Multimodal Transfer: A Hierarchical Deep Convolutional Neural Network for Fast Artistic Style Transfer

Xin Wang[1,2], Geoffrey Oxholm[2], Da Zhang[1], Yuan-Fang Wang[1]

[1]University of California, Santa Barbara [2]Adobe Research, San Francisco

IEEE 2017 Conference on Computer Vision and Pattern Recognition

CVPR July 21-26 2017 HONOLULU

Motivation

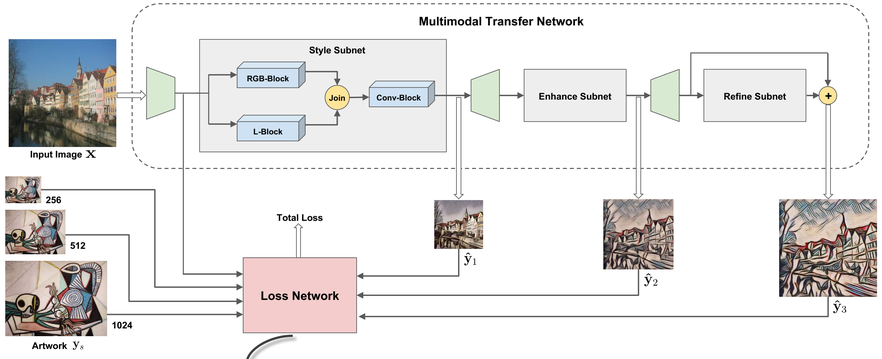
Our approach addresses the shortcomings of past work:

- ❑ Optimization-based methods (*e.g.* Gatys *et al.*):
  - Run very slow
  - Consume large GPU memory
- ❑ Feed-forward networks for fast style transfer (*e.g.* Johnson *et al.*, Ulyanov *et al.*):
  - Lower quality results
  - Capture only single scale of texture
  - Fail to capture small, intricate textures

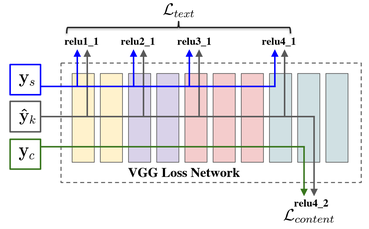
Contributions

- ❑ End-to-end hierarchical network for fast style transfer;
- ❑ Hierarchical training scheme allow us to combine multiple models into one network to handle increasingly larger image sizes;
- ❑ Novel way to utilize representations of both color and luminance channels;
- ❑ Hierarchical style transfer network that can better capture both coarse and intricate texture patterns
- ❑ Extensive experimentation with works of fine art

Framework



Loss Definition



For each output at scale k, the stylization loss is defined as:

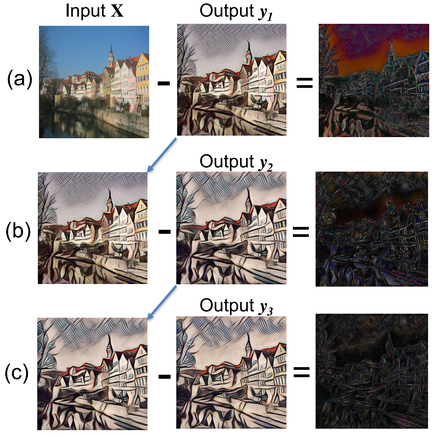
$$\mathcal{L}_S^k(\hat{\mathbf{y}}_k, \mathbf{y}_c^k, \mathbf{y}_s^k) = \alpha\mathcal{L}_{content}(\hat{\mathbf{y}}_k, \mathbf{y}_c^k) + \beta\mathcal{L}_{text}(\hat{\mathbf{y}}_k, \mathbf{y}_s^k)$$

Hierarchical Loss: $\mathcal{L}_H = \sum_{k=1}^{K} \lambda_k \mathcal{L}_S^k(\hat{\mathbf{y}}_k, \mathbf{y}_c^k, \mathbf{y}_s^k)$

Objective: $\Theta_k = \underset{\Theta_k}{\arg\min}\, E_{\mathbf{x}\sim\mathcal{X}}\left[\sum_{i\geq k}^{K} \lambda_i \mathcal{L}_S^k(f(\cup_{j=1}^{k}\Theta_j, \mathbf{x}), \mathbf{y}_c^i, \mathbf{y}_s^i)\right]$

$$\Delta\Theta_k = \begin{cases} f^{-1}(\lambda_k \mathcal{L}_S^k) & k = K \\ f^{-1}(\lambda_k \mathcal{L}_S^k, \Delta\Theta_{k+1}) & 1 \leq k < K \end{cases}$$

Visual Interpretation



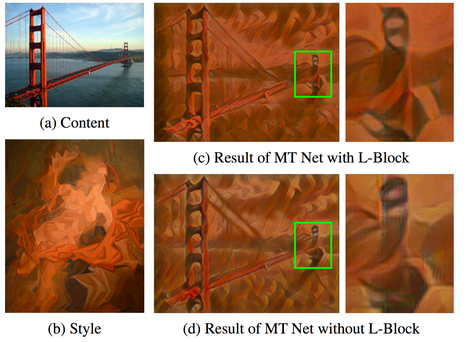
Each successive scale contributes more fine detail.

Experimental Results

Inspection of Scale

(a) Style
(b) Singular Transfer with style size 256
(c) Singular Transfer with style size 1024
(d) Multimodal Transfer (Ours)

Luminance Branch



(a) Content
(b) Style
(c) Result of MT Net with L-Block
(d) Result of MT Net without L-Block

Speed & Memory Usage

Network	Test Time	Memory Usage
Ours	0.54s	3100 MB
Johnson *et al.*	0.42s	2400 MB
DS Net*	0.63s	6700 MB

Fine Details

(a) Style (b) Gatys *et al.* (c) Johnson *et al.* (d) Ulyanov *et al.* (e) Ours (f) Content

*DS Net: Deep Singular Network, which has the same architecture with our multimodal transfer network but only has one single scale