Neural Multi-style Transfer with Attention Masking

Richard Li, Zhaoyi Zhang, Yizhou Liu

Current progress:

Implemented architecture and conditional instance normalization proposed by A Learned Representation For Artistic Style paper.
Implemented the following content loss and style loss evaluation metrics,

$$\mathcal{L}_s(p) = \sum_{i \in \mathcal{S}} \frac{1}{U_i} \mid\mid G(\phi_i(p)) - G(\phi_i(s)) \mid\mid_F^2$$

$$\mathcal{L}_c(p) = \sum_{j \in \mathcal{C}} \frac{1}{U_j} \mid\mid \phi_j(p) - \phi_j(c) \mid\mid_2^2$$

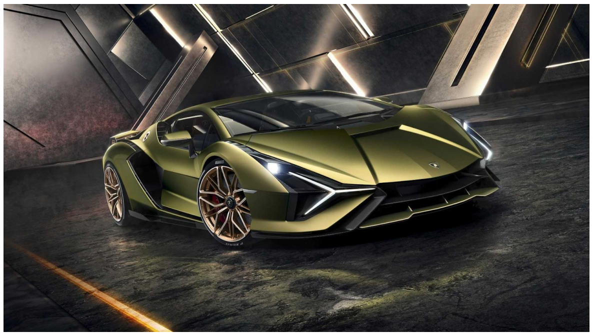
where we utilized a pre-trained vgg16 network.

Training phase:
Datasets: We trained on 5000 images sampled from coco and six(6) style images.
Main parameters: Epoch = 5 , Batch size = 10, Optimizer: Adam, Learning rate(lr): 3e-4

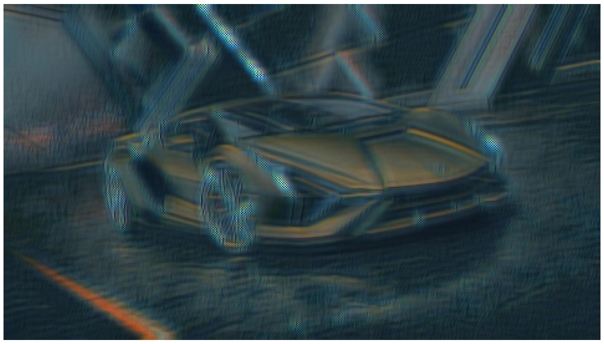
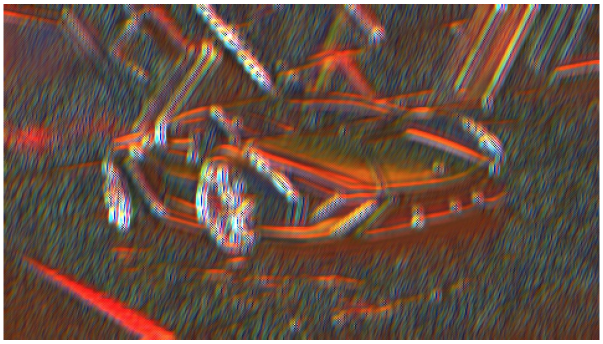
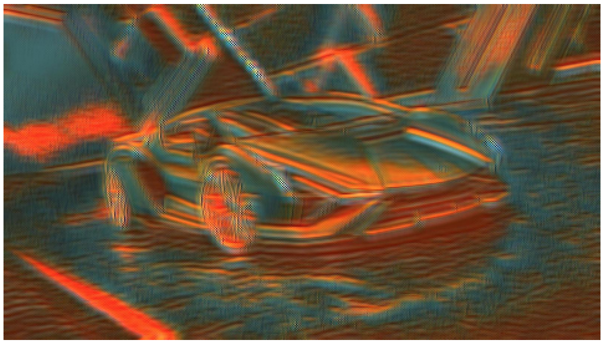
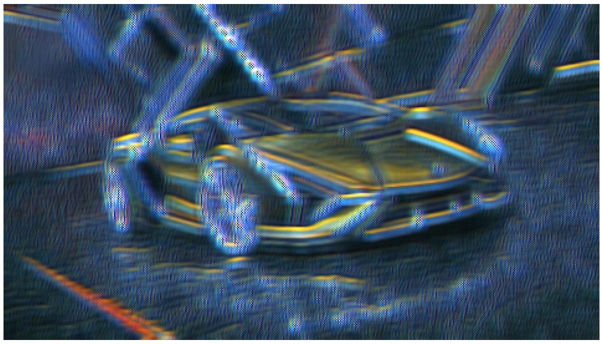
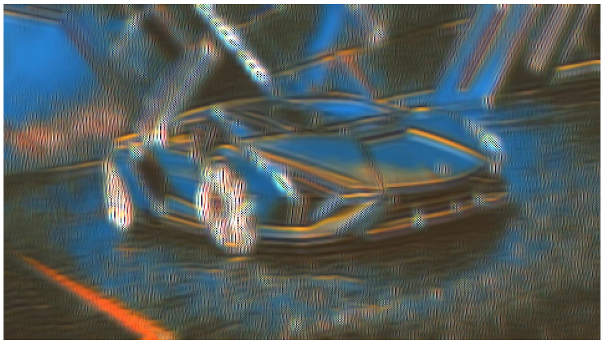
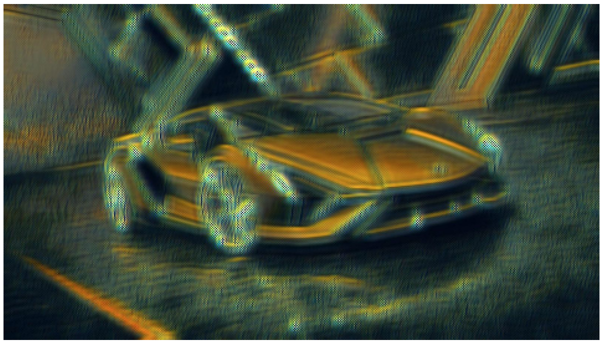
Inference phase:
Input: one content image;
Output: Six (6) stylized images.

Current results are shown below:

Content image:



Stylized images:

(Note : By modifying weighing parameters \alpha and \beta in L_total = \alpha * L_s + \beta * L_c, we can control whether we want our output to be more abstract or concrete.)

Difficulties:

To the best of our knowledge, it is still an unsolved problem for machines to accurately detect multiple “major” objects perceived by humans in an image, which may require physical knowledge. There are two potential major difficulties so far as we found:

1. Due to the limitation of training dataset (feature space learned), NN will fail on outlier detection.
2. For objects which are not in a perfect angle (eg. tilted or rotated), NN has a high change to fail. It is hard to achieve spatial robustness, since it will require physical knowledge for machines to “recognize” those are objects seen before, but just rotated in an angle. (A relevant paper we found about spatial robustness: Exploring the Landscape of Spatial Robustness by Tsipras,.et al(2018)))

Possible future changes in proposal:

1. Explore more schemes on object detection,(Our thought is that we only need the machine to know those which are our “attentions”, but not necessarily how to classify them.) and develop an algorithm which could achieve accuracy above a certain threshold. (We need to compare with others’ works as baselines and shoot for a higher one)
2. If the first one is hard to achieve, we may try the followings:
   a) Go for a lower standard (stay with a lower but acceptable accuracy).
   b) Limit our objects to a certain category (eg. Faces, animals, architectures…) to achieve a satisfactory accuracy.