



CLIP FOR IMAGE STYLE TRANSFER: EXPLORING TEXT-IMAGE CORRELATIONS

Lappeenranta-Lahti University of Technology LUT

Master's Program in Computational Engineering, Master's Thesis

2024

Nadine Bisanukuli Cyizere

Examiners: Professor Zhisong Liu

(Use \SecondExaminer to replace this text.)

ABSTRACT

Lappeenranta-Lahti University of Technology LUT
School of Engineering Science
Computational Engineering

Nadine Bisanukuli Cyizere

CLIP for Image Style Transfer: Exploring Text-Image Correlations

Master's thesis

2024

x pages, y figures, z tables, w appendices

Examiners: Professor Zhisong Liu and (Use \SecondExaminer to replace this text.)

Keywords: computer vision, machine vision, image processing, pattern recognition

ACKNOWLEDGEMENTS

The author can decide the contents of this page. Usually the place of work and related people (supervisors, collaborators, friends, relatives, etc.) are acknowledged.

I would like to thank my supervisors ...friends ... family ...

Lappeenranta, January 7, 2024

Nadine Bisanukuli Cyizere

Advice: All symbols and abbreviations are listed on this page in the alphabetical order. Remember to introduce the abbreviation when it is used in the text for the first time.

You may use the automated system, depending on your LaTeX environment:

```
\glsnogroupskiptrue
\setlength{\glsdescwidth}{1.0\hsize}
\printglossary[title=LIST OF ABBREVIATIONS,type=\acronymtype,
style=long, nonumberlist, nopostdot]
```

CONTENTS

| | | |
|----------|--|-----------|
| 1 | INTRODUCTION | 5 |
| 1.1 | Background | 5 |
| 1.2 | Objectives and delimitations | 6 |
| 1.3 | Structure of the thesis | 7 |
| 2 | RELATED WORK | 8 |
| 3 | PROPOSED METHODS | 10 |
| 3.1 | Results | 10 |
| 4 | DISCUSSION | 11 |
| 4.1 | Future work | 11 |
| 5 | CONCLUSION | 12 |
| | REFERENCES | 13 |

1 INTRODUCTION

1.1 Background

In the field of computer vision and Natural Language processing, the combination of textual description and visual elements has given rise to innovation in many real-world applications with image style transfer as the crucial and captivating one. Image style transfer gives the possibility to improve visual aesthetics and generation of artistic creations easier. The field of Image style transfer has had different advancements over the years with the involvement of different machine learning techniques mostly the use of deep learning. The main goal of Image style transfer is to apply the style of an image usually referred to as **style reference**, to other images while preserving the original content. This thesis will dive deep into the intersection of Contrastive Language-Image Pre-training (CLIP) and image style transfer to leverage the text-image correlations for more advancements in the domain. OpenAI's CLIP model, introduced by Radford et al, has demonstrated new and better possibilities in the field of image style transfer [1, 2] as shown in Figure 1. Unlike some traditional methods that require reference style image, CLIP can do style transfer by only using a text description of the desired style. This is very useful in case a user doesn't have reference style images but is interested in transferring styles based on their imagination. Several methods have been proposed to better the performance of

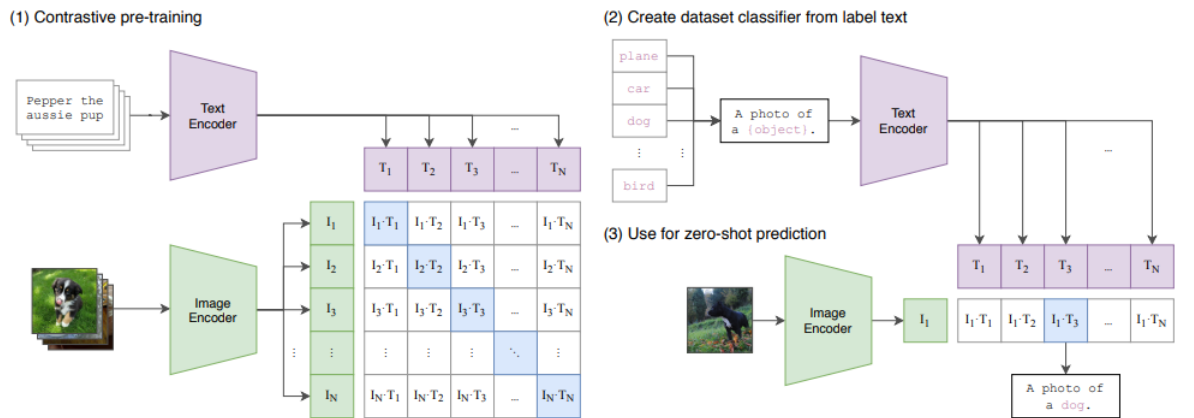


Figure 1. CLIP takes a different technique in label prediction, as opposed to training a linear classifier and an image feature extractor together. It focuses on predicting the correct pairings within a batch of (image, text) training examples, training both a text and an image encoder simultaneously. By embedding the names or descriptions of the classes in the target dataset, the trained text encoder creates a zero-shot linear classifier during testing. This novel method provides a more flexible and context-aware predictive model by improving the synthesis of meaningful links between images and related textual data.

image style transfer. Despite the advancements, challenges persist in the field of image style transfer. One of the challenges is the need for methods that can adapt to styles without loss of content. Traditional methods depend on reference style image which limits their applicability in scenarios where users might not have specific reference images but wish to transfer styles. Additionally achieving balance between style transfer and preservation of content remains an issue. This lies in more development of algorithms that can incorporate the desired style while preserving content of the image. Another challenge lies in the domain specificity since many existing methods lie to specific domain such as portraits, landscapes and industrial settings. Hence generalising these models to handle arbitrary styles is an ongoing challenge but has been tried to be solved. Lastly computational efficiency is a concern since as image style transfer models become more and more complex, there is a growing need of development of less complex models that can deliver results without much consumption of computational resources. There have been multiple solutions to the challenges such as development of PCA-based knowledge distillation method to distill lightweight models demonstrating its usability with different architectures hence improvement of efficiency and suitability for real time applications. Also, advancements in the Industrial Style Transfer method have shown promising results in creating new visual products with a nice appearance for industrial designers' reference.

The exploration of CLIP for image style transfer presents an exciting edge in the field of computer vision. The ability to manipulate and transfer styles using text descriptions could revolutionize how we interact with digital imagery. This thesis aims to contribute to this field by addressing the challenges of style transfer without explicit style references, achieving a balance between style and content.


1.2 Objectives and delimitations

The primary objective of this thesis is to explore the potential of CLIP for image style transfer with a particular focus on leveraging text-image correlations. The specific research questions that this work aims to address are:

1. **Can CLIP be effectively used for image style transfer?** This involves the development and evaluation of a method that uses CLIP to transfer the style of a text description to an image. The performance of this method will be evaluated based on the quality of the style transfer and the preservation of the original content.
2. **How can text-image correlations be leveraged to improve image style transfer?**

~~This involves investigating how the correlations between text descriptions and their corresponding images, as learned by CLIP, can be used to enhance the style transfer process. This could involve modifying the style transfer method based on the specific text-image correlations identified.~~



The  scope of this thesis is limited to the following delimitations:

- The study will focus on the use of OpenAI's CLIP model for image style transfer. Other models or methods for image style transfer or text-image correlation learning will not be considered.
- The performance of the proposed method will be evaluated using available datasets. The collection of new data is beyond the scope of this work.
- While the aim is to develop a method capable of high-quality style transfer, the computational efficiency of the method will not be a primary focus of this work.
- The thesis will not explore the use of CLIP for other tasks beyond image style transfer, such as image generation or text-to-image synthesis.

By addressing these research questions, this thesis aims to contribute to the ongoing efforts to leverage text-image correlations for image style transfer. However, it is important to note that the proposed methods are subject to the inherent limitations and uncertainties of machine learning techniques. Future work may be needed to refine the methods and address any limitations identified in this study.

1.3 Structure of the thesis

2 RELATED WORK

In the realm of Computer vision, style transfer has witnessed significant advancements with exploration of different techniques [3,4]. Style transfer involves the process of transforming the visual appearance of an image by adding artistic nuances found in a reference style image. Over years, researchers have developed multiple approaches to achieve this goal which led to the categorization of style transfer methods into three main types: text-driven style transfer, image-driven text transfer, and attention-based style transfer.

Text-Driven Style Transfer. It emphasizes using textual descriptions to guide the transformation of images. It has been a significant area of research in style transfer over the past years marked by different discoveries. Such discoveries include the discovery of a novel framework. This work introduces a novel neural style transfer framework addressing practical scenarios where users lack reference style images but desire style transfer through text descriptions. Leveraging CLIP's pre-trained text-image embedding model, the proposed method achieves style transfer without a style image, relying solely on a single text condition [1]. This was followed by another crucial discovery that introduced text-driven style transfer (TxST) as a flexible alternative to traditional image style transfer, leveraging advanced image-text encoders like CLIP [5]. The proposed method employs a contrastive training strategy and a novel cross-attention module, enabling arbitrary artist-aware style transfer with superior performance, demonstrating potential for future advancements in mimicking various artistic styles.

Image-Driven Text Transfer. It focuses on generating textual descriptions based on the visual content of an image. It has undergone significant advancements intending to discover better techniques to apply artistic styles to diversify visual content. The foundation was laid by Johnson et al. that introduced the use of the VGG network for style transfer emphasizing the effectiveness of feature extraction from pre-trained networks in capturing and applying artistic styles across images [6]. More research was done by Huang et al. who showed real-time abilities of arbitrary style transfer through adaptive instance normalization in the AdaIN framework [7]. The two papers showed the foundation of Image-driven style transfer across multiple domains. Later on, Image-driven style transfer extended to more complex scenarios such as indoor 3D scene constructions. Hollein et al. introduced styleMesh, emphasizing the ability to complex three-dimensional environments [8]. Additionally, the issue of unpaired cartoon images was addressed by introducing gated cycle mapping which showed the potential for image-driven style transfer

without explicitly paired training data [9]. This also gave rise to another style transfer that's known as Attention-based style transfer whose architectures emerged as a significant theme in image-driven style transfer with the introduction to StyTr2 that incorporated transformers to capture complex style patterns [10]. Additionally, efforts were made to handle the computational efficiency and quality of the styles issue as seen in [11]. Moreover, there was work done for industrial style transfer addressing industrial contexts for large-scale geometric warping and content preservation for preserving content while applying complex styles [12]. More efforts were made to arbitrary style transfer and domain generalization with exact feature distribution matching provided insights into handling diverse styles and applicability of models across different domains [13].

Attention-Based Style Transfer. It uses the power of attention mechanisms through the use of advanced architectures like transformers. This has become a focus recently due to its crucial application in image-driven style transfer since it employs sophisticated architectures, specifically transformers, to enhance the quality and diversify the style transfer results. A very early contribution that dives deep into the integration of transformers using attention mechanism to capture complex style patterns and elevate expression abilities of style transfer models [10]. The adaptability of attention-based architectures showcased in StyTr2 signifies a move from traditional methods while offering more possibilities to create visually exciting images. Although the paper [13] doesn't explicitly mention attention-based mechanisms, the use of transformers suggests the use of attention-based features. The paper hints at the importance of aligning feature distributions for arbitrary style transfer and domain generalization that showcase the ongoing development of methodologies in attention-based styles. Furthermore, the paper [14] talks about the importance of the attention mechanism in style transfer. This work introduces an attention-based approach allowing the model to create realistic styled images with multiple strokes, this ensures focus on specific regions enhancing the stylization and adding good characteristics to the output image. These papers show how attention-based mechanisms are transforming artistic style transfer.

3 PROPOSED METHODS

3.1 Results

4 DISCUSSION

4.1 Future work

5 CONCLUSION

REFERENCES

- [1] Suchen Wang, Yueqi Duan, Henghui Ding, Yap-Peng Tan, Kim-Hui Yap, and Jun-song Yuan. Learning transferable human-object interaction detector with natural language supervision. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 929–938, 2022.
- [2] Jimmy Lei Ba, Kevin Swersky, Sanja Fidler, and Ruslan Salakhutdinov. Predicting deep zero-shot convolutional neural networks using textual descriptions. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 4247–4255, 2015.
- [3] 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)*, pages 18041–18050, 2022.
- [4] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021.
- [5] Zhi-Song Liu, Li-Wen Wang, Wan-Chi Siu, and Vicky Kalogeiton. Name your style: text-guided artistic style transfer. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 3530–3534, 2023.
- [6] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 694–711, Cham, 2016. Springer International Publishing.
- [7] ManMan Peng and Zhongrui Zhu. Enhanced style transfer in real-time with histogram-matched instance normalization. In *2019 IEEE 21st International Conference on High Performance Computing and Communications; IEEE 17th International Conference on Smart City; IEEE 5th International Conference on Data Science and Systems (HPCC/SmartCity/DSS)*, pages 2001–2006, 2019.
- [8] Lukas Höllein, Justin Johnson, and Matthias Nießner. Stylemesh: Style transfer for indoor 3d scene reconstructions. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6188–6198, 2022.
- [9] Yifang Men, Yuan Yao, Miaomiao Cui, Zhouhui Lian, Xuansong Xie, and Xian-Sheng Hua. Unpaired cartoon image synthesis via gated cycle mapping. pages 3491–3500, 2022.

- [10] Yingying Deng, Fan Tang, Weiming Dong, Chongyang Ma, Xingjia Pan, Lei Wang, and Changsheng Xu. Stytr2: Image style transfer with transformers. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11316–11326, 2022.
- [11] Tai-Yin Chiu and Danna Gurari. Pca-based knowledge distillation towards lightweight and content-style balanced photorealistic style transfer models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7834–7843, 2022.
- [12] Jinchao Yang, Fei Guo, Shuo Chen, Jun Li, and Jian Yang. Industrial style transfer with large-scale geometric warping and content preservation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7824–7833, 2022.
- [13] Yabin Zhang, Minghan Li, Ruihuang Li, Kui Jia, and Lei Zhang. Exact feature distribution matching for arbitrary style transfer and domain generalization. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8025–8035, 2022.
- [14] Yuan Yao, Jianqiang Ren, Xuansong Xie, Weidong Liu, Yong-Jin Liu, and Jun Wang. Attention-aware multi-stroke style transfer. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1467–1475, 2019.