Arbitrary Style Transfer

with Parallel Self-Attention

Tiange Zhang*, Ying Gao*, Feng Gao, Lin Qi, Junyu Dong

Ocean University of China

TABLE OF CONTENTS

Introduction

Neural style transfer aims to create artistic images by synthesizing patterns from a given style image. Recently, the Adaptive Instance Normalization (AdaIN) layer is proposed to achieve real-time arbitrary style transfer. However, we observed that

- if crucial features based on AdaIN can be further emphasized during transfer, both content and style information will be better reflected in stylized images.
- Furthermore, it is always essential to preserve more details and reduce unexpected artifacts in order to generate appealing results.

Introduction

In this paper, we introduce an improved arbitrary style transfer method.

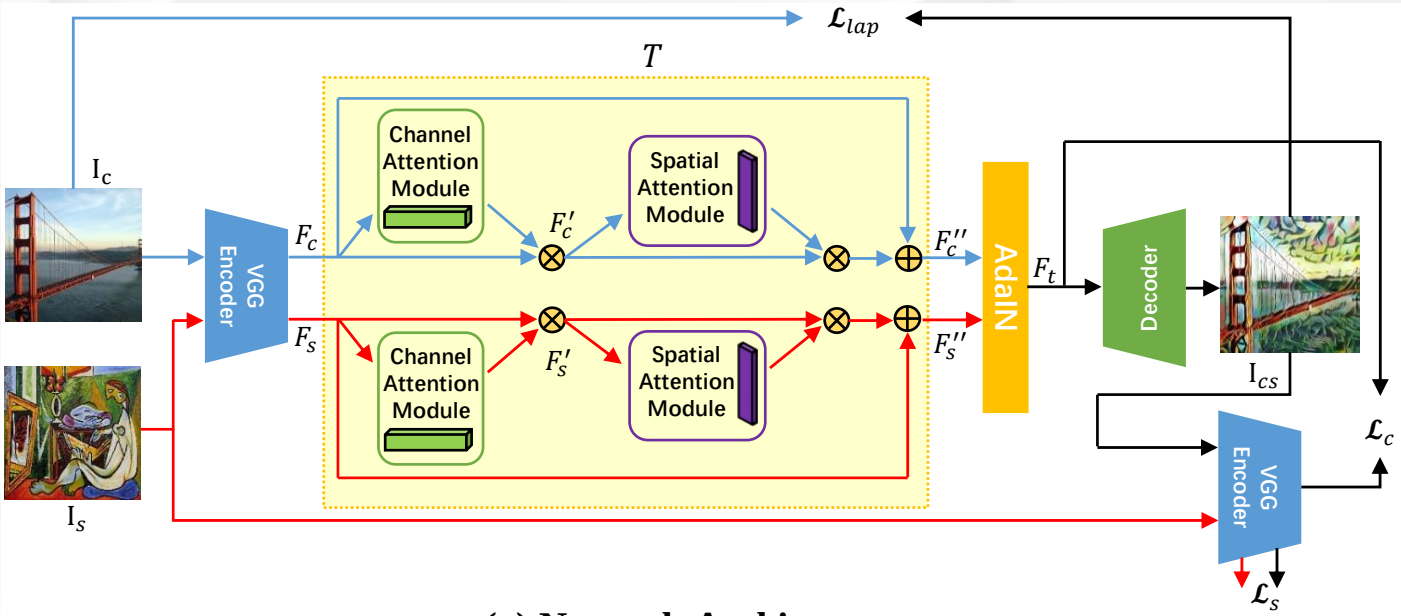
- A self-attention module is designed to learn what and where to emphasize in the input image.
- In addition, an extra Laplacian loss is applied to preserve structure details of the content while eliminating artifacts.
- Experimental results demonstrate that the proposed method outperforms AdaIN and can generate more appealing results.

Method

Network Architecture

Our network is composed of

- an encoder–decoder framework
- a parallel self-attention module *T*
- an AdaIN layer

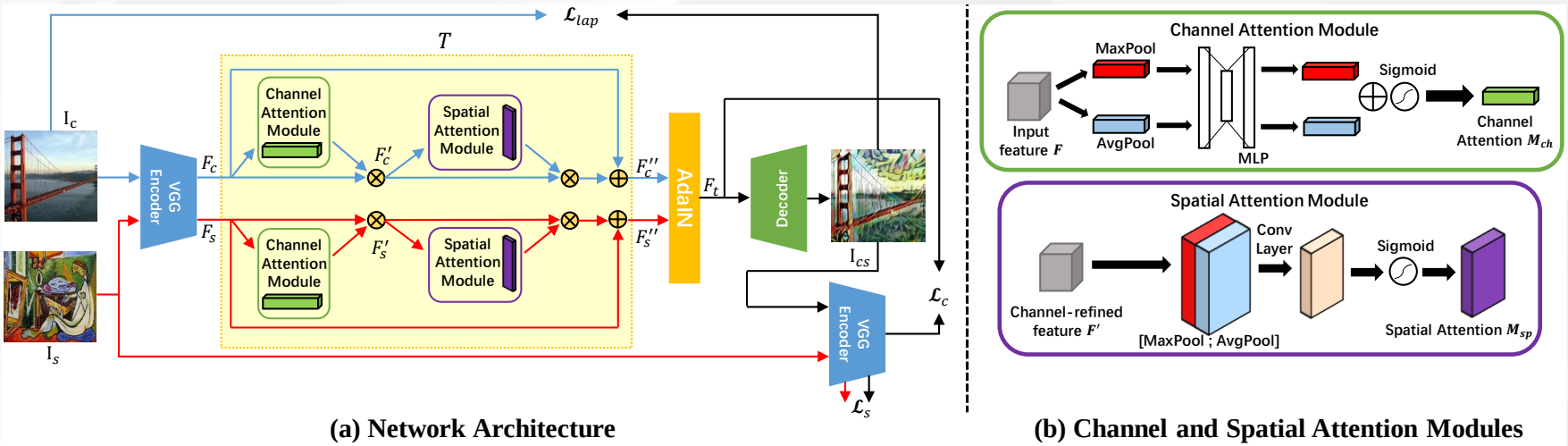


(a) Network Architecture

Method

The Attention Module

- Inspired by the Convolutional Block Attention Module, we desinged a similar attention module by adding residual values from the encoded feature maps to the refined feature maps for feature fusion.
- Given the content and style features, there is a parallel process where the attention module *T* sequentially infers the required attention maps in channel and spatial dimensions.



(a) Network Architecture **(b) Channel and Spatial Attention Modules**

Method

Adaptive Instance Normalization (AdaIN)

AdaIN receives a content input x and a style input y, and adaptively aligns the channel-wise mean and variance of x to match those of y:

$$AdaIN(x) = \sigma(y)(\frac{x - \mu(x)}{\sigma(x)}) + \mu(y),$$

where we just simply scale the normalized content input with $\sigma(y)$, and shift it with $\mu(y)$. These statistics are computed across the spatial locations.

Method

Training Overflow

The total loss function for training:

$$\mathcal{L} = \lambda_c \mathcal{L}_c + \lambda_s \mathcal{L}_s + \lambda_{lap} \mathcal{L}_{lap},$$

where $\mathcal{L}_c$, $\mathcal{L}_s$ and $\mathcal{L}_{lap}$ represent the content loss, style loss and Laplacian loss respectively.

Method

Content Loss

We set our content loss as the Euclidean distance between the AdaIN output features F_t and the features of the stylized output image $E(I_{cs})$ as:

$$\mathcal{L}_c = \|E(I_{cs}) - F_t\|_2.$$

Style Loss

The style loss restricts the features of stylized output image I_{cs}. It matches the mean and standard deviation of the input style features as:

$$\begin{aligned}\mathcal{L}_s &= \sum_{i=1}^{L} \|\mu(E_i(I_{cs})) - \mu(E_i(I_s))\|_2 \\ &+ \sum_{i=1}^{L} \|\sigma(E_i(I_{cs})) - \sigma(E_i(I_s))\|_2,\end{aligned}$$

where each E_i denotes a layer in VGG-19 used to compute the style loss.

Laplacian Loss

The Laplacian filter is defined as:

$$D = \begin{bmatrix} 0 & -1 & 0 \\ -1 & 4 & -1 \\ 0 & -1 & 0 \end{bmatrix},$$

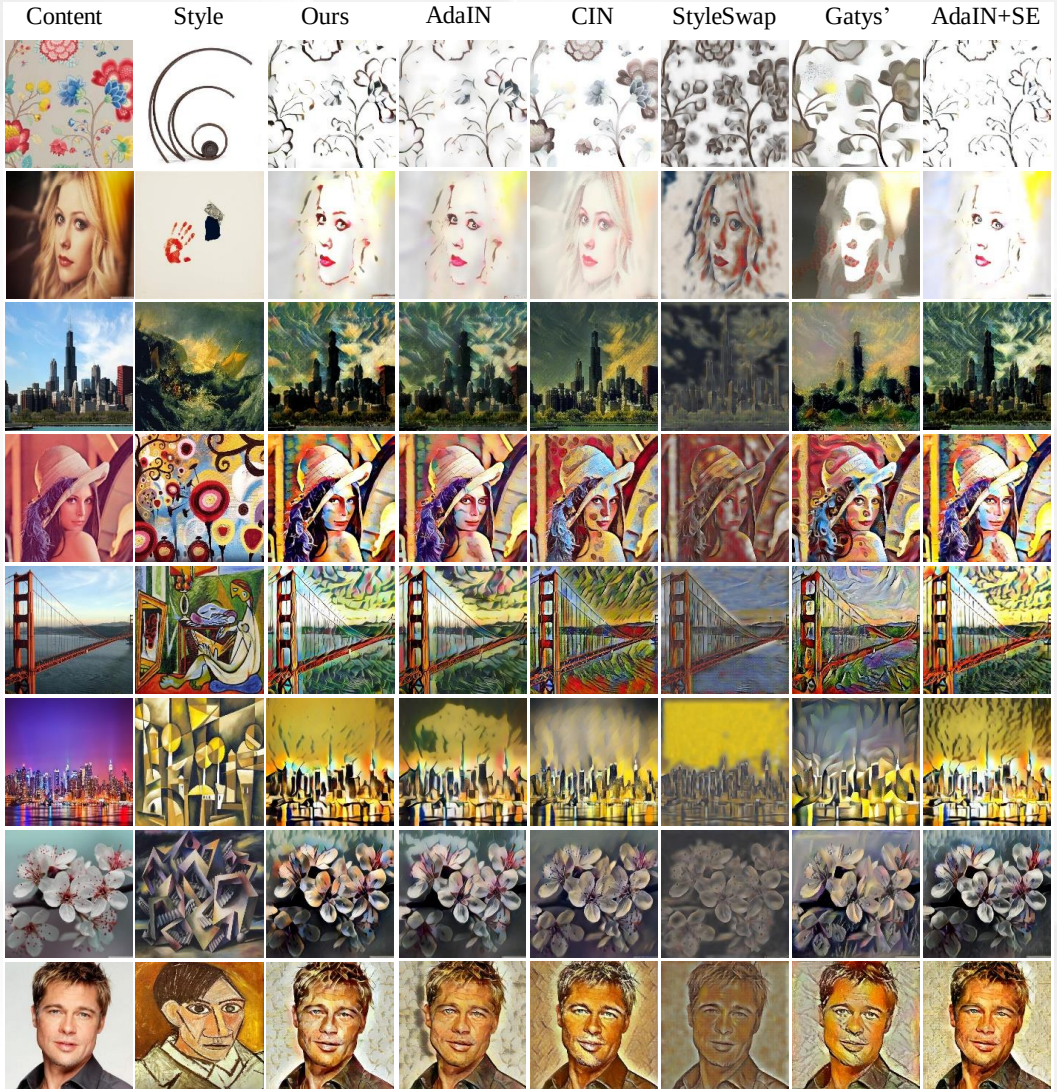
Therefore, the Laplacian matrix of an RGB image x can be obtained by convolving the 3 channels with D. Hence, the MSE loss between the Laplacians of stylized and content images can be measured in 3 channels separately:

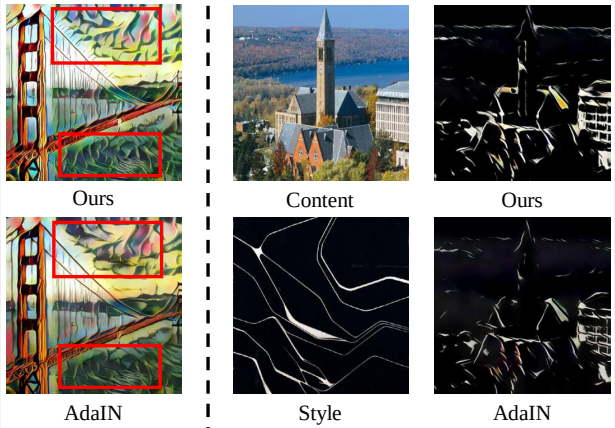
$$\begin{aligned} \mathcal{L}_{lap} &= \sum_{ij}(D(I^R_{cs}) - D(I^R_c))^2_{ij} + \sum_{ij}(D(I^G_{cs}) - D(I^G_c))^2_{ij} \\ &+ \sum_{ij}(D(I^B_{cs}) - D(I^B_c))^2_{ij}, \end{aligned}$$

where i and j mean the pixel of the image.

Experimental Results and Analysis

Qualitative Comparison





Example results on artifacts elimination and details.

Experimental Results and Analysis

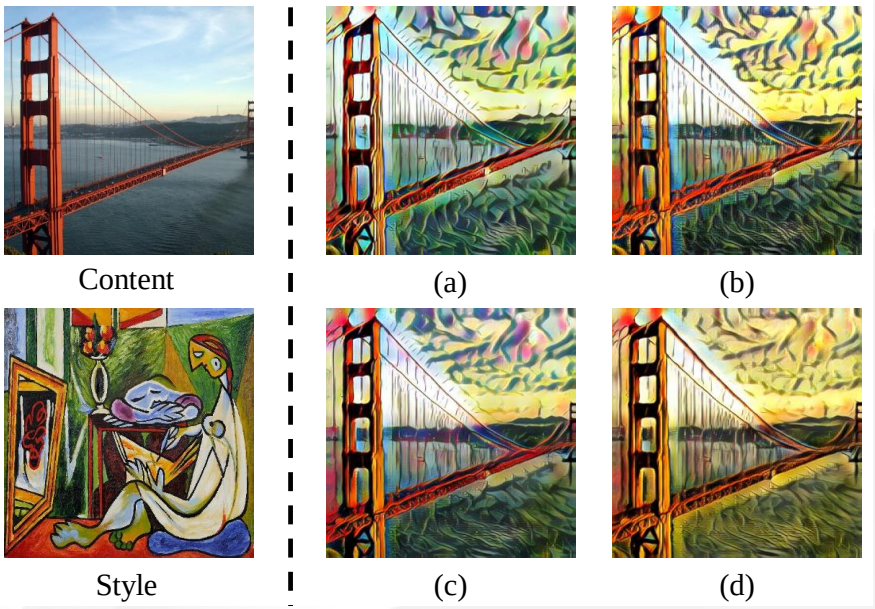
Quantitative Evaluations

For quantitative evaluations, we conduct an efficiency analysis on **processing speed**, and evaluate the **style loss** $\mathcal{L}_{\text{style}}$ defined by the method from Gatys'.

Method	Times(s) 256×256	Times(s) 512×512	Styles	ln($\mathcal{L}_{\text{style}}$)	Learning-free
Gatys'	15.024	34.110	∞	11.56	✓
CIN	0.2315	0.9033	Limited	11.81	×
StyleSwap	0.0202	0.1012	∞	13.04	×
AdaIN	0.0181	0.0384	∞	11.89	×
Ours	0.0204	0.0396	∞	11.72	×

Experimental Results and Analysis

Attention Module Analysis

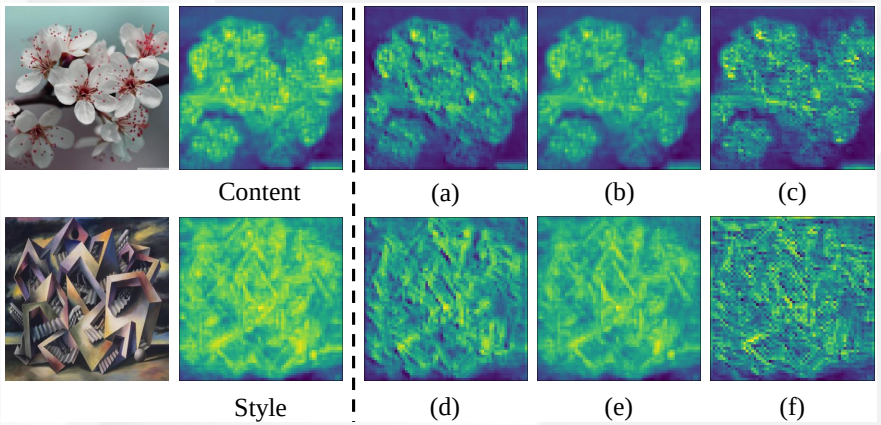


An example for the comparison of different attention utilization.

- (a) our method with both channel and spatial attention
- (b) with SE module
- (c) our method with spatial attention only
- (d) our method with channel attention only

Experimental Results and Analysis

Attention Module Analysis



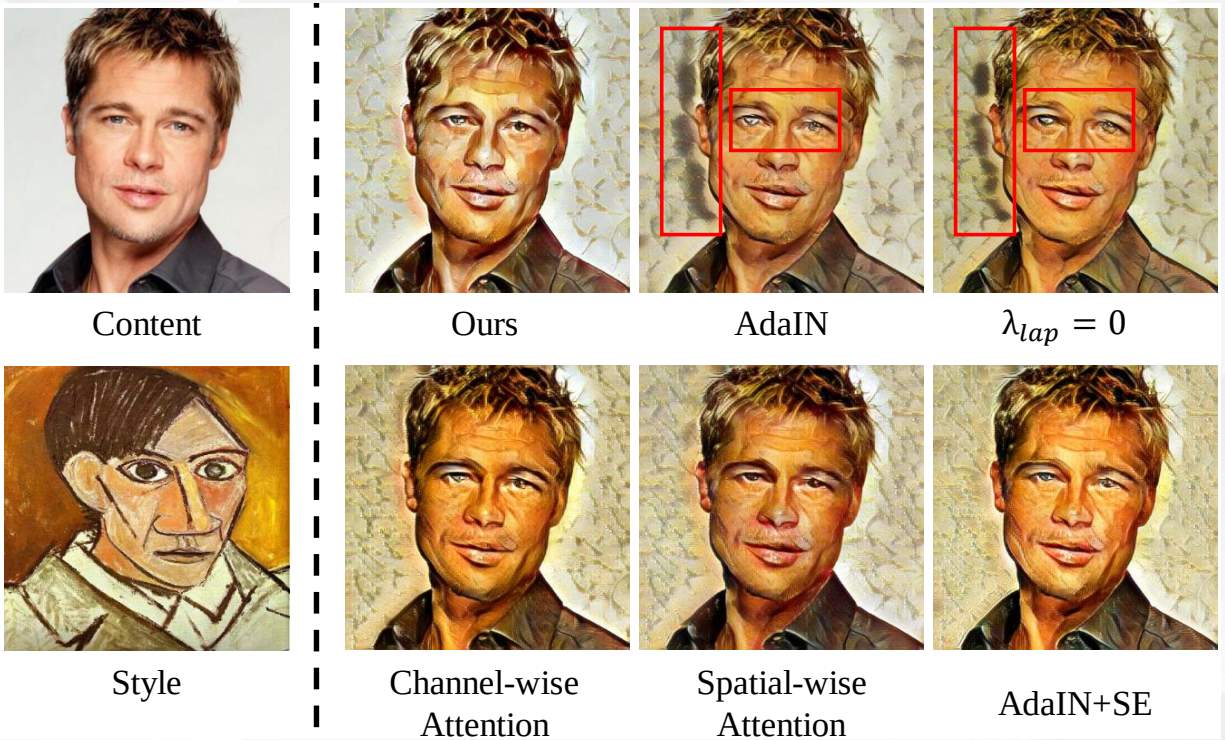
Visualization of the feature maps refined by different attention modules.

- (a) (d) with both channel and spatial attention
- (b) (e) with channel attention
- (c) (f) with spatial attention

Experimental Results and Analysis

Laplacian Loss Analysis

Figure: Ablation study of the Laplacian loss.



Methods imposing an Laplacian loss tend to generate less artifacts.

Conclusions

- In this paper, we present an improved arbitrary style transfer model based on the self-attention mechanism.
- By introducing an improved convolutional block attention module into the style transfer network and applying a Laplacian loss, more appealing stylized images can be generated.
- Experimental results demonstrate that the proposed method can eliminate unexpected artifacts, maintain more content structure details and transfer the most important style patterns properly. Meanwhile, we believe that there is still room for improvement.

Conclusions

THANK YOU !