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 Zhi-Song Liu
Doctor Jun Xiao

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 Zhi-Song Liu and Doctor Jun Xiao

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

ACKNOWLEDGEMENTS

This thesis includes two examiners: **Prof. Zhi-Song Liu** from Lappeenranta-Lahti University of Technology (LUT) and **Dr. Jun Xiao** from Hong Kong Polytechnic University.

I would also like to acknowledge that while I have utilized the Free version of Grammarly as a tool for identifying and correcting grammatical errors in my written work, I have not yet used Artificial Intelligence(AI) to generate any text. However, I am anticipating the use of AI technologies in my workflow to further improve efficiency and accuracy in the debugging process.

Lappeenranta, February 27, 2024

Nadine Bisanukuli Cyizere

LIST OF ABBREVIATIONS

AI	Artificial Intelligence
AttnGAN	Attentional Generative Adversarial Network
BERT	Bidirectional Encoder Representations from Transformers
CV	Computer Vision
CLIP	Contrastive Language-Image Pre-training
CIDEr	Consensus-based Image Description Evaluation
DNN	Deep Neural Networks
GAN	Generative Adversarial Network
GPT	Generative Pre-trained Transformer
MRF	Markov Field
MSG-Net	Multi-style Generative Network
NLP	Natural Language Processing
LDAST	Language Driven Artistic Style Transfer
PCA	Principal Component Analysis
SigCLIP	Sigmoid loss for Language-Image Pre-training
StackGANs	Stacked Generative Adversarial Networks
Txst	Text-driven Style Transfer
VGG	Visual Geometry Group

CONTENTS

1 INTRODUCTION

1.1 Background

In the field of Computer Vision(CV) and Natural Language Processing(NLP), 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 Specifically, deep learning for style transfer has shown great potential. The main goal of image style transfer is to apply the style of an image usually referred to as style reference Given one content image while preserving the original content. Imagine being able to paint your painting using Pablo Picasso's style without him being there as shown in Figure 1, the transfer of styles is transferred, and a new stylized image(Neural style transfer image) is got. 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. This will leverage the text-image correlations for more advancements in the domain as shown in Figure 2.

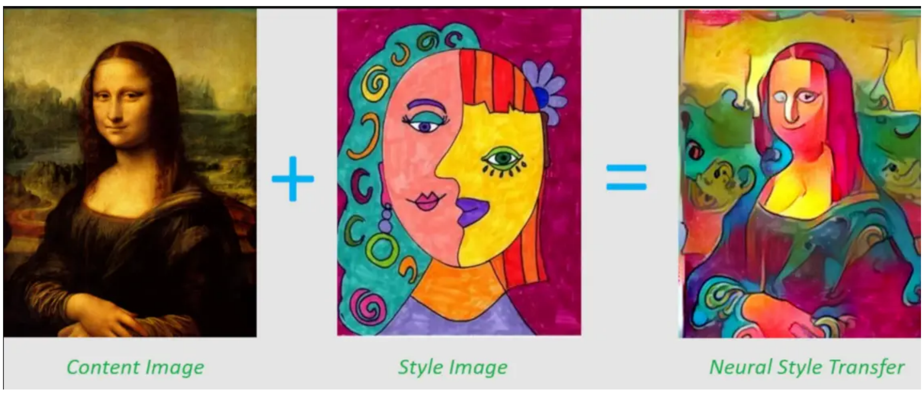

Figure 1. Style Transfer

The diagram in Figure 2 **(a)** illustrates the process of an Image-guided Style Transfer Network. The process starts with the input image(I_c) and the style reference image(I_s). These images are encoded using a VGG encoder to extract features from the input image and style information from the VGG feature maps. The information from the encoder is fed into an attention network that applies attention over features in the encoded image representation to focus on relevant regions for style transfer. followed by a Feature Fusion Layer that Combines the attended features with the style information encoded earlier. Lastly, the output is fed into the VGG decoder that reconstructs the stylized image and gives the final reconstructed stylized image(I_{cs}). This is followed by applying a loss function($\mathcal{L}_{I_c,I_s,I_{cs}}$) which aims to minimize the difference between the Input image, style image, and the Stylized output image. Figure 2(**b**) This part of the figure represents a Text-guided Style Transfer Network. It starts with an input image(I_c) and a text-embedded input(I_t). They are both encoded by CLIP Encoders whose outputs are later fed into an Attentional network. The outputs from the CLIP correlation are the inputs of the neural network which are in turn fed into an attention neural network, that calculates the global feature correlations between texts and images. This is followed by Image Feature Fusion which combines the output image with the style information encoded from the text embeddings. Lastly, the output is reconstructed and decoded using a CLIP decoder to give a styled output image(I_{ct}). As the last step, a loss function($\mathcal{L}_{I_c,I_t,I_{ct}}$) is applied to minimize the difference between the Input image, style image, and the Stylized output image.

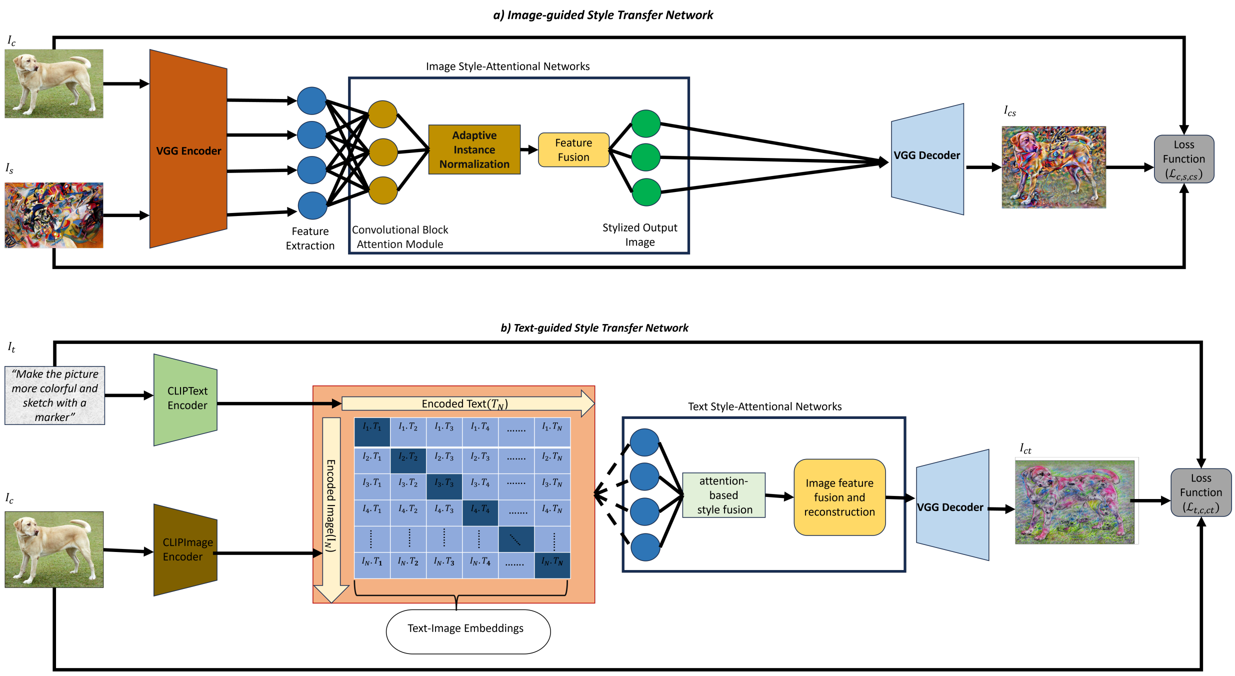

Figure 2. **a)**Image-guided style transfer with content image(I_c) and style image (I_s) which goes through a network and outputs a stylized image (I_{cs}) **b)**Text-guided style transfer with content image(I_c) and reference style text (I_t) which goes through a network and outputs a stylized text-image (I_{ct})

OpenAI's CLIP model was initially used for classification but was potentially used for style transfer. It has demonstrated new and better possibilities in the field of image style transfer. Unlike some traditional methods that require reference-style images [3–6] for example painting own paintings while looking at a reference image as shown in Figure 3. Due to Domain specificity which hinders generalization to arbitrary styles, CLIP overcame these challenges and 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.

Figure 3. Person painting his paint using the view as reference style transfer

Several methods [7–10] have been proposed to better the performance of image style transfer. While significant progress has been made, challenges continue to exist within the realm of image style transfer. These challenges include ensuring content preservation, achieving a visual balance between content and style, and addressing the crucial aspect of computational efficiency. 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. Furthermore, 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 generalizing these models to handle arbitrary styles is an ongoing challenge. 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 without much consumption of computational resources. There have been multiple solutions to the challenges. For instance, [11] proposes a PCA(Principal Component Analysis) -based knowledge distillation method to compress the original model to a lightweight version with fewer parameters, hence it can run in real-time. Also, advancements in the Industrial Style Transfer [12] method have shown promising results in creating new visual products with a nice appearance for industrial designers' reference. It involves applying style transfer techniques to various industrial settings, such as engineering designs [13], automotive design [14], and more. The goal is to enhance the visualization of elements in these settings 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 style similarity between results and texts 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 arbitrarily artistic styles to the target content images. Additionally, there will be an exploration of the new development on SigLIP for text-driven style transfer.

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. To demonstrate the generalization, Additionally, there will be an extension to the other domain-specific text-driven style transfer.

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

- The study focuses on the use of CLIP model [15] 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 is 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 is to 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

The chapter outlines the structure of the thesis, beginning with a review of Related Work, summarizing style transfer algorithms' evolution followed by Proposed Methods to discuss CLIP and/or SigCLIP analysis, text-driven style transfer architecture, dataset details, and evaluation methods. The Experiment section reports objective and subjective evaluations, extensions, challenges, failures, and problems. last but not least The Discussion section critically analyzes results, compares with literature, and assesses strengths and weaknesses. Finally, the Conclusion summarizes contributions, and findings, and suggests future research.

2 Related Work

2.1 CLIP and CLIPSTYLER

CLIP is a method that is used to match image-text pairs to continuously learn embeddings or alignments between image regions and textual concepts. CLIP was recently introduced by Radford et al. [16] who mainly highlighted how to represent images and text correlations. It was trained on over 400 million image-to-text pairs which was guided by contrastive unsupervised loss. Many works have used CLIP for computer vision tasks that require an understanding of text descriptions such as generating or editing image-based natural language conditions [17, 18]. In this thesis, there will be use of CLIP model for the task of style transfer. This model was trained on many images and textual descriptions using a contrastive loss. The goal is to check whether the images and textual descriptions are well correlated.

As shown in Figure 4, CLIP is used to create embeddings, which are numerical representations, after processing the input image and associated text descriptions. It's trained on a vast dataset containing pairs of images and their descriptions. Through training, CLIP learns to group similar concepts from both images and text closely together in a shared space. This training involves teaching CLIP to distinguish between correct pairs (where the image and text match) and incorrect pairs (where they don't). As a result, CLIP gains the ability to understand the meanings behind images and their related text. During use, CLIP can perform various tasks like classifying images or generating new ones, drawing from its understanding of both visual and textual information. Early work used LSTM [19, 20] but new ways were discovered [21]. BERT [22] was one of the crucial works that were constructed under the transformer, which demonstrated domination in the introduced method. The study opts for GPT-2(Generative pre-trained transformer) [23], an auto-regressive language model, considering the training loss term, while some recent methods also employ self-critical sequence training [24] for optimizing the CIDEr(Consensus-based Image Description Evaluation) metric. Similar works employ vision-and-language pre-training to establish a shared latent space for both modalities. For instance, Zhou et al. [25] utilize visual tokens extracted from object detectors in conjunction with BERT, while others like Li et al. [26] and Zhang et al. [27] require object tags for supervision, limiting their applicability to datasets with such annotations. Wang et al. [28] attempt to mitigate the need for supplementary annotations but require extensive pre-training with millions of image-text pairs, resulting in lengthy training times. This exhaustive pre-training aims to compensate for the lack of joint representation of language and vision, a limitation addressed by employing CLIP in the present study.

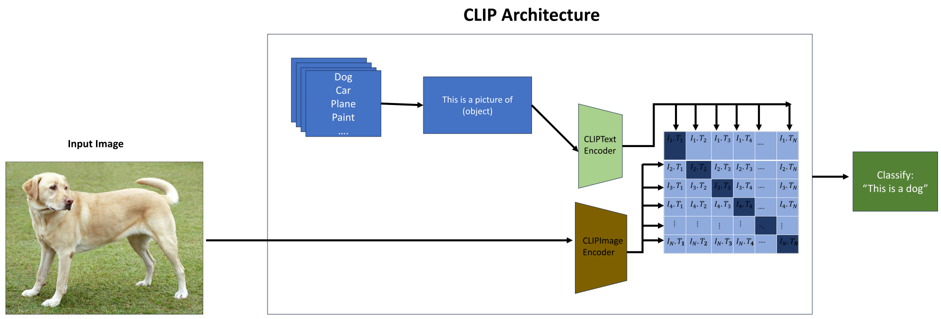

Figure 4. CLIP Architecture

Tentatively, while applying CLIP in this study, there will be an application of CLIPystyler [3] which is a framework that enables style transfer without a style image using a text description. While previous methods for artistic style transfer require a specific reference style image, which may not always be accessible, a new approach called CLIPstyler has been introduced to address this limitation. CLIPstyler utilizes CLIP as shown in Figure 5, an embedding model that maps both images and text into a shared embedding space. This enables the application of a textual prompt to stylize images instead of relying on a reference style image. This approach, known as Language Driven Artistic Style Transfer (LDAST) [29], allows users to generate stylized images based on textual input rather than requiring a specific visual reference. In this study, there will be exploitation of CLIP and CLIPstyler to better carry out style transfer consistently while trying different methods of style transfer.

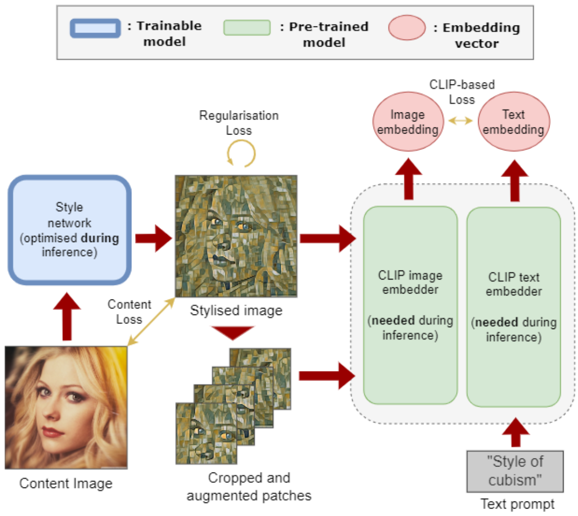

Figure 5. CLIPstyler Architecture

2.2 Style Transfer

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, 16]. Style transfer involves producing a content image by using the style from another image, which allows for adaption to arbitrary new styles via a feed-forward neural network. Executing real-time style transfer without being restricted to a preset range of styles, involves matching the mean and variance of the content features to those of the style features. Style transfer has been under study for a long time with it originating from unrealistic photo output [30] also it is closely related to texture synthesis and transfer [31–33]. 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 [32, 34] and histogram matching on linear filter responses [35]. These approaches are low-level statical methods and tend to fail to capture the underlying structures. This was followed by Gaty et al. [7] who demonstrated interesting style transfer methods using convolutional layers by matching features of pre-trained Deep Neural Networks(DNN).After this research, Li and Wand [36] developed a method to concentrate on local patterns in deep space using the Markov field (MRF). Later, Gatys et al. [37] suggested methods to maintain color, spatial placement, and style during style transfer.

Going deeply on Gatys et al. [7], 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 [37–39]. 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 [40]where they enhanced the method by making a multi-resolution architecture. Later on, Ulyanov et al [41] 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. [42] who introduced a network able to more styles and interpolations. The styles were up to 32 styles. More was done by Chen and Schmidt [43] who introduced a feed-forward method that can transfer arbitrary styles by the use of a style swap layer. The aforementioned approaches can be summarized and categorized into three main tasks: text-driven style transfer, image-driven style transfer, and attention-based style transfer.

2.2.1 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 as shown in Figure 6. 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 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 [32, 33, 44, 45]. 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. Dong et al. [46] proposed 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. Furthermore, Johson et al . introduced the use of the

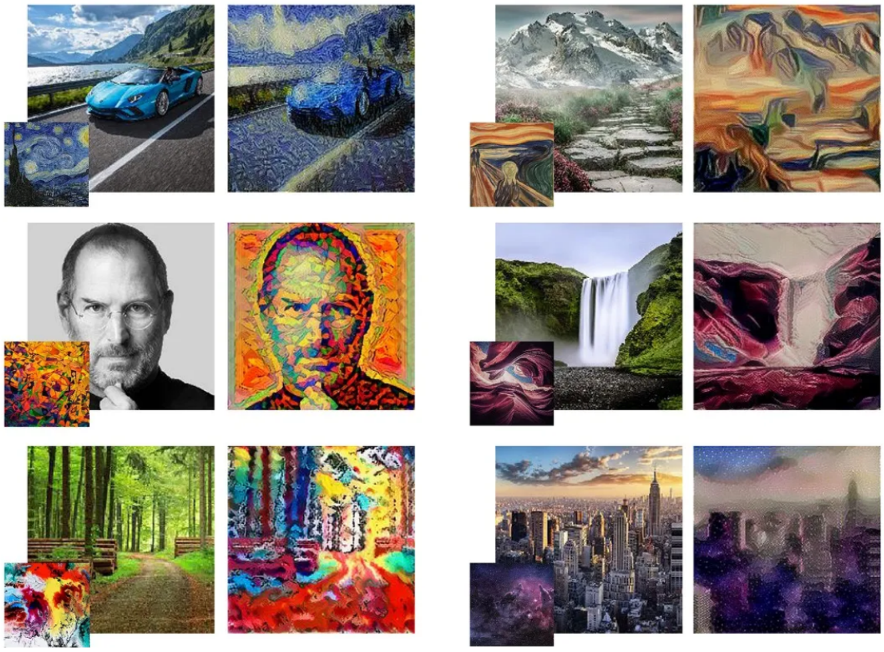

Figure 6. Image driven style transfer

VGG network for style transfer, emphasizing the effectiveness of feature extraction from pre-trained networks in capturing and applying artistic styles across images [9]. Huang et al. [47] showed real-time abilities of arbitrary style transfer through adaptive instance normalization in the AdaIN framework. These 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 [48]. 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 [49]. 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 [50] that incorporated transformers to capture complex style patterns. 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 [51]. More efforts were made to arbitrary style transfer and domain generalization with exact feature distribution matching providing insights into handling diverse styles and applicability of models across different domains [52]. More discoveries were done on Image-driven style transfer, with recent advancements focusing on transferring style from a single image. However, existing methods either suffer from slow processing or struggle to merge multiple styles effectively. Introducing ST-VAE, a Variational AutoEncoder by Liu et al [53], offers a solution by enabling efficient multiple style transfer through nonlinear style projection onto a linear latent space, outperforming other methods in speed, flexibility, and effectiveness, as demonstrated through experiments on the COCO dataset and case studies.

2.2.2 Loss function

To achieve effective style transfer, it's crucial to define an appropriate loss function that guides the optimization process. In this thesis section, Let's delve into the concept of content and style loss using VGG networks. The VGG network is built using small convolutional filters. The architecture of VGG-16, a variant of the VGG network, consists of thirteen convolutional layers and three fully connected layers as shown in Figure 7.

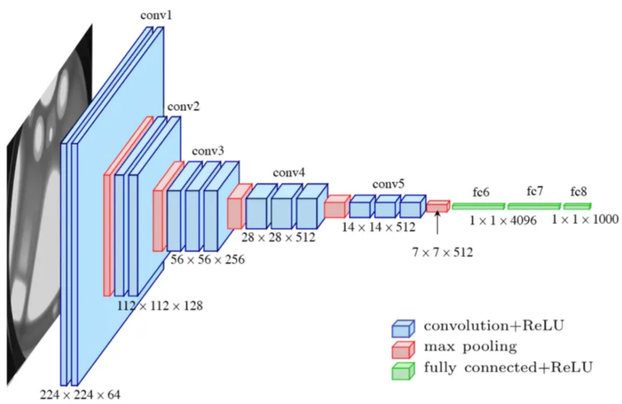

Figure 7. VGG-16 Network Architecture

VGG loss for style transfer

VGG network was initially used for image classification on ImageNet ILSVRC-2014 which gave a pretty low error rate which was 7.3% which was a breakthrough for the use of VGG. It uses five convolutional layers that are stacked accordingly for **feature extraction** which plays a crucial role in finding patterns and other complex features while classifying objects. As shown from Figure 8 [9], diagram consists of mainly two components: *Image transformation* f_W and *Loss network* ϕ that defines other loss functions. The Image transformation part gives a stylized image $\hat{y}$ and from this, there is a calculation of two-loss functions: Content loss and Style loss.

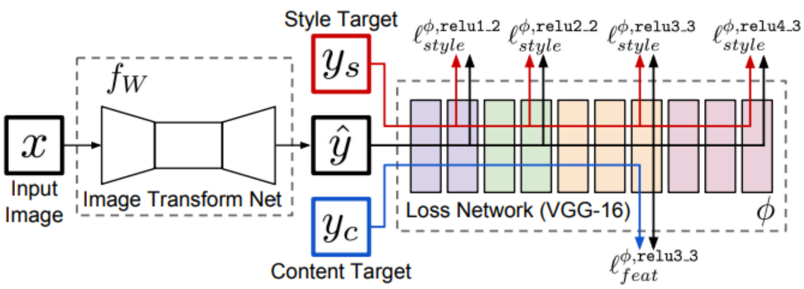

Figure 8. VGG Loss Function Network

Content loss also known as perceptual loss, involves extracting weights from different layers of the VGG16 network. By comparing feature representations from the stylized ($\hat{y}$) and content images(y_c), with Euclidian distance between features as follows define the content loss

$$l_{\text{feat}}^{\phi,j}(i, y) = \frac{1}{C_j H_j W_j} |\phi_j(\hat{y}) - \phi_j(y_c)|_2^2$$

where ($\phi_j(\hat{y})$) and ($\phi_j(y_c)$) are the feature representations of the generated image and the content image, respectively, and ($|| \cdot ||$) denotes the Euclidean norm. Deeper layers of the VGG network focus on general details and patterns, influencing the quality of reconstructed images.

Style Loss prioritize texture information, we employ methods like the Gram Matrix, based on VGG network features. The Gram Matrix captures style by computing inner products of flattened feature maps. The Gram matrix can be defined as $G_j^{\phi}(x)$ to be $C_i \times C_j$ which can be expressed as:

$$G_j^\phi(x)_{c,c'} = \frac{1}{C_j H_j W_j} \sum_{h=1}^{H_j} \sum_{w=1}^{W_j} \phi_j(x)_{h,w,c} \phi_j(x)_{h,w,c'}$$

By comparing Gram Matrices of stylized ($\hat{y}$ and style(y_s) images at different layers by use of Frobenius norm between Gram matrices of the two images, the discrepancy between the images can be evaluated as follows:

$$l_{\text{style}}^{\phi,j}(\hat{y}, y_s) = ||G_j^\phi(\hat{y}) - G_j^\phi(y_s)||_F^2$$

where $G_j^\phi(\hat{y})$ and $G_j^\phi(y_s)$ are the Gram matrices of the generated image and the style reference image, respectively, and $(|| \cdot ||_F^2)$ denotes the Frobenius norm.

By including content and style loss in our optimization process, we can achieve compelling artistic style transfer effects, leveraging the capabilities of VGG networks in a meaningful way.

2.2.3 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. Among these text-guided picture synthesis breakthroughs are encoders for generative models' text embedding tasks. Using Stacked Generative Adversarial Networks (StackGANs), a two-stage generative adversarial network architecture, Zhang et al. [54], for example, discovered a method to achieve higher resolution of images. StackGAN-v1 is used for text-to-image synthesis, and Stage-II GAN uses the text description and step I results to generate high-resolution images with photo-realistic details. Tu et al. [55] 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. [56] who proposed a novel framework for text-guided image manipulation method that introduced referring image segmentation using Mani-Generative Adversarial Network(GAN).

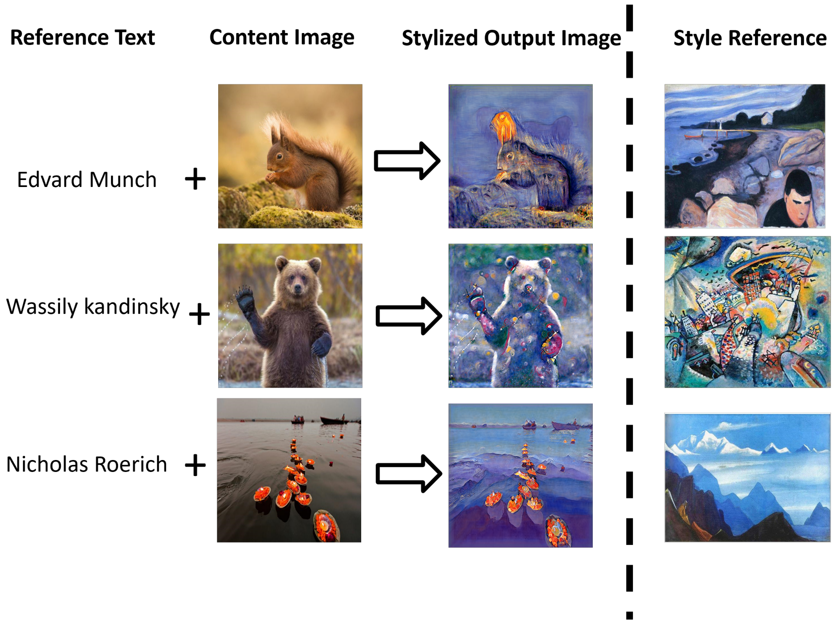

Figure 9. Text-driven Style transfer

Leveraging a recent discovery CLIP [57], 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. [58] 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. [18] 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. Given the ability of linear correction between text and images in the CLIP domain, Gihyun et al. [3] 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, 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. In addition to significant work by Liu et al. [59] in image-driven style transfer, traditional methods often require additional style images, limiting flexibility. They propose Text-driven Image Style Transfer (TxST), leveraging advanced image-text encoders for flexible style transfer without extra images. Through contrastive training and a style attention module, TxST aligns stylization with text descriptions, achieving superior performance and advancing image style transfer.

This paper will adopt the Sigmoid loss for Language-Image Pre-training (SigLIP) methodology, as introduced by Xiaohua et al. [15]. 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.

2.2.4 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 model in use can concentrate on the most significant elements or characteristics of a picture thanks to the attention mechanism. Several tasks, including image classification [60, 61], captioning [62,63], and visual question answering [61,64], have demonstrated the high efficiency of this mechanism. Since being proposed, the mechanism of attention has been used mostly in NLP [65, 66]and CV [67, 68]. More advancements were made to the attention mechanism to improve its abilities and optimize the use of computational resources, this was achieved with the development of transformers [69–71]. Attention mechanisms were further studied to determine their ability to scale and it being availability in computing heavy models, such studies include [8, 72–74] that adapt the use of attention mechanism.

StyTr2 [50] is one of the pioneers works that utilize an attention mechanism to capture key styles for style transfer. 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 [52] does not explicitly mention attention-based mechanisms, the use of transformers suggests the use of attention-based features. It showcases the ongoing development of methodologies in attention-based styles. Furthermore, [75] 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. Given the findings of these attention-based works, our work will consider the use of attention mechanisms to better capture the most important features during style transfer.

After reviewing existing research, it's clear that progress has been made in computer vision's style transfer. This paper aims to use recent developments, such as SigLIP and CLIP, using a novel strategy. This paper will determine how CLIP can preserve the original content while transferring text-driven styles onto images. In this paper, there will be extensive tests to evaluate the objective and subjective quality of the technique on various datasets, there will also be an exploration of multiple style transfers to fuse multiple artists' styles to the target images for stylization.

3 PROPOSED METHODS

3.1 Analysis of CLIP and/or SigCLIP

3.2 Architecture of text-driven style transfer

3.2.1 Overall pipeline

3.2.2 Optimization

3.2.3 Training strategy

3.3 Dataset and evaluation

3.3.1 Dataset: Training, testing and types

3.3.2 Evaluation: Objective and subjective

3.4 Experiment

3.4.1 Report on the objective evaluation and analysis

3.4.2 Report on the subjective evaluation and analysis

3.4.3 Extension and challenges

3.4.4 Failures and Problems(disadvantages of the model)

3.5 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] F. Luan, S. Paris, E. Shechtman, and K. Bala. Deep photo style transfer. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6997–7005, Los Alamitos, CA, USA, jul 2017. IEEE Computer Society.

[5] Yanghao Li, Naiyan Wang, Jiaying Liu, and Xiaodi Hou. Demystifying neural style transfer. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, IJCAI'17, page 2230–2236. AAAI Press, 2017.

[6] Haozhi Huang, Hao Wang, Wenhan Luo, Lin Ma, Wenhao Jiang, Xiaolong Zhu, Zhifeng Li, and Wei Liu. Real-time neural style transfer for videos. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7044–7052, 2017.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[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] Nicholas Kolkin, Jason Salavon, and Gregory Shakhnarovich. Style transfer by relaxed optimal transport and self-similarity. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10043–10052, 2019.

[14] Jie An, Haoyi Xiong, Jiebo Luo, Jun Huan, and Jinwen Ma. Fast universal style transfer for artistic and photorealistic rendering. *CoRR*, abs/1907.03118, 2019.

[15] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer. Sigmoid loss for language image pre-training. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11941–11952, Los Alamitos, CA, USA, oct 2023. IEEE Computer Society.

[16] 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.

[17] Tugrul Adiguzel, Ahmet Akbulut, and A. Egemen Yilmaz. Painting with words: one step beyond 'paint by numbers'. In *2009 Fifth International Conference on Soft Computing, Computing with Words and Perceptions in System Analysis, Decision and Control*, pages 1–4, 2009.

[18] 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.

[19] X. Chen, L. Ma, W. Jiang, J. Yao, and W. Liu. Regularizing rnns for caption generation by reconstructing the past with the present. In *2018 IEEE/CVF Conference on*

Computer Vision and Pattern Recognition (CVPR), pages 7995–8003, Los Alamitos, CA, USA, jun 2018. IEEE Computer Society.

[20] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3156–3164, 2015.

[21] Simao Herdade, Armin Kappeler, Kofi Boakye, and Joao Soares. *Image captioning: transforming objects into words*. Curran Associates Inc., Red Hook, NY, USA, 2019.

[22] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*, 2019.

[23] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

[24] J. Gao, S. Wang, S. Wang, S. Ma, and W. Gao. Self-critical n-step training for image captioning. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6293–6301, Los Alamitos, CA, USA, jun 2019. IEEE Computer Society.

[25] Luowei Zhou and. Unified vision-language pre-training for image captioning and VQA. *CoRR*, abs/1909.11059, 2019.

[26] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. Oscar: Object-semantics aligned pre-training for vision-language tasks, 2020.

[27] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5575–5584, 2021.

[28] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. Simvlm: Simple visual language model pretraining with weak supervision, 2022.

[29] Tsu-Jui Fu, Xin Eric Wang, and William Yang Wang. Language-driven artistic style transfer. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVI*, page 717–734, Berlin, Heidelberg, 2022. Springer-Verlag.

[30] 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.

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

[32] 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.

[33] 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.

[34] 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.

[35] 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.

[36] 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.

[37] 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.

[38] Chuan Li and Michael Wand. Precomputed real-time texture synthesis with markovian generative adversarial networks. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*, pages 702–716. Springer, 2016.

[39] Dmitry Ulyanov, Vadim Lebedev, Andrea Vedaldi, and Victor Lempitsky. Texture networks: feed-forward synthesis of textures and stylized images. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, page 1349–1357. JMLR.org, 2016.

[40] 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.

[41] 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.

[42] Vincent Dumoulin, Jonathon Shlens, and Manjunath Kudlur. A learned representation for artistic style. In *International Conference on Learning Representations*, 2017.

[43] Tian Qi Chen and Mark Schmidt. Fast patch-based style transfer of arbitrary style. *arXiv e-prints*, page arXiv:1612.04337, December 2016.

[44] 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.

[45] Vivek Kwatra, Arno Schödl, Irfan Essa, Greg Turk, and Aaron F. Bobick. Graphcut textures: image and video synthesis using graph cuts. *ACM SIGGRAPH 2003 Papers*, 2003.

[46] 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.

[47] 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.

[48] 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.

[49] Yifang Men, Yuan Yao, Miaomiao Cui, Zhouhui Lian, Xuansong Xie, and Xian-Sheng Hua. Unpaired cartoon image synthesis via gated cycle mapping. In *2022*

IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 3491–3500, 2022.

[50] 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.

[51] 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.

[52] 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.

[53] Zhi-Song Liu, Vicky Kalogeiton, and Marie-Paule Cani. Multiple style transfer via variational autoencoder. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 2413–2417, 2021.

[54] 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.

[55] 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.

[56] 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.

[57] 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.

[58] 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.

[59] 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.

[60] 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.

[61] 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.

[62] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2048–2057, Lille, France, 07–09 Jul 2015. PMLR.

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

[64] 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.

[65] 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.

[66] 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.

[67] 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, 2020.

[68] 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.

[69] 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.

[70] A. Vaswani, P. Ramachandran, A. Srinivas, N. Parmar, B. Hechtman, and J. Shlens. Scaling local self-attention for parameter efficient visual backbones. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12889–12899, Los Alamitos, CA, USA, jun 2021. IEEE Computer Society.

[71] Long Zhao, Zizhao Zhang, Ting Chen, Dimitris N. Metaxas, and Han Zhang. Improved transformer for high-resolution gans. In *Neural Information Processing Systems*, 2021.

[72] Pan Zhang, Bo Zhang, Dong Chen, Lu Yuan, and Fang Wen. Cross-domain correspondence learning for exemplar-based image translation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5142–5152, 2020.

[73] 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.

[74] 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.

[75] 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.