Unsupervised Image Style Embeddings for Retrieval and Recognition Tasks - Supplementary Material

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

This supplementary material includes the following additional results and information.

Qualitative results Figures 1,2, 3, 4, 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.

Confusion matrix Figures 7, 8, 9, 12, 10, 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.

t-SNE visualizations - Figures 15, 16, 17, 18, 19 show t-SNE [5] visualizations of BAM dataset images based on following feature representations: FC2 features and PCA-reduced Gram features (both 4096 and 256 dimensional) computed from pre-trained VGG19, 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.

Dataset Details Tables 2, 3, 4, 5, 6, 7 provide details of number of images per class for each dataset discussed in Section 4 of main paper.

Samples from Clustering Figure 20 shows randomly drawn images from different clusters formed using PCA reduced Gram features for BAM dataset.

Additional plots Figures 13, 14 show bar plots for retrieval and recognition mAPs for different feature representations.

Additional Tables Table 1 shows the recognition performance (in terms of mAP) of gram matrices computed across different layers ($Conv_1$ to $Conv_5$) of VGG19 [4] Networks for different datasets. A combination of all the layers performs the best.

References

[1] 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)*, pages 2679–2687, 2017.

[2] S. Karayev, M. Trentacoste, H. Han, A. Agarwala, T. Darrell, A. Hertzmann, and H. Winnemoeller. Recognizing image style. In *Proceedings of the British Machine Vision Conference*. BMVA Press, 2014.

[3] F. Perronnin. Ava: A large-scale database for aesthetic visual analysis. In *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, CVPR '12, pages 2408–2415, 2012.

[4] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

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

[6] M. J. Wilber, C. Fang, H. Jin, A. Hertzmann, J. Collomosse, and S. Belongie. Bam! the behance artistic media dataset for recognition beyond photography. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

*This work was done while the author was at IIIT Hyderabad.

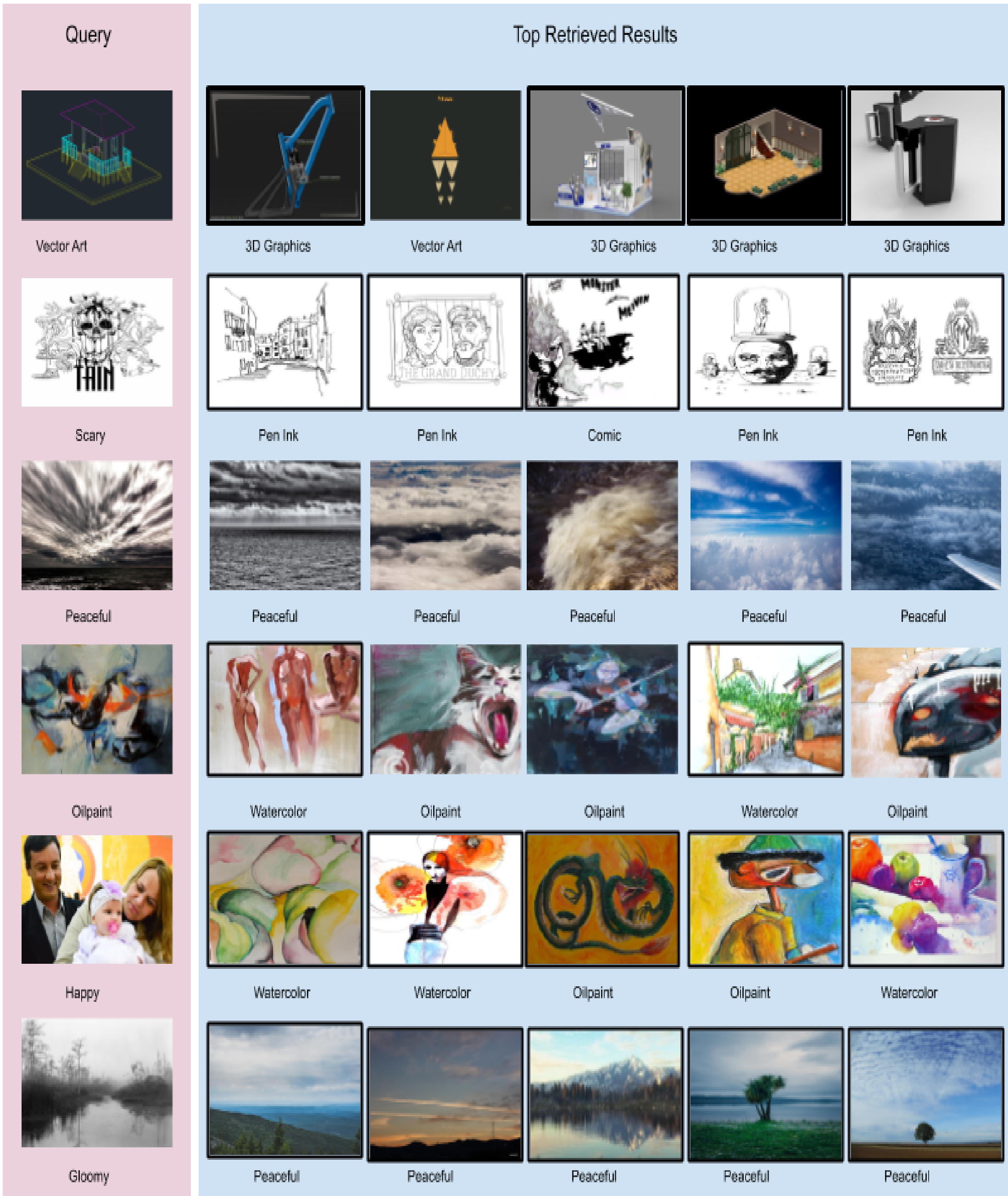

Figure 1. Nearest Neighbour retrieval results for select queries from BAM subset test split. Notice that for rows 1 and 2, the queries and neighbours are very similar looking but the labels do not match. This indicates the lower mAP scores for retrieval using unsupervised methods. ‘Oil Paint’ and ‘Water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

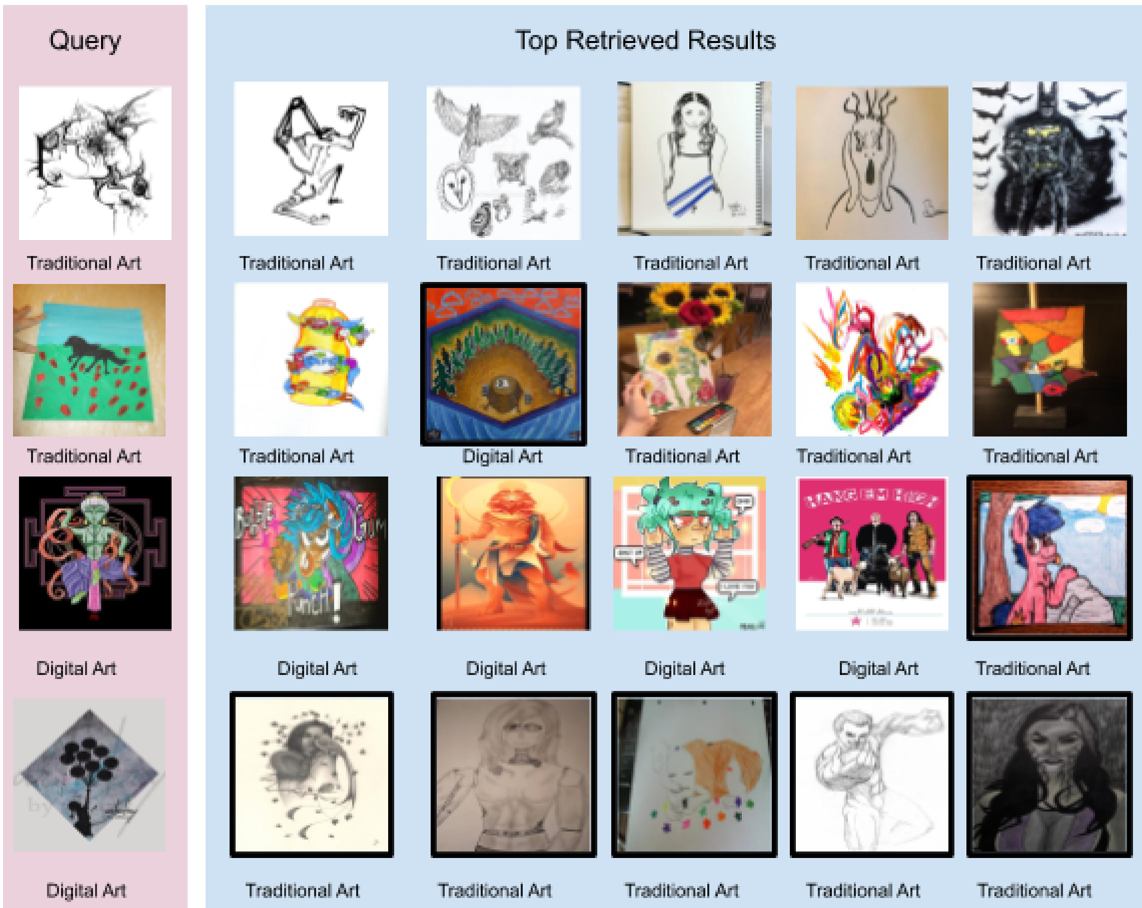

Figure 2. Retrieval Results for Query and Top Neighbours Deviantart dataset.

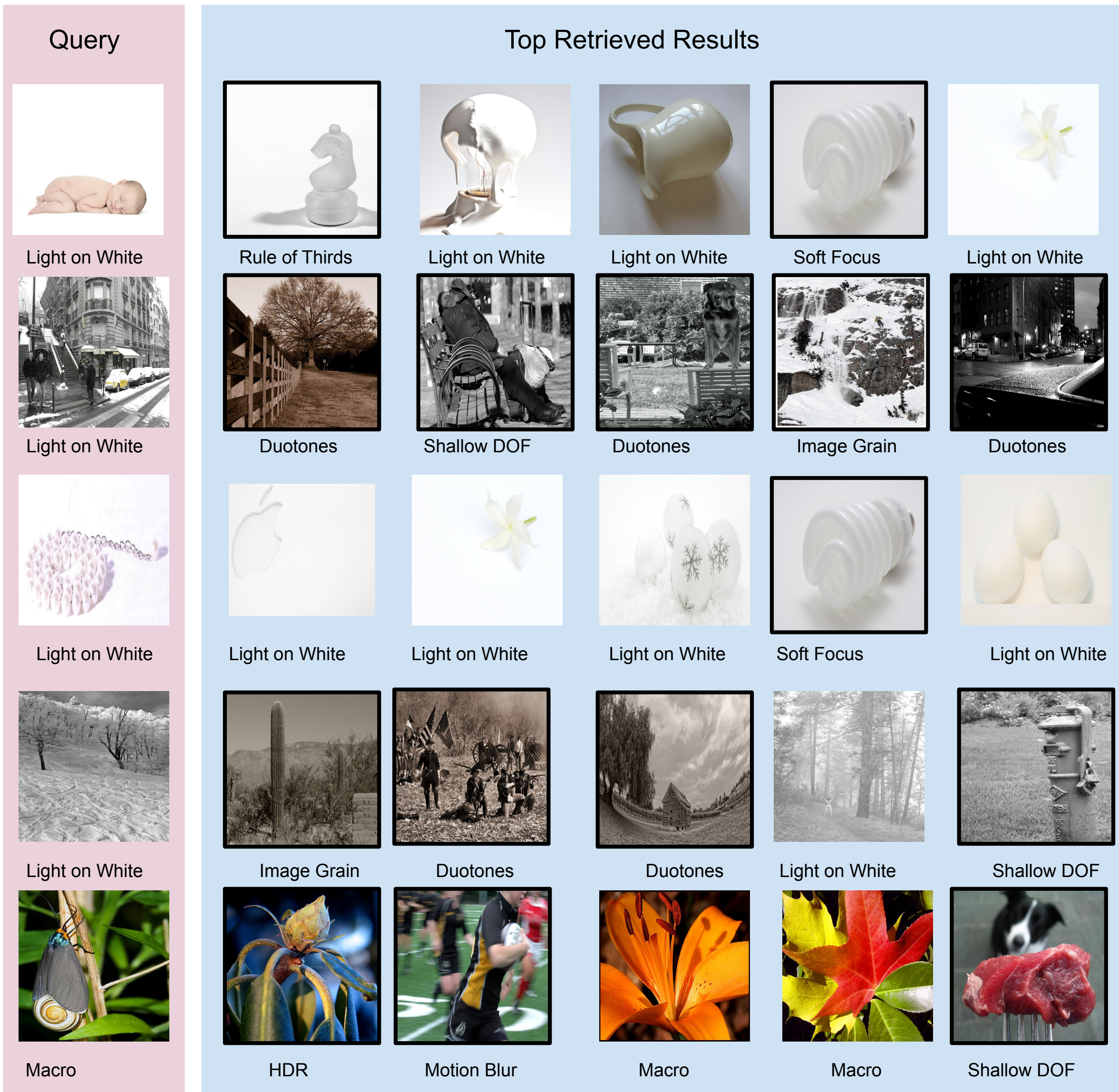

Figure 3. Retrieval Results for Query and Top Neighbours AVA Style dataset.

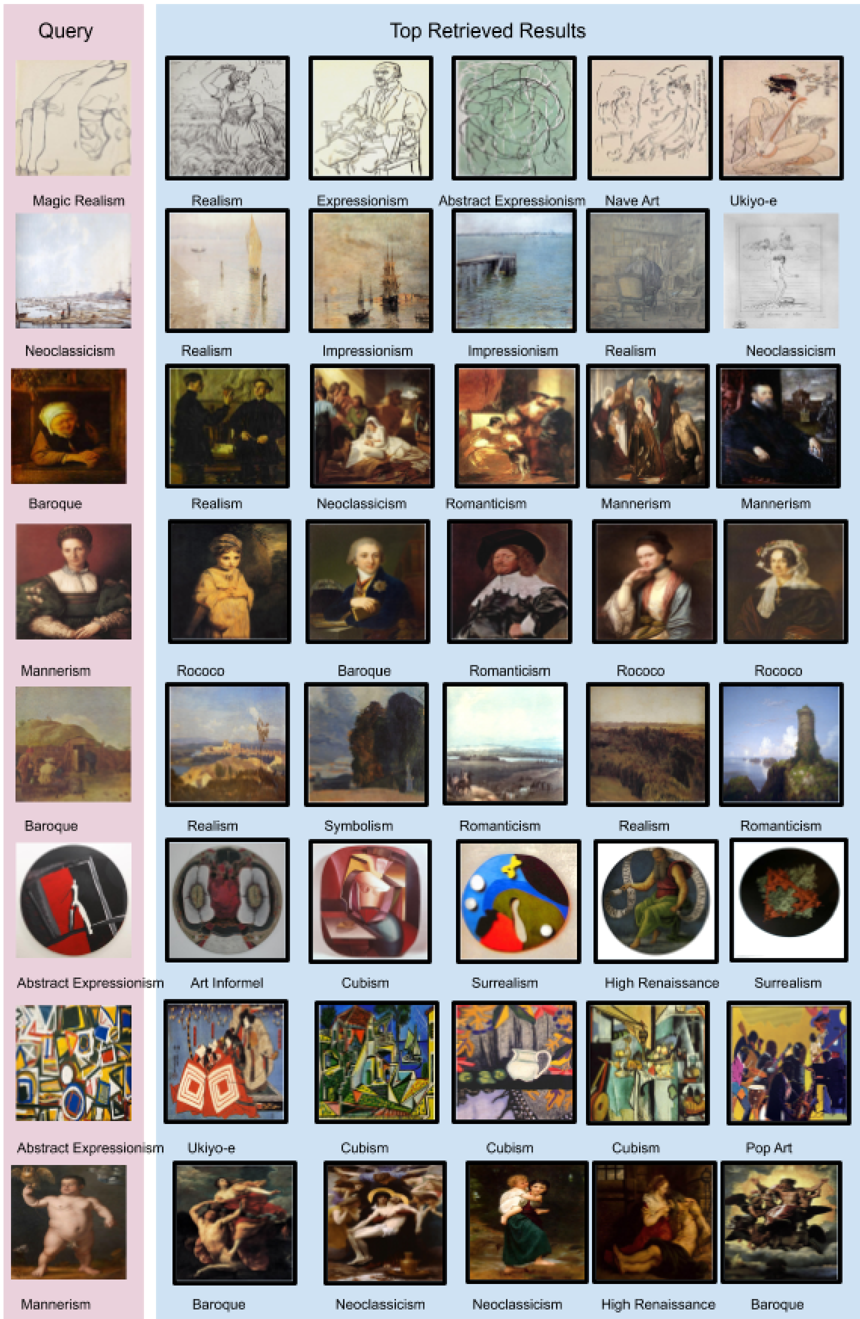

Figure 4. Retrieval Results for Query and Top Neighbours Wikipaintings Subset dataset.It is interesting to see the retrieved results and their relevance with respect to the query image. Notice row 7 where, ‘Abstract Expressionism’ labelled query retrieves ‘Ukiyo-e’, ‘Cubism’ and ‘Pop Art’ paintings.

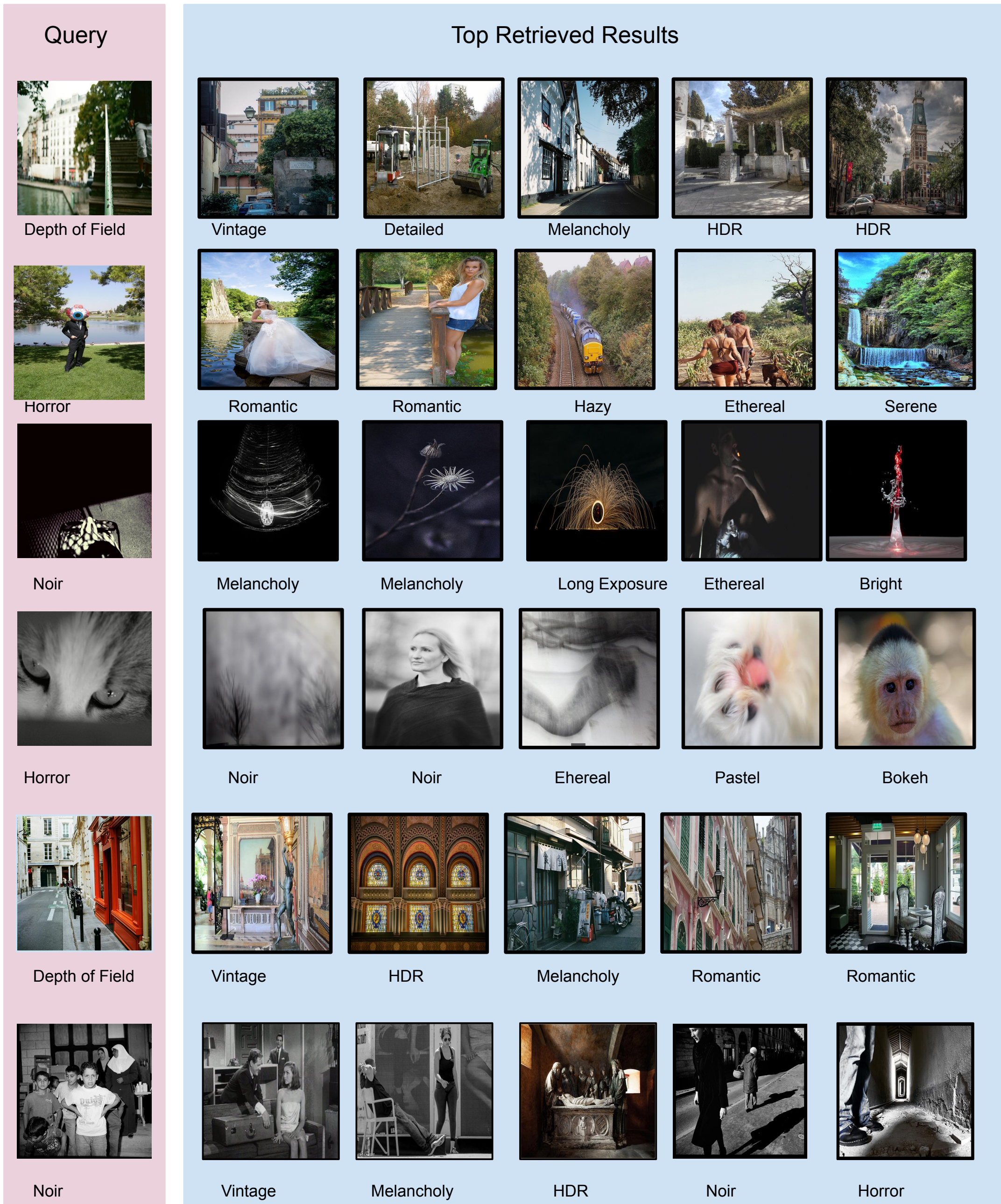

Figure 5. Retrieval Results for Query and Top Neighbours Flickr Test Set.

Figure 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.

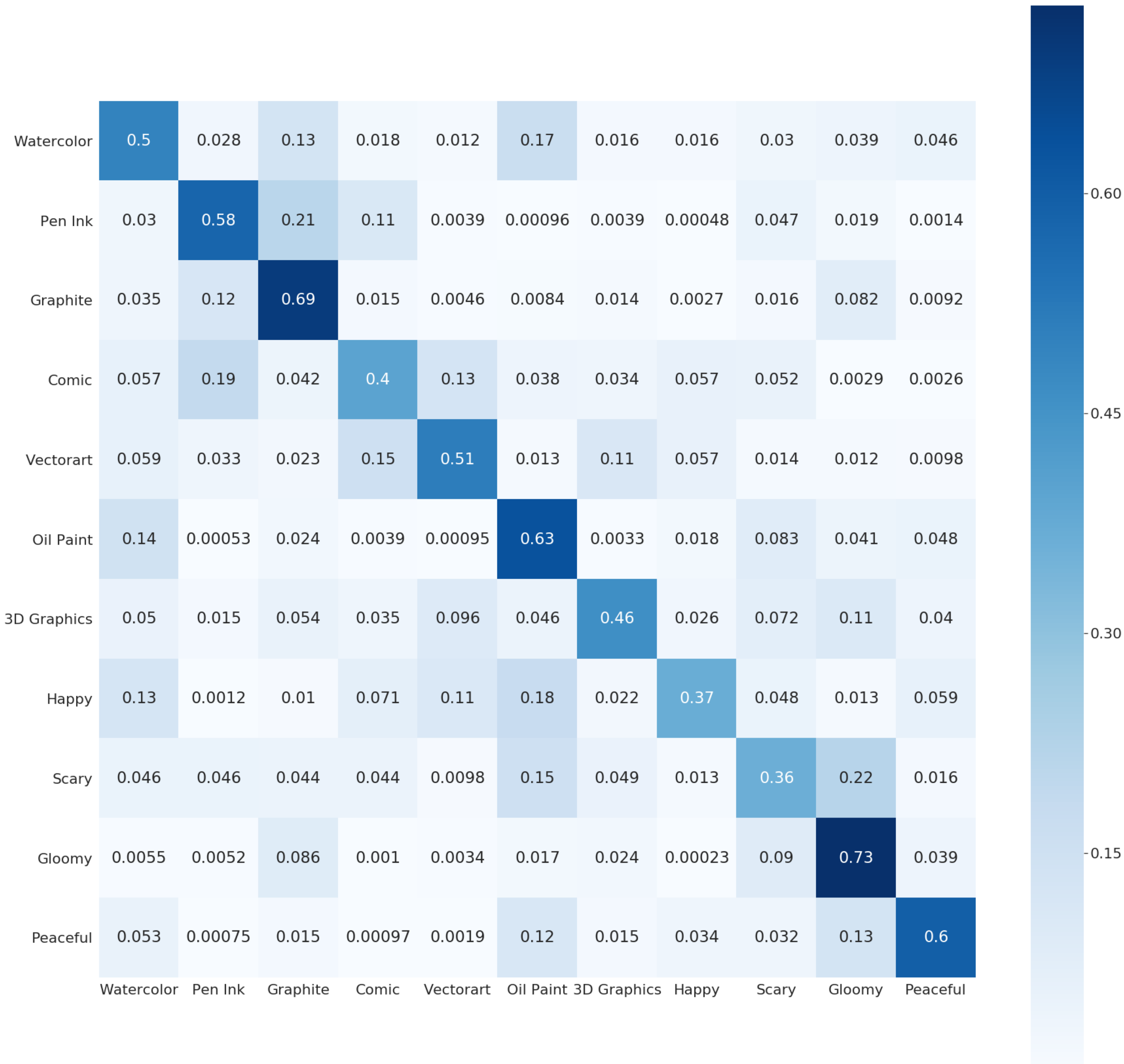

Figure 7. Confusion Matrix for Top 100 retrievals for 1000 Query images on Behance Subset Test set using learnt representations. Here we see the following pairs confusing with each other - ‘Watercolor’ with ‘Oilpainting’ since both are very colourful, ‘Graphite’ and ‘Pen Ink’ both are hand-drawn and dull, and ‘3D Graphics’ with ‘Vectorart’.

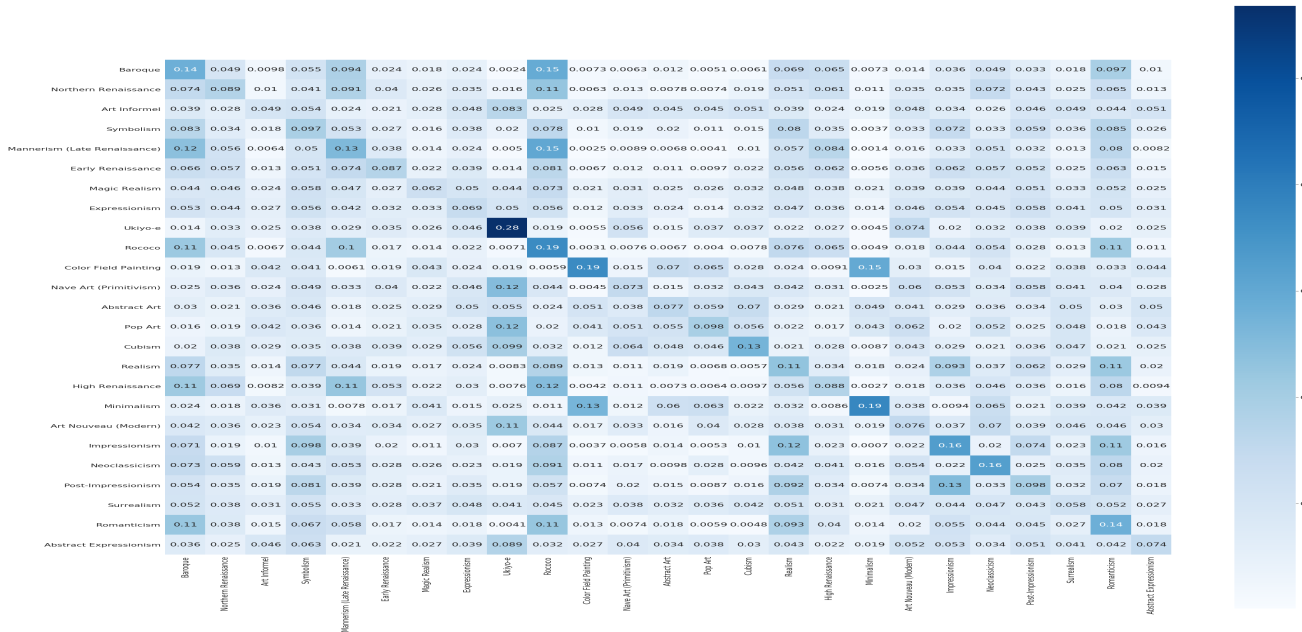

Figure 8. Confusion Matrix for Top 100 retrievals for 1000 Query images on Wikipaintings Subset Test set using learnt representations.

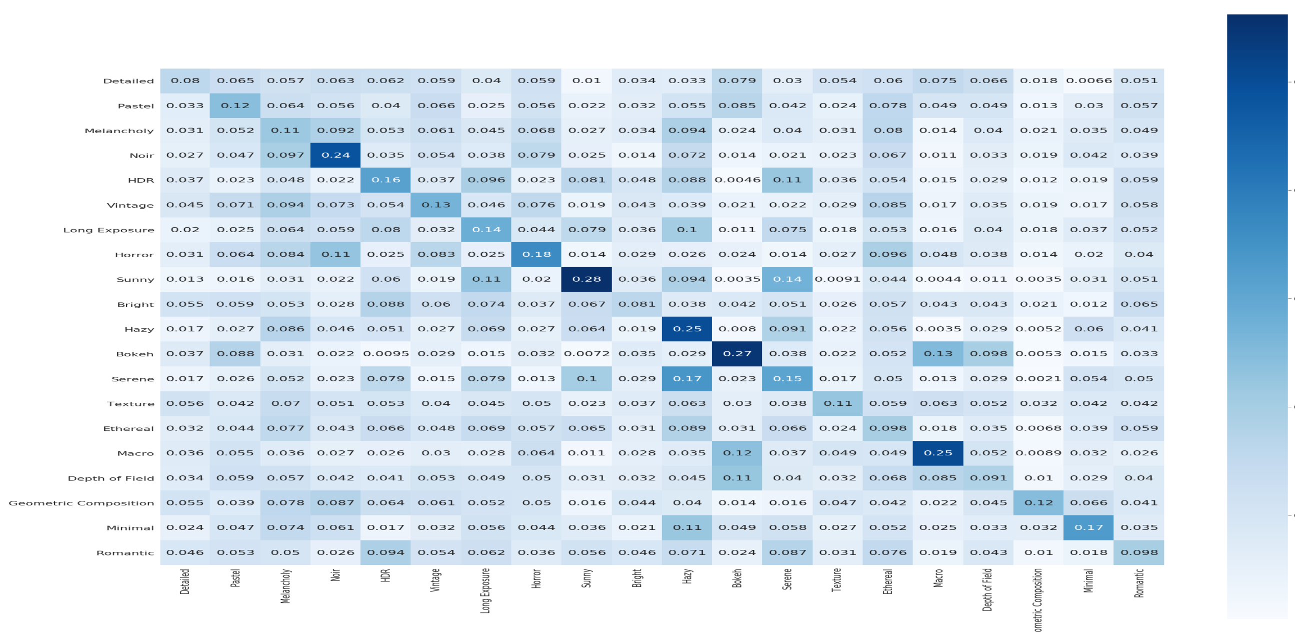

Figure 9. Confusion Matrix for Top 100 retrievals for 1000 Query images on Flickr Test set using learnt representations.

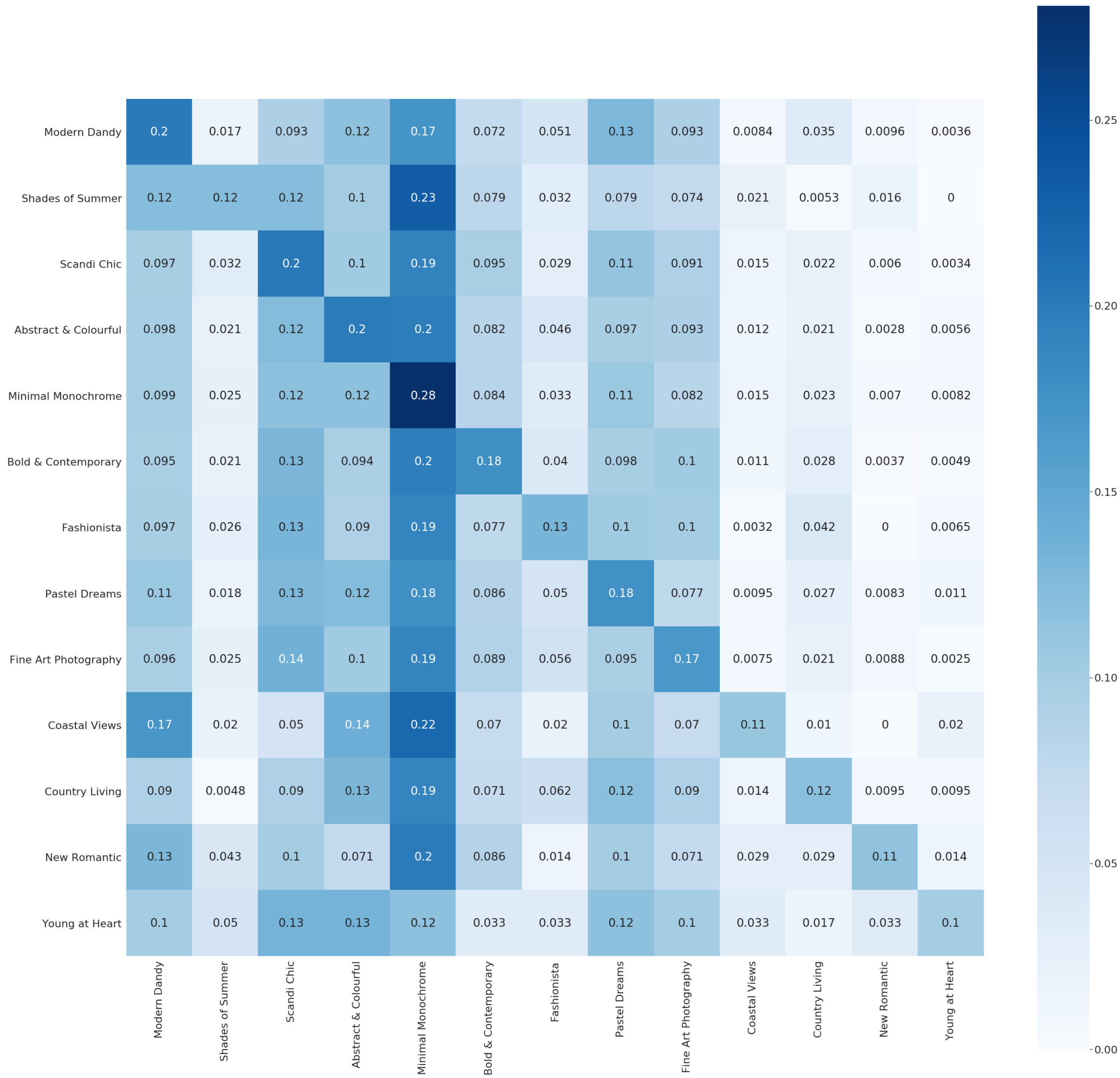

Figure 10. Confusion Matrix for Top 20 retrievals for 100 Query images on WallArt Test set using learnt representations for 13 style themes.

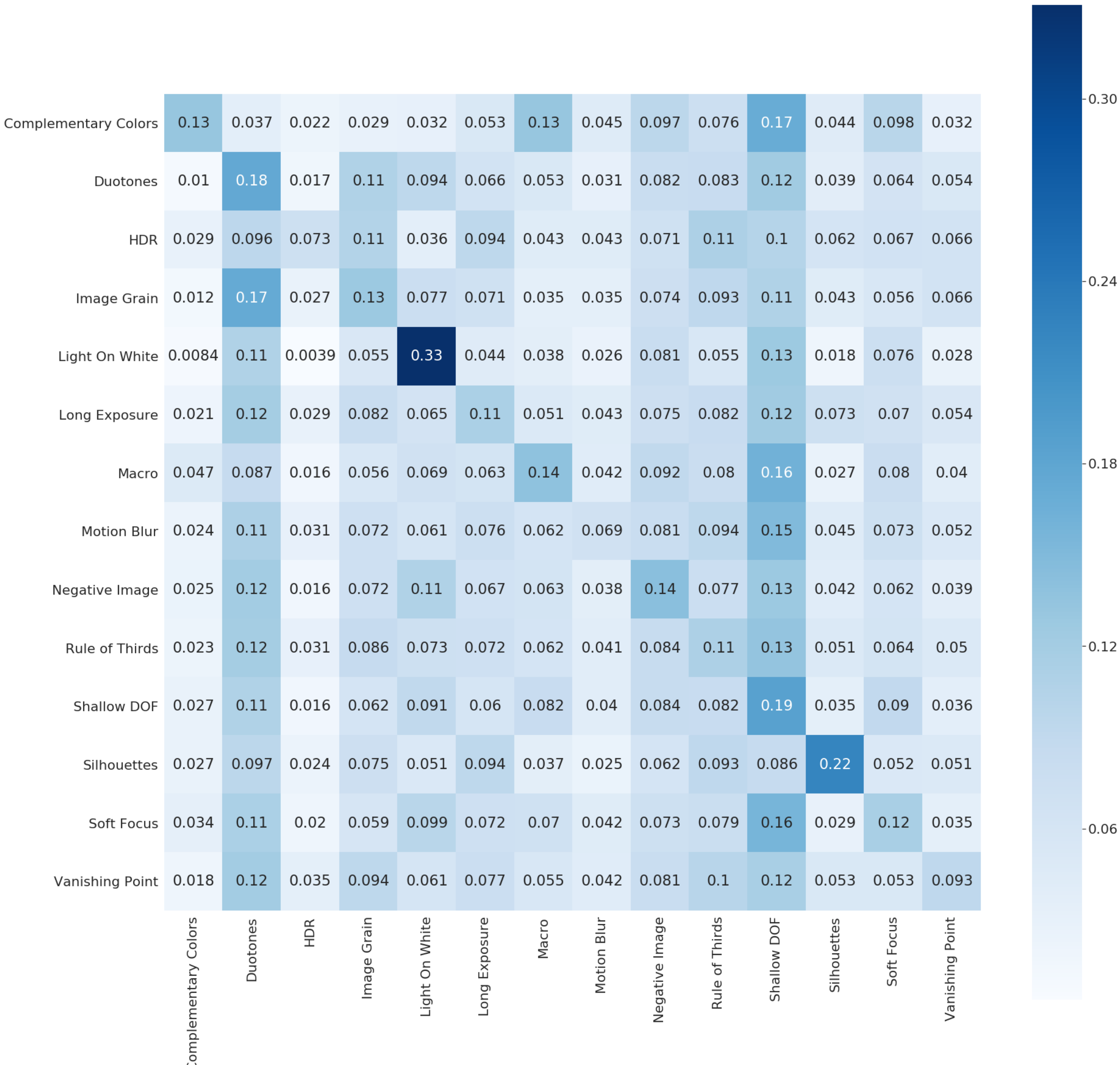

Figure 11. Confusion Matrix for Top 100 retrievals for 200 Query images on AVA Style Test set using learnt representations.

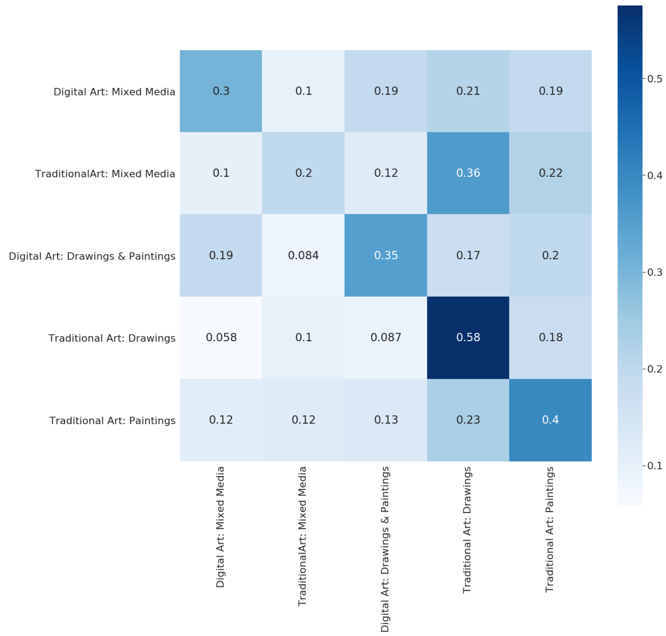

Figure 12. Confusion Matrix for Top 50 retrievals for 100 Query images on Deviant Art Test set using learnt representations for 5 labels.

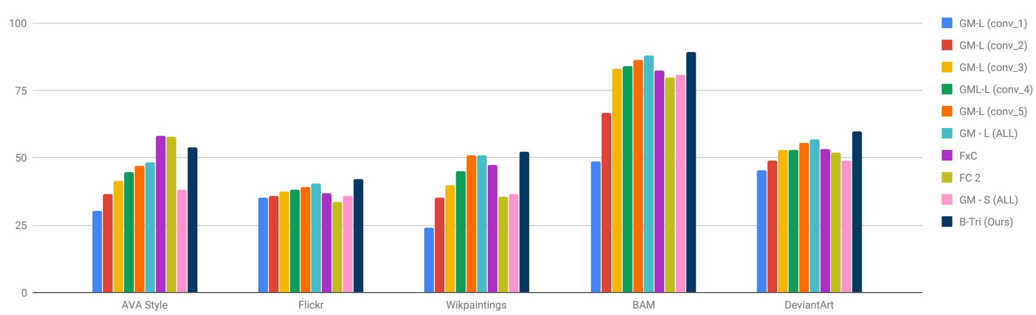

Figure 13. Dataset wide mAP scores for style based classification using different features. Notice that B-Tri features clearly show improvement over other features across most datasets.

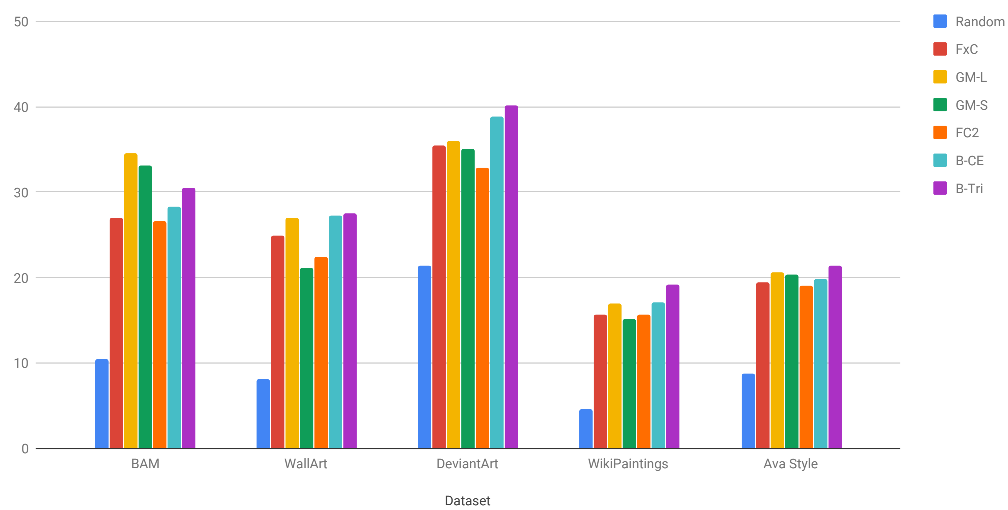

Figure 14. Dataset wide mAP scores for retrieval performance using different features. Notice that B-Tri and B-CE features clearly show improvement over other features across most datasets.

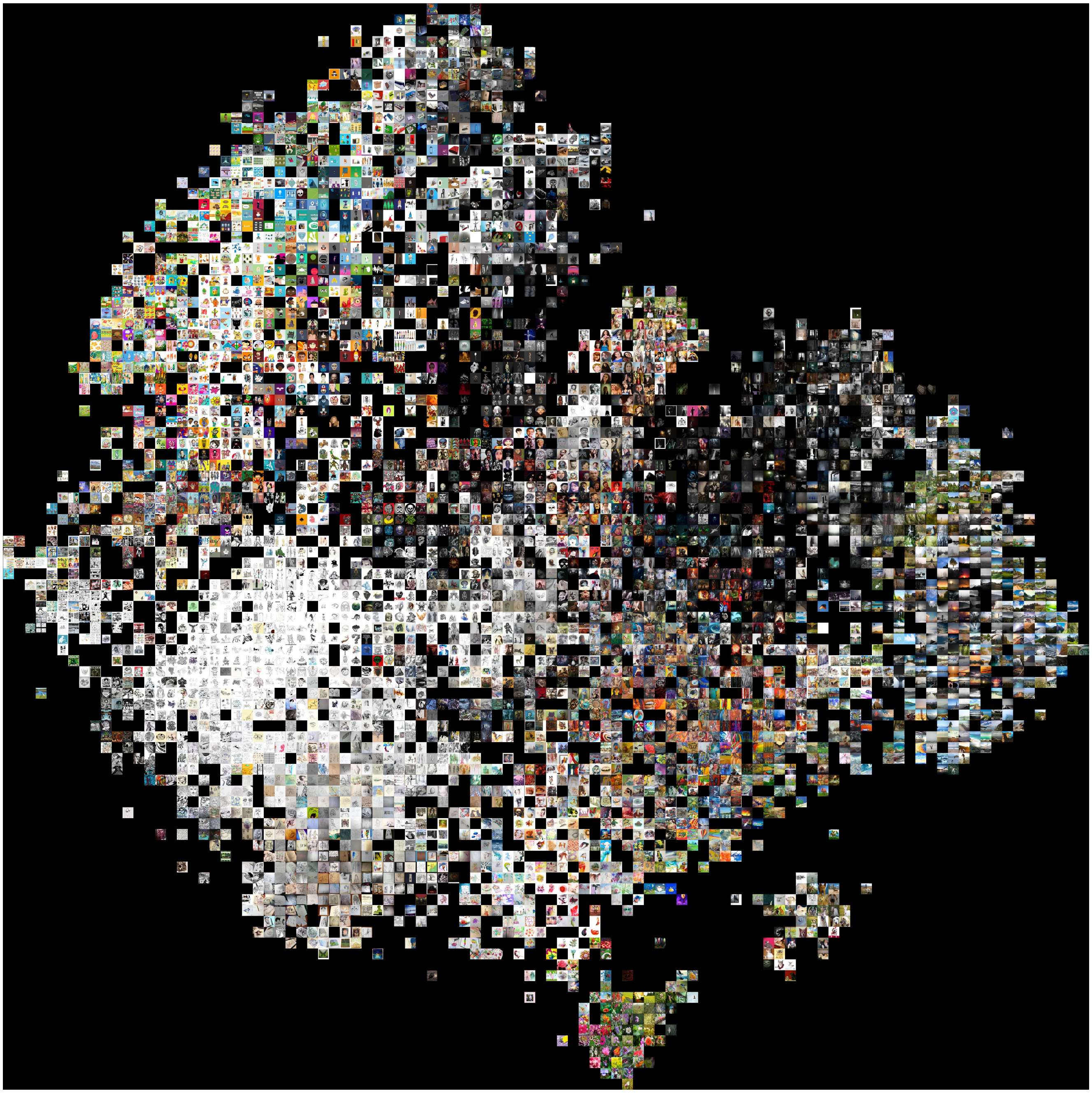

Figure 15. t-SNE visualization on BAM dataset for FC2 pre-trainined features (4096-D) from VGG19.

Figure 16. t-SNE visualization on BAM dataset for PCA-reduced Gram Matrix (4096-D) pre-trained features from VGG19.

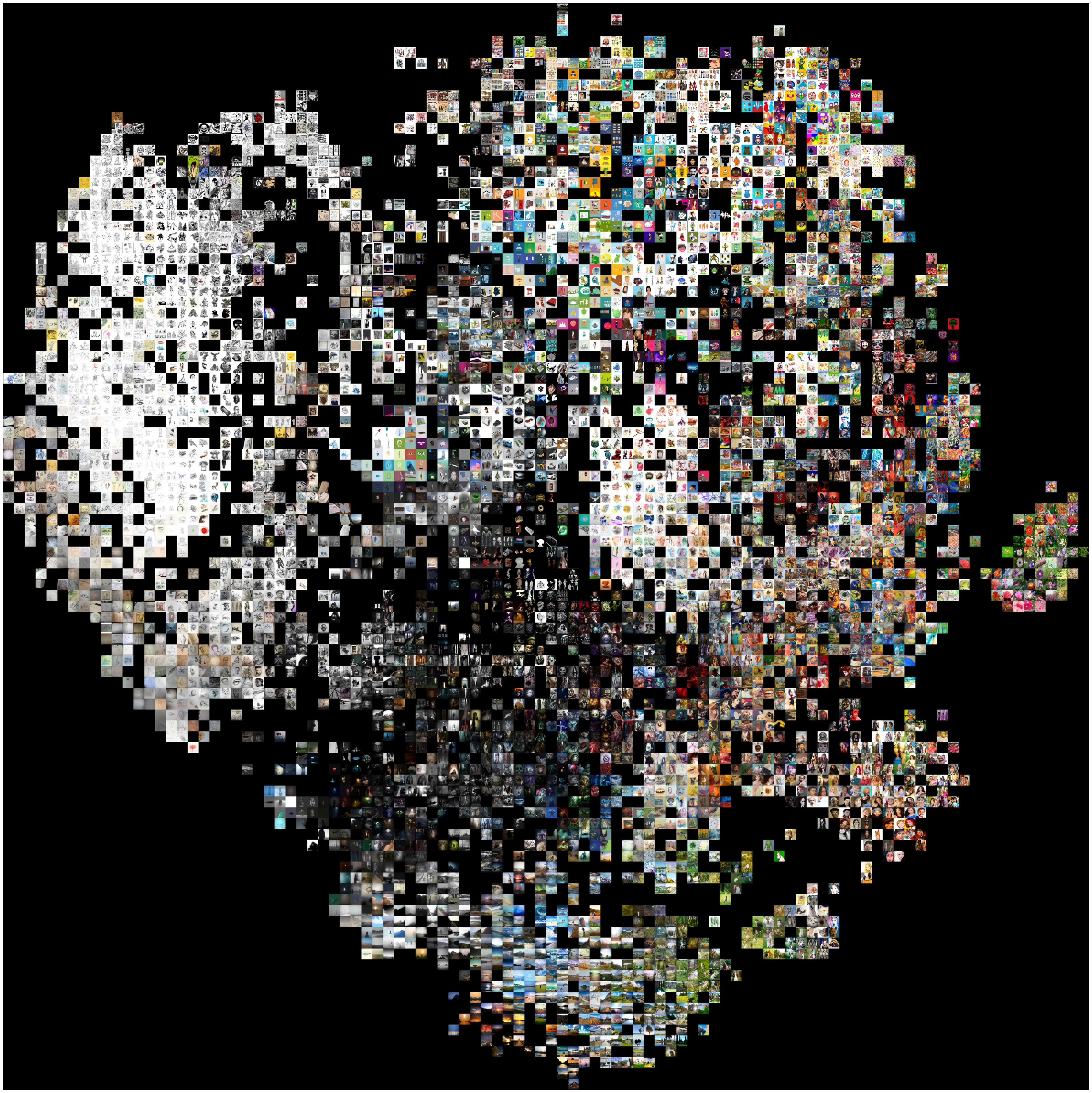

Figure 17. t-SNE visualization on BAM dataset for PCA-reduced Gram Matrix (256-D) pre-trained features from VGG19.

Figure 18. 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.

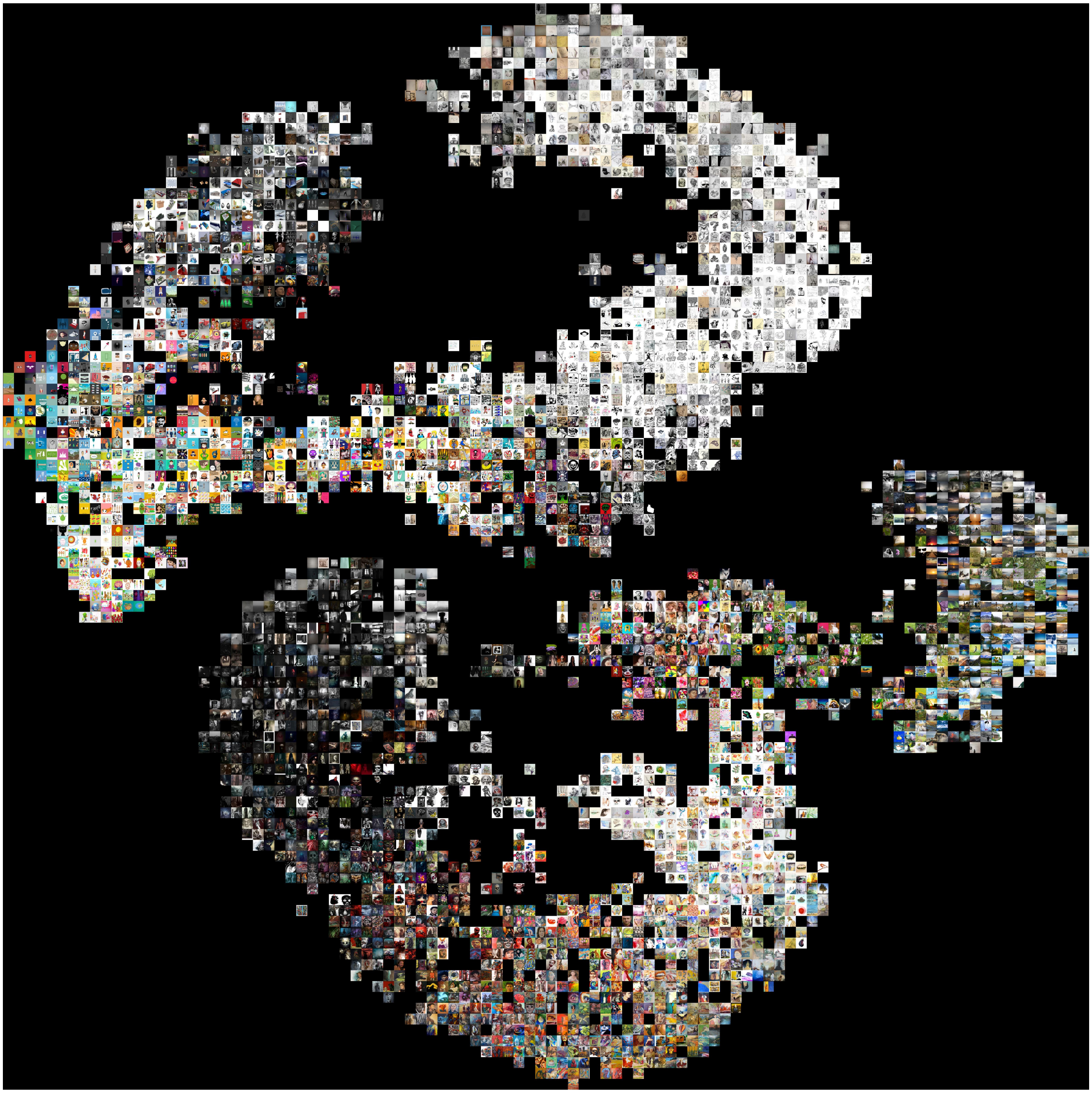

Figure 19. t-SNE visualization on BAM dataset for B-Tri (256-D) features learnt when training with triplet loss. Notice that using triplet loss (B-Tri) further reinforces the stylistic similarity in comparison to other features as can be seen from Figures 15, 16 and 17.

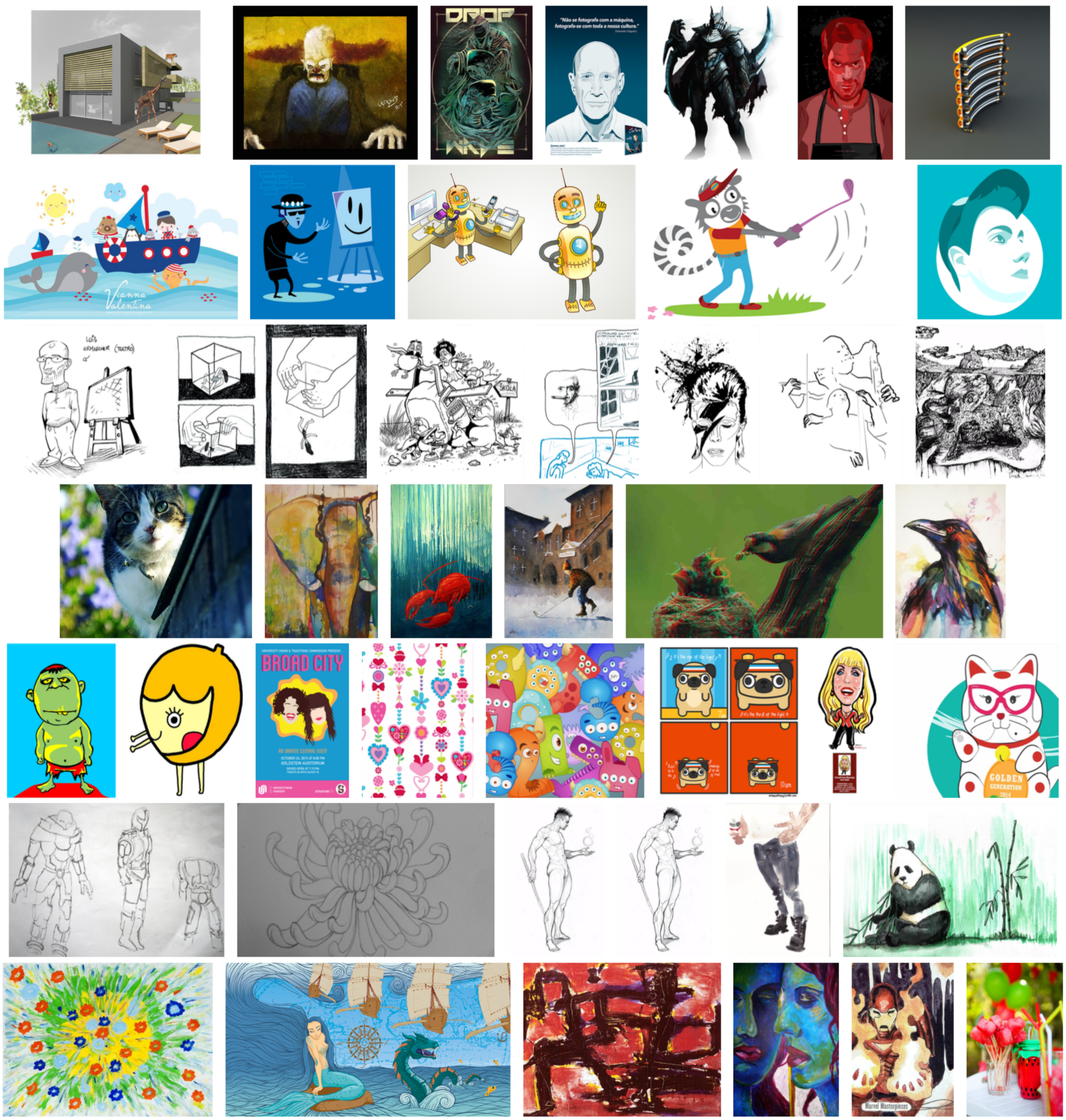

Figure 20. Each row shows examples drawn randomly from seven clusters, for clustering applied to BAM [6] subset. It can be seen that clustering in Gram matrix space groups stylistically similar images together.(Each row only contains samples from a single cluster)

Dataset	Feat. Dim : ~ 4096 Conv 1	Conv 2	Conv 3	Conv 4	Conv 5	All Conv	Feat. Dim : 256 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 1. mAPs for gram matrices computed for different layers (conv1-conv5) of VGG19 [4] Network for recognition using a softmax classifier on different datasets and features. Evidently a combination of all convolutional layers performs best.

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 2. Ava Style dataset (a subset of AVA dataset [3]) similar to [2] 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 3. Flickr dataset [2] 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 4. DeviantArt dataset 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. 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 6. Wikipaintings Subset dataset, which is a subset of Wikipaintings dataset [2] 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 7. 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 [6] very similar to the Behance-Net-TT used in [1].