IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

Multimodal Style Transfer via Graph Cuts

Groupe 1:
- Belet Antoine
- Santarelli Quentin

Groupe 2:
- Le Bihan Eustache
- Rouyer Pierre
- Savatier-Dupré Hélène

07/03/2023

Citation

This presentation is based on the work of :

Yulun Zhang, Chen Fang, Yilin Wang, Zhaowen Wang, Zhe Lin, Yun Fu, Jimei Yang

From :

Northeastern University, Adobe Research and ByteDance AI Lab

IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

PLAN

IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

CHAPTER 1
CONTEXT

IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

What is image style transfer ?

"Image style transfer (IST) is the process of rendering a content image with characteristics of a style image"

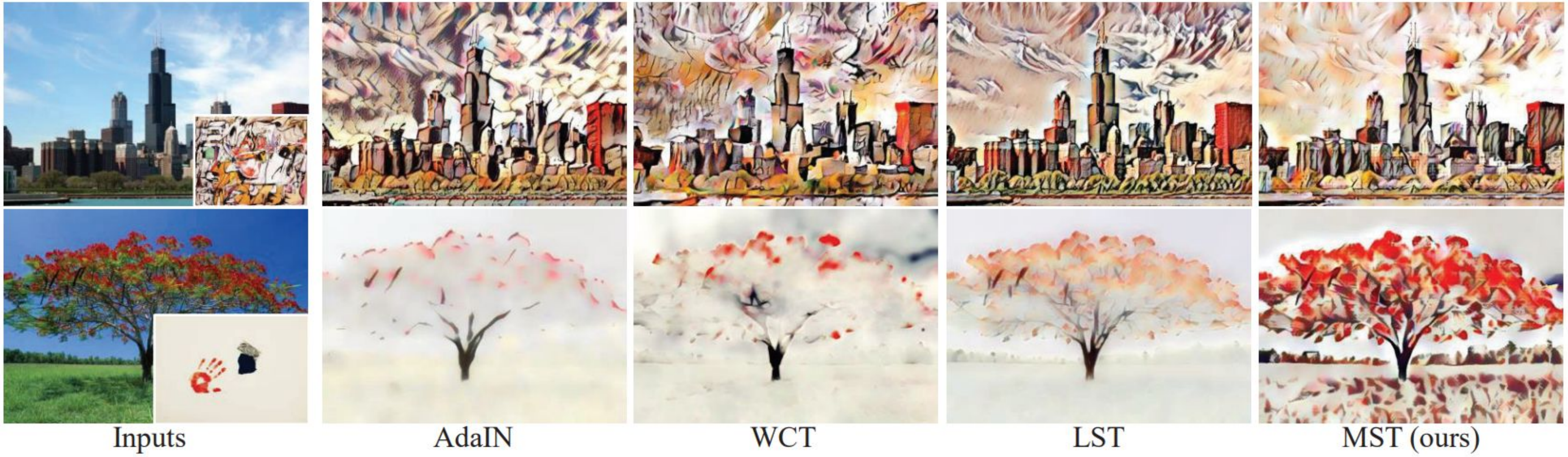


Figure: Gram matrix based style transfer methods (AdaIN [3], WCT[5], and LST[4]) and our MST method.

The work of Gatys *et al* [1] recently pushed further interest toward IST.
The discovery that the **correlation** between convolutional features of deep networks **can represent** image styles.

This IST method assume that **style can be represented as followed with a Gram matrix** [2].

Constructing an image that matches the style is a **minimization problem** solved with gradient descent algorithm.

$$G_{\mathrm{ij}} = F_{\mathrm{i}}^{\intercal} F_{\mathrm{j}} \qquad (2)$$

The Gram matrix determines the vectors F_i up to isometry and indicates the correlation between filters.

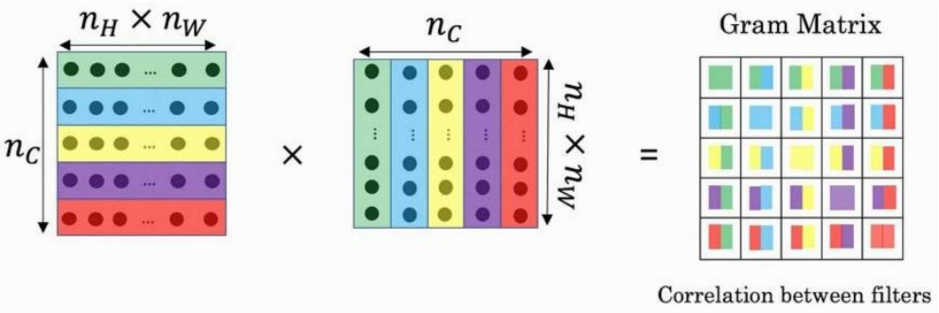


Figure 4. The Gram Matrix is created from a target image and a reference image.

Pros:

- Preserve content
- Match overall style

Cons:

- Distort local style pattern
- Unpleasing visual artifacts
- Fail to maintain content structure

Result: Unimodal representation such as Gram or covariance matrix may not be sufficient.

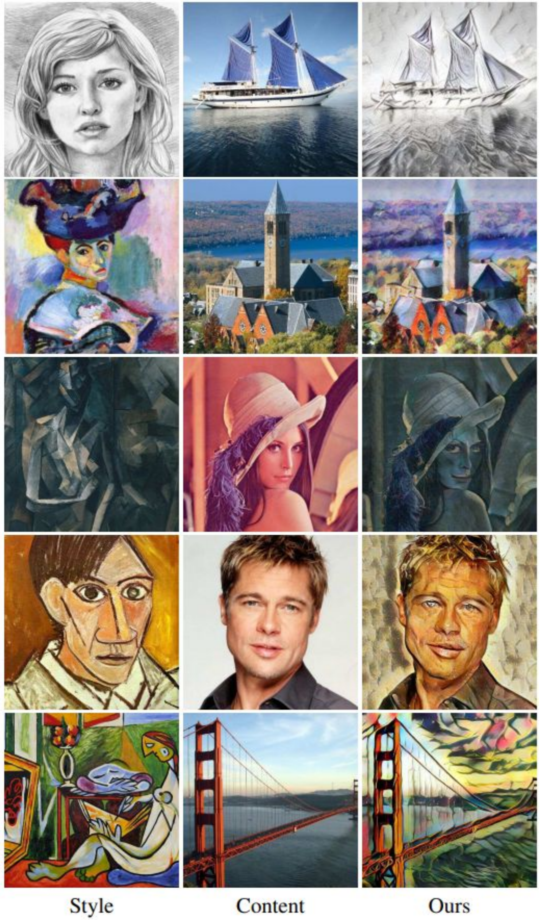
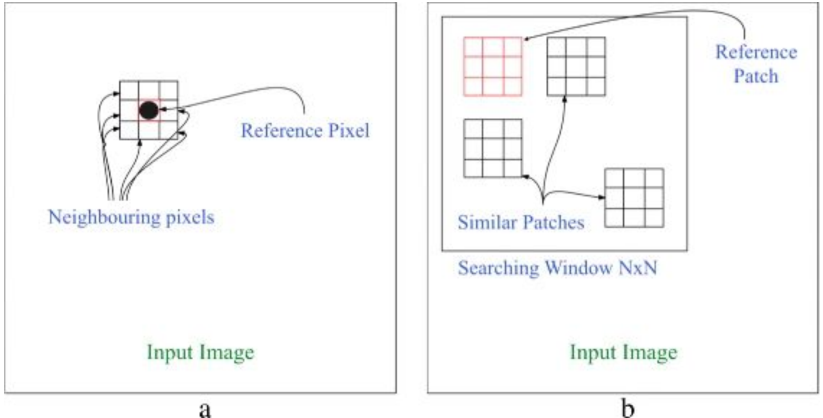


Figure: Examples for the AdaIN method [3].

Neural patch-based methods

"*An ideal style representation should respect the spatially-distributed style patterns*"

A solution to this: Patch-based methods.
They usually use **greedy** example matching for the style and content.



Pros:
- Visually pleasing results when content and style images have same structure
- Regularize / prevent over-exciting artifacts

Cons:
- Less desired style pattern
- Shape distortion

Result: Limit in the choice of style images

Figure: **a** filtering based on neighboring pixels located within a kernel in pixel-based denoising schemes and **b** filtering based on patches located within a search window in patch-based denoising schemes

Neural patch-based methods

Figure: Patch-swap based methods (CNNMRF [6], DFR [7], and AvatarNet [8]) may copy some less desired style patterns (labeled with red arrows) compared to MST.

Solution proposed:
Multimodal style representation with **graph based style matching** mechanism, to adaptively match the style patterns to a content image.

Why ?

- Robustness and flexibility.
- Better models the style feature distribution.
- The user can mix and match different styles to render diverse stylized results.
- Style clusters are adapted to content features with respect to the content spatial configuration.

How ?

- They we formulate the matching between content and style features as an energy minimization problem.
- Then they use **Graph cuts**, a powerful method for discrete optimization problem.

Let’s see this in details in our next chapter !

CHAPTER 2
MST METHOD

IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

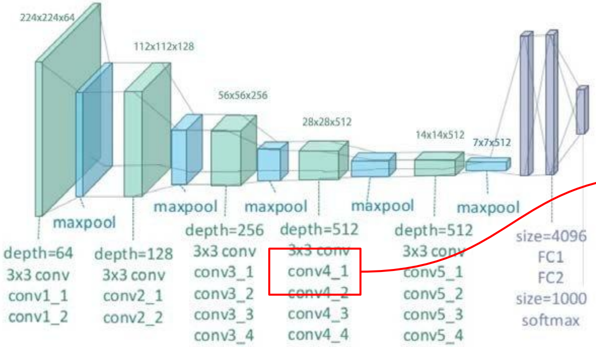
previous work

- features from whole image treated equally
- patch based methods

lack of flexibility

Solution: **“Multimodal style representation”**

Style image



Pretrained VGG19

F_s

28x28x512
=
28x28 features of 512 dimensions

28x28x512
=
28x28 features of 512 dimensions

“Multimodal style representation”

The idea is to find the principal modes (patterns) in this high dimension (512) feature space

K-means clustering on the 28x28 vectors.
In one cluster, features are likely drawn from the same distribution

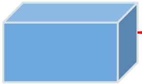
$$F_s = F_s^{l_1} \cup F_s^{l_2} \cup \cdots \cup F_s^{l_k} \cup \cdots \cup F_s^{l_K}$$

BUT is it relevant to have such expectations ?

Is it relevant to expect that features are likely drawn from same distributions ?

VGG 19 Conv4_1

F_s

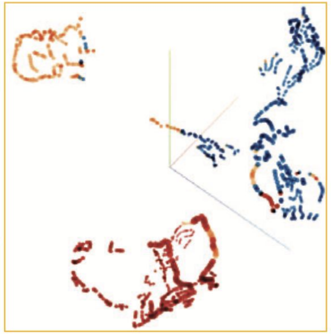


28x28x512

K-means clustering with K=3

t-SNE dimension reduction
28x28x512 -> 28x28x3
Projection of the points in a 3 dimensional space keeping distances

Represent points in 3 dim space, coloring the points with its cluster

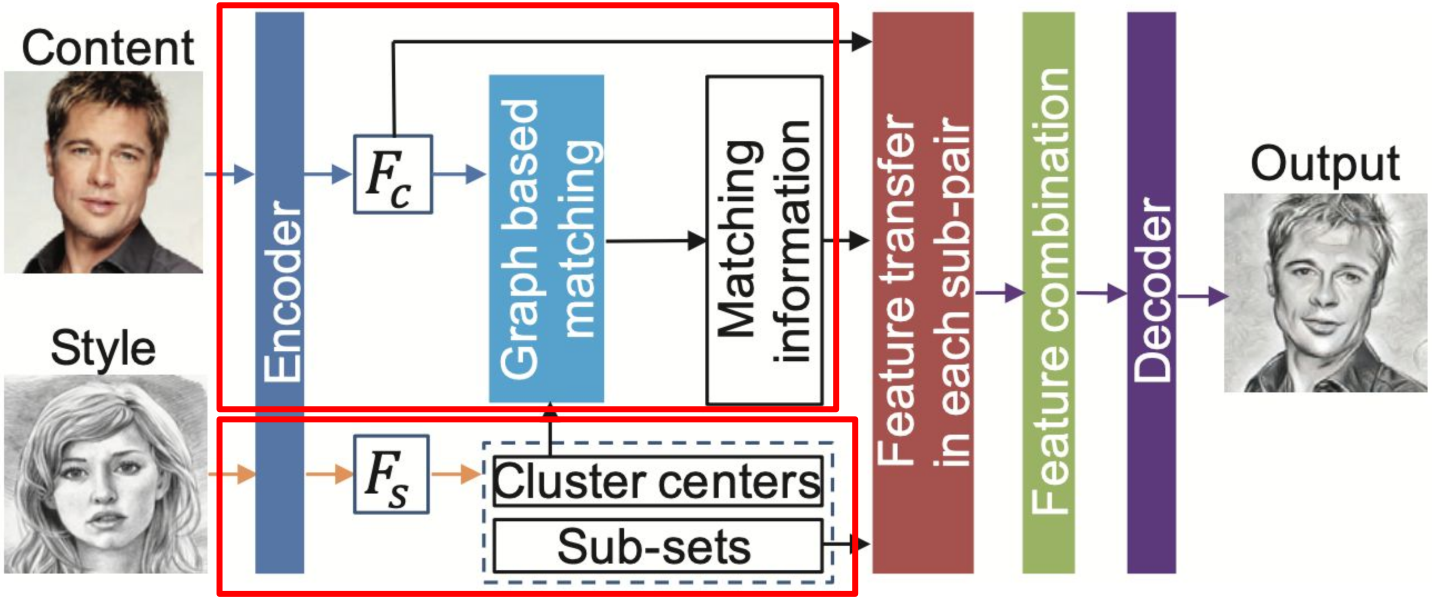


Interpretation:
Nearby points in feature space (512 dim) tend to be in the same cluster
-> there are tendencies, **modes**, in features

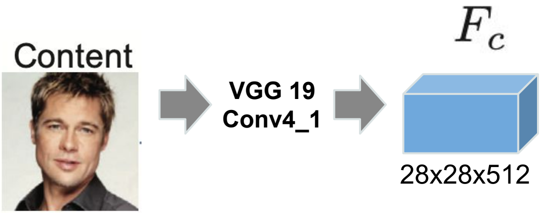
K-means cluster centroids

style features (principal modes)

2.1 Encoding content and style features



2.1 Encoding content and style features



For each feature vector, find closest style feature

"Closest" ?

$$D\left(F_{c,p}, F_{s,l_k}\right) = 1 - \boxed{\frac{{F_{c,p}}^T F_{s,l_k}}{\|F_{c,p}\| \, \|F_{s,l_k}\|}}$$

cosine distance
similarity degree between vectors

Energy minimization problem:
find labelling f that minimizes

$$E(f) = E_{data}(f) + E_{smooth}(f)$$

$$E_{data}(f) = \sum_{p=1}^{HcWc} D\left(F_{c,p}, F_{s,f_p}\right)$$

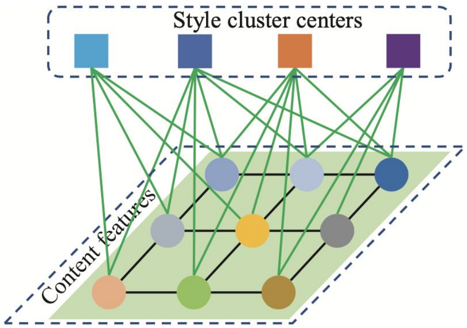
To guarantee that nearby features in content image get same style label for smoothness

Problem: NP-hard !

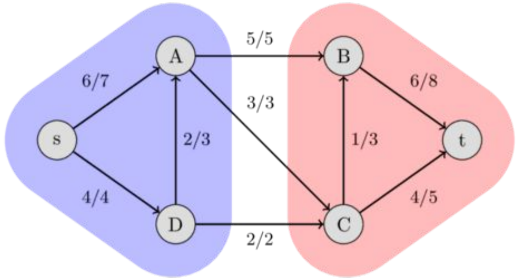
How to solve it ?

Graph cut !

Graph formulation

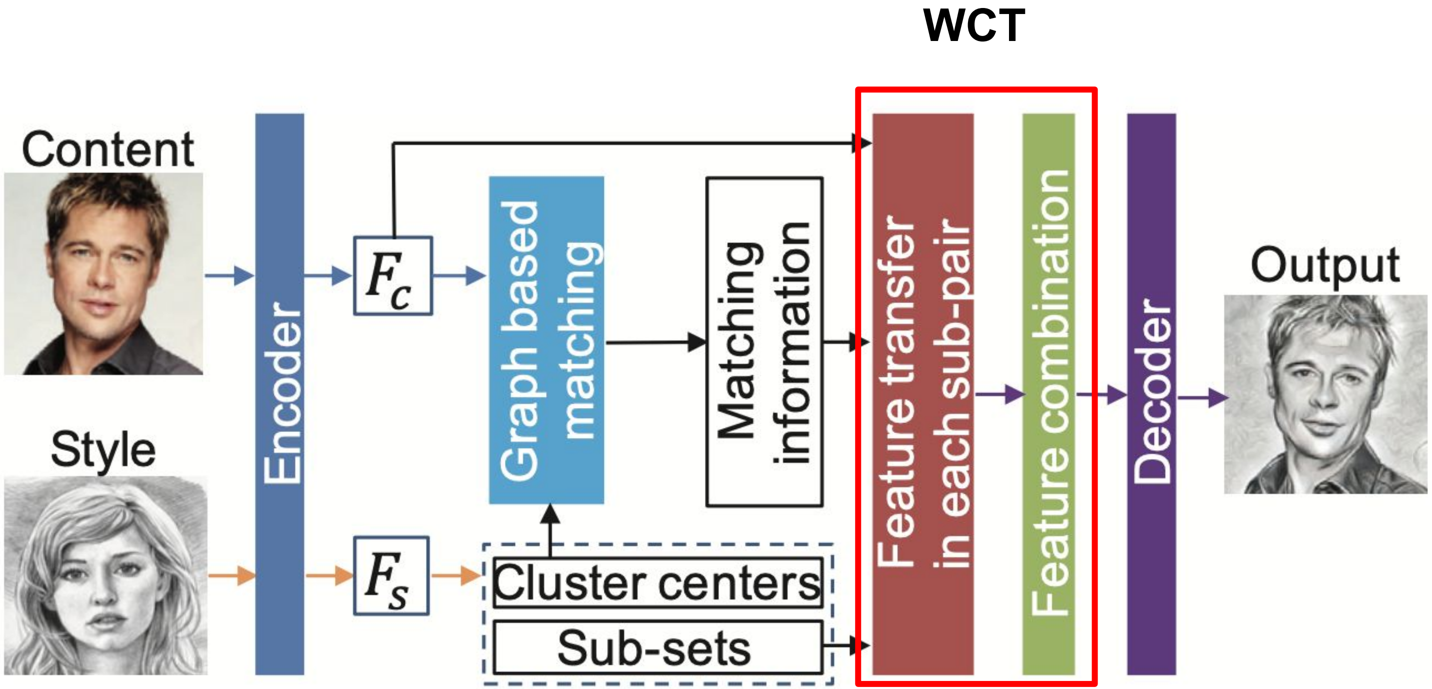


Graph cut : max-flow min cut



algorithm for minimization
pros: finds the global minimum of the energy formulation

2.2 Feature transfert & combination



Whitening :

step 1 : center Fc

step 2 : operate linear transformation of Fc so that the features maps are uncorrelated $(\hat{f}_c \hat{f}_c^{\top} = I)$

$$\hat{F}_c = E_c D_c^{-\frac{1}{2}} E_c^T F_c$$
$$with \;\; F_c^T . F_c = E_c D_c E_c^T$$

Coloring :

step 1 : operate linear transformation of Fc so that the features maps get the desired correlation $(f_{cs} f_{cs}^{\top} = f_s f_s^{T})$

$$\tilde{F}_{cs} = E_s D_s^{+\frac{1}{2}} E_s^T \tilde{F}_c$$
$$with \;\; F{s}^T . F_s = E_s D_s E_s^T$$

step 2 : de-center Fcs with mean of Fs

Figure 2: Inverting whitened features. We invert the whitened VGG Relu_4_1 feature as an example. Left: original images, Right: inverted results (pixel intensities are rescaled for better visualization). The whitened features still maintain global content structures.

IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

For each content style pair group :

$$F_{cs}^{l_k} = C_s W_c F_c^{l_k} + \mu\left(F_s^{l_k}\right)$$

Why WCT ?!?

its robustness and efficiency

Blending for better performance

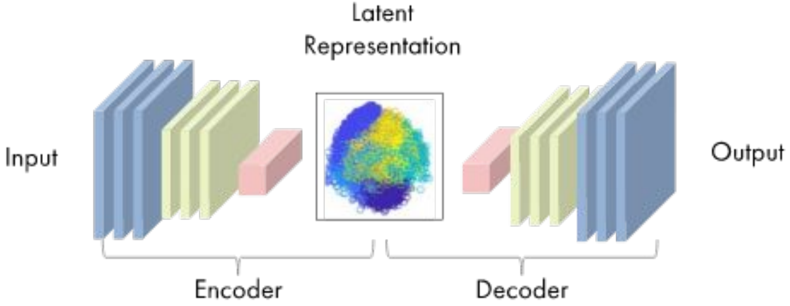
$$F_{cs}^{l_k} = \alpha_k F_{cs}^{l_k} + (1 - \alpha_k)\, F_c^{l_k}$$

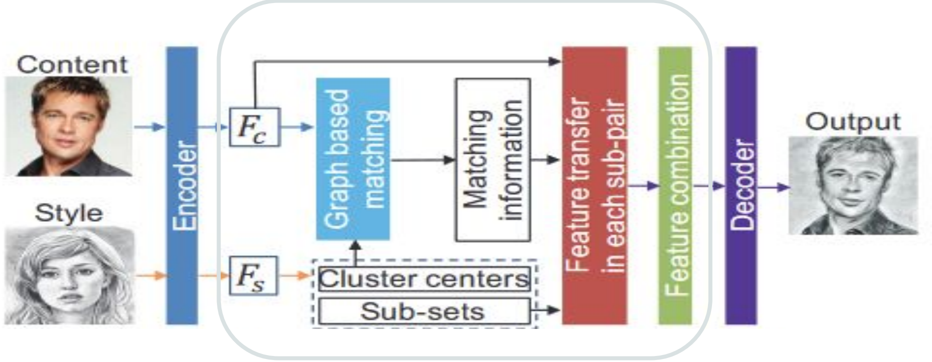
Ponderate mean of the new style and the original image

CHAPTER 3
IMPLEMENTATION

IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

Autoencoder :
The decoder is obtained by mirroring the encoder and replacing the max pooling by “Nearest Up scaling layers”

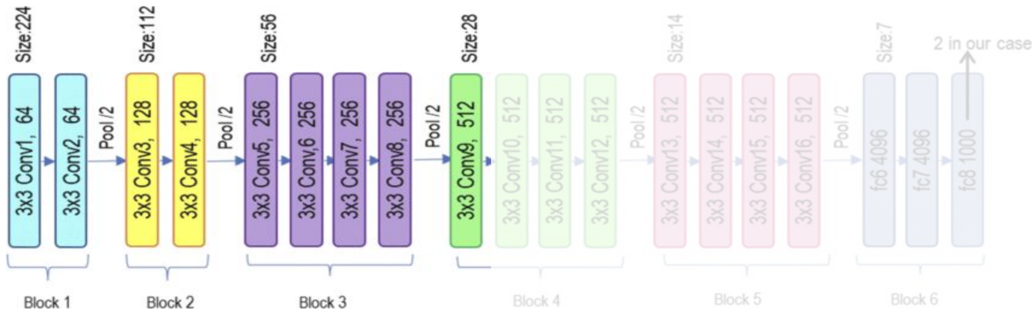




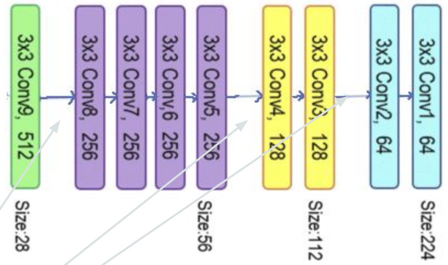
Hand Crafted process : Clustering, graph cut, WCT

Base CNN : VGG19

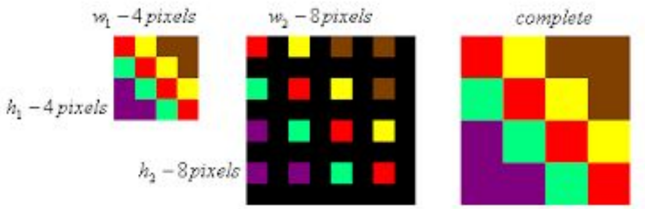
Encoder



Decoder



Nearest Up scaling



IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

Encoder :

- VGG19 trained on ImageNet
- Weights are frozen

Decoder :

- Train on COCO and WikiArt (40k images), around 80k images in total, randomly cropped at 256x256
- lr=10e-4

Used loss :

$$l_{total} = l_c + \gamma l_s$$

γ = 0.01

$$l_c = \|\phi_{4\_1}(I_c) - \phi_{4\_1}(I_{cs})\|_2$$

$$l_s = \sum_{i=1}^{4} (\|\mu(\phi_{i\_1}(I_s)) - \mu(\phi_{i\_1}(I_{cs}))\|_2) + \sum_{i=1}^{4} (\|\sigma(\phi_{i\_1}(I_s)) - \sigma(\phi_{i\_1}(I_{cs}))\|_2)$$

Ic : Content image
Is : Style image

φi_1 : Feature map of block i layer 1

CNNMRF :

- Extracts a pool of neural patches from style images, with which patch matching is used to match content
- Minimizes energy function to synthesize the results

MST :

- Clusters style features into multiple sub-sets and matches style cluster centers with content feature points via graph cuts
- MST generates stylization results with a decoder

WCT :

- The decoder is trained by using only content data and loss
- Uses multiple layers of VGG features and conducts multi-level coarse-to-fine stylization, which costs much more time and sometimes distorts structures

MST :

- Introduces additional style images for training
- Only transfers single-level content and style features.

CHAPTER 4
EXPERIMENTS

IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

In graph building, we compute distance between style features

Inputs MST-3 (Euclidean) MST-3 (Cosine)

Figure 7: Distance measurement investigation.

$$D(F_{c,p}, F_{s,l_k}) = 1 - \frac{{F_{c,p}}^T F_{s,l_k}}{||F_{c,p}||\ ||F_{s,l_k}||}$$

- ▶ No normalization in the euclidean distance

Evaluation the effectiveness of the smooth term of energy measurement :

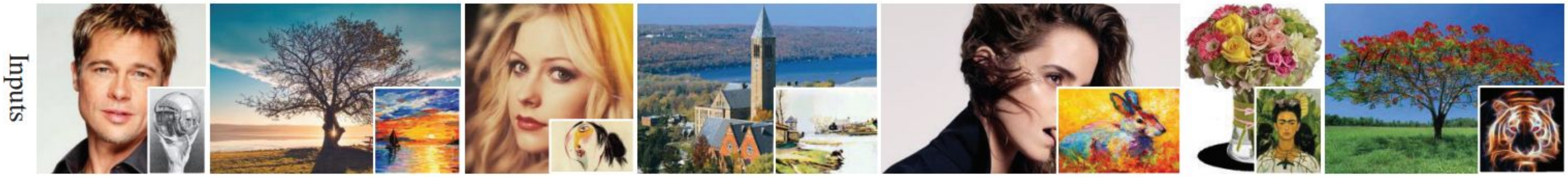
$$E(f) = E_{data}(f) + E_{smooth}(f)$$

$$E_{smooth}(f) = \sum_{\{p,q\}\in\Omega} V_{p,q}(f_p, f_q),$$

$$V_{p,q}(f_p, f_q) = \boxed{\lambda} \cdot T(f_p \neq f_q),$$

Figure 8: Discontinuity preservation investigation.

4.3 Qualitative comparisons to prior work



4.4 User study

- 20 pairs of content-style/user
- the user select its favourite among the 6 methods
- we obtain 2000 votes from 100 users

Table 1: Percentage of the votes that each method received.

Method	Gatys	AdaIN	WCT	DFR	AvatarNet	MST
Perc./%	21.41	11.31	12.67	11.55	9.61	**33.45**

4.5 Efficiency

Table 2: Running time (s) comparisons.

Method	Gatys	AdaIN	WCT	DFR	AvatarNet
Time (s)	116.46	0.09	0.92	54.32	0.33
Method	MST-1	MST-2	MST-3	MST-4	MST-5
Time (s)	0.20	1.10	1.40	1.97	2.27

Average time on 100 image pairs of 512x512px

4.6 Style cluster number

MST offers :

- a more adaptive and distinctive style representation
- various relevant stylization with different K, providing multiple selection for the user

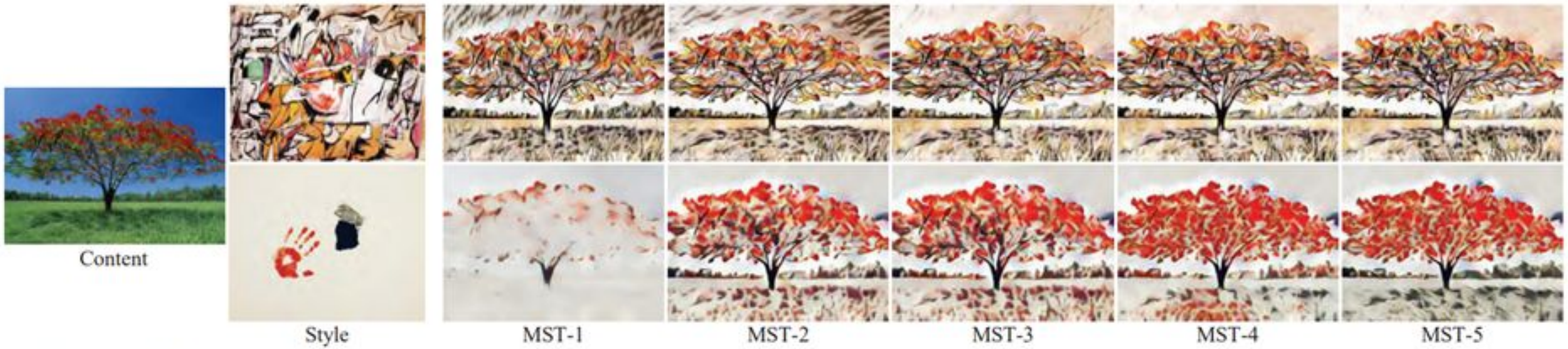


Figure 11: Style cluster number investigation. Same content image with complex and simple style images.

Figure 12: Multi-style transfer. MST treats patterns from different style images distinctively and transfers them adaptively.

- Reveal the importance of weight and style matching

4.8 Generalization of MST

Figure 13: Generalization of MST to AdaIN [11].

CONCLUSION

IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

CONCLUSION

Take home messages

Image Style transfer and Deep learning (neural style transfer methods) :

- Recovering the style of an image with :
  - correlation of the different features map (parametric)
  - patch based and local information (non parametric)
- Style matching is a minimization problem

Multimodal style représentation and graph cut:

- Better modelisation of the style feature distribution
- Graph representation allow style - content matching w.r.t content spatial configuration
- Energy minimization problem solved with graph cuts algorithm

Sources

- [1] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In CVPR, 2016.
- [2] Le Huy Hien, Ngo & Huy, Luu & V.H., Nguyen. (2021). Artwork Style Transfer Model using Deep Learning Approach. Cybernetics and Physics. https://www.researchgate.net/publication/356667127_Artwork_Style_Transfer_Model_using_Deep_Learning_Approach
- [3] Xun Huang and Serge J Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In ICCV, 2017
- [4] Xueting Li, Sifei Liu, Jan Kautz, and Ming-Hsuan Yang. Learning linear transformations for fast arbitrary style transfer. In CVPR, 2019
- [5] Yijun Li, Chen Fang, Jimei Yang, Zhaowen Wang, Xin Lu, and Ming-Hsuan Yang. Universal style transfer via feature transforms. In NIPS, 2017
- [6] Chuan Li and Michael Wand. Combining markov random fields and convolutional neural networks for image synthesis. In CVPR, 2016
- [7]Shuyang Gu, Congliang Chen, Jing Liao, and Lu Yuan. Arbitrary style transfer with deep feature reshuffle. In CVPR, 2018.

Sources

- [8] Lu Sheng, Ziyi Lin, Jing Shao, and Xiaogang Wang. Avatarnet: Multi-scale zero-shot style transfer by feature decoration. In CVPR, 2018

Questions

1) How can the MST algorithm benefit other existing style transfer methods? (Wassim CHAKROUN)

2) Do you think that in a certain (something similar to a PCA) decomposition all masterpieces will have a similar cluster however the style, meaning that there is a “structure of greatness” that does exist in our human minds? (Michel TARLIN)

3) In the algorithm, they extract features from the conv_4_1 layer of VGG-19. Do you have any idea why they choose this layer and how the performance of the algorithm may change by modifying the feature layer? (Thibaud ETEVENARD)

4) In the graph based style matching part, can you explain to me how the difference in scale between the content and style features was taken into account? (Achraf JENZRI)

5) Why your model does not copy some less desired style patterns like other models (ex : eyes) ?(Robin Armingaud)

Questions