
LUT
University
LAND
OF THE
CURIOUS

Computational Engineering School
Lappeenranta–Lahti University of Technology LUT

LUT University

CLIP FOR IMAGE STYLE TRANSFER: EXPLORE TEXT-IMAGE CORRELATIONS

CVPR SEMINAR PRESENTATION

Presented by Nadine Cyizere Bisanukuli

February 23, 2024

Examiners:

- Prof. Zhi-Song Liu, Lappeenranta-Lahti University of Technology (LUT)
- Doctor Jun Xiao, Hong Kong Polytechnic

OUTLINE

LUT University

1 Introduction and Background
- Introduction
- Background and Challenges
- Architecture
- Objectives

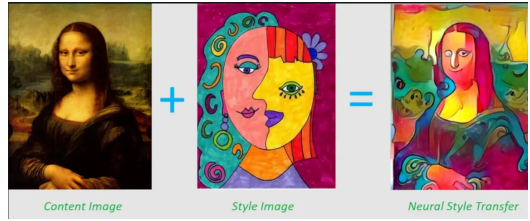
2 Related Work
- Image-Driven Style Transfer
- Text-Driven Style Transfer
- Attention-Driven Style Transfer
- Loss Function

3 Proposed Methods
- CLIP for text-image correlation
- Clip for style transfer

4 Next Steps

INTRODUCTION

Style transfer: the process of applying artistic styles to images.



Applications: Film production, Fashion design, Virtual reality, social media

Film Production

Fashion Design

BACKGROUND AND CHALLENGES

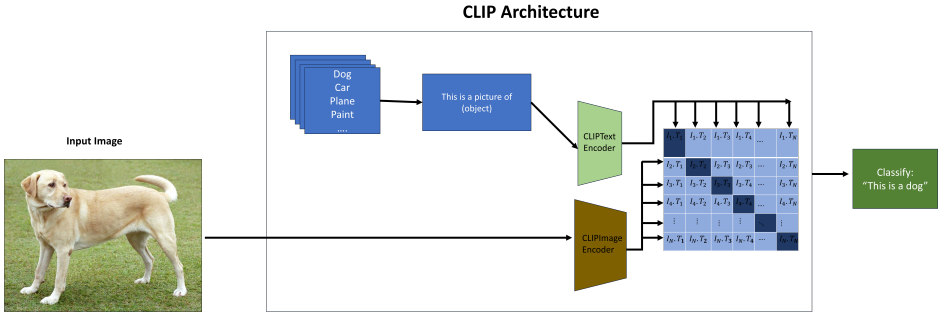
Traditional methods rely on reference images, limiting creativity and applicability

Domain specificity hinders generalization to arbitrary styles.

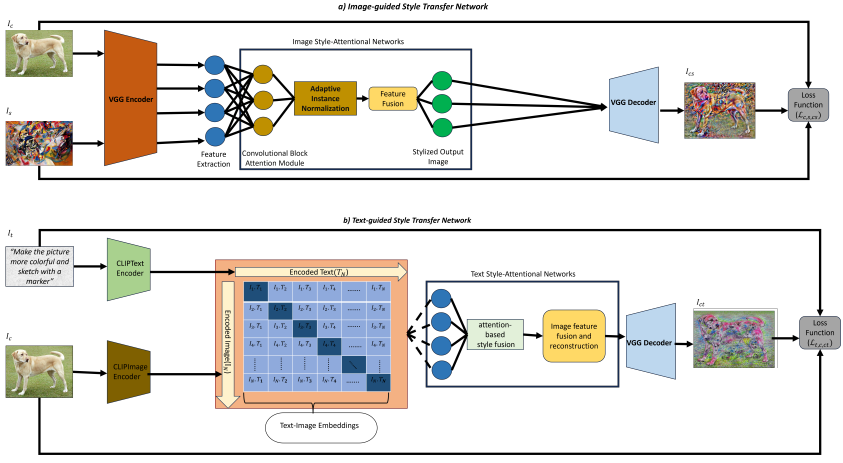
- Computational efficiency.
- CLIP(*Contrastive Language–Image Pre-training*) [3] presents new possibilities for manipulating styles using text descriptions

CLIP ARCHITECTURE

This thesis will dive deep into the intersection of CLIP of the original target image and image style transfer. This will leverage the text-image correlations.



STYLE TRANSFER ARCHITECTURE



MOTIVATION AND OBJECTIVES

LUT University

Objectives:

1. Explore the effectiveness of CLIP for image style transfer.
2. Leverage text-image correlations to enhance image style transfer.
3. Conduct extensive experiments and analysis on text-driven style transfer.

Scope:

- Focus on using the CLIP model for image style transfer, excluding other models or methods.
- Evaluate the proposed method using available datasets.
- Computational efficiency is not the primary focus.
- Limited exploration of CLIP for tasks beyond image style transfer.

OUTLINE

LUT University

1 Introduction and Background
- Introduction
- Background and Challenges
- Architecture
- Objectives

2 Related Work
- Image-Driven Style Transfer
- Text-Driven Style Transfer
- Attention-Driven Style Transfer
- Loss Function

3 Proposed Methods
- CLIP for text-image correlation
- Clip for style transfer

4 Next Steps

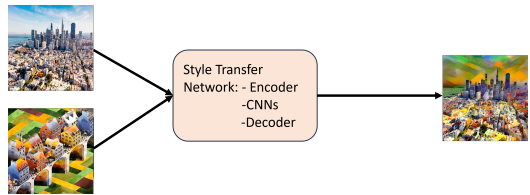
RELATED WORK

Style transfer in Computer Vision has seen significant advancements. This was done through the Exploration of various techniques and methodologies.

1. Image-Driven Style Transfer:

Advantages:

- Utilizes of CNN.
- Easy transfer of styles across multiple domains which makes creation and alteration easy [6].



Disadvantages:

- Fixed styles in feed-forward methods [1].
- computational efficiency and quality.
- Limit generalization.

CONT'D RELATED WORK

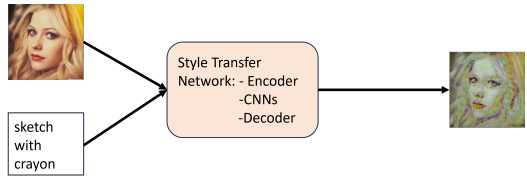
2. Text-Driven Style Transfer:

Advantages:

- Textual descriptions for image transformation.
- CLIP for text-image correlations [4].
- Flexibility for arbitrary styles.

Disadvantages:

- Limitations in manipulating images beyond trained domains.
- Content preservation for complex styles.
- Computational complexity



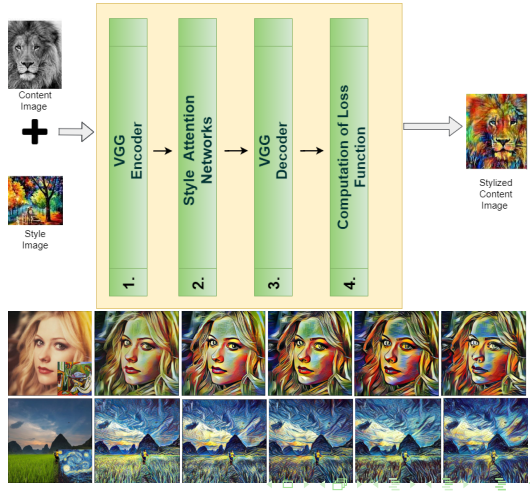
CONT'D RELATED WORK

3. Attention-Driven Style Transfer: [5]
Advantages:

- Attention mechanisms to capture key styles.
- Quality and diversification of output.
- Adaptability and scalability of features.

Disadvantages:

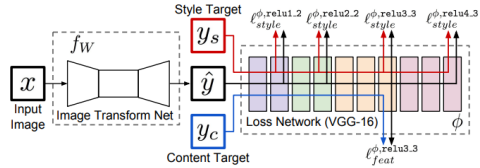
- Optimizing attention mechanisms



CONT'D RELATED WORK

3. VGG Loss Function

- Content Loss: takes the transformed image (y) and the content target (y_c) to compute the loss.
- Style Loss: takes the transformed image (y) and the style target (y_s) to compute the loss.



VGG Loss network: is a pre-trained network (usually VGG-16) used to calculate the loss functions.

Goal: Minimize the total loss, which is a weighted combination of the style and content losses. The weights of the Image Transform Network are updated iteratively to minimize this loss.

OUTLINE

1 Introduction and Background
- Introduction
- Background and Challenges
- Architecture
- Objectives

2 Related Work
- Image-Driven Style Transfer
- Text-Driven Style Transfer
- Attention-Driven Style Transfer
- Loss Function

3 Proposed Methods
- CLIP for text-image correlation
- Clip for style transfer

4 Next Steps

DATA DESCRIPTION

1. **WikiArt Dataset**
   - **Source**: Hugging Face Datasets Repository
   - **Purpose**: The WikiArt dataset is a collection of artwork images along with associated metadata.
2. **Images:** Over 11300 high-resolution images of artworks spanning various styles, genres, and periods.
3. **Metadata** Each artwork image is accompanied by rich metadata including:
   - Artist Name
   - Genre
   - Style
   - Image
4. The dataset covers a wide range of artistic styles, including but not limited to:
   - Impressionism
   - Cubism
   - Surrealism
   - etc.

CLIPSTYLER OUTPUT

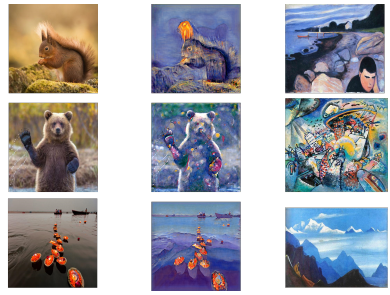
ClipStyler [2] for Style transfer:

1. Input Text Reference and content Image
2. Text and images are encoded into a shared space.
3. Images are iteratively generated to match text.
4. Optimization
5. Loss function

Edvard Munch

Wassily kandinsky

Nicholas Roerich

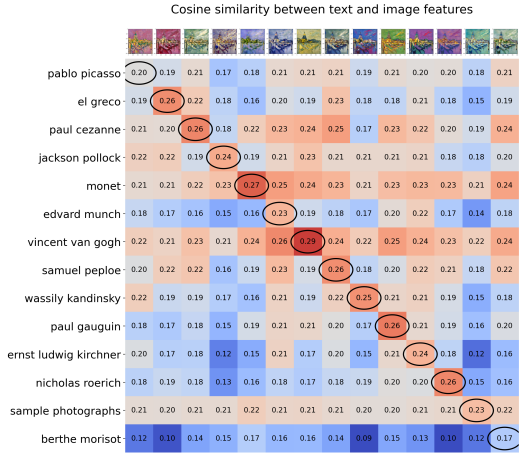


PROPOSED METHOD

CLIP for text-image correlation:

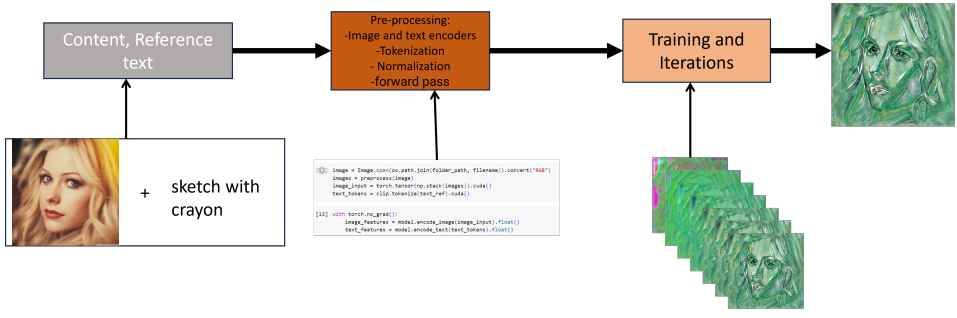
1. CLIP model: "ViT-B/32"
2. Preprocess Images
3. Tokenize Text
4. Encode Features:
5. Normalize Features
6. Calculate Cosine Similarity

Goal: Make sure the style transfer results is correlated to the text, Use clip as the most efficient and common model for the task.



CONT'D PROPOSED METHOD

2. Clip for style transfer



MULTIPLE STYLE TRANSFER

3. Multiple Style Transfer

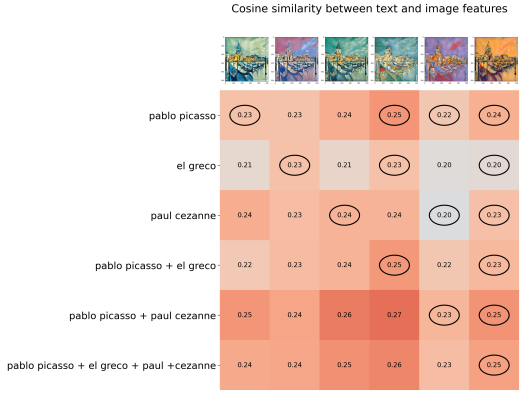
- Combines multiple artistic styles on an input image.
- Creates composite images with various artistic characteristics.

Reading the Similarity Matrix:

- Higher values indicate stronger text-image correspondence.
- Rows represent textual descriptions; columns represent images.

Acknowledging Inaccuracies:

- Results may not always be correct so Continued exploration is needed for refinement.



OUTLINE

1 Introduction and Background
- Introduction
- Background and Challenges
- Architecture
- Objectives

2 Related Work
- Image-Driven Style Transfer
- Text-Driven Style Transfer
- Attention-Driven Style Transfer
- Loss Function

3 Proposed Methods
- CLIP for text-image correlation
- Clip for style transfer

4 Next Steps

NEXT STEPS

1. Ablation experiments on attention for text-driven style transfer
2. Full evaluation, including VGG loss and CLIP text-image correlation
3. Exploration on Multiple Style Transfer

CONCLUSION

LUT University

THANK YOU!

References

[1] Xiancai Ji, Yao Lu, and Li Guo. “Image Super-Resolution with Deep Convolutional Neural Network”. In: *2016 IEEE First International Conference on Data Science in Cyberspace (DSC)*. 2016, pp. 626–630. DOI: 10.1109/DSC.2016.104.

[2] Gihyun Kwon and Jong Chul Ye. “CLIPstyler: Image Style Transfer with a Single Text Condition”. In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022, pp. 18041–18050. DOI: 10.1109/CVPR52688.2022.01753.

[3] Or Patashnik et al. “StyleCLIP: Text-Driven Manipulation of StyleGAN Imagery”. In: *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. 2021, pp. 2065–2074. DOI: 10.1109/ICCV48922.2021.00209.

[4] Alec Radford et al. “Learning Transferable Visual Models From Natural Language Supervision”. In: *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*. Ed. by Marina Meila and Tong Zhang. Vol. 139. Proceedings of Machine Learning Research. PMLR, 2021, pp. 8748–8763. URL: http://proceedings.mlr.press/v139/radford21a.html.

[5] Yuan Yao et al. “Attention-Aware Multi-Stroke Style Transfer”. In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019, pp. 1467–1475. DOI: 10.1109/CVPR.2019.00156.

[6] Yabin Zhang et al. “Exact Feature Distribution Matching for Arbitrary Style Transfer and Domain Generalization”. In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.


LUT
University