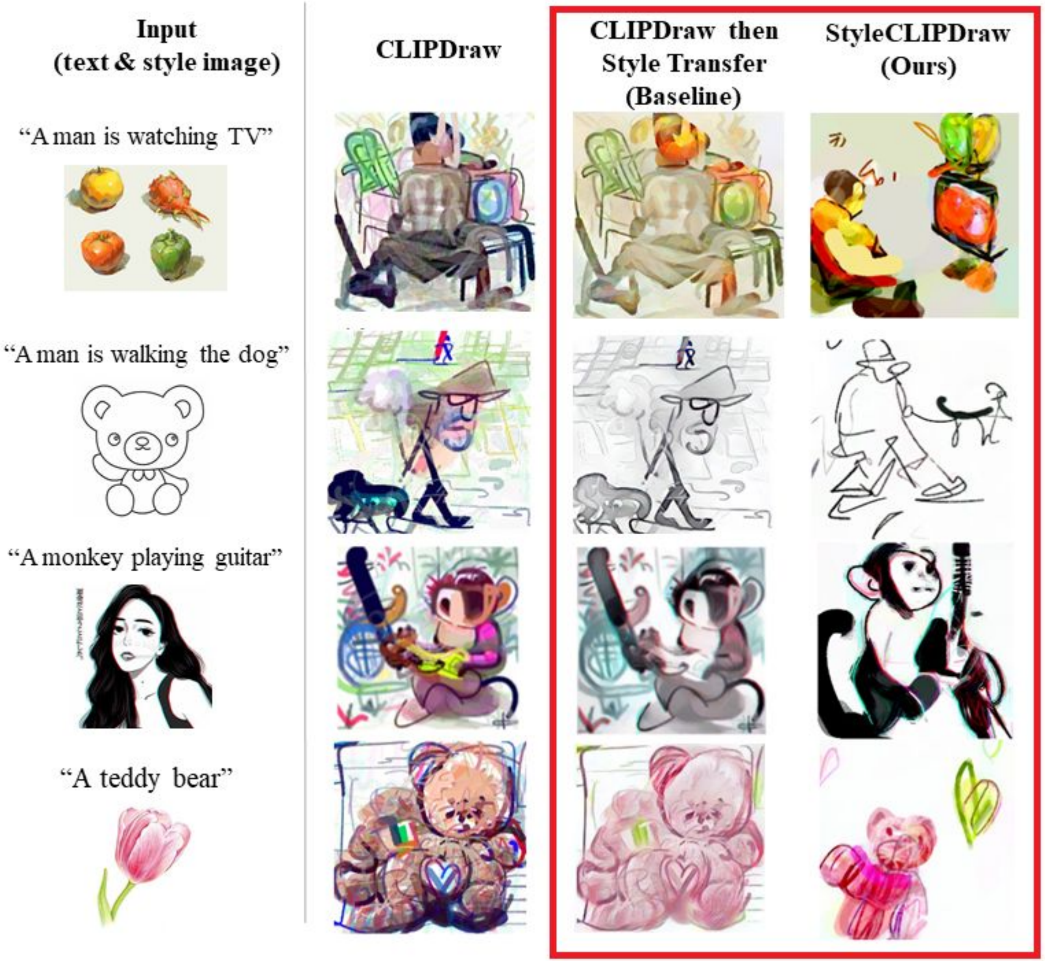
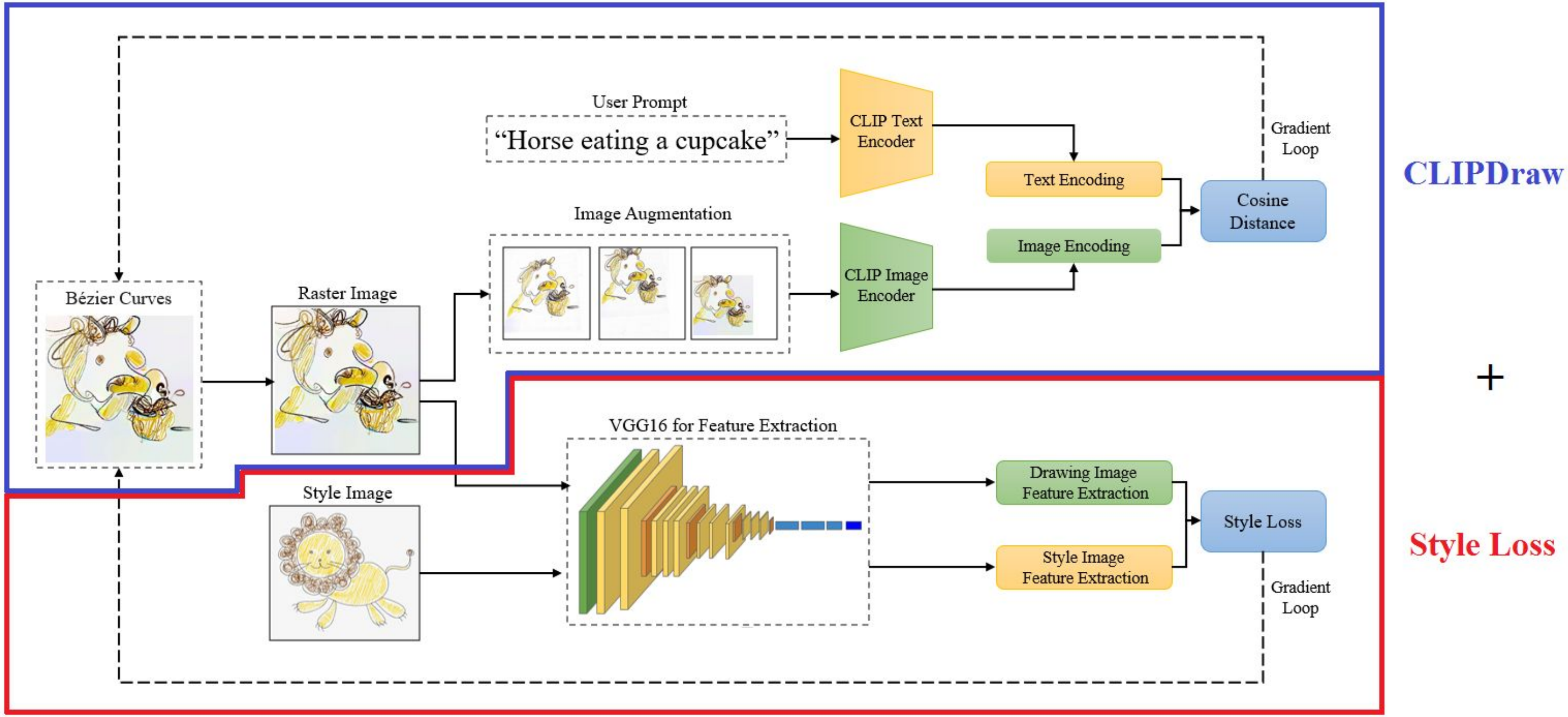
StyleCLIPDraw: Coupling Content and Style in Text-to-Drawing Synthesis

Peter Schaldenbrand
Robotics Institute
Carnegie Mellon University
Pittsburgh, PA 15213
pschalde@andrew.cmu.edu

Zhixuan Liu
School of Data Science
The Chinese University
of Hong Kong
Shenzhen, China
zhixuanliu@cuhk.edu.cn

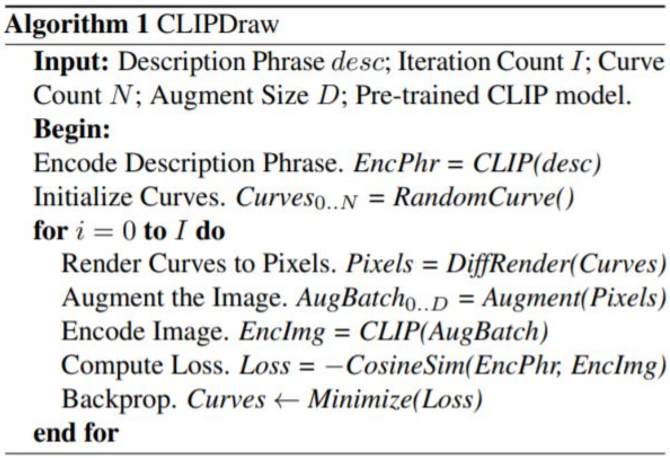
Jean Oh
Robotics Institute
Carnegie Mellon University
Pittsburgh, PA 15213
jeanoh@nrec.ri.cmu.edu


Input
(text & style image)
CLIPDraw
CLIPDraw then Style Transfer (Baseline)
StyleCLIPDraw (Ours)
"A man is watching TV"
"A man is walking the dog"
"A monkey playing guitar"
"A teddy bear"



The StyleCLIPDraw model architecture

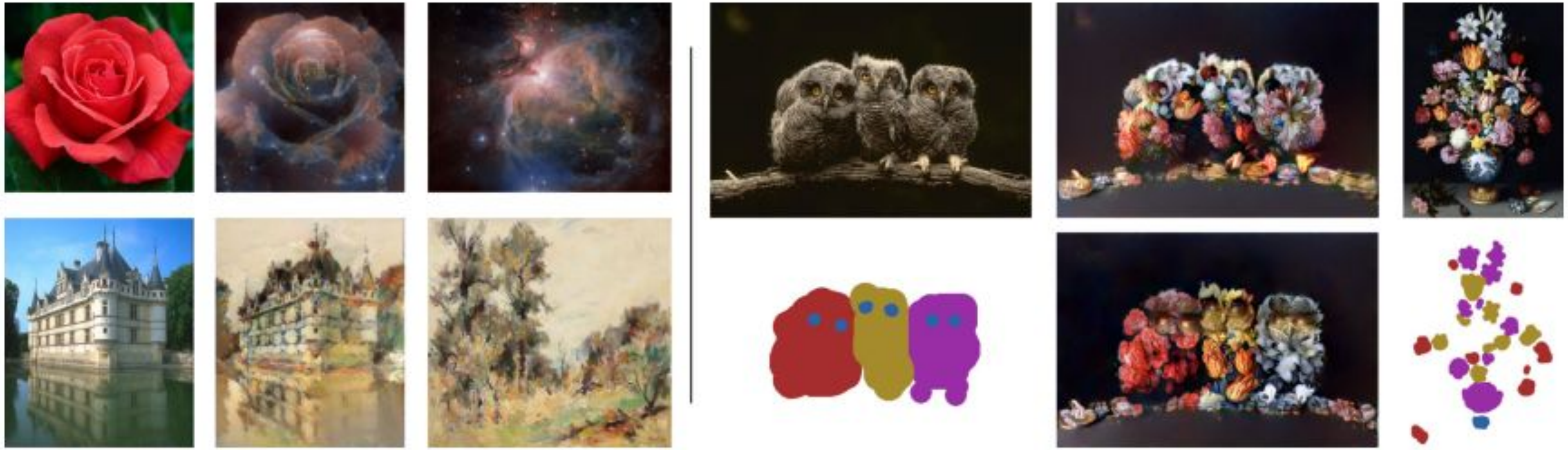
The drawing begins as randomized Bézier curves on a canvas and is optimized to fit the given style and text:



```
Algorithm 1 CLIPDraw
  Input: Description Phrase $desc$; Iteration Count $I$; Curve Count $N$; Augment Size $D$; Pre-trained CLIP model.
  Begin:
  Encode Description Phrase. $EncPhr = CLIP(desc)$
  Initialize Curves. $Curves_{0..N} = RandomCurve()$
  for $i = 0$ to $I$ do
      Render Curves to Pixels. $Pixels = DiffRender(Curves)$
      Augment the Image. $AugBatch_{0..D} = Augment(Pixels)$
      Encode Image. $EncImg = CLIP(AugBatch)$
      Compute Loss. $Loss = -CosineSim(EncPhr, EncImg)$
      Backprop. $Curves \leftarrow Minimize(Loss)$
  end for
```

- The CLIPDraw produces drawings consisting of a series of Bézier curves defined by a list of coordinates, a color, and an opacity.
- The brush strokes are rendered into a raster image via differentiable model.
- The image is augmented to avoid finding shallow solutions to optimizing through the CLIP model.
- The text input and the augmented raster drawing are fed the the CLIP model and the difference in embeddings are compared using cosine distance to compute a loss that encourages the drawing to fit the text input.
- The raster image and the style image are fed through early layers of the VGG-16 model (per the STROTSS style-transfer algorithm) and the difference in extracted features form the loss that encourages the drawings to fit the style of the style image.

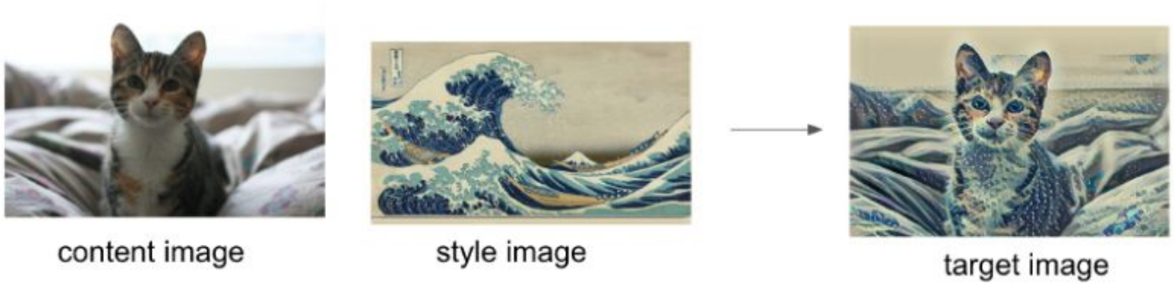
STROSS - Style Transfer by Relaxed Optimal Transport and Self-Similarity



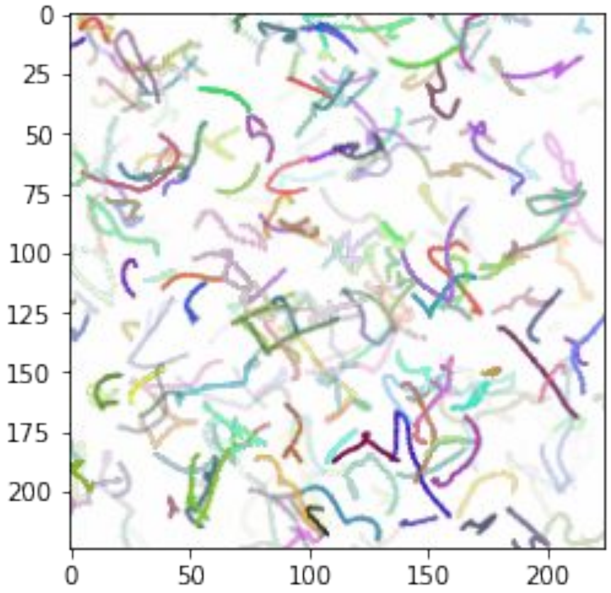
Images are arranged in order of content, output, style

Reimplementation

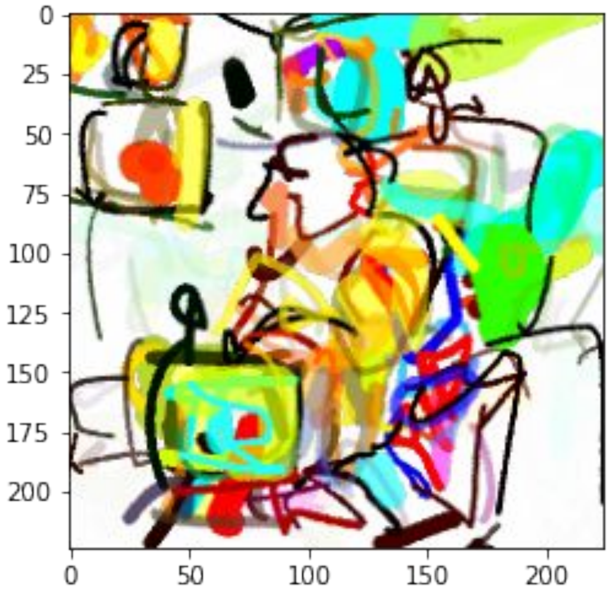
Neural Style Transfer



From Bézier curves to final output

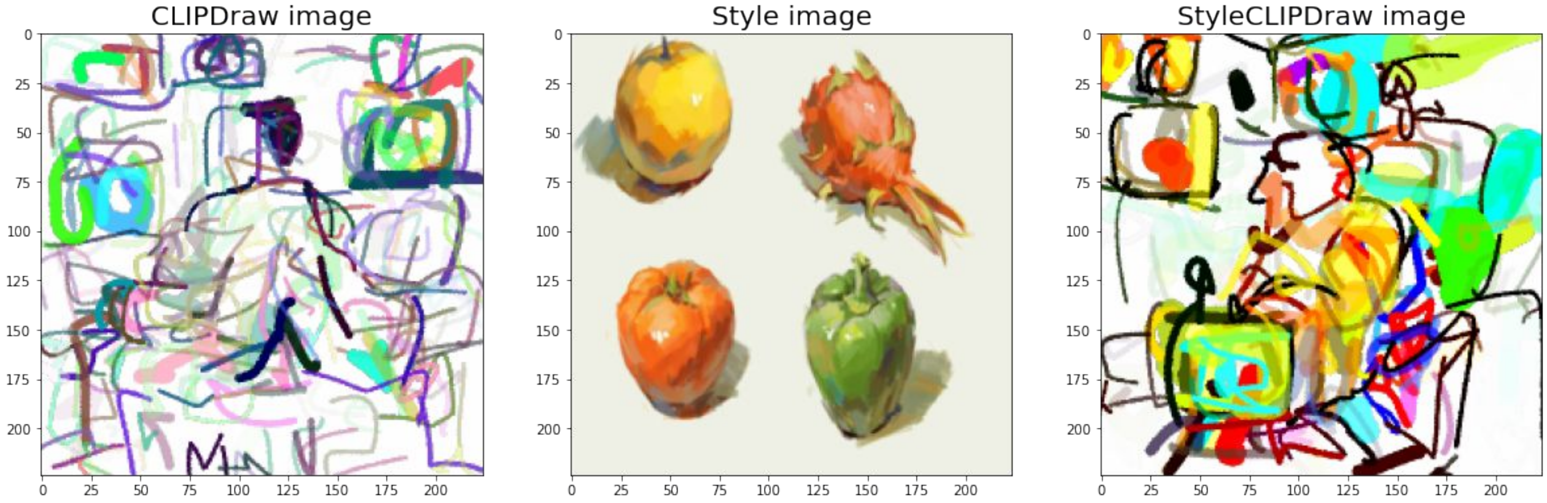


Bézier curves



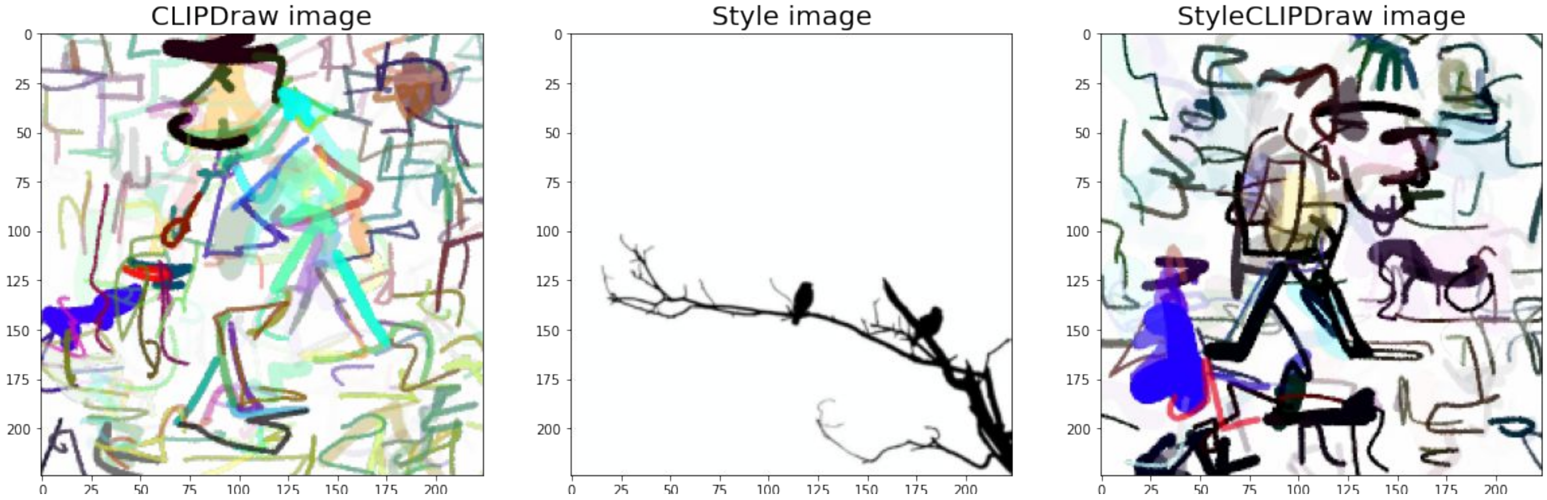
StyleCLIPDraw image

Result



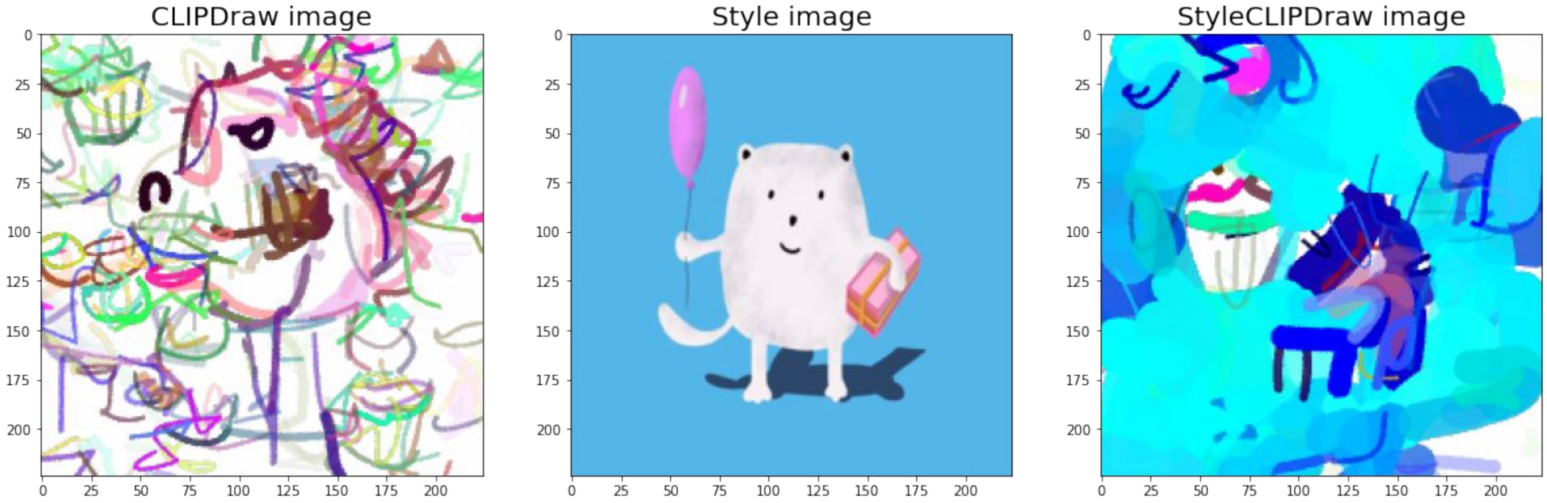
A man is watching TV

Result



A man is walking the dog

Result

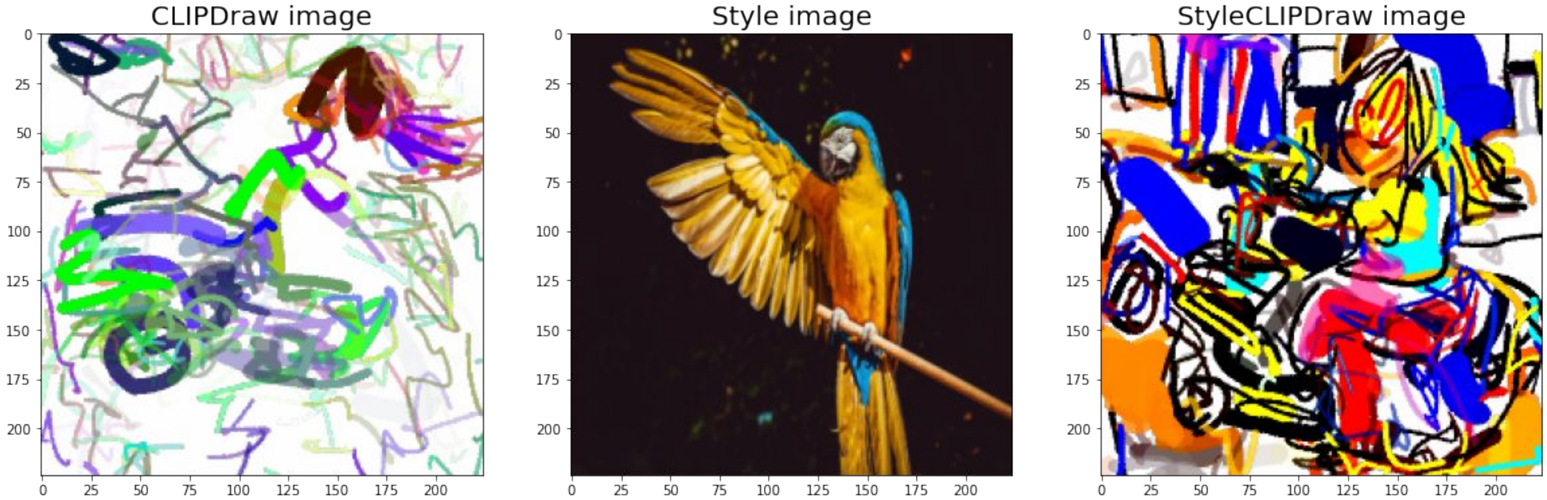

CLIPDraw image

Style image

StyleCLIPDraw image

A horse eating a cupcake

Result



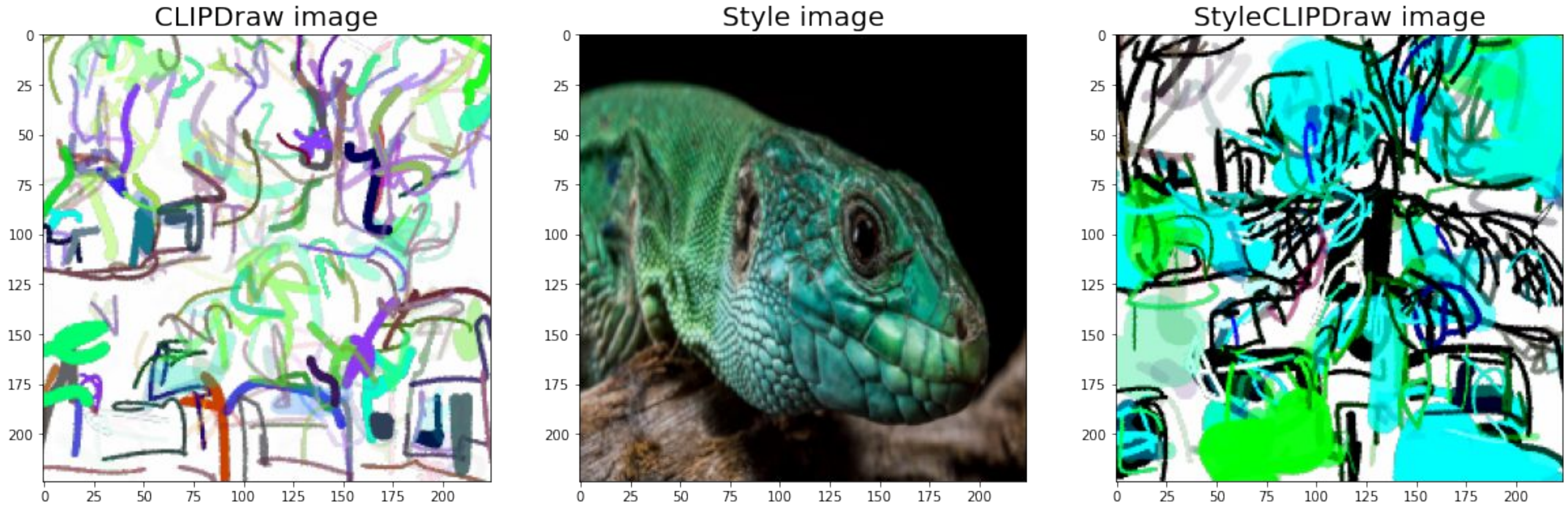
CLIPDraw image

Style image

StyleCLIPDraw image

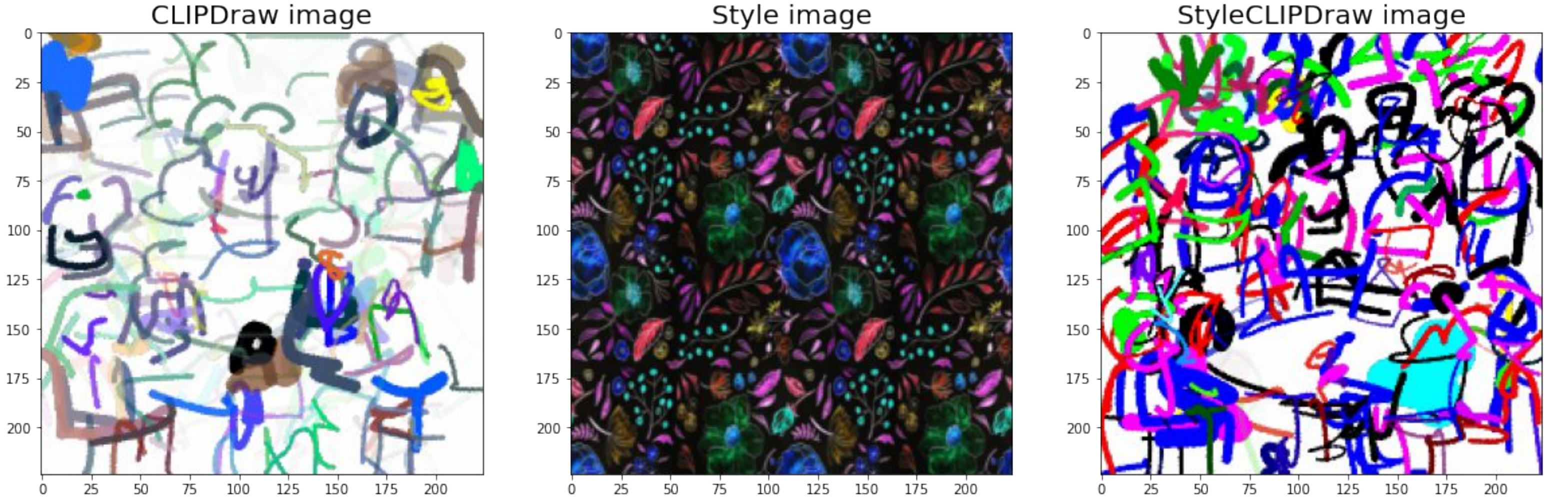
A girl riding a motorcycle

Result



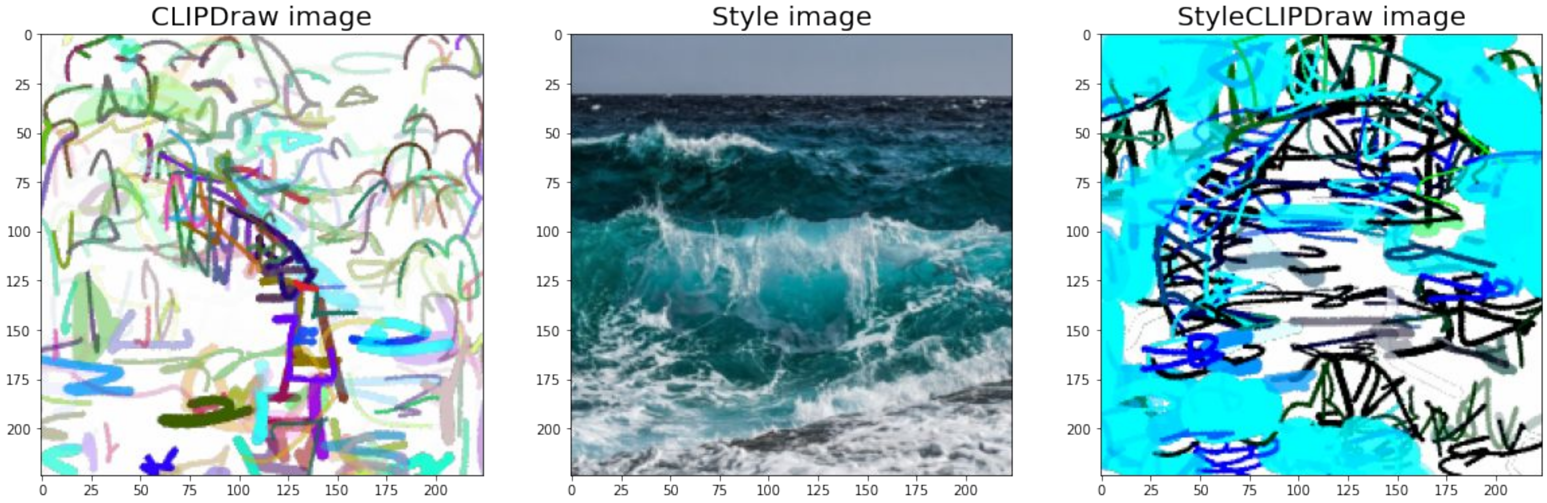
The trees across the street

Result



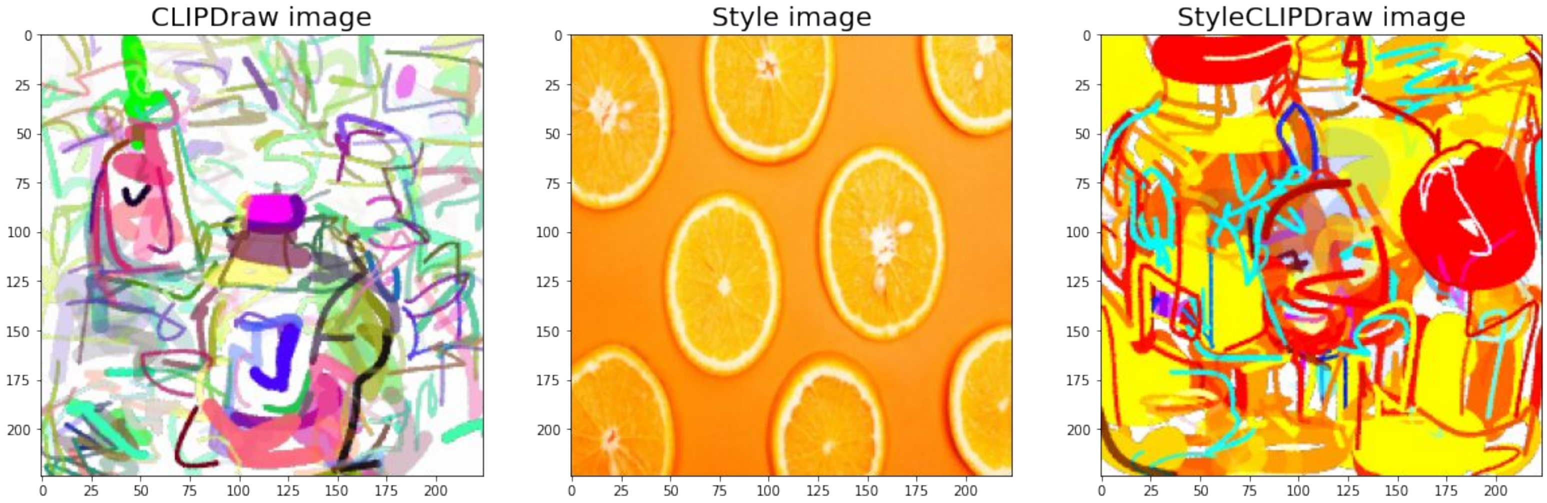
A group of people at the meeting

Result



A bridge over the lake

Result



A bottle of juice

Thank you for your attention!