Paper 5 -
Sketching with Style: Visual Search with Sketches and Aesthetic Context (2017 ICCV)

A triplet network is used to learn a feature embedding capable of measuring style similarity independent of structure.
This model is incorporated within a hierarchical triplet network to unify and learn a joint space from two discriminatively trained streams for style and structure.
Use the Behance Artistic Media Dataset to learn a model of aesthetic style.
Propose a hierarchical triplet convolutional neural network (convnet) architecture to learn a low-dimensional joint embedding for structure and style.

Contributions of the Work -

1. A triplet convnet to learn an embedding for aesthetic style, showing this novel model to outperform by a large margin, previous attempts to use deep convnets for measuring visual style similarity.
2. Build upon our model, incorporating a state of the art convnet for sketch photo similarity [4] to develop a hierarchical triplet convnet for learning a joint space for structural and style similarity over a diverse domain of digital artwork (comprising not only photos, but also paintings, 3D renderings, hand-drawn and vector-art drawings, in a variety of media).
3. Demonstrate and evaluate the performance of our model within a novel SBIR framework that uniquely accepts a set of contextual images alongside the sketched query shape, enabling stylistic constraint of the visual search.

#cross-domain learning

Method -

The annotations in BAM include

1. semantic categories (bicycle, bird, cars, cat, dog, flower, people, tree)
2. Seven labels - (3D renderings, comics, pencil/graphite sketches, pen ink, oil paintings, vector art, watercolor)
3. Four emotion labels of images likely to induce certain emotions in the viewer (happy, gloomy, peaceful, scary)
4. Short textual captions for a small subset of images.

The work uses images labeled at a precision of > 90%

BAM is adopted due to high diversity of content spanning drawings, paintings, graphics and vector art in contemporary and classic styles.
AVA consists of largely photographic content proposed for aesthetic attribute mining.

Hierarchical Network Architecture

The overall triplet model integrates the style stream and the triplet stream to learn a joint feature embedding within which we measure visual similarity for search.

1. Style Network

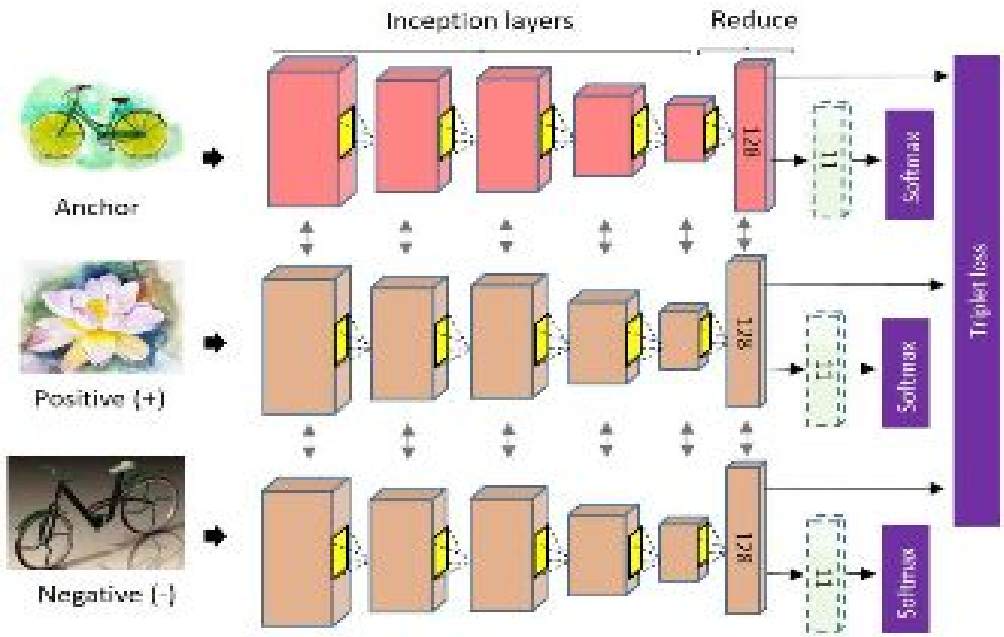
3 branches, each augments GoogLeNet by an addition of a **128-D** inner product layer which acts as a bottleneck after pool5 layer and prior to drop-out.

Trained on 88k artwork training set (Behance-Net-TT)
Evenly partioned into -
11 style categories (S)
Each balanced across the 8 semantic categories (Z)

First trained initially via classification loss (soft-max loss, 30 epochs)
Then Refinement under triplet loss (50 epochs)

Triplet refinement improves the decorrelation between semantics and style, discouraging learned correlations with objects(e.g - trees-> peaceful, skulls->scary scenes)

This refinement is shown to provide significant performance gain.



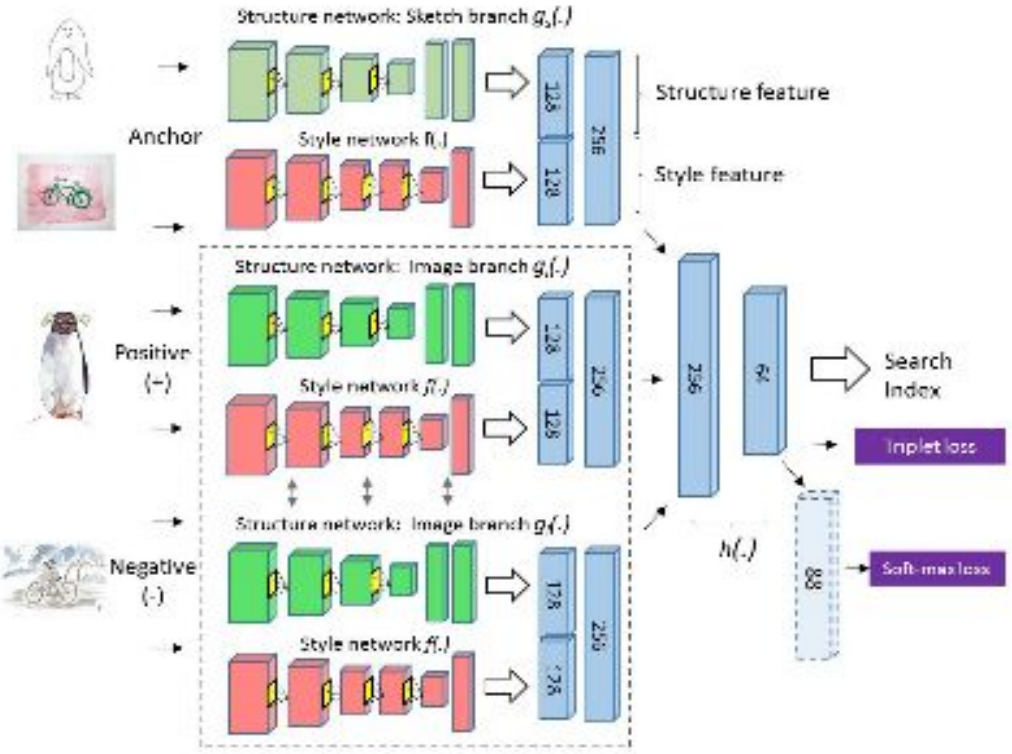
2. Structure Network

The triplet model is fine-tuned over BAM. This is from the work of ***Generalisation and Sharing in Triplet Convnets for Sketch based Visual Search.***
.The network learns a joint embedding from exemplar triplets comprising query sketches, positive photos, that match those sketches and negative photos that do not.

This process utilises the TU-Berlin sketch dataset (for the anchor) augmented with social media sourced photographs (for the positive/negative pair).

Final step, involves fine-tuning the network using triplets sampled from representative imagery. The work uses random artwork images sampled from BAM with 480 Tu-Berlin sketches having category overlap.

3. Final Hierarchical (Multi-modal) network



4. Visual Search with Aesthetic Context