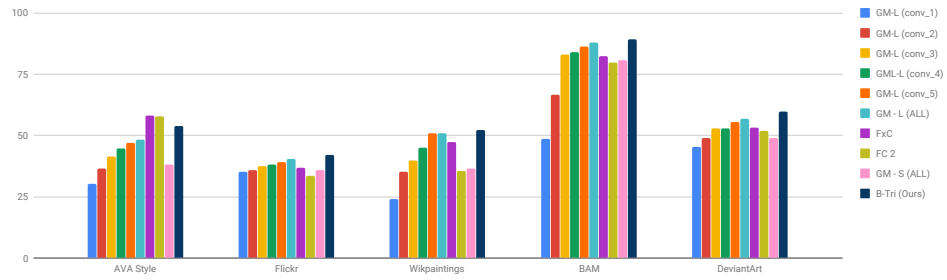


Figure 5.19: Dataset wide mAP scores for style based classification using different features. Notice that B-Tri features clearly show improvement over other features across most datasets.

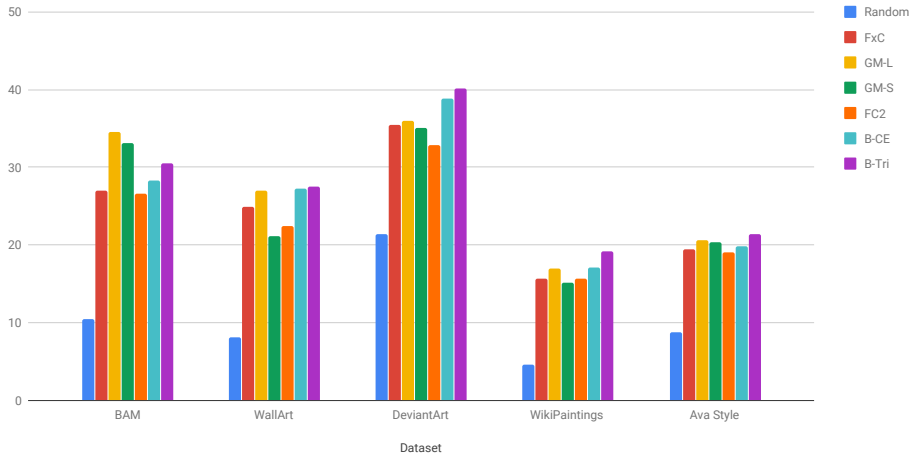


Figure 5.20: Dataset wide mAP scores for retrieval performance using different features. Notice that B-Tri and B-CE features clearly show improvement over other features across most datasets.

	Feat. Dim : $\sim$ 4096						Feat. Dim : 256
Dataset	Conv 1	Conv 2	Conv 3	Conv 4	Conv 5	All Conv	All Conv
AVA Style	30.20	36.40	41.68	44.74	46.96	**48.32**	38.19
Flickr	35.02	35.74	37.62	37.77	39.25	**40.47**	35.80
WikiPainting	25.36	34.35	39.81	44.70	50.92	**51.02**	36.47
BAM	48.68	66.59	83.03	83.81	86.20	**87.81**	80.76
Deviant Art	43.51	49.03	52.60	53.57	55.39	**56.77**	49.03

Table 5.7: mAPs for gram matrices computed for different layers (conv1-conv5) of VGG19 [33] Network for recognition using a softmax classifier on different datasets and features. Evidently a combination of all convolutional layers performs best.

Chapter 6

Conclusions and Future Work

6.1 Conclusions

In this thesis we explored and studied various image representations and their applications in understanding visual style, image retrieval, image recognition and background replacement.

We discuss our proposed data-driven method that given a query image produces interesting and realistic composites with different skies without using color transfer as a post-processing step. To achieve interesting replacements, we curated a new dataset of outdoor images with interesting skies. To achieve realism without color transfer, we proposed a foreground similarity hypothesis and validated it using a realism prediction model. We also experimented with a variety of image based features for this task and observed color statistical features to be very effective. We further showed a re-ranking technique to achieve both realism and diversity in the final subset presented to the user. The effectiveness of our method is evaluated by conducting a thorough user study.

We describe in detail our proposed 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. 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.

6.2 Future Work

In future, the applications of our proposed unsupervised visual style representations learning protocol may be explored with other proxy measures for style-aware grouping, e.g. semantic descriptions for fashion image search. The unsupervised framework described in chapter 4 has been effective in learning style representations well, and might prove useful in understanding hierarchies of styles or capturing multiple notions of style.

With the recent advent of generative algorithms [13], the sky replacement procedure may be replaced with a generative module to generate compatible skies for a given outdoor scene.

Related Publications

- **S Gairola**, R Shah, PJ Narayanan (2020), "Unsupervised Image Style Embeddings for Retrieval and Recognition Tasks", **Winter Conference on Applications of Computer Vision *(WACV '20)***.
- S Rawat*, **S Gairola***, R Shah, PJ Narayanan (2018), "Find Me a Sky: A Data-Driven Method for Color-Consistent Sky Search and Replacement", **International Conference on Multimedia Modeling**, Pages 216-228.

[0]*Both the authors had equal contribution

Bibliography

[1] Peter J. Burt and Edward H. Adelson. A multiresolution spline with application to image mosaics. *ACM Trans. Graph.*, 2(4), October 1983.

[2] Alex J. Champandard. Semantic style transfer and turning two-bit doodles into fine artworks. *ArXiv*, abs/1603.01768, 2016.

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

[4] 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.

[5] 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.

[6] Vincent Dumoulin, Jonathon Shlens, and Manjunath Kudlur. A learned representation for artistic style. *ArXiv*, abs/1610.07629, 2016.

[7] 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.

[8] 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.

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

[10] 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.

[11] 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.

[12] Sreenivas Gollapudi and Aneesh Sharma. An axiomatic approach for result diversification. In *Proc. ACM WWW*, 2009.

[13] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*, 2014.

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

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

[16] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*, 2016.

[17] 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.

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

[19] J.-F. Lalonde and A.A. Efros. Using color compatibility for assessing image realism. In *Proc. IEEE ICCV*, 2007.

[20] Jean-François Lalonde, Alexei A. Efros, and Srinivasa G. Narasimhan. Estimating natural illumination from a single outdoor image. In *Proc. IEEE ICCV*, 2009.

[21] Yanghao Li, Naiyan Wang, Jiaying Liu, and Xiaodi Hou. Demystifying neural style transfer. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, IJCAI'17. AAAI Press, 2017. ISBN 978-0-9992411-0-3.

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

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

[24] Noboru Ohta and Alan R. Robertson. Colorimetry: Fundamentals and applications. *Colorimetry: Fundamentals and Applications*, 05 2006. doi: 10.1002/0470094745.ch2.

[25] Aude Oliva and Antonio Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *Int. J. Comput. Vision*, 42(3), 2001.

[26] 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.

[27] Saumya Rawat, Siddhartha Gairola, Rajvi Shah, and P. J. Narayanan. Find me a sky : a data-driven method for color-consistent sky search & replacement. In *The 24th International Conference on Multimedia Modeling (MMM 2018), Bangkok, Thailand*, 2018.

[28] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. "grabcut": Interactive foreground extraction using iterated graph cuts. *ACM Trans. Graph.*, 23(3), 2004.

[29] Manuel Ruder, Alexey Dosovitskiy, and Thomas Brox. Artistic style transfer for videos. In *GCPR*, 2016.

[30] 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.

[31] 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.

[32] I-Chao Shen and Wen-Huang Cheng. Gestalt rule feature points. *IEEE Transactions on Multimedia*, 17:526–537, 2015.

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

[34] Josef Sivic and Andrew Zisserman. Video google: A text retrieval approach to object matching in videos. In *Proc. IEEE ICCV*, 2003.

[35] Michael Stokes, Matthew Anderson, Srinivasan Chandrasekar, and Ricardo Motta. A standard default color space for the internet — srgb, 1996.

[36] Ryosuke Tanno, Shin Matsuo, Wataru Shimoda, and Keiji Yanai. Deepstylecam: A real-time style transfer app on ios. volume 10133, pages 446–449, 01 2017. ISBN 978-3-319-51813-8. doi: 10.1007/978-3-319-51814-5_39.

[37] Litian Tao, Lu Yuan, and Jian Sun. Skyfinder: Attribute-based sky image search. *ACM Trans. Graph.*, 28(3), 2009.

[38] Y.-H. Tsai, X. Shen, Z. Lin, K. Sunkavalli, and M.-H. Yang. Sky is not the limit: Semantic-aware sky replacement. *ACM Trans. Graph.*, 35(4), 2016.

[39] Dmitry Ulyanov, Vadim Lebedev, Andrea Vedaldi, and Victor S. Lempitsky. Texture networks: Feed-forward synthesis of textures and stylized images. In *ICML*, volume 48 of *JMLR Workshop and Conference Proceedings*, pages 1349–1357. JMLR.org, 2016.

[40] Dmitry Ulyanov, Andrea Vedaldi, and Victor S. Lempitsky. Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 4105–4113. IEEE Computer Society, 2017.

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

[42] A. Vedaldi and B. Fulkerson. VLFeat - an open and portable library of computer vision algorithms. In *ACM MM*, 2010.

[43] J. Wågberg. *OptProp: Matlab Toolbox for Calculation of Color Related Optical Properties : Version 2.1*. FSCN-rapport. 2007.

[44] Dong Wang, Weijia Jia, Guiqing Li, and Yunhui Xiong. Natural image composition with inhomogeneous boundaries. In Yo-Sung Ho, editor, *Advances in Image and Video Technology: Pacific Rim Symposium (PSIVT)*, 2012.

[45] 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.

[46] Jue Wang and Michael F. Cohen. Image and video matting: A survey. *Found. Trends. Comput. Graph. Vis.*, 3(2), 2007.

[47] Y. Wang, S. Li, and A. C. Kot. On branded handbag recognition. *IEEE Transactions on Multimedia*, 18(9):1869–1881, Sep. 2016. ISSN 1520-9210. doi: 10.1109/TMM.2016.2581580.

[48] 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.

[49] Bing-Yi Wong, Kuang-Tsu Shih, Chia-Kai Liang, and Homer H. Chen. Single image realism assessment and recoloring by color compatibility. *IEEE Trans. Multimedia*, 14, 2012.

[50] 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.

[51] Su Xue, Aseem Agarwala, Julie Dorsey, and Holly Rushmeier. Understanding and improving the realism of image composites. *ACM Trans. Graph.*, 31(4), 2012.

[52] Yufeng Zheng, Clifford Yang, and Aleksey Merkulov. Breast cancer screening using convolutional neural network and follow-up digital mammography. page 4, 05 2018. doi: 10.1117/12.2304564.

[53] Bolei Zhou, Aditya Khosla, Àgata Lapedriza, Antonio Torralba, and Aude Oliva. Places: An image database for deep scene understanding. *CoRR*, abs/1610.02055, 2016.

[54] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proc. IEEE CVPR*, 2017.

[55] Jun-Yan Zhu, Philipp Krähenbühl, Eli Shechtman, and Alexei A. Efros. Learning a discriminative model for the perception of realism in composite images. In *Proc. IEEE ICCV*, 2015.