Photorealistic Image Stylization:

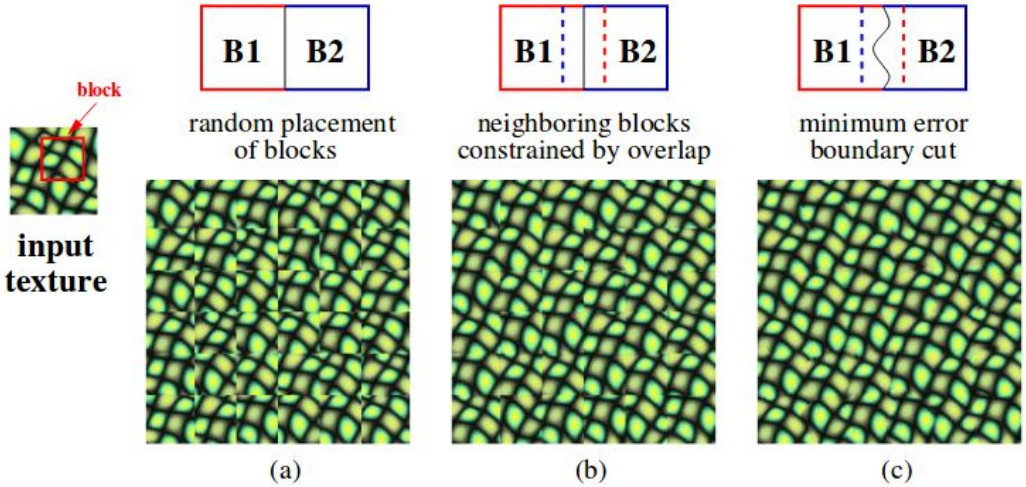
Image Stylization using Convolutional Neural Networks

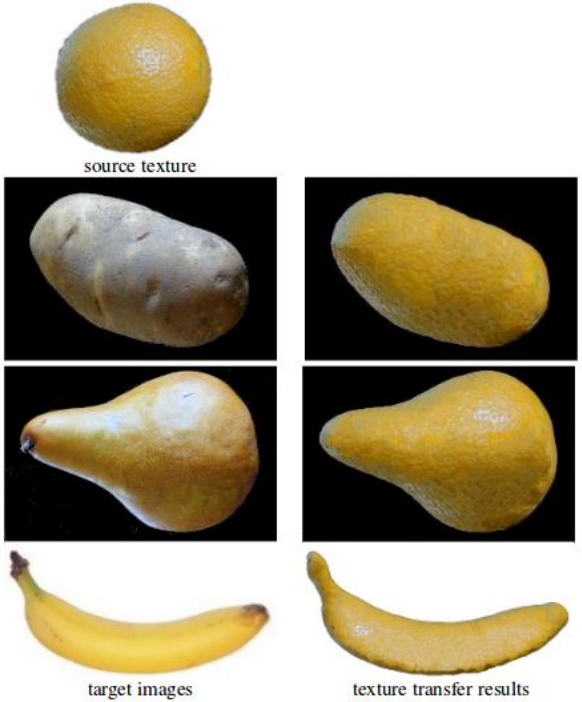
Leon A. Gatys, Alexander S. Ecker, Matthias Bethge

CVPR 2016

Related Work

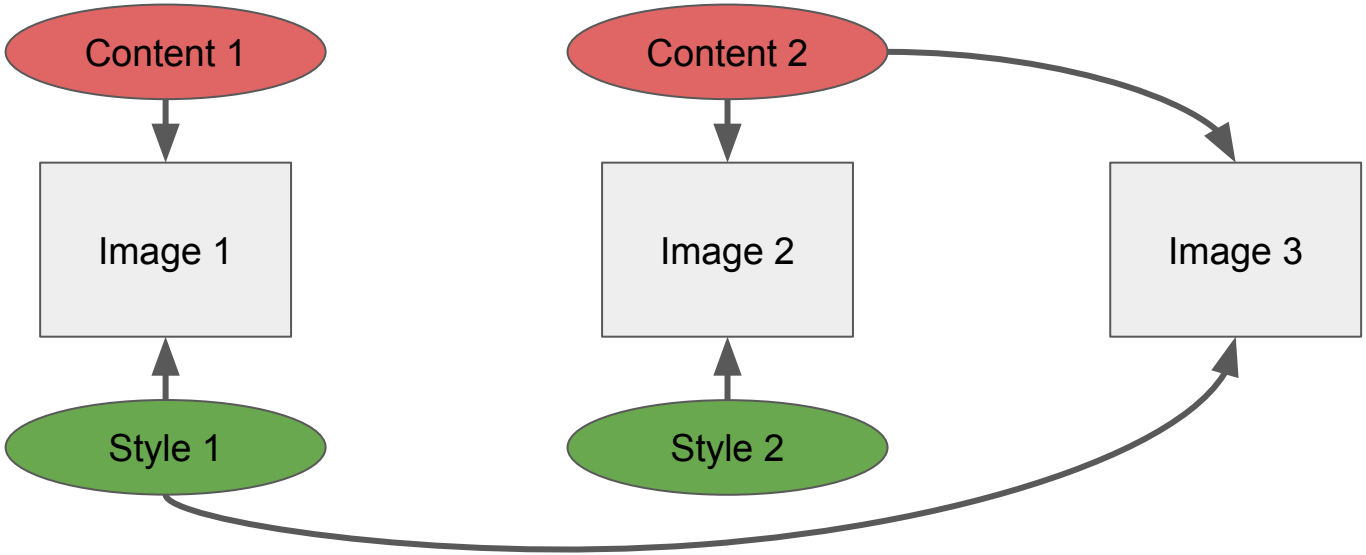
Efros and Freeman, 2001.





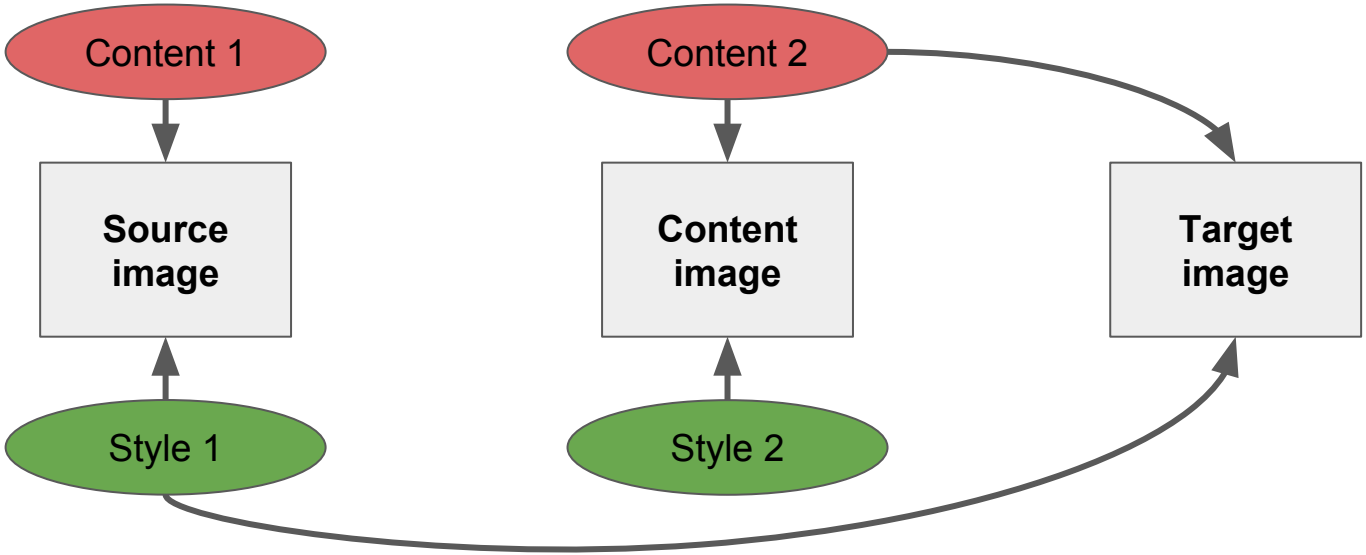
Motivation

- Any image can be factorize into content and style.
- Content provides the semantics of the entities present in the image.
- Style provide the textures corresponding to those entities.

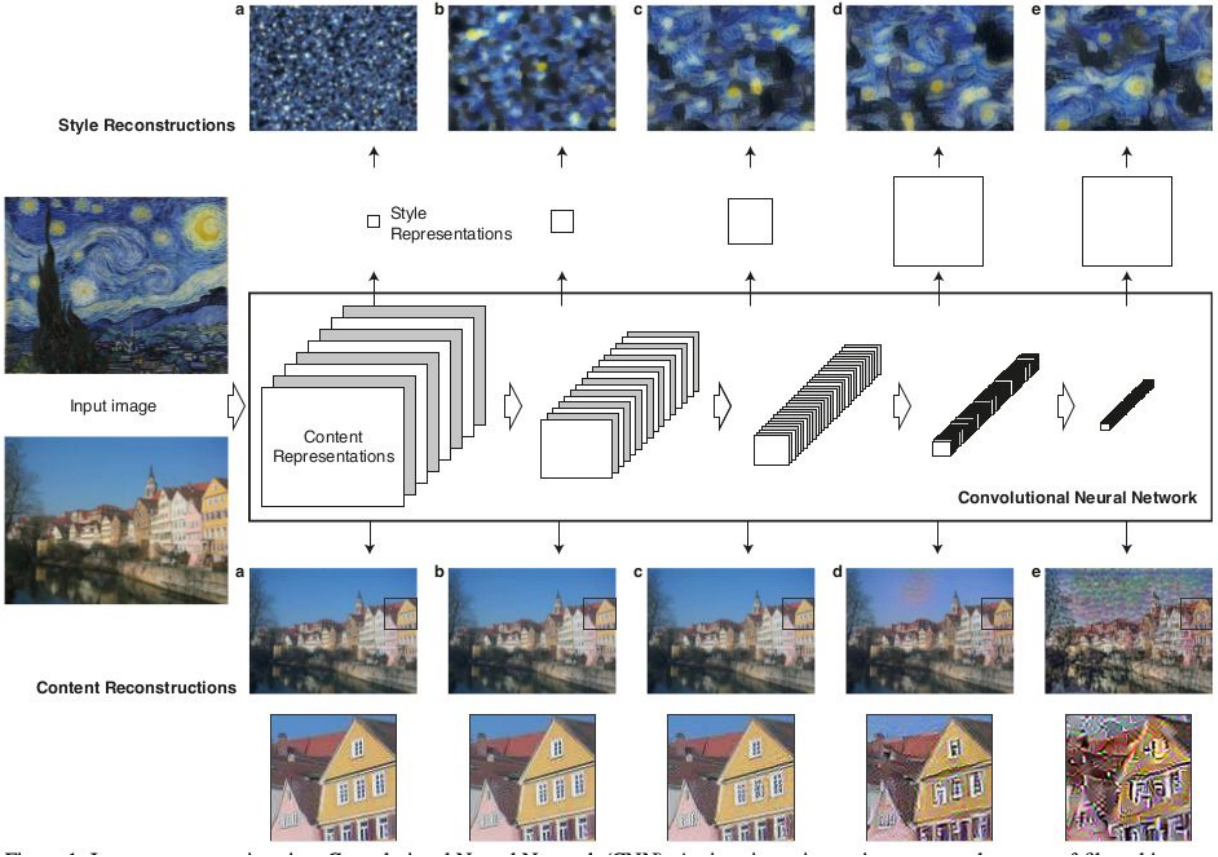


Motivation

- Any image can be factorize into content and style.
- Content provides the semantics of the entities present in the image.
- Style provide the textures corresponding to those entities.



Feature selection



Depth from which the features are taken

Pixel level detail of the reconstructed image

Content Representation

For any image x; the filter response in layer l: $F^l \in \mathcal{R}^{N_l \times M_l}$
Where N_l is number of filters in layer l and M_l = HxW of the feature map.

Content Loss:
If F^l_{ij} is the activation of *i-th* filter in the position *j* in the layer l.

Let p → original image

Let x → generated image

S.t. P^l and F^l are their respective feature response in layer l.

$$\mathcal{L}_{\text{content}}(\vec{p}, \vec{x}, l) = \frac{1}{2} \sum_{i,j} \left(F^l_{ij} - P^l_{ij}\right)^2 . (1)$$

$$\frac{\partial \mathcal{L}_{\text{content}}}{\partial F^l_{ij}} = \begin{cases} \left(F^l - P^l\right)_{ij} & \text{if } F^l_{ij} > 0 \\ 0 & \text{if } F^l_{ij} < 0 , \end{cases} \quad (2)$$

Style Representation

For any image x; the Gram matrix for layer l: $G^l \in \mathcal{R}^{N_l \times N_l}$
is the inner product between the vectorised feature maps i and j in layer l.

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

Style Loss:

If F^l_{ij} is the activation of *i-th* filter in the position *j* in the layer l.

Let a → original image 

Let x → generated image 

S.t. A^l and G^l are their respective feature response in layer l.

$$E_l = \frac{1}{4N_l^2 M_l^2} \sum_{i,j} \left(G^l_{ij} - A^l_{ij}\right)^2 \qquad (4)$$

$$\mathcal{L}_{\text{style}}(\vec{a}, \vec{x}) = \sum_{l=0}^{L} w_l E_l, \qquad (5)$$

Style Representation

For any image x; the Gram matrix for layer l: $G^l \in \mathcal{R}^{N_l \times N_l}$
is the inner product between the vectorised feature maps i and j in layer l.

Styl

If F

Let

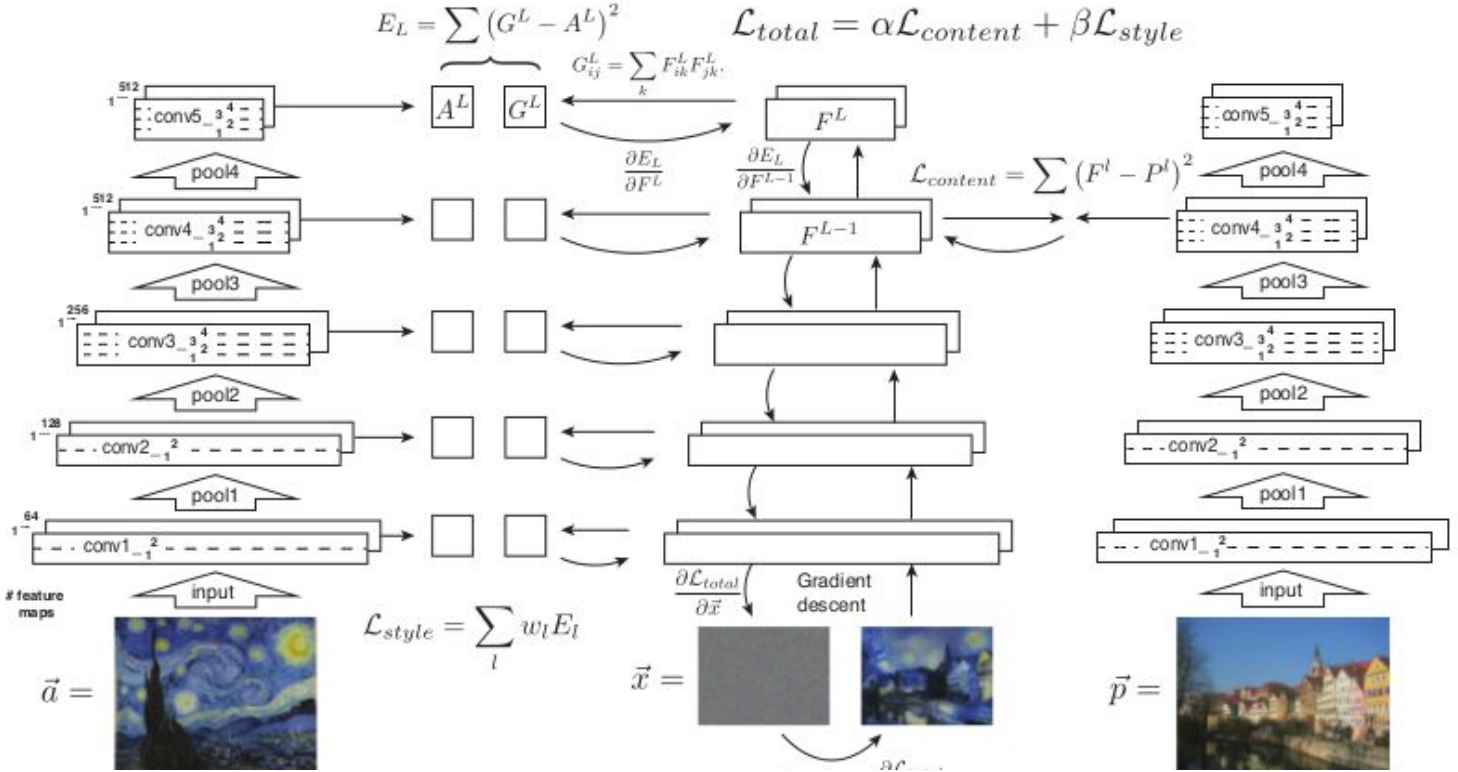
Let

$$\frac{\partial E_l}{\partial F_{ij}^l} = \begin{cases} \frac{1}{N_l^2 M_l^2} \left((F^l)^{\mathrm{T}} \left(G^l - A^l \right) \right)_{ji} & \text{if } F_{ij}^l > 0 \\ 0 & \text{if } F_{ij}^l < 0 \,. \end{cases} \quad (6)$$

$$E_l = \frac{1}{4N_l^2 M_l^2} \sum_{i,j} \left(G_{ij}^l - A_{ij}^l \right)^2 \quad (4)$$

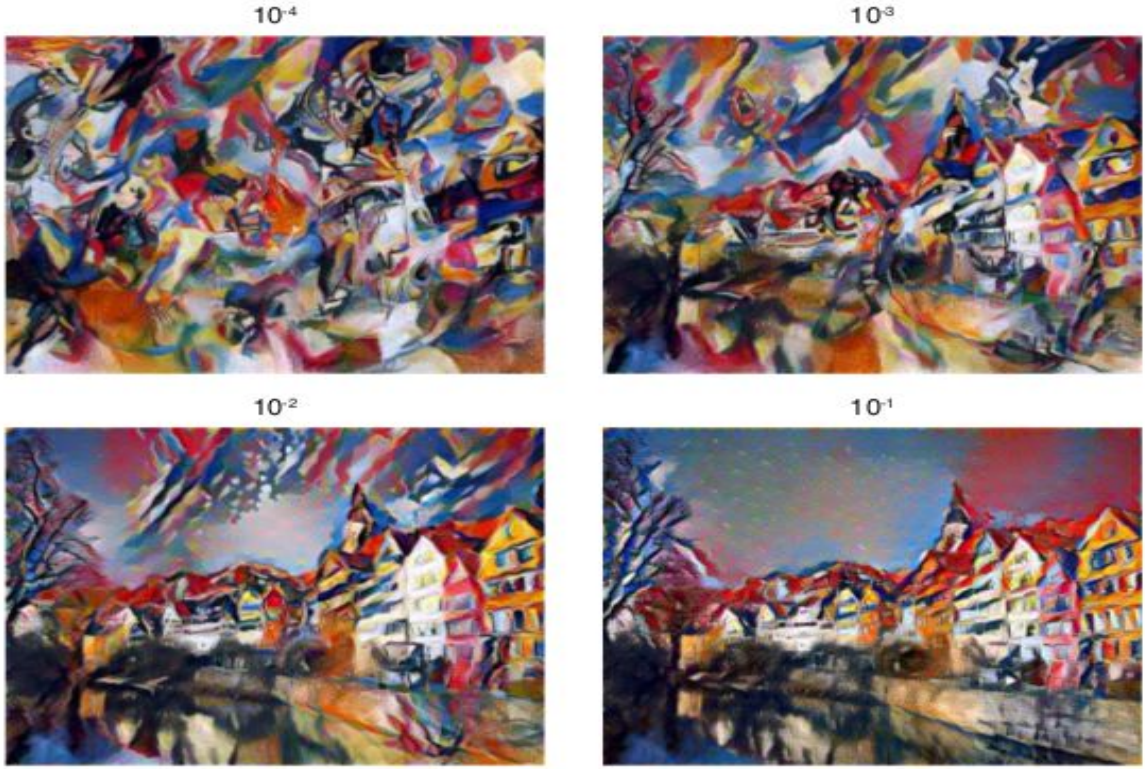
$$\mathcal{L}_{\text{style}}(\vec{a}, \vec{x}) = \sum_{l=0}^{L} w_l E_l, \quad (5)$$

Style Transfer Algorithm

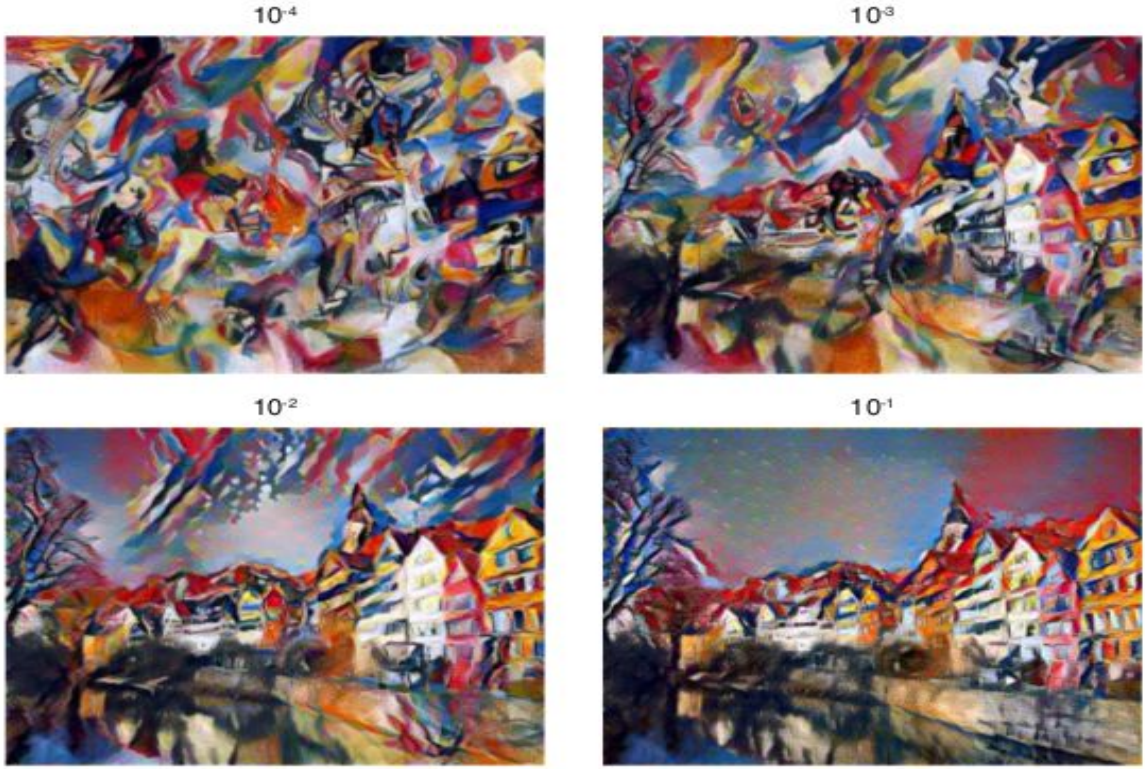


$$\mathcal{L}_{\text{total}}(\vec{p}, \vec{a}, \vec{x}) = \alpha \mathcal{L}_{\text{content}}(\vec{p}, \vec{x}) + \beta \mathcal{L}_{\text{style}}(\vec{a}, \vec{x}) \qquad (7)$$

Trade-off between Content and Style: α / β



Trade-off between Content and Style: α / β



Selection of layer for content matching

Effect of different initialization

A

Content image

B

Style image

Effect of different initialization (different white noises)

Thanks!