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 31, 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

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 is an accommodating tool that facilitates the development of artistic works and enhances the quality of visual aesthetics. This technique is useful in many fields, such as graphic design, virtual reality, photo editing, film production, and even social media optimization. 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) [1, 2] of the original target image and image style transfer to leverage the text-image correlations for more advancements in the domain as shown in **Figure 1**. OpenAI's CLIP model has demonstrated new and better possibilities in the field of image style transfer. Unlike some traditional methods that require reference-style images, CLIP can find the correlations between texts and images. Benefiting from this ability, CLIPstyler [3] uses CLIP to achieve text-driven style transfer, using style description text only to transfer desirable styles to the content image. This is very useful in case a user does not have reference style images but is interested in transferring styles based on their imagination.

Several methods have been proposed to better the performance of 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 images which limits their applicability in scenarios where users might not have specific reference images but wish to transfer styles. Additionally achieving a 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 the content of the image. Another challenge lies in the dominant specificity since many existing methods lie in specific domains such as portraits, landscapes, and industrial settings. Hence generalising these models to handle arbitrary styles is an ongoing challenge [4] that has been tried to solve. Lastly, computational efficiency is a concern since as image style transfer models become more and more complex, there is a growing need for the development of less complex models that can deliver results

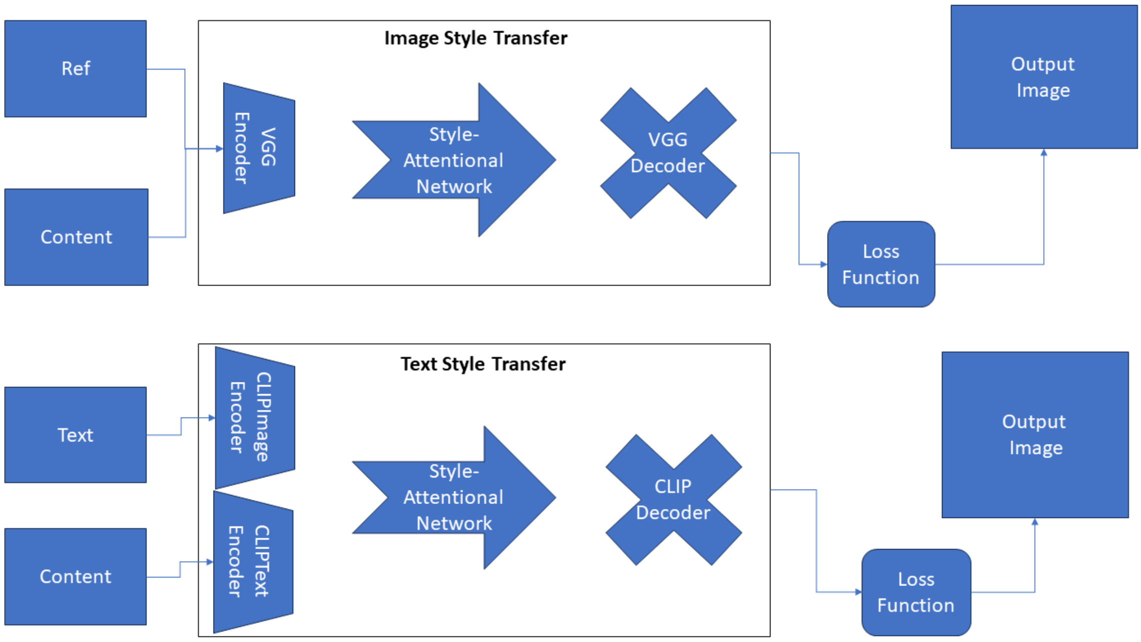

Figure 1. *CLIP takes a different technique in label prediction, as opposed to training a linear classifier and an image feature extractor together. It focuses on predicting the correct pairings within a batch of (image, text) training examples, training both a text and an image encoder simultaneously. 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.*

without much consumption of computational resources. There have been multiple solutions to the challenges such development of a PCA-based knowledge Distillation method to distill lightweight models demonstrating its usability with different architectures [5] hence improving efficiency and suitability for real-time applications. Also, advancements in the Industrial Style Transfer [6] method have shown promising results in creating new visual products with a nice appearance for industrial designers' reference. Industrial Style Transfer involves the application of style transfer techniques to industrial settings, aiming to better visualize elements with unique and aesthetically pleasing characteristics.

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 to 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 is proposed to fully analyze the ability of CLIP for general image style transfer. This will be applied to find the linear or sub-linear correlations between texts and images to demonstrate that CLIP can generally map arbitrary patterns to the target images.

3. **Extensive experiments and analysis on text-driven style transfer.** This will be done to conduct experiments on several datasets to apply our text-driven style transfer and analyze its subjective and objective quality.

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

This section will introduce the roadmap to be used to navigate the whole paper, it will be written chapter by chapter to give the reader a clear map on the paper will be structured.

1. Introduction

1.1 Background. This section will introduce the topic and terms that will be used throughout the paper.

1.2 Objectives and Delimitation. This section will introduce the goal of this paper and what will be covered in this paper.

2. Related Work. This will give an overview of what has been done on the algorithms and methods being used in this paper. This will encompass three categories of style transfer and how it evolved.

3. Proposed Methods

3.1 Analysis of CLIP and/or SigCLIP. This subsection dives into the characteristics of CLIP and/or SigCLIP, exploring their functionalities, strengths, and potential shortcomings.

3.2 Architecture of text-driven style transfer

3.2.1 Overall pipeline. This part breaks down the sequential steps involved in the style transfer process based on textual input

3.2.2 Optimization. It discusses the techniques employed to enhance the efficiency and performance of the text-driven style transfer model.

3.2.3 Training strategy. Details about the strategy used to train the model, including hyperparameters and data augmentation, are explained in this subsection.

3.3 Dataset and Evaluation. This section focuses on the dataset used for training and testing, including its composition and sources. It then discusses the evaluation process, encompassing both objective metrics and subjective evaluations gathered from human assessors.

3.3.1 Dataset: Training, testing, and Types. This subsection provides insights into the datasets used, their division into training and testing sets, and any specific characteristics that make them suitable for the proposed style transfer task.

3.3.2 Evaluation: Objective and Subjective. This will focus on the objective evaluation metrics employed to quantify the model's performance. Additionally, it outlines the subjective evaluation process, where human assessors provide feedback on the output images.

4. Experiment

4.1 Report on the objective evaluation and analysis. This subsection presents the results of the objective evaluation, analyzing the model's performance based on predefined metrics.

4.2 Report on the subjective evaluation and analysis. Here, the findings from the subjective evaluation, which involves human assessments, are reported and analyzed.

4.3 Extension and challenges. This subsection discusses any extensions made to the proposed model during the experiments and outlines the challenges faced in implementing these extensions.

4.4 Failures and Problems(disadvantages of the model). This part highlights any failures, shortcomings, or problems encountered during the experimentation, providing a candid assessment of the model's limitations.

4. Discussion. The discussion section provides an in-depth analysis and interpretation of the experimental results. It explores the implications of the findings, compares them with existing literature, and critically examines the strengths and weaknesses of the proposed approach.

5. Conclusion. The conclusion section summarizes the key contributions of the research, highlights significant findings, and outlines avenues for future work. It serves as a concise wrap-up of the entire thesis.

2 RELATED WORK

In the realm of Computer vision, style transfer has been an important yet interesting field. Style transfer has witnessed significant advancements with an exploration of different techniques [3, 7]. Style transfer involves producing a content image in the style of another image, which allows for adaption to arbitrary new styles via a feed-forward neural network. It entails matching the mean and variance of the content features to those of the style features to perform real-time style transfer without being limited to a predefined range of styles. Style transfer has been under study for a long time with it originating from unrealistic photo output [8] also it is closely related to texture synthesis and transfer [9–11]. There have been multiple approaches that were put into practice to help solve the issue of unrealistic photo output, such approaches include Non-parametric sampling [10, 12] and histogram matching on linear filter responses [13]. These approaches are low-level statical methods and tend to fail to capture the underlying structures. This was followed by Gaty et al. [14] who demonstrated interesting style transfer methods using convolutional layers by matching features of a DNN. As time went by, more improvements were made to [14]. Li and Wand [15] came up with an approach based on the Markov field(MRF) in the deep space to focus on local patterns. Gatys et al. [16] later proposed ways to preserve color, spatial location, and the style of style transfer. This was followed by Ruder et al. [17] where they improved the quality of video style transfer by imposing constraints.

Going deeply on Gatys et al. [14], it is based on a slow optimization process that updates the image while minimizing loss of content. It takes minutes to converge with modern GPUs. A common solution is to replace the optimization process with a feed-forward neural network trained to minimize the objective function [16, 18, 19]. These approaches are three orders of magnitude faster than the optimization alternative since they use a feed-forward style transfer approach. More was done on the feed-forwards style transfer by Wang et al [20]where they enhanced the method by making a multi-resolution architecture. Later on, Ulyanov et al [21] proposed an improved way to improve the quality and diversity of generated samples. Nonetheless, the feed-forward methods were limited since each network was limited to a fixed style. The problem was addressed by Dumoulin et al. [22] who introduced a single network able to encode 32 styles and interpolations. More was done by Chen and Schmidt who introduced a feed-forward method that can transfer arbitrary styles by the use of a style swap layer. Many more problems were addressed such as the loss function to be used, and more effective loss functions were introduced such as MRF [15], Adversarial loss [18], histogram loss [23], CORAL loss [24], MMD loss [25]. Style transfer methods can be divided into three main types:

text-driven style transfer, image-driven style transfer, and attention-based style transfer.

Image-Driven Style Transfer. It is a task in image processing that involves displaying a picture's semantic content in several styles. Convolutional Neural Networks facilitate the creation and alteration of high-quality images by separating and combining the image's content and style. Transferring styles from one image to another is an issue of texture transfer. In texture transfer, the goal is to produce texture from a source image while of course preserving the content of the expected resulting image. There exists a wide range of powerful non-parametric algorithms that can produce realistic photos by resampling pixels of a source texture [10, 11, 26, 27]. There have been multiple approaches for solving image transformation tasks, such approaches include training a feed-forward convolutional neural network which is done in a supervised manner by use of a per-pixel loss function to keep track of the differences between output and the true/expected images. The approach was used by Dong et al. [28] to increase the resolution of many types of images and create better artistic-looking images from photographs while having several orders of magnitude faster than many state-of-the-art techniques. Additionally, the approach was used by Cheng et al [29, 30] for colorization also by Long et al. for segmentation.

Furthermore Johson et al . 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 [31]. 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 [32]. 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 [33]. 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 [34]. 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 [35]. Additionally, efforts were made to handle the computational efficiency and quality of the styles issue as seen in [5]. 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 [36]. 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 [37].

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 novel frameworks. Such discoveries of text-guided image synthesis include encoders for text embedding work for generative models. For instance, Zhang et al. [38] found a way to generate high-resolution photo-realistic images using -Stacked Generative Adversarial Networks (StackGANs) which proposes a two-stage generative adversarial network architecture, StackGAN-v1, for text-to-image synthesis followed by Stage-II GAN that takes the text description as inputs and results from step I, and generates high-resolution images with photo-realistic details. Tu et al. [39] further improved the text-to-image mechanism using the Attentional Generative Adversarial Network (AttnGAN) that allows the use of attention-driven multi-stage modifications for text-to-image generation models. More was done by Watanabe et al. [40] who proposed a novel framework for text-guided image manipulation method that introduced referring image segmentation using Mani-Generative Adversarial Network(GAN).

Leveraging a recent discovery by OpenAI known, CLIP [41], which is a high-performing text-image model that was trained on $400M$ text-image paired images. CLIP can achieve a state-of-the-art performance while connecting text and image domains. More was done on CLIP, Patashnik et al. [42] who introduced an optimization scheme that uses CLIP-based loss to modify input latent vector regarding user text prompt and was named StyleCLIP. However, StyleCLIP has limitations in manipulating images that are only within the trained domain. To overcome the challenge, Rinon et al. [43] proposed a model modification that used text condition only to modulate the trained model into a novel domain without further training images and this was known as StyleGAN-NADA. CLIP's pre-trained text-image embedding model achieves style transfer without a style image, relying solely on a single text condition as introduced by Gihyun et al. [3] who found a way to transfer the texture of text condition conditions to the image regardless of the image's domain which was not put into consideration by other models. 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 [44]. 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.

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. The attention mechanism allows the model employed to focus on the most important parts of an image or the features of an image. Such a mechanism has been proven to be very efficient in many tasks such as image classification [45, 46], captioning [47, 48], and visual answering of questions [46, 49]. Since being proposed, the mechanism of attention has been used mostly in NLP [50, 51]and CV [52, 53]. More advancements were done on attention mechanismto improve its abilities and optimise the use of computational resources, this was achieved with the development of transformers [54–56]. Attention mechnaism were further studied to dicover its ability to scale and it being available in computing heavy models, such studies include [57–60] that adapts the use of attention mechanism.

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 [35]. 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 [37] does not 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 [61] 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, our work will consider the use of attention mechanisms to better capture the most important features during style transfer.

In the course of this paper, we will adopt the Sigmoid loss for Language-Image Pre-training (SigLIP) methodology, as introduced by Xiaohua et al. [62]. Their novel approach involves the implementation of a straightforward pairwise sigmoid loss during image-text pre-training operations. Unlike traditional methods that rely on a global view of pairwise similarities and softmax normalization, SigLIP predominantly focuses on image-text pairs. This unique approach not only enhances performance at smaller batch sizes but also facilitates the scalability of batch sizes, presenting a significant advancement in the field of language-image pre-training.

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 Junsong 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] Shuai Yang, Liming Jiang, Ziwei Liu, and Chen Change Loy. Pastiche master: Exemplar-based high-resolution portrait style transfer. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7683–7692, 2022.

[5] 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.

[6] 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.

[7] 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.

[8] Jan Eric Kyprianidis, John Collomosse, Tinghuai Wang, and Tobias Isenberg. State of the "art": A taxonomy of artistic stylization techniques for images and video. *IEEE Transactions on Visualization and Computer Graphics*, 19(5):866–885, 2013.

[9] Michael Elad and Peyman Milanfar. Style transfer via texture synthesis. *IEEE Transactions on Image Processing*, 26(5):2338–2351, 2017.

[10] Alexei A. Efros and William T. Freeman. Image quilting for texture synthesis and transfer. In *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '01, page 341–346, New York, NY, USA, 2001. Association for Computing Machinery.

[11] A.A. Efros and T.K. Leung. Texture synthesis by non-parametric sampling. In *Proceedings of the Seventh IEEE International Conference on Computer Vision*, volume 2, pages 1033–1038 vol.2, 1999.

[12] Oriel Frigo, Neus Sabater, Julie Delon, and Pierre Hellier. Split and match: Example-based adaptive patch sampling for unsupervised style transfer. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 553–561, 2016.

[13] D.J. Heeger and J.R. Bergen. Pyramid-based texture analysis/synthesis. In *Proceedings., International Conference on Image Processing*, volume 3, pages 648–651 vol.3, 1995.

[14] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2414–2423, 2016.

[15] Chuan Li and Michael Wand. Combining markov random fields and convolutional neural networks for image synthesis. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2479–2486, 2016.

[16] Leon A. Gatys, Alexander S. Ecker, Matthias Bethge, Aaron Hertzmann, and Eli Shechtman. Controlling perceptual factors in neural style transfer. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3730–3738, 2017.

[17] Sun Fengxue, Sun Yanguo, Lan Zhenping, Wang Yanqi, Zhang Nianchao, Wang Yuru, and Li Ping. Image and video style transfer based on transformer. *IEEE Access*, 11:56400–56407, 2023.

[18] Chuan Li and Michael Wand. Precomputed real-time texture synthesis with markovian generative adversarial networks. volume 9907, pages 702–716, 10 2016.

[19] Dmitry Ulyanov, Vadim Lebedev, Andrea Vedaldi, and Victor Lempitsky. Texture networks: Feed-forward synthesis of textures and stylized images. ICML'16, page 1349–1357. JMLR.org, 2016.

[20] Xin Wang, Geoffrey Oxholm, Da Zhang, and Yuan-Fang Wang. Multimodal transfer: A hierarchical deep convolutional neural network for fast artistic style transfer. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7178–7186, 2017.

[21] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4105–4113, 2017.

[22] Vincent Dumoulin, Jonathon Shlens, and Manjunath Kudlur. A learned representation for artistic style. page 9, 04 2017.

[23] Pierre Wilmot, Eric Risser, and Connelly Barnes. Stable and controllable neural texture synthesis and style transfer using histogram losses. 01 2017.

[24] Xingchao Peng and Kate Saenko. Synthetic to real adaptation with generative correlation alignment networks. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1982–1991, 2018.

[25] Yanghao Li, Naiyan Wang, Jiaying Liu, and Xiaodi Hou. Demystifying neural style transfer. pages 2230–2236, 08 2017.

[26] Li-Yi Wei and Marc Levoy. Fast texture synthesis using tree-structured vector quantization. In *Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '00, page 479–488, USA, 2000. ACM Press/Addison-Wesley Publishing Co.

[27] Vivek Kwatra, Arno Schödl, Irfan Essa, Greg Turk, and Aaron Bobick. *Graphcut Textures: Image and Video Synthesis Using Graph Cuts*. Association for Computing Machinery, New York, NY, USA, 1 edition, 2023.

[28] 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)*, pages 626–630, 2016.

[29] Zezhou Cheng, Qingxiong Yang, and Bin Sheng. Deep colorization. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 415–423, 2015.

[30] Richard Zhang, Phillip Isola, and Alexei A. Efros. Colorful image colorization. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 649–666, Cham, 2016. Springer International Publishing.

[31] 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.

[32] 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.

[33] 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.

[34] 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.

[35] 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.

[36] 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.

[37] 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.

[38] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N. Metaxas. Stackgan++: Realistic image synthesis with stacked generative adversarial networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(8):1947–1962, 2019.

[39] T. Xu, P. Zhang, Q. Huang, H. Zhang, Z. Gan, X. Huang, and X. He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In

2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 1316–1324, Los Alamitos, CA, USA, jun 2018. IEEE Computer Society.

[40] Yuto Watanabe, Ren Togo, Keisuke Maeda, Takahiro Ogawa, and Miki Haseyama. Text-guided image manipulation via generative adversarial network with referring image segmentation-based guidance. *IEEE Access*, 11:42534–42545, 2023.

[41] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 2021.

[42] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2065–2074, 2021.

[43] Rinon Gal, Or Patashnik, Haggai Maron, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. Stylegan-nada: Clip-guided domain adaptation of image generators. *ACM Trans. Graph.*, 41(4), jul 2022.

[44] 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.

[45] Tianjun Xiao, Yichong Xu, Kuiyuan Yang, Jiaxing Zhang, Yuxin Peng, and Zheng Zhang. The application of two-level attention models in deep convolutional neural network for fine-grained image classification. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 842–850, 2015.

[46] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2921–2929, 2016.

[47] Kelvin Xu, Jimmy Lei Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. Show, attend and tell: neural image caption generation with visual attention. ICML'15, page 2048–2057. JMLR.org, 2015.

[48] Deema Abdal Hafeth, Stefanos Kollias, and Mubeen Ghafoor. Semantic representations with attention networks for boosting image captioning. *IEEE Access*, 11:40230–40239, 2023.

[49] Qiang Sun and Yanwei Fu. Stacked self-attention networks for visual question answering. In *Proceedings of the 2019 on International Conference on Multimedia Retrieval*, ICMR '19, page 207–211, New York, NY, USA, 2019. Association for Computing Machinery.

[50] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. *XLNet: generalized autoregressive pretraining for language understanding*. Curran Associates Inc., Red Hook, NY, USA, 2019.

[51] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA, 2017. Curran Associates Inc.

[52] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.

[53] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7794–7803, 2018.

[54] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9992–10002, 2021.

[55] Ashish Vaswani, Prajit Ramachandran, Aravind Srinivas, Niki Parmar, Blake Hechtman, and Jonathon Shlens. Scaling local self-attention for parameter efficient visual backbones. 03 2021.

[56] Long Zhao, Zizhao Zhang, Ting Chen, Dimitris Metaxas, and Han Zhang. Improved transformer for high-resolution gans. 06 2021.

[57] Pan Zhang, Bo Zhang, Dong Chen, Lu Yuan, and Fang Wen. Cross-domain correspondence learning for exemplar-based image translation. In *2020 IEEE/CVF Con-*

ference on Computer Vision and Pattern Recognition (CVPR), pages 5142–5152, 2020.

[58] Wentao Jiang, Si Liu, Chen Gao, Jie Cao, Ran He, Jiashi Feng, and Shuicheng Yan. Psgan: Pose and expression robust spatial-aware gan for customizable makeup transfer. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5193–5201, 2020.

[59] Dae Young Park and Kwang Hee Lee. Arbitrary style transfer with style-attentional networks. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5873–5881, 2019.

[60] Songhua Liu, Tianwei Lin, Dongliang He, Fu Li, Meiling Wang, Xin Li, Zhengxing Sun, Qian Li, and Errui Ding. Adaattn: Revisit attention mechanism in arbitrary neural style transfer. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6629–6638, 2021.

[61] 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.

[62] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training, 2023.