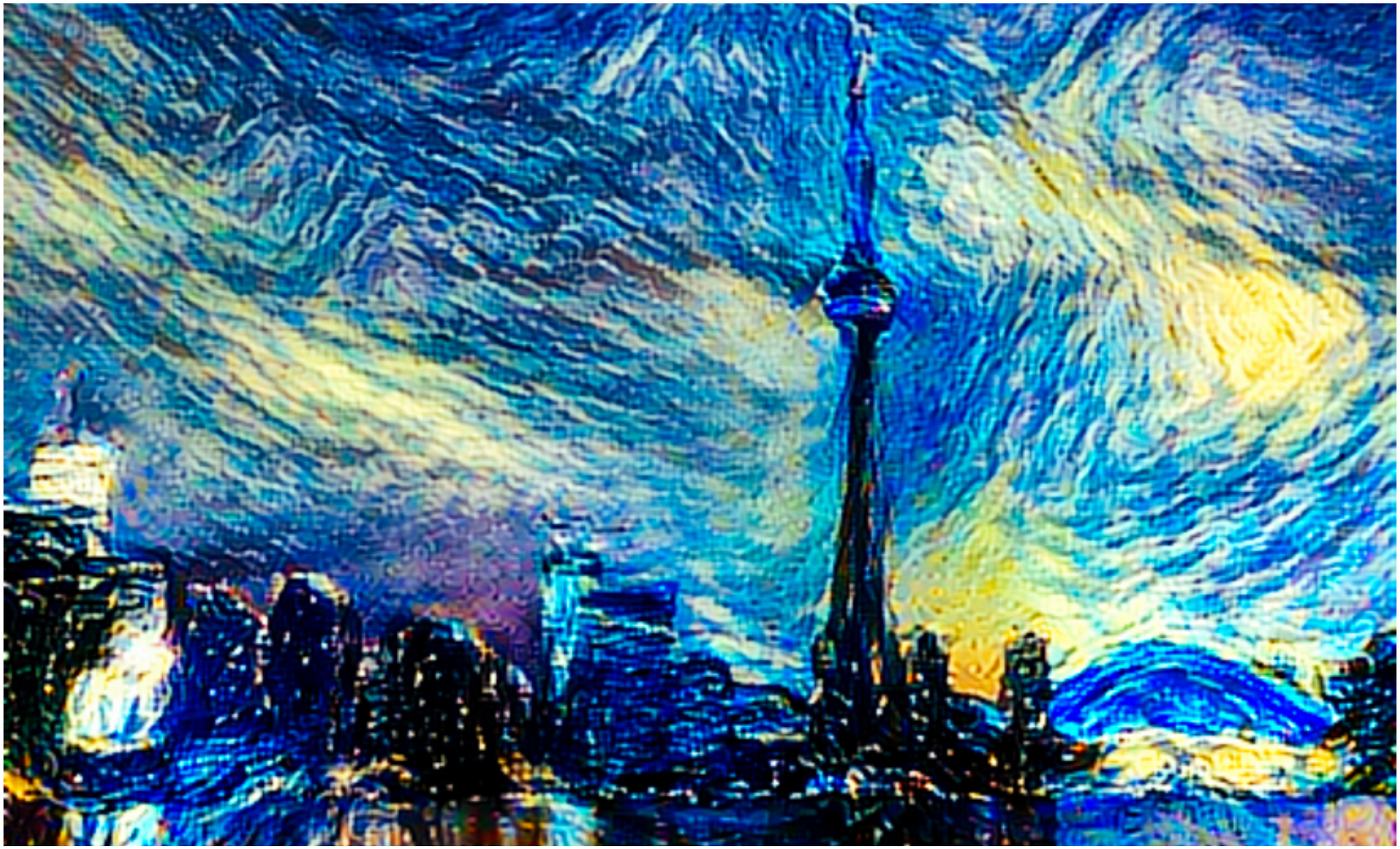


# A Neural Algorithm of Artistic Style (2015)

Leon A. Gatys, Alexander S. Ecker, Matthias Bethge

Nancy Iskander (niskander@dgp.toronto.edu)

# Overview of Method

- *Content*: Global structure. *Style*: Colours; local structures
- Use CNNs to capture style from one image and content from another image.
- Each convolutional layer outputs differently filtered versions of the input. Those layers are used in both content and style reconstructions.
- Images are transformed to representations (in convolutional layers) that emphasize content and de-emphasize specific pixel values.
- Content is reconstructed using those representations, and style is represented as correlations between them.

# Motivation for method

- NPR style/texture transfer methods are typically applied to pixel representations directly.
- By using Deep Neural Networks trained on object recognition (VGG), manipulations are carried out in feature spaces that explicitly represent the high level content of an image.

# Reconstructing an image from a convolutional layer

- Representation function:

$$\Phi : \mathbb{R}^{H \times W \times C} \to \mathbb{R}^d$$

- Representation:

$$\Phi_0 = \Phi(\mathbf{x}_0)$$

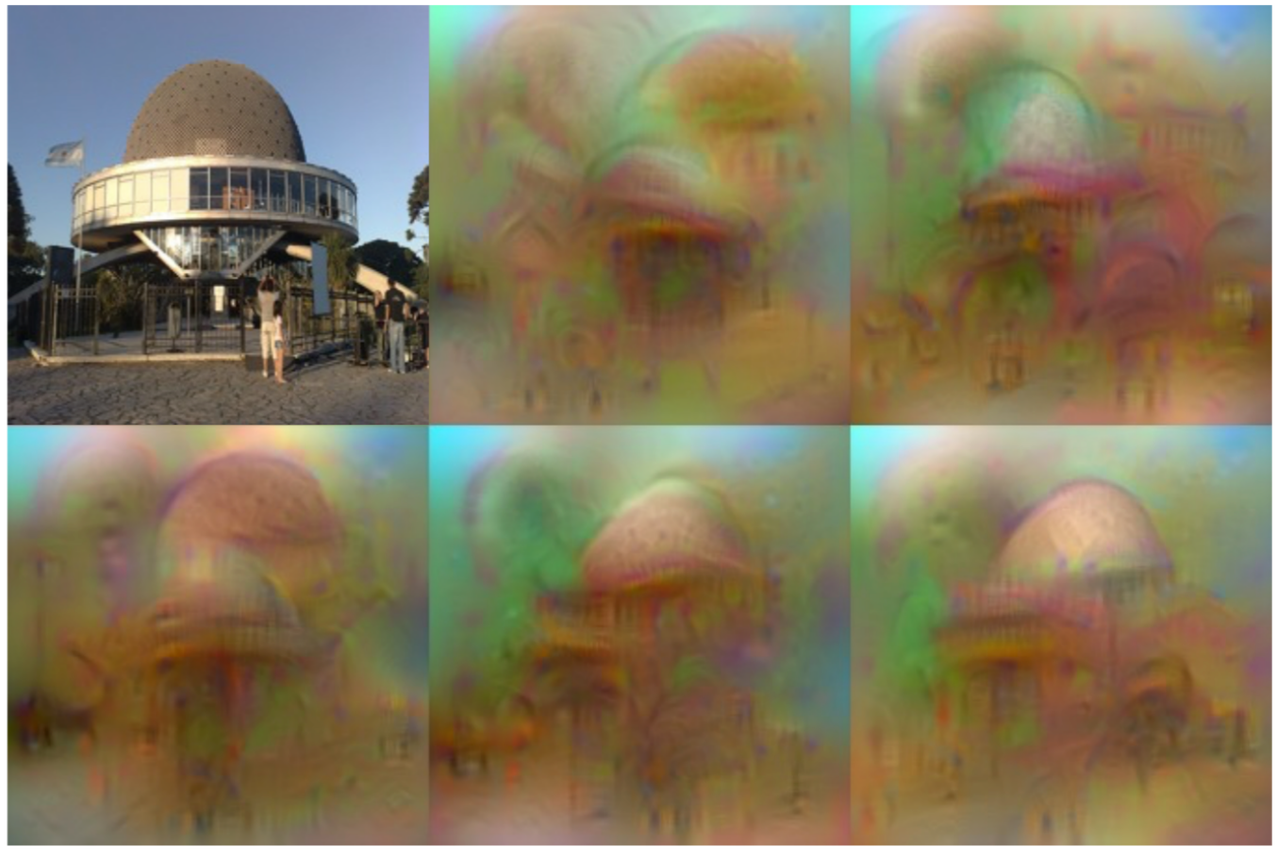
- We need to find:

$$\mathbf{x} \in \mathbb{R}^{H \times W \times C}$$

- By minimizing:

$$\mathbf{x}^* = \underset{\mathbf{x} \in \mathbb{R}^{H \times W \times C}}{\operatorname{argmin}} \ \ell(\Phi(\mathbf{x}), \Phi_0) + \lambda \mathcal{R}(\mathbf{x})$$

Results in an image $x*$ that "resembles" x0 from the viewpoint of the representation.



Possible reconstructions obtained from a convolutional layer of a CNN

# Content Reconstruction


Image reconstructed from layers ‘conv1_1’ (a), ‘conv2_1’ (b), ‘conv3_1’ (c), ‘conv4_1’ (d) and ‘conv5_1’ (e) of the original VGG-Network

- Filters at layer l: $N_l$
- Size of receptive field at layer l: $M_l$
- Response at layer l: $F^l \in \mathcal{R}^{N_l \times M_l}$

  $F^l_{ij}$ represents the ith filter at position j in layer l

- Given image: $\vec{p}$
- We generate image: $\vec{x}$
- Squared-error loss: $\mathcal{L}_{content}(\vec{p}, \vec{x}, l) = \frac{1}{2} \sum_{i,j} \left(F^l_{ij} - P^l_{ij}\right)^2$

We change the generated image until it produces the same response at a certain layer of the CNN as the original image

# Style Reconstruction


Style representations compute correlations between the different filter responses. Representations from: ’conv1_1’ (a), ‘conv1_1’ and ‘conv2_1’ (b), ‘conv1_1’, ‘conv2_1’ and ‘conv3_1’ (c), ‘conv1_1’, ‘conv2_1’, ‘conv3_1’ and ‘conv4_1’ (d), ‘conv1_1’, ‘conv2_1’, ‘conv3_1’, ‘conv4_1’ and ‘conv5_1’ (e). The representations match the style of the given image on an increasing scale.

- Filter correlations are given by the Gram matrix

$$G^l \in \mathcal{R}^{N_l \times N_l}$$

- $G^l_{ij}$ is the inner product between the filters i and j in layer l

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

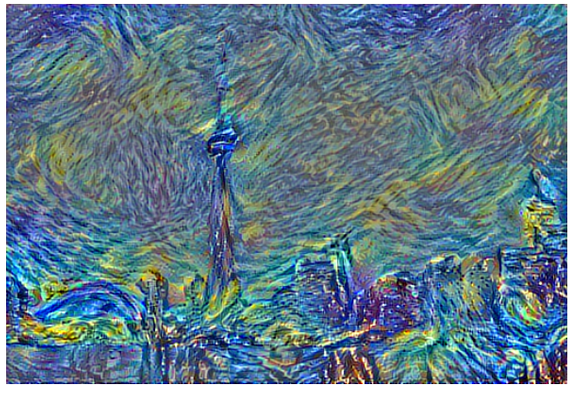
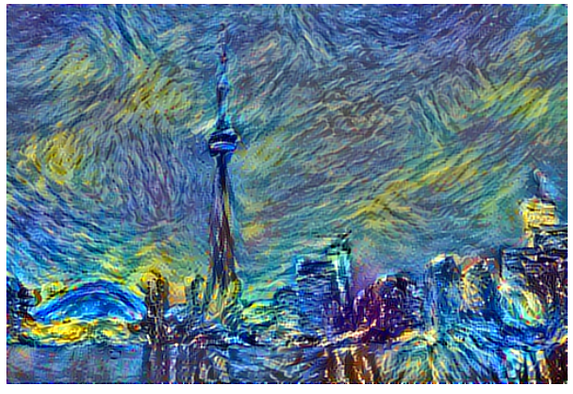
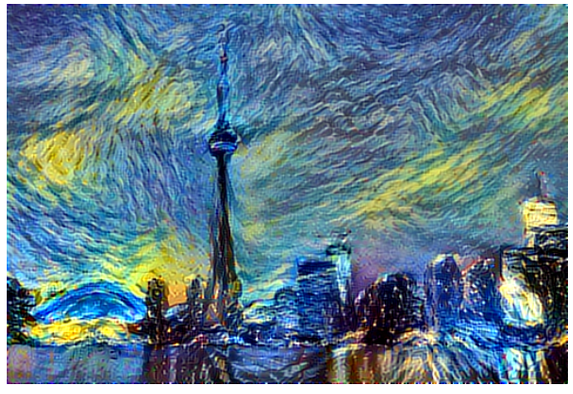
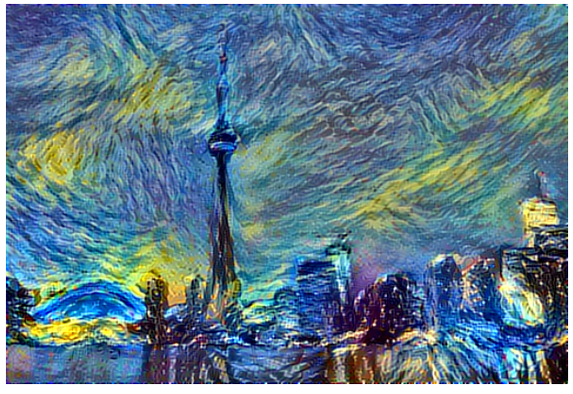
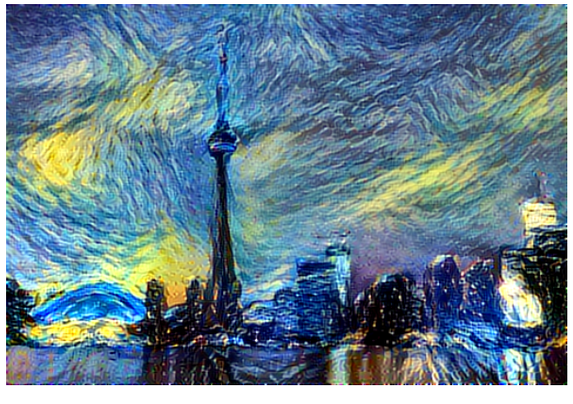
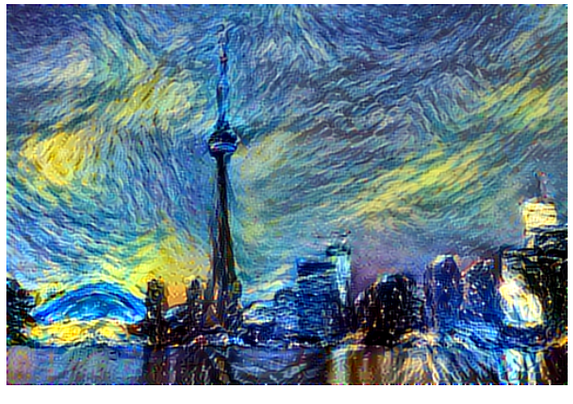
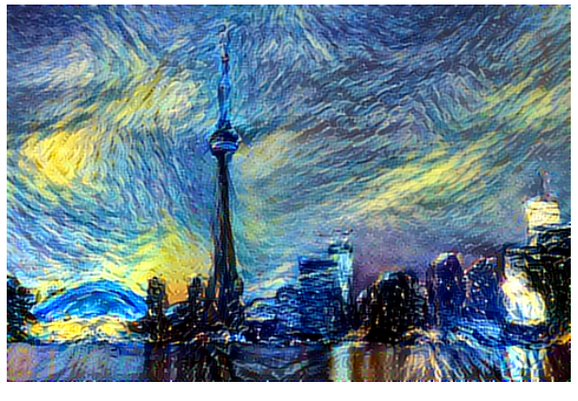
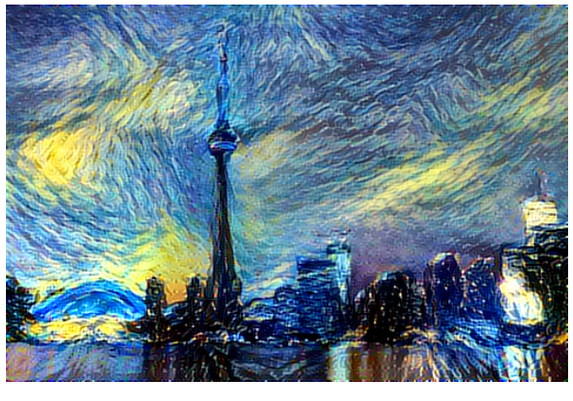
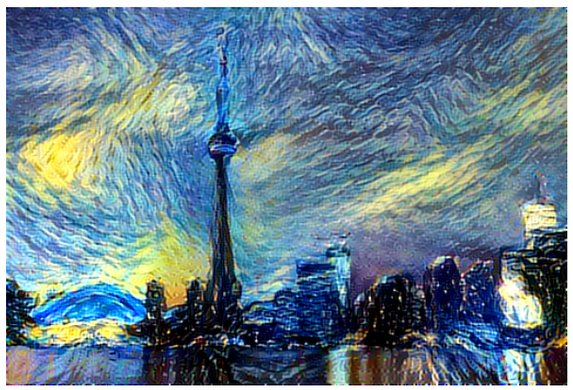
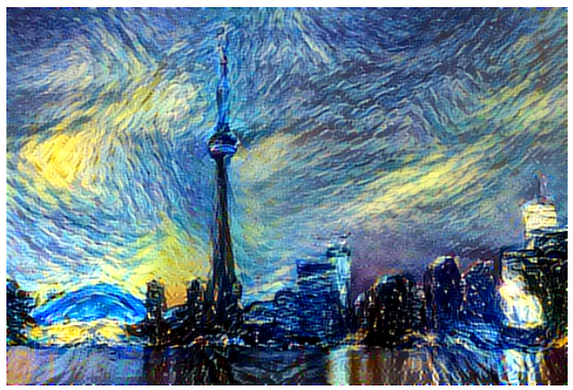
We generate an image by minimizing the mean-squared distance between the entries of the Gram matrix from the original image and the Gram matrix of the image to be generated.

Main contribution: content and style are separable.

We can mix the content and the style by starting with a white noise image and jointly minimizing both losses.

$$\mathcal{L}_{total}(\vec{p}, \vec{a}, \vec{x}) = \alpha \mathcal{L}_{content}(\vec{p}, \vec{x}) + \beta \mathcal{L}_{style}(\vec{a}, \vec{x})$$

Extracting correlations between neurons is a biologically plausible computation that is, for example, implemented by so-called complex cells in the primary visual system (V1)

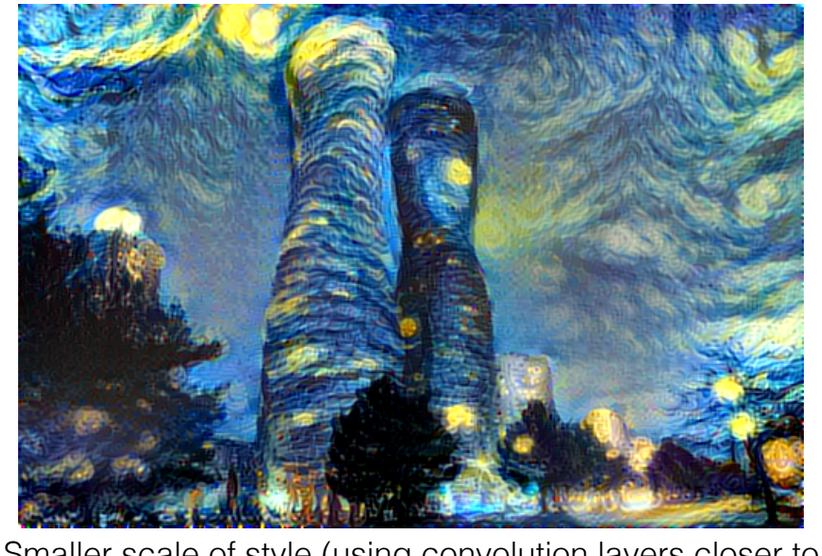
Outputs at intervals of a 100 iterations, using white noise for initialization


Content image

Large scale of cropped Starry Night as style image (emphasizes dark foreground)

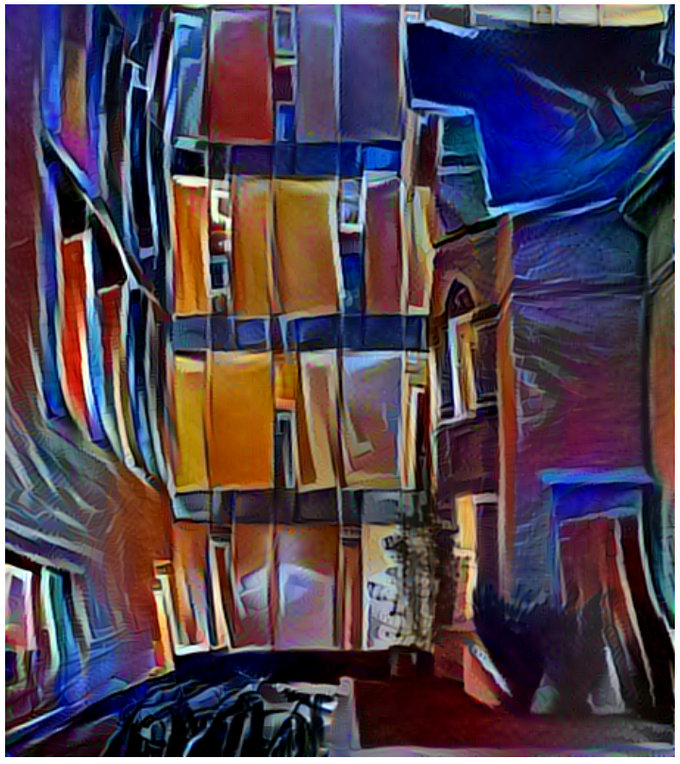
Large scale of full Starry night as style image, initialized with content image

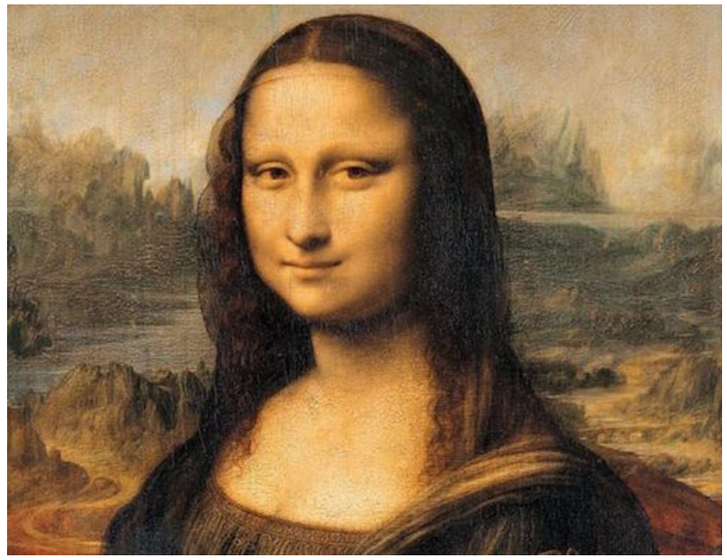
Using Leonid Afremov painting as style image

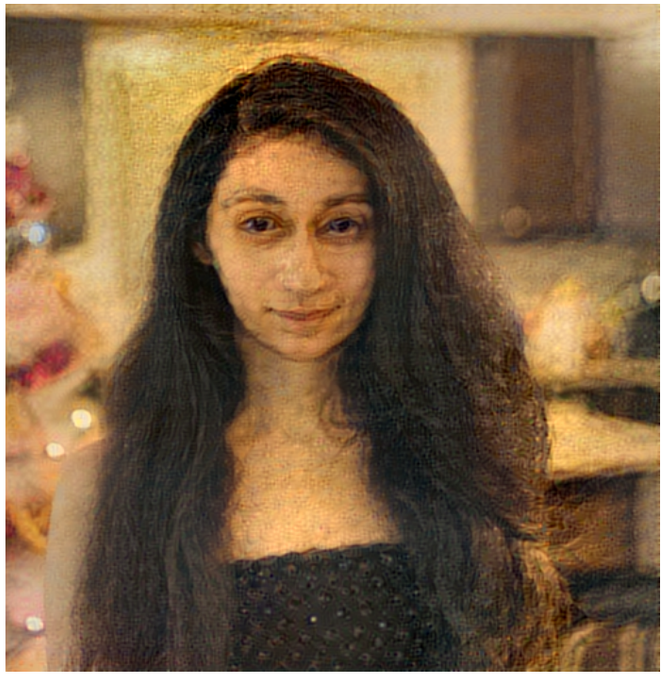


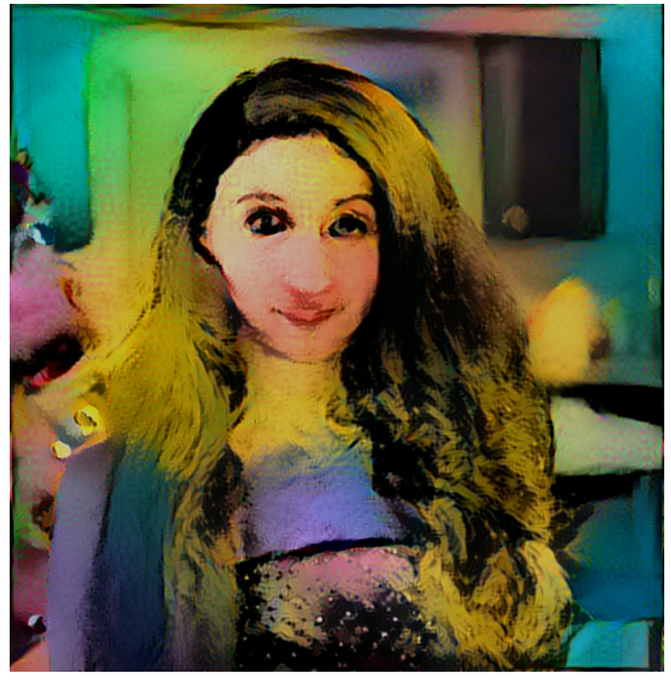
Smaller scale of style (using convolution layers closer to the input layer)

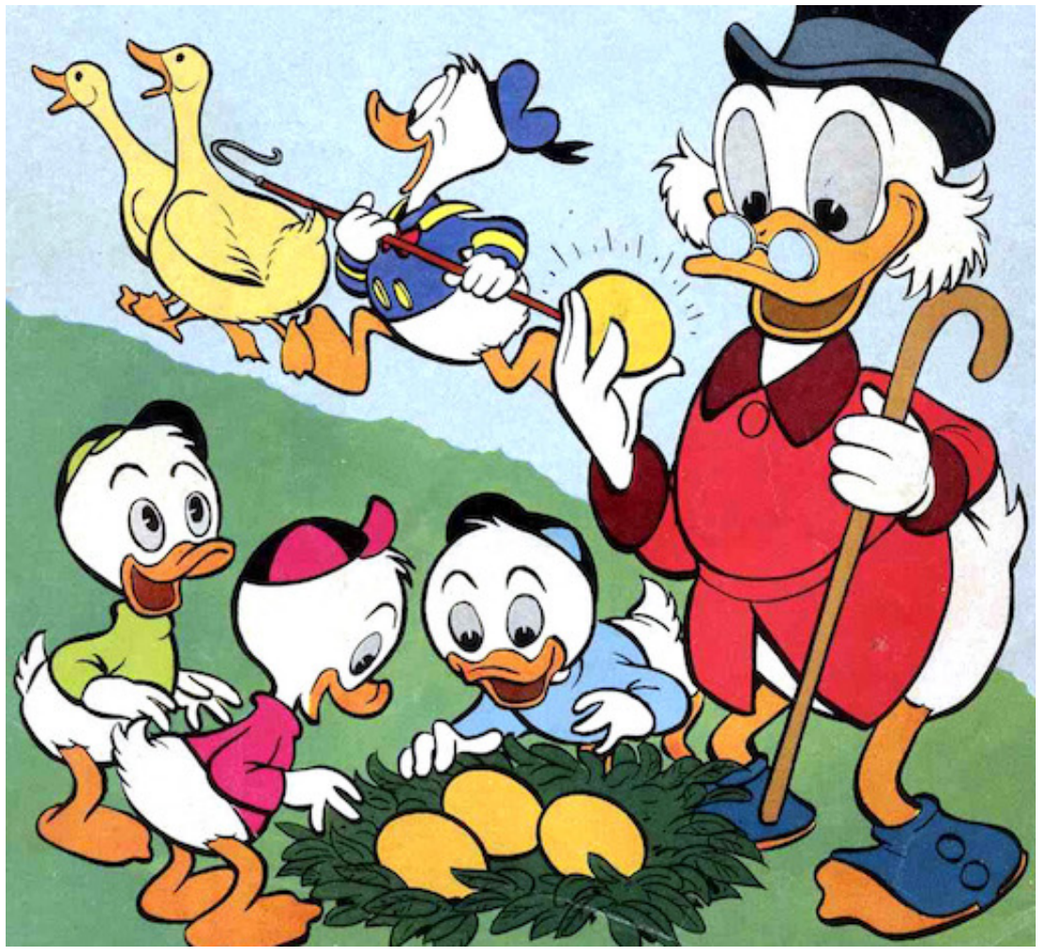
Large scale of full Starry night as style image, initialized with white noise












A
$10^{-5}$
$10^{-4}$
$10^{-3}$
$10^{-2}$
Conv1_1
B
Conv2_1
C
Conv3_1
D
Conv4_1
E
Conv5_1

# Discussion

- Evaluation: None. However, the method appears to work very well and is easy to implement.

- New method of mixing content and style from different sources.

- Useful for studying the neural representation of art, style and content-independent image appearance.

# Bibliography

Mahendran, Aravindh, and Andrea Vedaldi. "Understanding deep image representations by inverting them." *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*. IEEE, 2015.

Gatys, Leon A., Alexander S. Ecker, and Matthias Bethge. "A neural algorithm of artistic style." *arXiv preprint arXiv:1508.06576* (2015).

Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." *arXiv preprint arXiv:1409.1556* (2014).