Rapid Content-aware Image Style Transfer using Attention Map Guidance and Diffusion Model

by

Jungmin Hwang

Thesis submitted to the University of Ottawa in partial fulfillment of the requirements for the Master of Applied Science in Electrical and Computer Engineering

School of Electrical Engineering and Computer Science

Engineering

University of Ottawa

Abstract

Despite the considerable interest and successful progress in image generation and editing applications using diffusion models, a critical challenge persists in balancing style transfer and content preservation. Effectively addressing this challenge is crucial for enhancing the overall success and usability of image editing tools that leverage diffusion methodologies, especially text-driven or image-driven.

One approach to tackle this challenge is to determine areas with high level of content in a given image. It involves preserving certain areas, containing more information than the rest of the image. To address this, we propose a method that anchors representative points in these areas for both source image and generated image. Self-attention mechanism intentionally selects queries and produces features at these anchor points. Then we employ contrastive learning in a self-supervised manner. This approach enables our method to generate an image that maintains the important content in the given source image while transferring the style. Our proposed method eliminates the need for additional fine-tuning or auxiliary networks.

Our method uses conventional diffusion model, but without fine running for content preservation. Normally fine tunning is required additional network therefore ours results speeding up the inference process compared to other diffusion methods. Our experiments showcase the superior performance of our approach, particularly in preserving image content during editing, along with a notable superiority when compared to both other diffusions models and GAN-based models.

Acknowledgements

First, I would like to appreciate Jesus for granting me wisdom and guiding me throughout my studies. Additionally, I am deeply thankful to Dr. WonSook Lee for her various support during my master's program. Her insightful feedback and valuable suggestions have greatly advanced my research. During difficult times, she provided significant help and motivation as advisor. Lastly, I want to thank my wife, Ian, Ethan, and Cappu for their encouragement and support.

Table of Contents

List of Figure

List of Tables

Abbreviations

GAN	Generative Adversarial Network
CLIP	Contrastive Language Image Pre-Training
QS	Query-Selection
Attn	Attention Map
ViT	Vision Transformer
CUT	Contrastive Unpaired image-to-image Translation
CNN	Convolution Neural Network
RELU	Rectified Linear Unit
AdaIN	Adaptive Instance Normalization
VQGAN	Vector Quantized Generative Adversarial Network
DDPM	Denoising Diffusion Probability Model
DDIM	Diffusion Implicit Model

Chapter 1 Introduction

1.1 Motivation and Objective

Image generation and style transfer have been very popular and attractive subjects. Style transfer involves applying a specific artistic style to an original image. For instance, consider a photograph of a dog that a user wishes to transform into a painting in the style of Van Gogh. Here, the dog photo serves as the content image, while the brush characteristics of Van Gogh's art represent the style. The Van Gogh painting can be provided as input either in the form of an image or a text. In this scenario, the photo of the dog is the content image, and the Van Gogh painting serves as the style image.

Among numerous methods developed for these, Generative Adversarial Network (GAN) and Diffusion have been most successful while Convolution Neural Network (CNN) based style transfer model is not. Thus, style transfer method interpolates stylized image's features extracted from CNN to content image's features from the model to follow the artistic characteristics. However, the method takes a long time by using two times CNN models, comparing to GAN and diffusion models.

With the advent of high-quality face generation, GAN models have progressed from simple face generation to image manipulation and editing, including style transfer. GAN inversion [1] is a very much used technique for image manipulation when an image is given as an input. The combination of GAN inversion models with Contrastive Language Image Pre-Training (CLIP) [2] suggests a powerful method for manipulating images with text prompts and one of popular uses is style transfer. CLIP is a method making connection using probability between text captions and images. However, challenges arise, especially in handling diverse images with unexpected artifacts.

Existing GAN inversion methods [3] are known to encounter difficulties in reconstruction, particularly for images with high variance.

The emergence of diffusion models has transformed the landscape, drawing significant attention away from GANs. In particular, Stable Diffusion [4] has become highly popular for utilizing physics inspired diffusion models in image generation. Diffusion models have demonstrated success in image generation, surpassing other generative models such as GAN and Variational Autoencoder (VAE) [5]. Conditional diffusion models make interesting applications as they use constrains such an image input or text input. Unconditional diffusion models generate samples without specific conditioning information, making images with various attributes. However, conditional diffusion models for image-to-image style transfer require paired datasets, and unconditional diffusion models face challenges in maintaining content due to the stochastic nature of reverse sampling.

To address these issues, Kim et al. propose DiffusionCLIP [6], a CLIP-guided image manipulation method in diffusion models. The method involves converting an input image to latent noises through forward diffusion and inverting them back using reverse diffusion, fine-tuning the process with a CLIP loss on the given text prompts. DiffusionCLIP addresses the limitations in the GAN inversion methods, showcasing effectiveness in various scenarios and outperforming existing baselines. However, fine-tuning requires a significant amount of time to generate stylized images, and it also demands moderate effort for an unseen domain style transfer.

In another approach, DiffuseIT [7] introduces semantic contrastive loss in Vision Transformer (ViT) [8] for zero-shot style transfer while preserving semantic content. The process is similar to DiffusionCLIP. DiffuseIT separately processes semantic and structural information of

the image using the pretrained DINO-ViT [9] with keys extracted from Multi-head Self-Attention (MSA) layers and Class token taken from the last layer under a self-supervised manner.

ZeCon [10], differently, introduces a patch-wise contrastive loss calculated using keys and a query selected on input and generated images for zero-shot style transfer while preserving semantic content, without additional training. ZeCon achieves effective content preservation and provides more accurate texture modification compared to DiffusionCLIP and DiffuseIT. However, ZeCon is based on Contrastive Unpaired image-to-image Translation (CUT) [11], which involves randomly extracting patches from images, not considering the important parts of content images.

By analyzing the aforementioned methods, we aim to achieve the following items. Main aim is that we enhance the quality of image style transfer with the content is well-kept without destroying details. Meanwhile we preserve the content of an image with lighter and simpler model just with one image input (content image) plus a text prompt (for style information) instead of using two image inputs (content and style images). Note that two image input models require twice processing. We also direct our method to generate stylized images rapidly by optimizing hyper parameters on gradient loss under reverse process.

As a summary, in pursuit of a content-aware style transfer model with text prompt, we have modified the ZeCon model by employing a Query-Selection (QS) module and contrastive loss under self-supervised learning. This approach aims to address the challenges associated with random feature selection and limited receptive fields in ZeCon. Our plan involves evaluating the proposed model both using a user study and CLIP score, demonstrating its superiority over the aforementioned models with their limitations.

1.2 Proposed Pipeline

The recent model, ZeCon, excels at preserving semantic content. However, relying on CUT loss, it may generate stylized images without focusing on content awareness. ZeCon is a pretrained diffusion model inherently embeds spatial information, facilitating content preservation through patch-wise contrastive loss. This loss is computed with a query and positive keys randomly selected from the source image and negative keys and an anchor from the generated images. Despite not requiring additional training, ZeCon struggles to precisely match the contents between source and output images due to the randomness in the anchor and keys.

In contrast, our proposed model leverages an attention map to deliberately select keys from source and queries as anchors in generated images. Note that we use several anchors rather than an anchor. The contrastive loss is then specifically computed on salient features containing more domain-specific information in the source image. Unlike a simple random strategy of ZeCon, the QS-attn module involves comparing a specified query with keys and selecting the most significant query as depicted in the orange box of QS-module in Figure 1-1.

In the QS-module, feature significance is measured by calculating the distribution entropy of an attention matrix using keys and queries from the encoder. The QS-Attn matrix is updated by retaining only features with smaller entropy values. Our proposed model aims to quantitatively assess the significance of each anchor feature, resulting in more effectively preserving content compared to ZeCon. Notably, we depart from the typical attention approach in diffusion model but avoid the use of distinct projection heads for query, key, and value. As a result, no additional model parameters are introduced into the diffusion model, making the model running rapid.

In the pipeline of our whole network, we follow the conventional diffusion method. The generation process starts with a source image, and the image is iteratively updated through a series

of diffusion steps. In each step, the image is perturbed with distributed gaussian noise. Following the forward process, the pretrained model can be used in a reverse process during the generation phase. Starting from a fully perturbed image, the model gradually reduces the noise through a series of steps, generating a realistic image. In the diffusion reverse process, our proposed method computes patch-wise contrastive loss by intentionally selecting a significant anchor in the QS-module and is supplemented with CLIP loss. Ultimately, our model generates more domain-specific patch information for image style transfer.

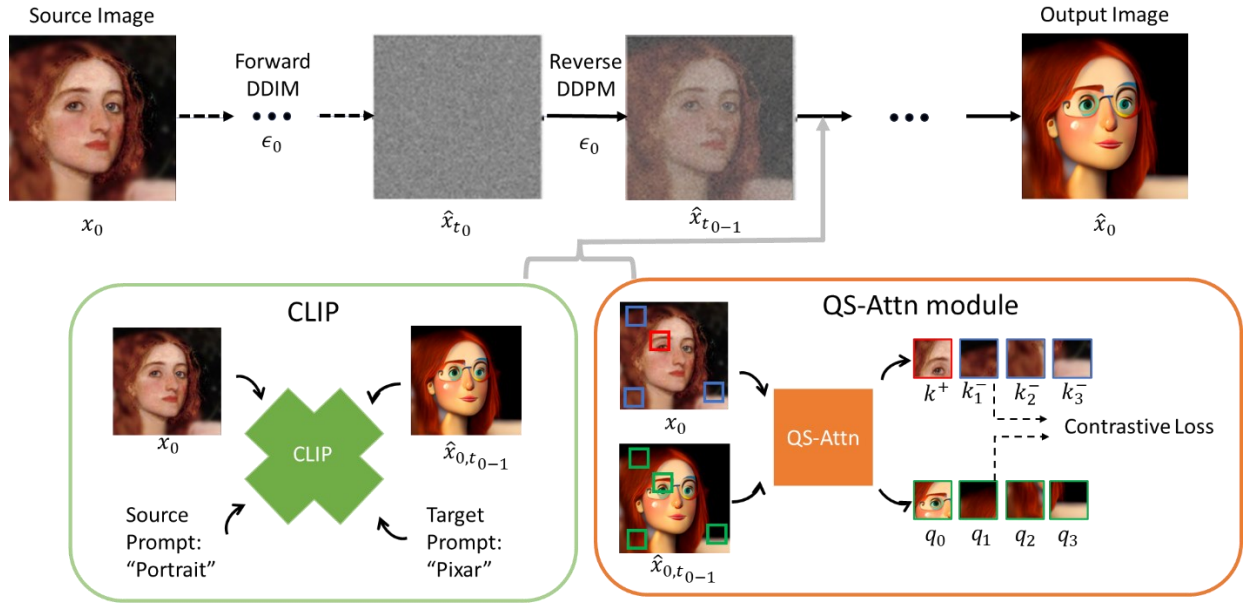

Figure 1-1 Conceptual process of Query-Selected Attention map in Diffusion model's reverse process. The example shows the style transfer from unseen domain to Pixar style.

1.3 Main Contributions

We present a novel architecture based on the diffusion model, coupled with CLIP and attention models, for achieving style transfer in a self-supervised manner.

The achievements of our study are summarized as follows:

- Our proposed model, guided by a text prompt, demonstrates the capability to generate diverse synthetic images while preserving the contents of the source images.
- Achieving high-quality images is attained without the need for additional training or fine-tuning as ZeCon. Our proposed method is similar to ZeCon's architecture but incorporates attention mechanisms, resulting in superior performance compared to ZeCon.
- The proposed model supports zero-shot image generation, facilitating rapid interference for user-based tasks.
- We experiment in generating synthesized images for various image editing purposes, such as modifying facial attributes and changing identities of animals.
- Two diffusion methods are employed in our approach to produce high-quality images.

Our technical innovation is highlighted as follows:

- It tackles the challenges of style transfer by integrating contrastive loss with an attention map into the diffusion model. This integration aims to better preserve the contents of source images compared to State of the Art models.

Our results demonstrate the superior performance compared to other diffusion models and GAN and it exhibits faster inference speed compared to the other diffusion methods.

1.4 Thesis Organization

The following sections of the thesis are structured as outlined below:

Chapter 2 provides a comprehensive introduction to fundamental concepts of Neural Network, including other network models as well as self-supervised learning. We introduce style transfer and other models such as GAN and diffusion. We explain GAN based method with CLIP, and Diffusion based method with CLIP for style transfer.

Chapter 3 explains our methodology, which leverages advanced contrastive learning techniques for the creation of stylized images while preserving the content of source images. Our architectural design incorporates elements such as Global, Local, and Local & Global attention map under diffusion process with other inventive techniques such CLIP Losses.

Chapter 4 explores the experimental phase of our research, centering on a variety of experiments conducted to assess the qualitative and quantitative performance on our proposed methodology with comparison of other models, demonstrating how our architecture enhances the quality on style transfer. Through meticulous experimentation, our objective is to substantiate the capabilities and advantages inherent in our approach to style transfer.

Chapter 5 concludes and discusses future works.

Chapter 2 Literature Review

2.1 Neural Network and Deep Learning

A neural network [13] is a computational model that computes functions from input nodes to output nodes. The network is structured as a directed acyclic graph, where nodes represent computations, and edges are parameterized with weights as Figure 2-1 [14]. Each node contains a variable resulting from a function computation or is fixed externally in the case of input nodes.

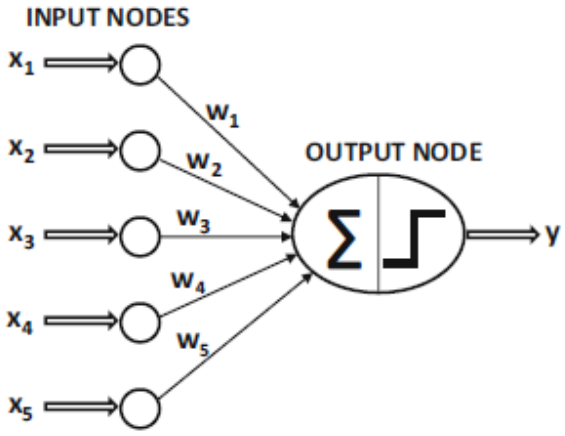

Figure 2-1 The fundamental structure of the Neural Network excluding the bias. Reprinted from [14].

The functions computed at individual nodes are influenced by the weights of incoming edges and the variables in the nodes at the tails of these edges. The overall function computed by the network is the result of summation function computations at individual nodes. By adjusting the weights on the edges, one can approximate almost any function from the input to the output.

In a single-layer neural network, a set of inputs is directly connected to one or more output nodes. The output nodes compute a function of their inputs and the weights on the edges. In multi-

layer neural networks, neurons are organized in layers, with input and output layers separated by hidden layers. The network learns to adjust the weights during training to improve its ability to perform specific tasks. Thus, the primary goal of a neural network is to learn a function that maps one or more inputs to one or more outputs using training examples. This construction is achieved by adjusting the weights on the edges of the network in a way that the computations from inputs to outputs align with the observed data. The process of setting the edge weights in a data-driven manner is at the core of all neural network learning, and this iterative adjustment of weights is commonly referred to as training. During training, the network learns to generalize from the provided examples, enabling it to make accurate predictions or perform desired tasks on new or unseen data.

2.1.1 **Convolutional Neural Networks (CNNs)**

Each layer in Convolutional Neural Networks (CNNs) [15] are structured in a spatial grid pattern, where is crucial as each feature value is derived from a small local spatial region in the hidden layer. Maintaining these spatial relationships is essential because operations like convolution and the transition to the next hidden layer heavily rely on these relationships. Each layer in a convolutional network is represented as a 3-dimensional grid structure, consisting of height, width, and depth. The term "depth" at the layer level refers to the number of channels in that layer, such as the primary color channels in an input image or the feature maps in hidden layers.

The functioning of a convolutional neural network closely resembles that of a traditional feed-forward neural network [16], with the distinction that operations in its layers are spatially organized, featuring sparse and carefully designed connections between layers. Common layer types in a CNN include convolutional layers, pooling layers, and Rectified Linear Unit (ReLU)

activation layers [17]. Typically, a final set of layers in a CNN is fully connected, mapping in an application-specific manner to a set of output nodes. The convolutional neural network structure involves interleaving these different layer types in a specific sequence.

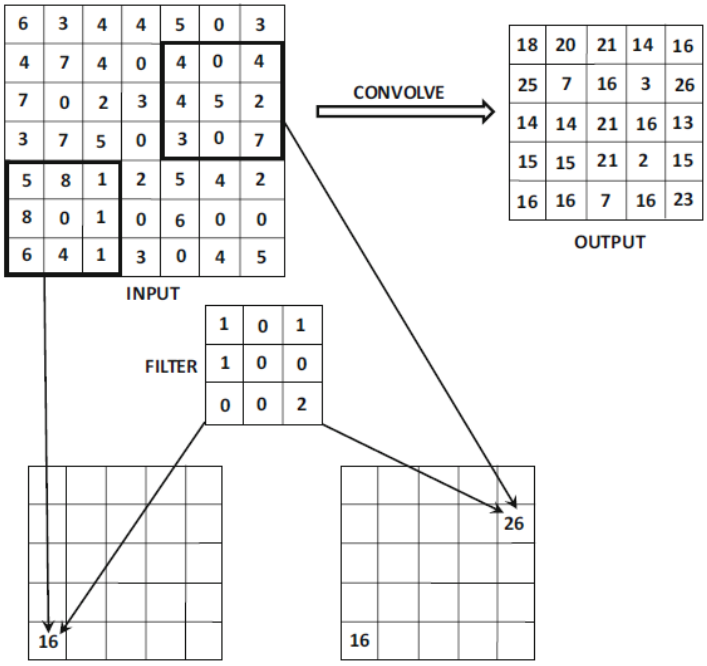

Figure 2-2 An instance of a convolution involving a 7x1x1 input and a 3x1x1 filter with a stride of 1. To maintain simplicity, a depth of 1 is selected for both the filter and input. If the depth were greater than 1, the contributions from each input feature map. Reprinted from [14].

As shown in Figure 2-2 [14], we illustrate an example of an input layer and a filter with a depth of 1 for simplicity, which is common in the case of grayscale images with a single color channel. It's important to note that the depth of a layer must exactly match that of its filter or kernel. In the general case, the contributions of the dot products over all the feature maps in the corresponding grid region of a particular layer need to be aggregated to create a single output feature value in the next layer. Figure 2-2 displays two specific instances of convolution operations, where a layer of size (7x7x1) interacts with a (3x3x1) filter in the bottom row. Additionally, the

entire feature map of the next layer is depicted on the upper right-hand side of Figure 2-2. The examples show two convolution operations, resulting in output values of 16 and 26, respectively.

2.1.2 **U-Net**

The U-Net architecture [18] is a convolutional neural network specifically designed for semantic image segmentation tasks, where the objective is to classify each pixel in an image into specific classes. The key characteristic of the U-Net is its U-shaped structure in Figure 2-3 [18], comprising contracting path bottleneck, and expansive path. Contracting Path as Encoder captures context and reduces spatial resolution. It consists of multiple convolutional and max-pooling layers, progressively diminishing the size of the input image. Following the contracting path, Bottleneck captures high-level features. Expansive Path as Decoder reconstructs the segmentation map from the high-level features obtained in the contracting path. It involves up-sampling the feature map and concatenating it with the corresponding feature map from the contracting path. The incorporation of skip connections between the contracting and expansive paths is a notable feature. These connections aid in preserving spatial information and facilitating precise localization during segmentation. The final layer typically employs a softmax activation function to generate pixel-wise probability maps for each class. U-Net has demonstrated efficacy in tasks such as medical image segmentation, where accurate delineation of structures is essential. Its architecture allows for capturing both local and global context, and the skip connections address challenges like the vanishing gradient problem. The success of U-Net has led to various modifications and adaptations for computer vision. Especially, this U-Net's structure is evolved diffusion's main architecture.

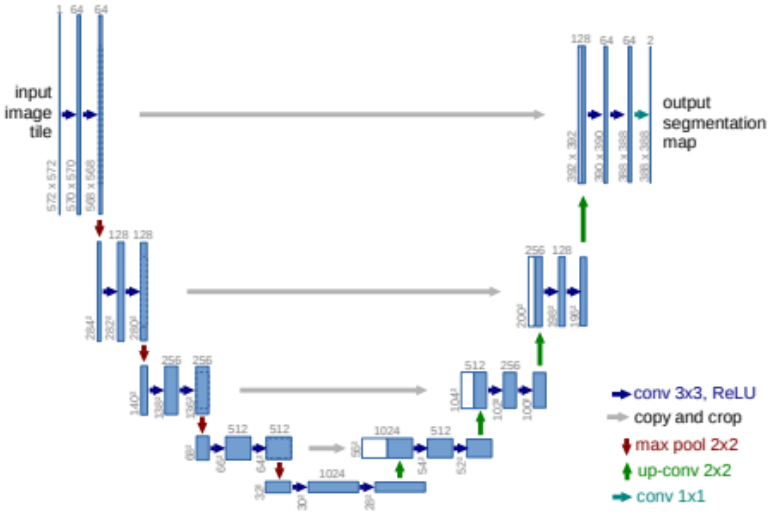

Figure 2-3 The U-Net architecture is illustrated with a focus on a 32x32 pixel input at the lowest resolution. Each blue box represents a multi-channel feature map, with the number of channels indicated on top. The x-y size is specified at the lower left corner of each box. White boxes represent duplicated feature maps. Arrows indicate the various operations within the network. Reprinted from [18].

2.2 Self-supervised Learning

Self-supervised learning [19] is a machine learning paradigm where a model learns from the inherent structure or information present in the input data itself, without the need for explicit external labels or annotations. In traditional supervised learning, a model is trained on a labeled dataset, where each input is paired with a corresponding output label. However, in self-supervised learning, the model generates its own labels or supervisory signals from the input data.

The process involves creating a pretext task, which is a task designed to generate artificial labels or annotations from the data. The model is then trained to solve this pretext task, and the knowledge gained during this process is later transferred to the actual target task of interest. Self-supervised learning is particularly useful in scenarios where obtaining labeled data is challenging or expensive. There are common techniques such as autoencoders, contrastive learning, and generative models.

Autoencoders [20] are trained to encode input data into a lower-dimensional representation and then decode it back to the original input. The objective is to minimize the reconstruction error.

Contrastive Learning model [21] learns to differentiate between positive and negative pairs of data instances. Positive pairs are different views of the same data, while negative pairs are instances from different data points. This helps the model learn representations that capture meaningful features.

Generative Models [22] is training models to generate parts of the input data or predict missing portions, such as predicting the next word in a sentence or filling in missing pixels in an image.

Self-supervised learning has shown success in various domains, including natural language processing, computer vision, and speech recognition. It leverages the abundance of unlabeled data

to pre-train models, which can then be fine-tuned on smaller labeled datasets for specific tasks, often leading to improved performance compared to training from scratch. Our solution utilizes patch-wise contrastive learning method as shown in Figure 2-4 [23].

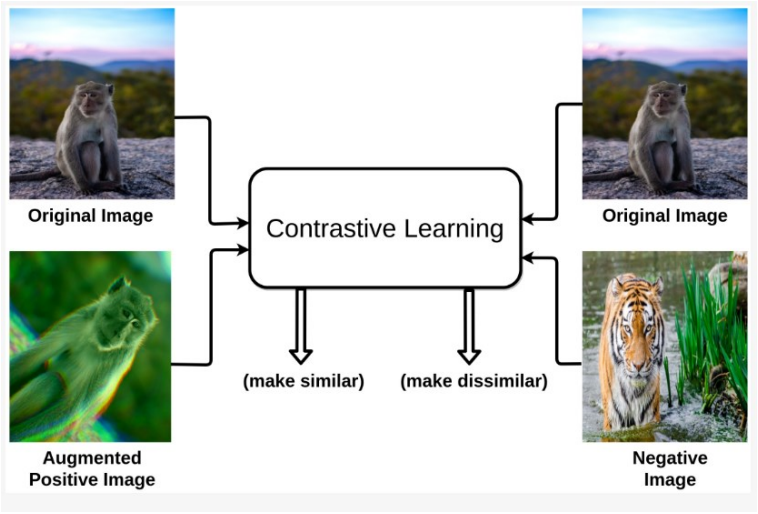

Figure 2-4 Basic intuition behind contrastive learning paradigm: push original and augmented images closer and push original and negative images away. Reprinted from [23].

2.2.1 **Contrastive Language-Image Pre-training**

CLIP [2] is a powerful deep learning model developed by OpenAI. CLIP is designed to understand images and text in a unified manner. Unlike traditional models that are often specialized for either image or text tasks, CLIP is trained to learn a shared representation for both modalities as shown in Figure 2-5 [2]. CLIP is trained using a contrastive learning approach. It learns to associate similar representations for images and their corresponding textual descriptions while maintaining a distinct representation for dissimilar pairs. This enables the model to understand the relationships between different images and their associated text. CLIP is capable of zero-shot learning, meaning it can perform tasks for which it has not been explicitly trained. This is achieved by leveraging the shared representation learned during pre-training. For example, CLIP can be

used for image classification or object detection by providing textual prompts without needing specific labeled data for those tasks. CLIP generates embeddings for both images and text in such a way that semantically similar content is placed close together in the embedding space. This allows for easy retrieval of relevant information across modalities. CLIP is pre-trained on a massive dataset that includes images and associated text from the internet. This extensive pre-training enables the model to learn rich and generalizable representations.

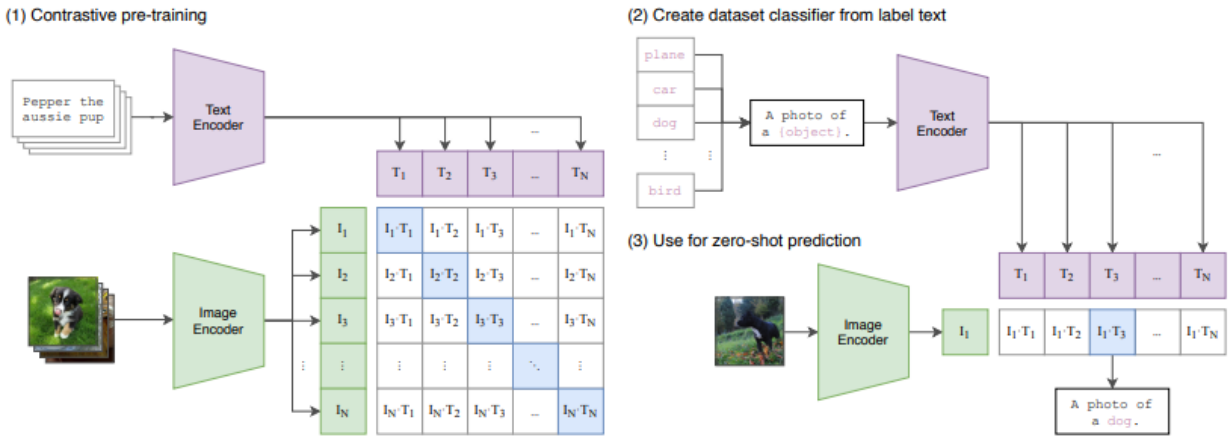

Figure 2-5 CLIP involves the simultaneous training of an image encoder and a text encoder in the CLIP model. It is designed to predict the correct pairings of (image, text) examples within a batch during training. At test time, the text encoder learned during training is employed to create a zero-shot linear classifier by embedding names or descriptions of classes from the target dataset. This innovative approach enhances the model's ability to generalize across different tasks and domains. Reprinted from [2].

To perform style transfer using text prompts with CLIP, our proposed method employs the Vision Transformer (ViT) model, which is a pre-trained CLIP model based on the ImageNet dataset.

2.3 Style Transfer

The process of style transfer involves generating images by identifying an image that aligns with both the content features of the original photograph and the stylistic features of a particular piece of art using CNN as illustrated in Figure 2-6 [24]. While maintaining the overall layout of the original photograph, the colors and local structures defining the overall scene are borrowed from the chosen artwork. In essence, this transforms the photograph into the artistic style of the chosen artwork, giving the synthesized image the visual characteristics of the art piece, despite depicting the same content as the original photograph.

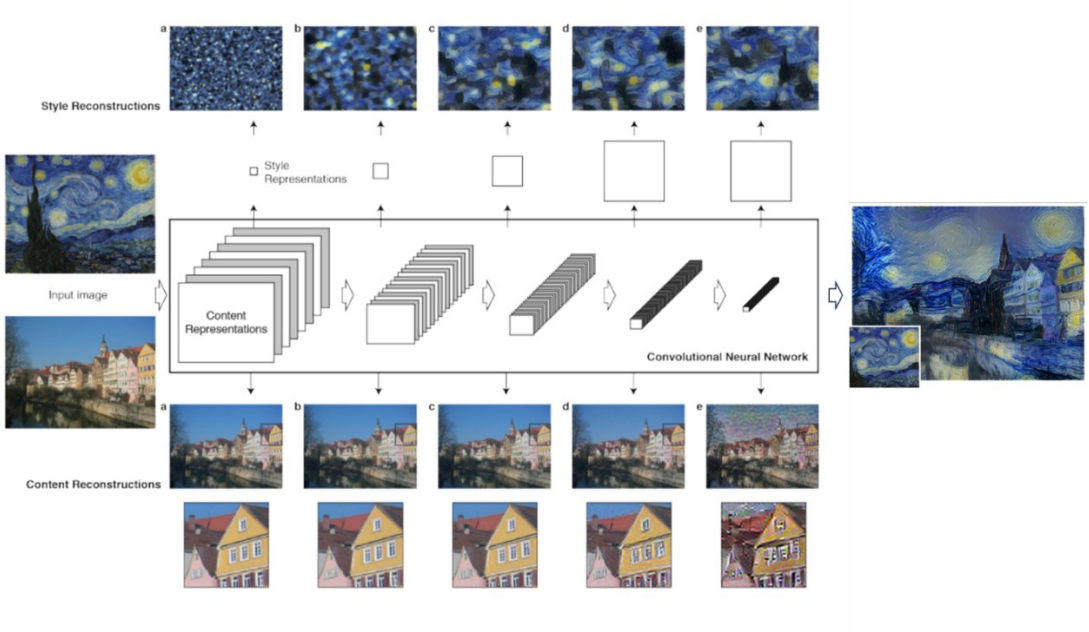

Figure 2-6 Style transfer with Convolution Neural Network. Reprinted from [24].

2.4 Generative Adversarial Network

Generative Adversarial Network (GAN) [25] is a class of generative models designed to generate new data samples that resemble a given dataset. The key innovation behind GANs is the adversarial training framework, where two neural networks, known as the generator and the

discriminator, are trained simultaneously through a competitive process as shown in Figure 2-7. The generator network takes random noise as input and transforms it into data samples that ideally resemble the true data distribution. It essentially creates fake data. On the other hand, the discriminator network evaluates input samples and tries to distinguish between real data from the dataset and fake data generated by the generator. Its goal is to correctly classify the source of the input as real or fake.

The training process on GAN involves a continual interplay between the generator and discriminator. The generator aims to improve its ability to generate realistic data to fool the discriminator, while the discriminator strives to become more accurate in distinguishing real from fake data. This adversarial dynamic leads to a balanced competition where both networks improve over time using loss function. Thus, the training process continues iteratively until a satisfactory equilibrium is reached, where the generator produces realistic samples that the discriminator finds challenging to distinguish from real data.

GANs have been successful in generating high-quality synthetic data, such as images, text, and even more complex outputs like art and music. They have applications in image synthesis, style transfer, data augmentation, and various creative domains. However, training GANs can be challenging, and issues such as mode collapse (where the generator produces limited diversity) and training instability are areas of ongoing research in the field.

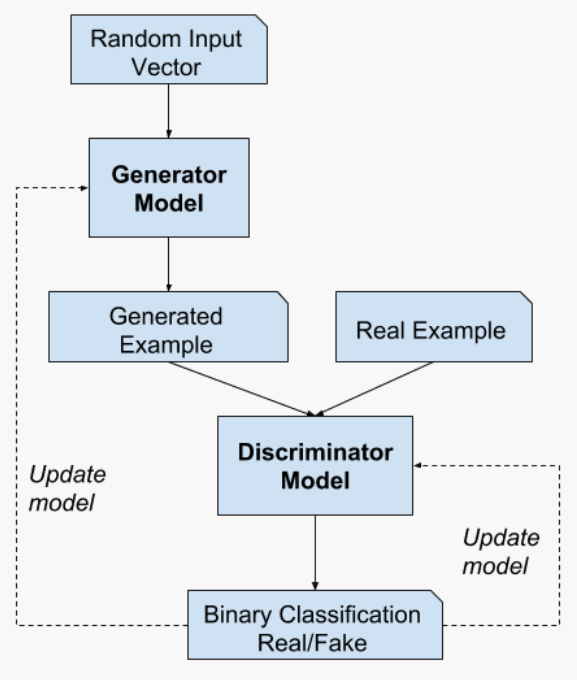

Figure 2-7 Example of Generative Adversarial Network Model Architecture

2.4.1 **Style Transfer using GAN Models**

CNN based the style transfer [24] involves an iterative process to align the content image with the style image, making it a time-consuming method. In contrast, adaptive instance normalization (AdaIN) of GAN [26] achieves style transfer by matching the feature statistics of a source image to a target image to limited range of style transfer. On the other hand, content preservation is approached differently by pix2pix [27], CycleGAN [28], and CUT [29]. CycleGAN's cycle consistency is often considered overly restrictive. In contrast, CUT focuses on maximizing mutual information between content and stylized images within a patch-based feature

space, ensuring the maintenance of structural information while altering appearance as shown Figure 2-8 [11]. Our research follows a similar method as CUT.

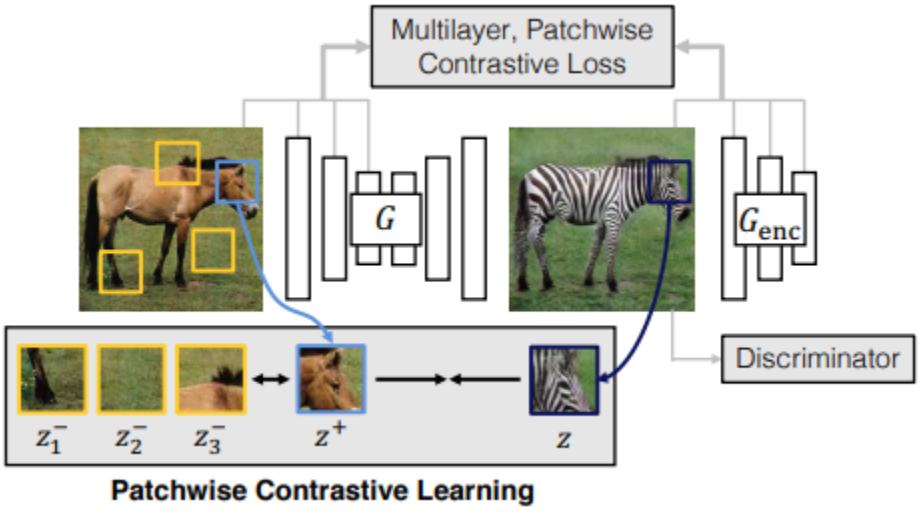

Figure 2-8 The objective is to ensure that a generated output patch closely resembles its corresponding input patch more than any other randomly chosen patches through Patch-wise Contrastive Learning. This is achieved through the application of a multilayer, patch-wise contrastive loss, aiming to maximize mutual information between corresponding input and output patches. This approach facilitates the translation in scenarios where pairs of input and output data are not explicitly matched or paired, allowing for effective transformation in unpaired settings. Reprinted from [11].

2.4.2 **CLIP using GAN Models**

The CLIP demonstrates semantic representative power derived from a vast dataset of 400 million image and text pairs, enabling text-driven image manipulation. StyleCLIP [29] utilizes CLIP and pretrained StyleGAN [26] to optimize the latent vector of the content input based on a text prompt, but its image modification is confined to the domain of the pretrained generator.

StyleGAN-NADA [30] introduces an out-of-domain image manipulation technique by shifting the generative model to new domains. VQGAN-CLIP [31] showcases the capability of using VQGAN [32] as a pretrained generative model for generating or editing high-quality images without the need for additional training. CLIPstyler [33] proposes a CNN encoder-decoder model that learns both content and style properties through patch wise CLIP loss. This approach facilitates image generation and manipulation beyond the limitations of pretrained generators in various domains.

2.5 Diffusion Models

The diffusion model operates by incrementally introducing Gaussian noise in a Markov chain [34] forward process. Subsequently, a trained noise estimation model is employed to iteratively denoise and generate clean samples from the latent noise as illustrated in Figure 2-9 [35].

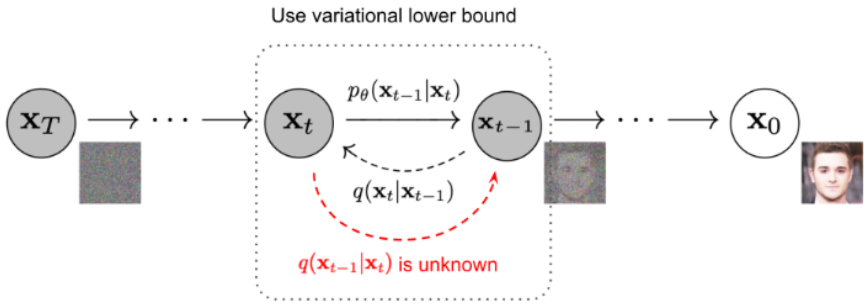

Figure 2-9 The Markov chain of forward (reverse) diffusion process of generating a sample by slowly adding (removing) noise. Reprinted from [35].

2.5.1 **Denoising Diffusion Probabilistic Models**

Denoising Diffusion Probabilistic Models (DDPM) [36] directly samples x_t from x_0 by adding Gaussian noise with $\beta\_t \in (0, 1)$ at time $t \in [1, ..., T]$,

$$x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t} \quad \textbf{Eq. 1}$$

where $\varepsilon \sim N(0, I)$, $\alpha_t = 1 - \beta_t$, and $\bar{\alpha}_t = \prod_{i=0}^{t} \alpha_i$. The reverse sampling process is then given by:

$$x_{t-1} = \frac{1}{\sqrt{1-\beta_t}} \left(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}} \varepsilon_0(x_t, t)\right) + \sigma_t \varepsilon \quad \textbf{Eq. 2}$$

where $\varepsilon_0(x_t, t)$ is used to estimate the noise as a score function. While the noise can enhance sample diversity in DDPM, it may introduce a challenge in maintaining content during style transfer. The iterative application of stochastic operations might produce images with substantially unsimilar content, despite sharing the same intermediate latent space.

2.5.2 **Denoising Diffusion Implicit Models**

Denoising Diffusion Implicit Models (DDIM) [37] addresses this issue by adopting a sampling process that ensures the preservation of content:

$$x_{t-1} = \sqrt{\bar{a}_{t-1}} \hat{x}_{0,t}(x_t) + \sqrt{1 - \bar{a}_{t-1} - \sigma_t^2} \varepsilon_0(x_t, t) + \sigma_t^2 \quad \textbf{Eq. 3}$$

where σ_t is the variance of noise that controls to process stochastic sampling, and $\hat{x}_{0,t}$ is denoising image given by:

$$\hat{x}_{0,t}(x_t) := \frac{x_t - \sqrt{1-\bar{a}_t} \varepsilon_0(x_t, t)}{\sqrt{\bar{a}_t}} \quad \textbf{Eq. 4}$$

2.5.3 **Style Transfer using Diffusion Models through CLIP**

The diffusion models using text guided to image translation maintains a balance between introducing noise for style modification and preserving the essential content of the input image to

be successful results. By carefully controlling the diffusion process, these models can achieve a wide range of stylized effects while maintaining the structure and content of the original image.

As representative model, DiffusionCLIP [6] needs fine-tuning that captures the desired style characteristics to be effective style transfer. Through the process, the model learns to generate images with similar statistical properties as the attribute. Moreover, DiffuseIT [7] adopts a disentangled approach to style and content representation, drawing inspiration from the slicing Vision Transformer. While DiffuseIT has demonstrated its effectiveness in content preservation, it grapples with the challenge of balancing the transformation of image textures and the retention of content. Additionally, the implementation of DiffuseIT demands an extra network for the computation of content losses. In order to solve this problem, ZeCon [10] is proposed new loss function into diffusion models to facilitate style transfer on a given image while keeping its semantic content in a zero-shot manner. The approach is that a pre-trained diffusion model inherently encapsulates spatial information within its embedding, enabling the maintenance of content through patch-wise contrastive loss with anchor and keys taken from the input image and the generated images. Even though ZeCon does not require additional training as well as achieves more precise texture modification while upholding content integrity, ZeCon employs a random selection process for the anchor (q), positive ($k+$) and negatives $(k-)$ in calculating the contrastive loss. This approach is potentially inefficient as their associated patches may not be derived from domain-relevant regions, such as the horse body in the “Horse to Zebra” task. It is important to note that certain features may not accurately capture domain characteristics and tend to persist during translation. As a result, the imposed contrastive loss on these features is not crucial for encoder within diffusion model.

2.6 Summary

In Chapter2, we focus on the core principles of neural network, other network models, self-supervised learning, GAN, and Diffusion. This chapter's primary objective is to introduce these concepts to generate content aware image for style transfer. We extensively explore self-supervised learning, a set of machine learning techniques that aim to uncover patterns and structures in data without explicit supervision or labels. Our detailed introduction to convolution neural network models and U-Net emphasizes their ability to generate various image through diffusion model, which add noise into source image and denoise target image, enabling meaningful representations without explicit labels.

Within Image style transfer application on GANs, we examine notable works such as StyleCLIP, StyleGAN-NADA, VQGAN-CLIP, and CLIPStyler. These works inspired our own architecture, where we integrate their techniques to generate human faces in various style transfer using CLIP. By leveraging advancements in GAN-based approaches, we aim to improve the realism and diversity of the generated facial images in our research.

Additionally, we provide a thorough overview of the diffusion model method and its associated techniques, including DDPM and DDIM for generating high-quality image profiles. Our introduction of DiffusionCLIP, DiffuseIT, ZeCon showcase the use of diffusion model to create stylized images keeping on input's identity, complemented by the utilization of prompt engineering using CLIP. These models are compared to our approach to show our model's superior performance.

Chapter 3 Methodology

Inspired by the CLIP guidance provided in [6], there exist several approaches to enhance the quality and realism of image style transfer. However, most diffusion-based models tend to produce similar images when manipulated [38], [39], [40], [41], [42], [43]. Addressing the challenge of maintaining content in unconditional diffusion models is not straightforward. Unlike GAN-based methods that explicitly incorporate reconstruction loss, such as content losses, diffusion models lack this loss during training, potentially resulting in degraded image quality due to the absence of semantic constraints.

Fortunately, DDIM [37] suggests a solution by advocating for reconstruction on the source image with zero noise variance, although it may sacrifice style preservation. To overcome this limitation, we explore various methods for generating images with CLIP guidance, with a primary focus on attention maps [44] using self-supervised learning techniques. Our research reveals a gap in the existing literature, with limited studies on the integration of CLIP and diffusion models for image style transfer. Motivated by this gap, we utilize CLIP, diffusion methods, and attention map guidance to generate images that depict content-aware stylized content, leading to promising outcomes.

Our proposed architecture seamlessly combines diffusion model methods and CLIP methods to generate content-aware stylized images without the need for additional training or fine-tuning. Leveraging self-supervised learning and attention maps, we produce high-quality, content-aware stylized images. Additionally, we apply data augmentation techniques to further enhance the realism of stylized images. This comprehensive approach demonstrates superior performance, as indicated by user preferences and CLIP scores [45].

3.1 Pipeline Overview

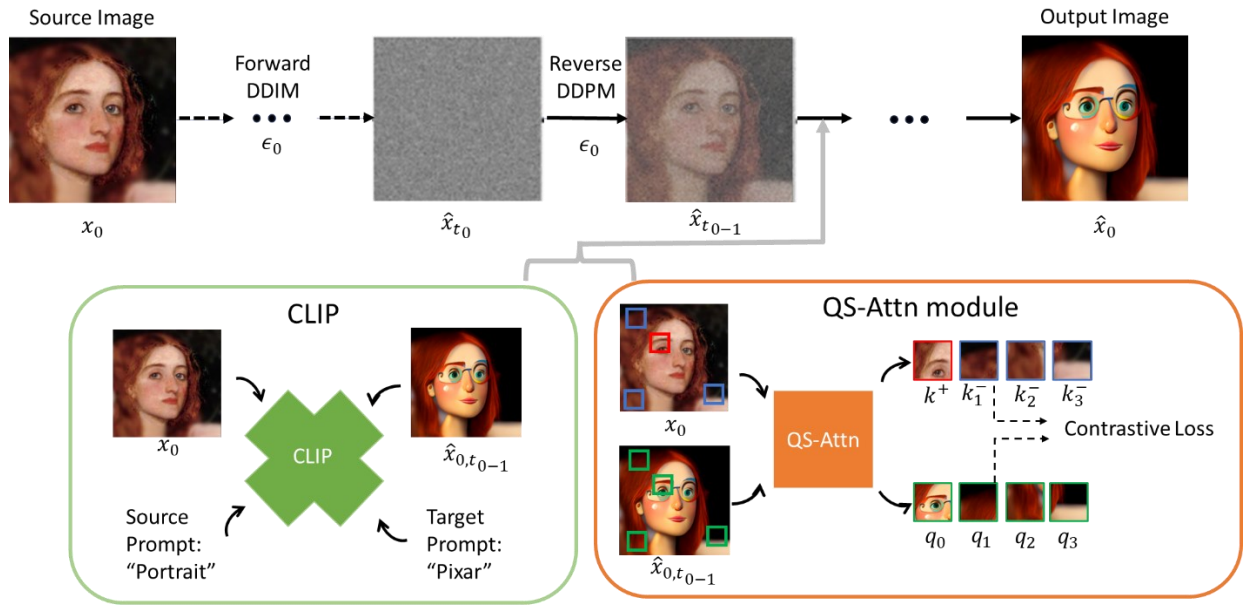

Figure 1-1. The CLIP and attention map within the pipeline are depicted in the accompanying in Figure 3-1. The system has two parts:

- Forward Process:

  - DDPM or DDIM model to generate noised image of source image.

- Reverse Process or Sampling Process:

  - CLIP and attention map guidance to generated content aware image stylization during denoising process

  - Core patches extracted from Query-Selection attention map

  - Calculate total loss using CLIP loss, and Contrastive loss with core patches

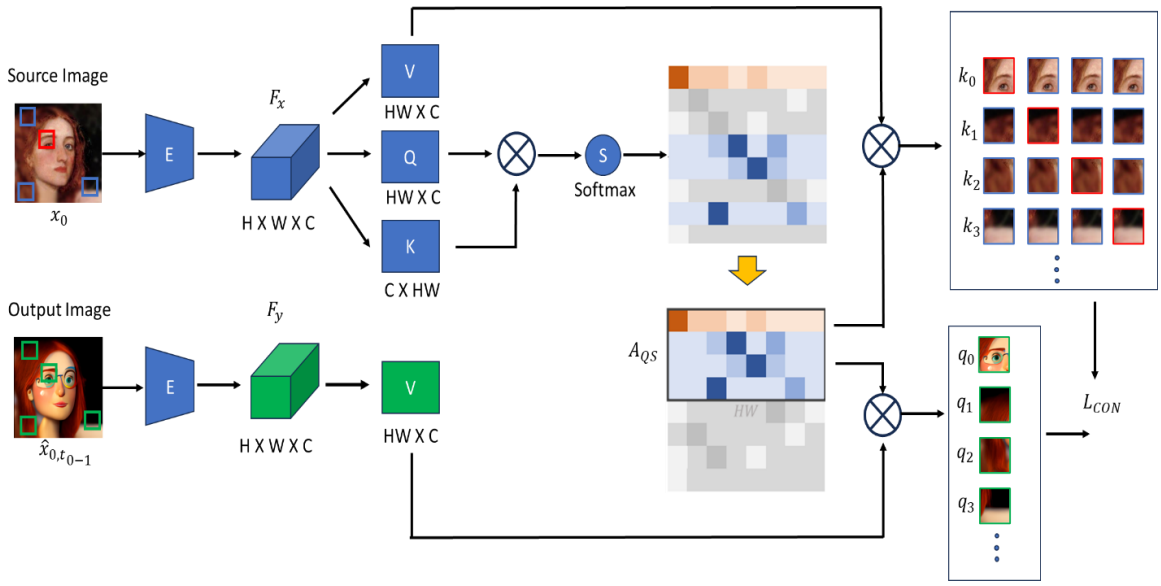

Figure 3-1 The details structure of QS-Attn module. The encoder E extracts each feature from source image and Target image. First, Fx is reshaped and computed to derive the attention matrix. Each row in the matrix is sorted by its metric of the significance, and the selected N rows forming the A_{QS}. We further apply A_{QS} to route both source and target domain value features, and obtain positive, negative and anchor features to construct the contrastive loss Lcon. Positive and negatives are from source image, while anchors are from target image. The patches in orange, blue and green indicate the positive, negative and anchor, respectively.

3.2 Sampling Strategies

To generate a stylized image, we focus on the reverse process of diffusion. Thus, our attention is directed towards the theory of the reverse process of diffusion. Similar to DDIM's sampling equation of Eq. 3, DDPM can be also represented as

$$x_{t-1} = \sqrt{\bar{a}_{t-1}}\hat{x}_{0,t}(x_t) + \sqrt{1 - \bar{a}_{t-1} - \sigma_t^2}\varepsilon_0(x_t, t) + \sigma_t^2 \qquad \textbf{Eq. 5}$$

If σ_t is given by

$$\sigma_t = \sqrt{(1-\bar{a}_{t-1})/(1-\bar{a}_{t)}}\sqrt{1-\bar{a}_t/\bar{a}_{t-1}} \quad \textbf{Eq. 6}$$

The total loss function is subsequently defined according to Eq. 7.

$$\ell_{total}(x) = \ell_{content}(x) + \ell_{CLIP}(x) \quad \textbf{Eq. 7}$$

where $\ell_{content}(x)$ and $\ell_{CLIP}(x)$ denotes the content and style loss, respectively. Then the denoised image estimation $\hat{x}_{0,t}(x_t)$ is augmented by incorporating the gradient of the total loss function, represented by Eq. 8**Error! Reference source not found.**

$$\hat{x}_{0,t}(x_t) = \hat{x}_{0,t}(x_t) + \nabla_x \ell_{total}(x)|_{x=\hat{x}_{0,t}(x_t)} \quad \textbf{Eq. 8}$$

3.2.1 **Content Loss Function**

Content loss function is defined as follows:

$$\ell_{content}(x) = \ell_{Lcon}(x) + \ell_{VGG}(x) + \ell_{MSE}(x) \quad \textbf{Eq. 9}$$

We incorporate with other loss functions such as VGG loss [46] and MSE loss [47]. The VGG loss quantifies the mean-squared error between the VGG feature maps. The MSE loss represents the *L2* norm of the pixel difference between them.

Lcon loss function is defined as follows:

$$\ell_{Lcon}(\hat{x}_{0,t}, x_0) = \mathbb{E}_{X_0}[\sum_l \sum_s \ell(\hat{z}_l^s, z_l^s, z_l^{S/s})] \quad \textbf{Eq. 10}$$

where $\hat{z}_l$ and z_l denote the l th layer features from $\hat{x}_{0,t}$ and x_0, respectively. s represents a spatial location in $1, \ldots, S_l$, where is the number of spatial locations in feature z_l

$\ell()$ is cross entropy loss function on Eq. 11.

$$\ell(\mathrm{q}, \mathrm{k}^+, \mathrm{k}^-) = -\log\left[\frac{\mathrm{e}^{\mathrm{q}\cdot\mathrm{k}^+/\tau}}{\mathrm{e}^{\mathrm{q}\cdot\mathrm{k}^+/\tau}+\sum_{\mathrm{i}=1}^{\mathrm{N}}\mathrm{e}^{\mathrm{q}\cdot\mathrm{k}_\mathrm{i}^-/\tau}}\right] \quad \textbf{Eq. 11}$$

where q, k^+, k^- denote query, positive part, and negative part respectively. τ is temperature.

The content loss function is similar to CUT loss [11], which maximizes the mutual information between positive pairs from the same location and minimizes the mutual information between negative pairs from different locations to preserve the content of source image through patch-wise contrastive loss. Since U-Net noise predictor has spatial information, it is possible to get spatial information such as keys, anchors from diffusion model without additional training. In order to apply the loss, significant features are selected by attention map and used to cross-entropy loss during diffusion reverse process.

3.2.2 **Style Loss Function**

The extensive training on language and image datasets endows the CLIP model with significant semantic power [48]. This semantic capability enables the generation of images in various styles using only a textual prompt.

$$\ell_{CLIP} = \ell_{global}(\hat{x}_{0,t}, P_{target}) + \ell_{dir}(\hat{x}_{0,t}, x_0, P_{target}, P_{source}) \quad \textbf{Eq. 12}$$

The global CLIP loss calculates the cosine distance in the CLIP embedding space between the denoising image $\hat{x}_{0,t}$ and the specified style prompt P_{target} [29] by

$$\ell_{global}(\hat{x}_{0,t}, P_{target}) = D_{CLIP}(\hat{x}_{0,t}, P_{target}) \quad \textbf{Eq. 13}$$

Due to mode collapse and degraded image quality in the global loss, a novel approach called directional CLIP loss (ℓ_{dir}) was introduced [30]. This method focuses on aligning the direction within the CLIP embedding space for text and image pairs, and its formulation can be expressed as follows:

$$\ell_{dir}(\hat{x}_{0,t}, x_0, P_{target}, P_{source}) = 1 - \frac{\Delta I \cdot \Delta T}{\|\Delta I\| \, \|\Delta T\|} \quad \textbf{Eq. 14}$$

In this context, P_{source} represents the source text prompt, and the differences (ΔI and ΔT) are defined as the disparities in the embeddings produced by CLIP's image encoder and text encoder. Specifically, ΔI correspond to the difference between the embeddings of the original image x_0 and the denoising image $\hat{x}_{0,t}$ with respect to the given text prompt, while ΔT represents the distinction between the embeddings of the source text prompt and the target text prompt. To improve the quality of generated images, a patch-based CLIP loss was introduced [30]. To enhance both the ℓ_{global} and ℓ_{dir} aspects, we incorporate the patch-based approach into our methodology.

3.3 QS-Attn for Contrastive Learning

Rather than relying on a straightforward random strategy, we utilize an attention-based approach [12]. This method involves initially comparing a specified query with keys and subsequently choosing the query based on the comparison results. Notably, we diverge from the conventional attention approach by abstaining from employing distinct projection heads for query, key, and value. Consequently, no supplementary model parameters are introduced in diffusion model. The specifics of the QS-Attn approach are explained in the subsequent two subsections. According to [12], some features do not reflect the domain characteristics due to random selection of patches. We propose three attention map methods to select intentionally anchor q, and then, it computes the loss with more domain-specific patch information during diffusion reverse process. As shown in Figure 3-1, the sequences of the QS-Attn module are as follows:

First, the encoder E extracts features from both the source and target images. Second, the feature set Fx is reshaped and computed to derive the attention matrix. Each row in this matrix is sorted based on its significance metric, and the top N rows are selected to form the matrix A_{QS}. Subsequently, $A_{QS}$is applied to route both source and target domain value features. By the process,

positive, negative, and anchor features are then obtained to construct the contrastive loss. The positive and negative features originate from the source image, while the anchor features are from the target image.

3.3.1 **Global Attention**

We aim to define a quantitative value for each potential location, which reflects the significance of the feature. The quadratic attention matrix is adopted, since it exhaustively compares each feature with all other locations, it accurately reflects the similarities with others. Based on Eq. 15 [12], To select all the significant queries, the rows of A_g are sorted by the entropy Hg in the ascending order, and the smallest N rows are selected as the QS-Attn matrix $A_{QS} \in \mathbb{R}^{NxHW}$. Note that A_{QS} is fully determined by the features.

$$H_g(i) = -\sum_{j=1}^{HW} A_g(i,j) log A_g(i,j) \quad \textbf{Eq. 15}$$

Where H_g, A_g represents entropy, global attention map. The indices $i\ and\ j$ refer to the positions of the query and key in the global attention map, indicating the specific row and column.

3.3.2 **Local Attention**

While non-local attention can capture the global context, it tends to blur the intricate details in the proximity of the queries. In contrast, local attention gauges the resemblance between a query and its nearby keys within a fixed window of $w \times w$ and stride of 1. This enables the capture of spatial interactions in local regions, concurrently diminishing the computational overhead. Based on [12], select the smallest N rows in A_l by sorting H_l in the ascending order to form A_{QS}. For the

value routing, we also locate the N indexes in local value matrix $V_l \in \mathbb{R}^{HWx\, w^2 x\, C}$ and get the selected value matrix $V_{ls} \in \mathbb{R}^{N\, x\, w^2 x\, C}$

$$H_l(i) = -\sum_{j=1}^{w2} A_l(i,j) log A_l(i,j) \qquad \textbf{Eq. 16}$$

Where H_l, A_l represents local entropy, local attention map. The indices $i\ and\ j$ refer to the positions of the query and key in the local attention map, indicating the specific row and column.

3.3.3 **Cross Domain Attention**

A_{QS} presents global or local relation by comparing the queries with keys so that gets value features from both source and target image, keeping valuable high-level descriptions of the shape and texture of source image. Through A_{QS}, It is possible to enhance the receptive field of selected queries and to get improved features from source image. In order to establish a self-supervised contrastive loss, one positive feature and negative features originate from source image, while N anchors are derived from the generated image.

3.4 Summary

In this chapter, we focus on the core principles of our solution, specifically patch wise contrastive learning under self-supervised learning, utilizing diffusion model. Our primary objective is to apply these concepts to generate content aware image for style transfer. We extensively explored two loss functions, a set of content losses that aim to keep content of image, a set of style losses using CLIP. Our detailed introduction to query selection attention map emphasizes decision of significant features to generate content aware stylized image through diffusion model. With our solution, we will examine notable works such as StyleCLIP, StyleGAN-

NADA, VQGAN-CLIP, and CLIPStyler to compare our model. Additionally, we will provide a thorough oh comparison of the other diffusion model method such as DiffusionCLIP, DiffuseIT, Zecon to showcase the use of our solution to create stylized images keeping on input identity, complemented by the utilization of prompt engineering using CLIP.

Chapter 4 Experiments

4.1 Datasets

The content reference images used in our research are sourced from CelebA-HQ [49], FFHQ [26], LSUN-Bedroom [50], AFHQ-Dog dataset [51], and ImageNet [52]. These datasets encompass a variety of visuals, including human faces, objects, scenes, animal, and churches. To assess our proposed model's performance on images from previously unseen domains, we include the Wikiart dataset [53]. All images are uniformly resized to 256 × 256 for compatibility with the diffusion models. For patch-based guidance, we randomly extract 96 patches from a source image, followed by the application of perspective augmentation and affine transformation.

4.1.1 **CelebFaces Attributes Dataset (CelebA)**

CelebA [49] is a widely used dataset in the field of computer vision and machine learning. It contains 200,000 collection of celebrity images, with each image labeled with 40 attribute annotations, including smile, wearing glass, bald. Celebrities in the dataset cover a wide range of ages, ethnicities, and occupations. The dataset is commonly utilized for tasks such as face recognition, attribute expression, and generative modeling. The images in the CelebA dataset vary in quality, capturing diverse facial expressions, poses, and backgrounds. This diversity makes the dataset suitable for training models to handle real-world variations as shown in Figure 4-1.

Figure 4-1 Examples of CelebA dataset, containing various attributes. Reprinted from [49].

4.1.2 **Flickr-Faces-HQ Dataset (FFHQ)**

FFHQ [26] refers to a high-quality dataset of human faces collected from Flickr. The dataset contains over 70,000 high-quality images of faces. The images cover a wide range of ages, ethnicities, and facial expressions, making it suitable for various research tasks. FFHQ captures variations in facial poses, lighting conditions, and backgrounds, enhancing its suitability for training and evaluating models that need to handle real-world scenarios. This dataset is commonly used in the field of computer vision and machine learning, particularly for tasks related to face generation, face recognition, and other facial analysis tasks. FFHQ is commonly used in research related to generative models, especially in the development and evaluation of GAN for realistic face generation as shown in Figure 4-2.

Figure 4-2 Examples of FFHQ dataset, containing various ethnicity. Reprinted from [26].

4.1.3 **LSUN Dataset (LSUN – bedroom)**

LSUN [50] focuses on providing a diverse and extensive collection of images that capture scenes rather than individual objects. The dataset is designed to promote research in scene understanding, including tasks such as scene classification and scene parsing. LSUN includes a wide range of scene categories, covering indoor scenes such as bedrooms, kitchens, living rooms, and outdoor scenes such as church. LSUN images are of high resolution and aim to represent real-world scenes accurately. The dataset includes images of varying sizes and qualities, reflecting the challenges of scene understanding in different contexts. We only utilized LSUN-Bedroom data as shown in Figure 4-3.

Figure 4-3 Examples of LSUN Bedroom dataset, containing various environment. Reprinted from [50].

4.1.4 **AFHQ-DOG Dataset**

AFHQ-DOG [51] contains high-resolution images of dog faces, capturing various breeds, poses, and expressions. The images aim to be visually appealing and of high quality, suitable for training and evaluating advanced computer vision models. The dataset covers a diverse range of dog breeds, ensuring variability in appearance, fur patterns, and facial features. This diversity is valuable for training models that can generalize well to different types of dogs as shown in Figure 4-4.

Figure 4-4 Examples of AHHQ dog dataset, containing various breed. Reprinted from [51].

4.1.5 **ImageNet**

ImageNet [52] is one of the largest publicly available image datasets, consisting of over 14 million labeled images. The dataset covers a wide range of object categories, including animals, plants, everyday objects, and scenes. Each image in ImageNet is labeled with a specific object category from a hierarchical structure of over 20,000 classes as shown in Figure 4-5. ImageNet serves as a benchmark dataset for training and evaluating image classification, object detection, and image segmentation models. Many pre-trained deep learning models are trained on subsets of ImageNet, and researchers often fine-tune these models for specific tasks.

Figure 4-5 Examples of ImageNet dataset, containing complexity of images. Reprinted from [52].

4.2 Implementation Details

In our experimental setup, we employ a system consisting of an i9-11300k CPU, a GTX 4090 graphics card, and 64GB RAM. The operating environment encompassed at Linux, with Anaconda installed to facilitate our work.

Our pipeline is based on the Python language, utilizing the PyTorch framework. Within our code, we integrate diffusion model to generate content aware stylized image into attention map and CLIP. Additionally, we implement various attention map versions in our pipeline to enhance the overall results. We utilize pretrained model for various dataset to generate various source images as well.

Our model can leverage both types, namely DDIM or DDPM, for the diffusion process. In our experiments, we employed the DDIM method as the forward process, followed by the adoption of the DDPM method as the reverse process. The diffusion process comprises a total of 1000 time steps, with a specified step size for manipulation time set at (50, 25) to achieve high-quality stylized images. We will explain the temporal aspects of this process in the ablation study, emphasizing the preservation of spatial features from the source image during manipulation of the stylized generated image. Thus, our approach allows for the reduction of inference time.

4.3 Comparative Studies

We assess our model by comparing it with GAN-based models with face image dataset under qualitative and quantitative methods. The quantitative evaluation uses CLIP scores and Face ID similarity.

In the qualitative approach, we conduct a user study involving 20 adult participants in their 30ties comprising 10 men and 10 women, all of whom had no prior knowledge of style transfer.

Initially, we explain the concept of style transfer, including the meaning of stylization, the use of text prompts to modify stylized image, and provide an example for image style transfer. Subsequently we provide two images in the user study form; the original image is left image, and the stylized image by text prompt is right image. We ask them two questions; (i) how well right image preserves content of left image and (ii) how well right image stylizes of left image.

In case of diffusion-based models, there is no error metrics to be used to compare different methods so that we only use qualitative method.

4.3.1 **Comparison with GAN Models**

In the evaluation of GAN-based models, we have conducted a comparison involving four cutting-edge techniques: StyleCLIP [29], StyleGAN-NADA [30], VQGAN-CLIP [31], and CLIPstyler [33]. The results, as depicted in Figure 4-6, highlight the superiority of our proposed model, particularly in terms of content retention. Our approach excels in producing outputs with realistic textures, distinguishing it from StyleCLIP, which tends to generate outputs resembling regular photos and may not work well with other style transfer methods. Additionally, StyleGAN-NADA's stylization differs from user-preference prompts, even though the content is well preserved. Moreover, VQGAN-CLIP does not maintain the contents of the source image when transforming images into various styles. CLIPstyler struggles to generate images in styles like Pixar, Gogh, and Ukiyo-e, whereas our method consistently delivers high-quality samples faithfully transferred into the styles specified by the target prompts. The superiority of our proposed method is further substantiated through quantitative evaluation. Table 4-1illustrates that our method achieves the highest scores in both user studies and CLIP scores. While StyleCLIP exhibits the smallest face identity loss, implying strong preservation of semantic information, it falls short in adequately transforming styles. Conversely, VQGAN-CLIP tends to overmodulate images, leading to significant content alterations, as indicated by the Face ID loss. In

contrast, even though StyleGAN-NADA has the second-highest score, our method's score indicates better preservation of content and stylization than StyleGAN-NADA.

Table 4-1 illustrates that our method achieves the highest scores in both user studies and CLIP scores. While StyleCLIP exhibits the smallest face identity loss, implying strong preservation of semantic information, it falls short in adequately transforming styles. Conversely, VQGAN-CLIP and StyleGAN-NADA tend to overmodulate images, leading to significant content alterations. In contrast, our method strikes a balance between preserving content and effectively transferring styles. The CLIP score is computed globally in a patch-based manner, as outlined in Table 4-1, while face identity loss (Face ID) is measured using ArcFace [54]. The same set of images used in the user study is employed for conducting the quantitative experiments.

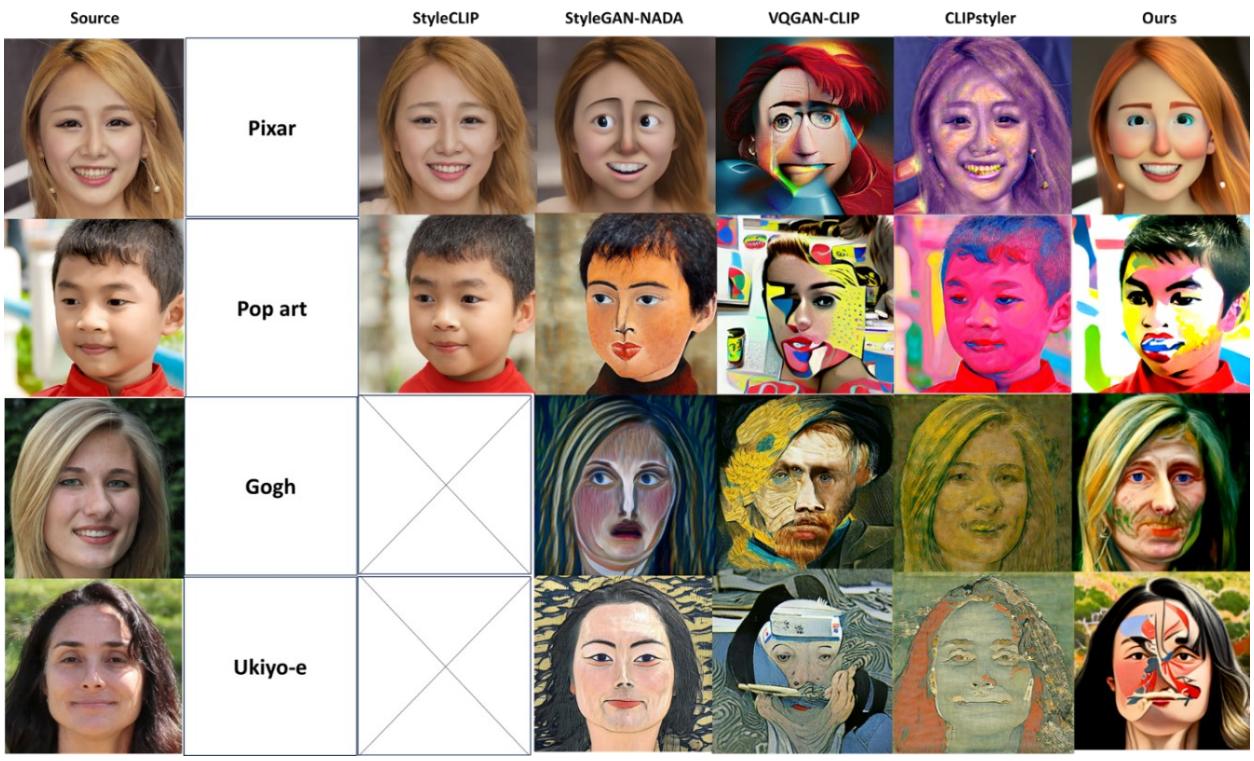

Figure 4-6 Comparison our method with other GAN model for Image Style Transfer. our approach stands out by delivering superior outcomes in both style transformation and content preservation.

Method	User Study		CLIP Score ↑	Face ID ↓
	Content ↑	Style ↑		
StyleCLIP	88.4	56.2	0.092	**0.375**
StyleGAN-NADA	83.2	76.5	0.122	0.495
VQGAN-CLIP	68.1	74.3	0.138	0.755
CLIPstyler	75.2	79.2	0.135	0.663
Ours	**93.2**	**94.5**	**0.148**	<u>0.399</u>

Table 4-1 Findings from user studies and quantitative assessments in the context of comparing our style transfer method with GAN-based approaches are highlighted. The bold text denotes the highest scores, and the underline text denotes the second scores. Each content and style score in the user study is average score about 40 images.

4.3.2 **Comparison with Diffusion Models**

We compare our proposed method with three diffusion-based models such as DiffusionCLIP, DiffuseIT, Zecon. The qualitative and quantitative results of the comparison are presented in Figure 4-7 and Table 4-2. The third row of Figure 4-7 shows that DiffusionCLIP suffers from identity loss, where the Butterfly and Bird identity of the images is destroyed in the translation. Moreover, the color of all translated images is less colorful than others even if maintaining overall contents of images. Additionally, the diffusion model must be trained for each new domain to be style through fine tuning method. On the other hand, DiffuseIT shows the trade-off between style transfer and content preservation as illustrated in Figure 4-7. While changing the style of the source image, the identity is also modified so that it does not keep contents of images. Similar our methods to ZeCon, it makes some issue when translating using guiding text. the identity on translated images is modified. In contrast, our proposed method can stylize images while maintaining it's identity. These results are confirmed with user study results presented in Table 4-2, where the scores between

the photo domain and unseen domains are highly similar. This means that our method can modulate images even from unseen domains. Especially, given computational time, our method is significantly faster than other diffusion models, e.g., our method takes 30s, DiffusionCLIP does 30 min (including fine-tune), and DiffuseIT does 48s.

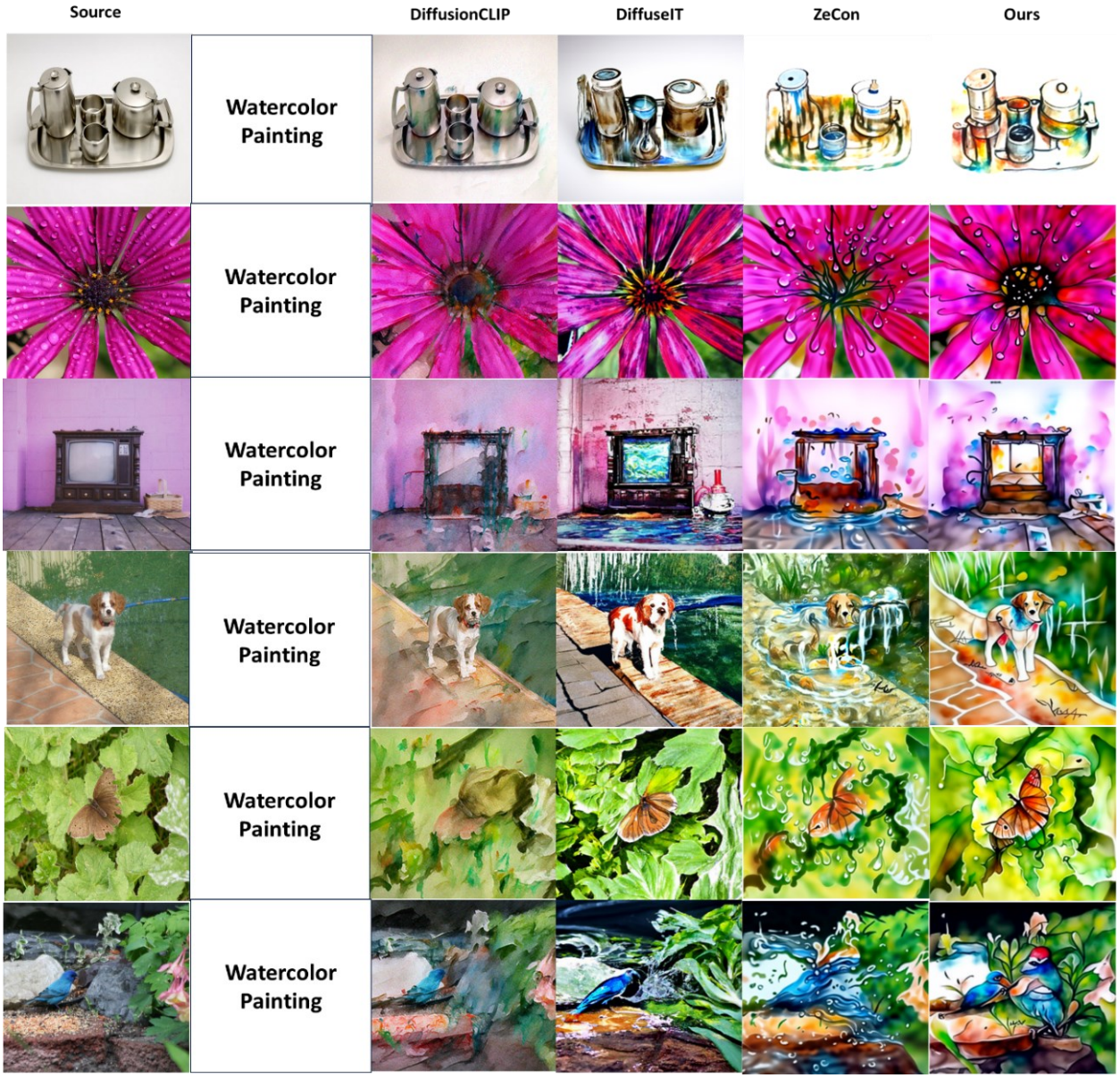

Figure 4-7 Comparison our method with other diffusion model for Image Style Transfer. our approach stands out by delivering superior outcomes in both style transformation and content preservation.

Method	Photo Domain		Unseen Domain	
	Content ↑	Style ↑	Content ↑	Style ↑
DiffusionCLIP	83.4	88.2	81.4	86.1
DiffuseIT	75.2	87.5	72.2	84.2
ZeCon	85.1	89.2	83.1	86.1
Ours	**91.2**	**90.5**	**90.1**	**88.3**

Table 4-2 Comparison our method with other diffusion model for Image Style Transfer in the user study. The bold text denotes the highest scores.

Our method can translate various style using text guiding method while maintaining contents of source images as Figure 4-8, e.g., “A sketch with crayon”, “Golden”, “Water Painting”, “Wooden”, “Red bricks”, and Figure 4-9, e.g., “Clay”, “Pop art”, “Stone wall”.

Figure 4-8 Image Style Transfer through Query-Selected Attention Diffusion model's reverse process, e.g., "A sketch with crayon", "Golden", "Water Painting", "Wooden", "Red bricks".

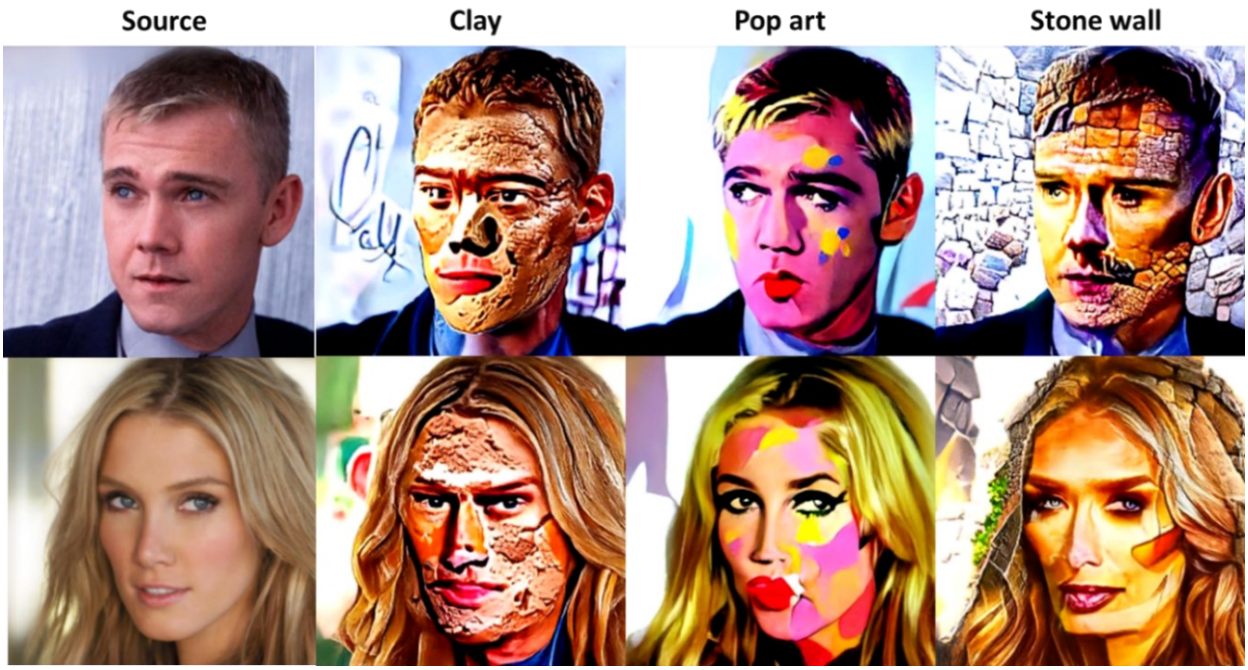

Figure 4-9 Image Style Transfer through Query-Selected Attention Diffusion model's reverse process, e.g., "Clay", "Pop art", "Stone wall".

Our proposed method showcases not only excels in image style transfer but also demonstrates promising potential for diverse tasks such as image translation and manipulation. In Figure 4-10, qualitative results for image translation illustrate the capability of the method to seamlessly transform different color while preserving details and overall coherence in the images. Beyond image translation, our method proves versatile in image manipulation tasks showcase the ability to change the attributes of person as illustrated in Figure 4-11. Moreover, Figure 4-12 showcases the ability to change the identity of animals. These results highlight the broad applicability and potential of the proposed method for various image manipulation tasks.

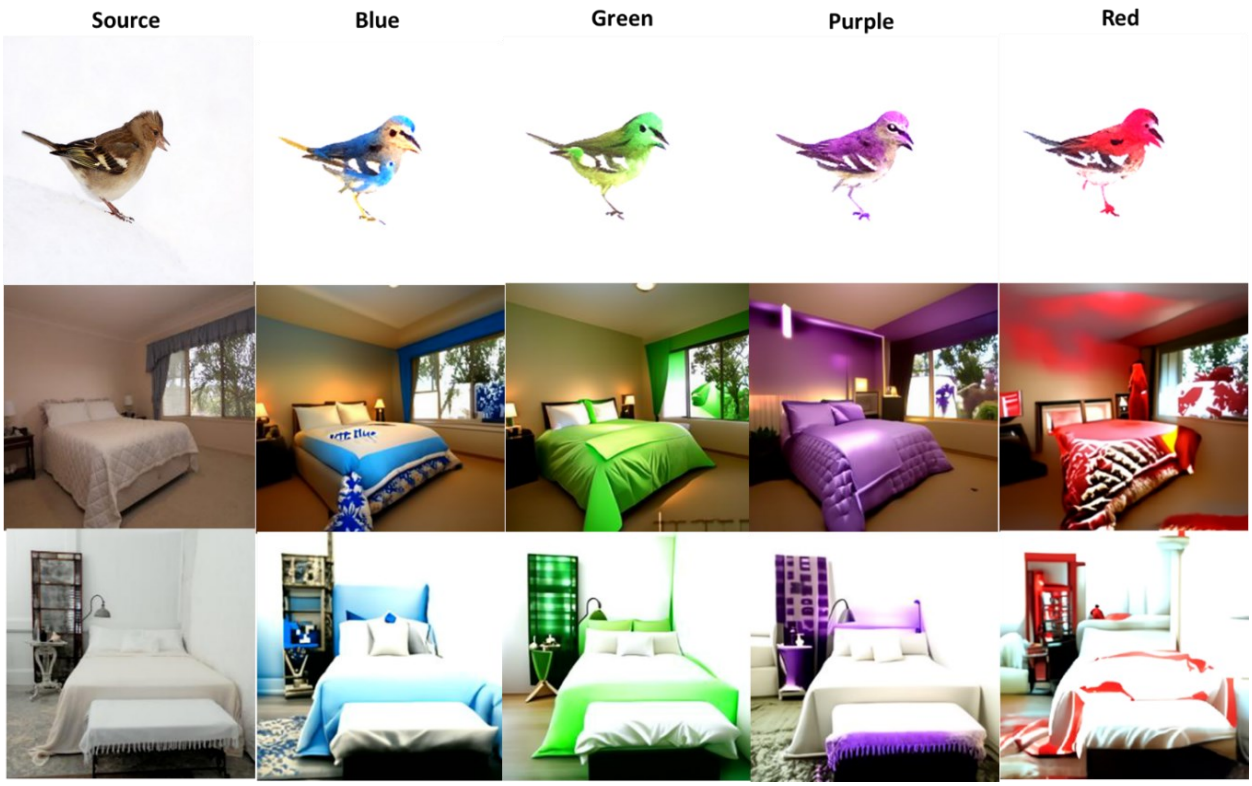

Figure 4-10 Additional results on various color style prompts

Figure 4-11 Additional results on image manipulation with human faces by prompts, e.g., "No make up", "Anrgy", "Makeup and Curly Hair".

Figure 4-12 Additional results on various image translations by prompts, e.g., "from horse to zebra", "from dog to wolf or bear".

4.4 Ablation Studies

4.4.1 **Style Transfer with Unseen Domain**

Our diffusion model can manipulate unseen domain in image style transfer. The qualitative results for image manipulation are shown in Figure 4-13, where our method can translate Pixar or Zombie images while preserving contents of the images without additional work like DiffusionCLIP, which need to change person and then stylization.

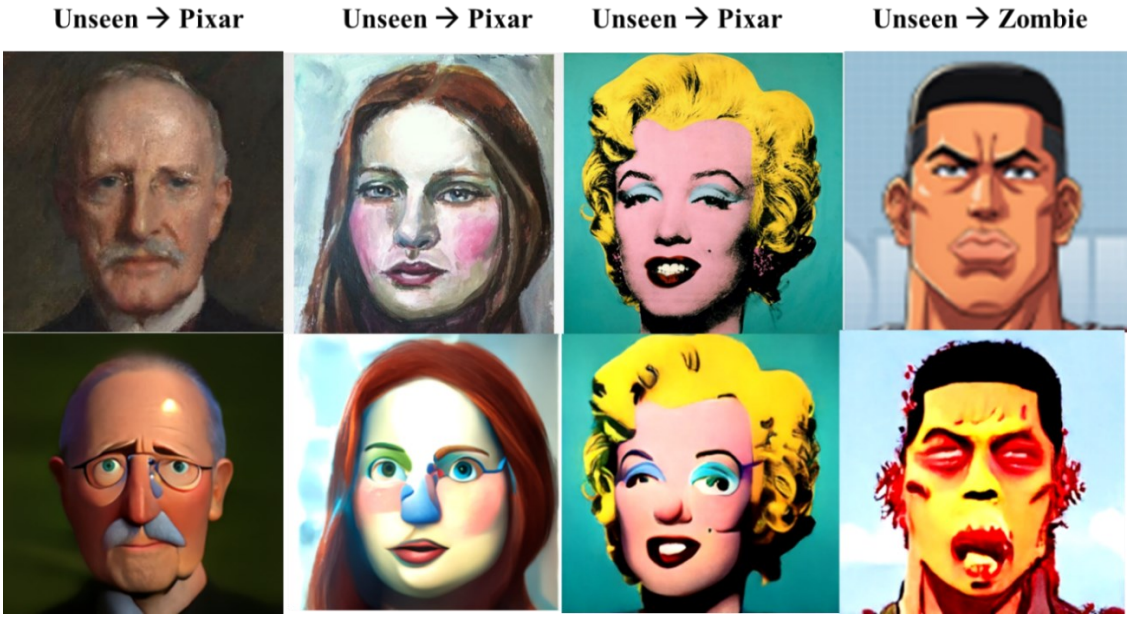

Figure 4-13 Image Style Transfer for unseen domain, e.g., "from unseen domain to Pixar or Zombie".

4.4.2 **Role of Attention Map Guidance**

To assess the effectiveness of attention map guidance, we conduct ablation studies on our proposed attention map guidance approach in mentioned methodology section, which comprises three distinct maps: Global, Local, and Local & Global. To evaluate the impact of each attention maps, we proceed each attention map and compared the outcomes. Figure 4-14 illustrates that global attention map led to a loss of fine structural details, such as wall parts in building outlines, despite maintaining the overall shape. Local attention map led to a little bit of loss of fine structural details, such as wall pattern parts in building outlines, despite maintaining the overall shape. Finally, local and global attention maps lead to show fine structural details, such as wall pattern in building outlines, maintaining inner shape and the overall shape. This indicates that relying solely on each global or local is insufficient for preserving intricate content details. Conversely, local and global attention map combined with two methods produced the most favorable results in terms of content preservation. The user study is presented in Table 4-3, and some samples of comparing results are shown in Figure 4-14.

Figure 4-14 Ablation study focusing on three attention maps – Global, Local , and Local & Global. The results show that, in each row of translated images, which are transformed into the styles of "golden", the local & global attention map is effective in preserving content information.

Method	Photo Domain	
	Content ↑	Style ↑
Global	84.6	92.2
Local	86.4	92.3
Local & Global	**91.5**	**92.4**

Table 4-3 Ablation study focusing on three attention maps with user study – the Local & Global attention map is effective in preserving content information. The bold text denotes the highest scores.

4.4.3 **Role of Contention Loss**

To assess the effectiveness of content guidance losses, we conducted ablation studies on our proposed content loss approach in Eq. 9, which comprises three distinct losses: Con, VGG, and MSE. To evaluate the impact of Con, we excluded it from the overall content loss and compared the outcomes with the complete content loss. illustrates that omitting Con led to a loss of fine structural details, such as building outlines, despite maintaining the overall shape. This indicates that relying solely on VGG and MSE is insufficient for preserving intricate content details. Conversely, incorporating all three losses produced the most favorable results in terms of content preservation. The results of a user study, presented in Table 4-4, further validate the superiority of Con over MSE, VGG, and even their combination. This suggests that Con effectively retains structural details while preventing overfitting image. In summary, the ablation studies underscore the critical role of Con in our proposed method for maintaining the structural properties of input images as shown in Figure 4-15.

Figure 4-15 Ablation study focusing on three losses for content guidance - MSE, VGG, and Con. The results show that, in each row of translated images, which are transformed into the styles of "golden, the proposed patch-wise content preservation loss is effective in preserving content information.

Method	Content ↑	Style ↑
Con	88.3	89.1
MSE	80.1	85.3
VGG	76.2	84.1
MSE, VGG	82.3	87.3
MSE, VGG, Con	**91.5**	**92.4**

Table 4-4 Ablation study focusing on content losses with user study – Con, MSE, VGG, MSE & VGG, Con & MSE & VGG. The Con & MSE & VGG is effective in preserving content information. The bold text denotes the highest scores.

4.4.4 **Role of Style Loss**

Ablation studies are conducted to investigate the role of each loss function in our style loss formulation, which comprises two components: CLIP-Global and CLIP-Directional loss, as depicted in Figure 4-16. To assess the impact of each loss, we individually applied them and assessed the outcomes across a distinct style – "Oil painting of flowers". The qualitative results, which showcased in Figure 4-16, revealed that incorporating the CLIP-Directional loss alongside the CLIP-Global loss resulted in more stylized images compared to using the CLIP-Global loss alone. This suggests that the CLIP-Directional loss is more effective in influencing the style of images.

Additionally, we examine the significance of patch-based guidance in our proposed method. We observe that utilizing whole-image guidance tended to stylize localized parts of the image, whereas patch-based guidance transformed the entire image into the specified style by covering a larger area. Figure 4-16 illustrates this, where patch-based guidance is employed to stylize the entire background while preserving the foreground object.

In summary, our ablation studies indicate that both CLIP-Global and CLIP-Directional loss play crucial roles in achieving high-quality style transfer results, and the use of patch-based guidance proves effective in transforming images into a desired style.

Figure 4-16 Ablation study focusing on two losses for style guidance – CLIP-Global and CLIP-Directional loss. The results show that, translated images are transformed into the styles of "Oil painting of flowers". the proposed patch-wise style loss with both method is effective in changing stylization.

4.4.5 **Choice on Diffusion Model's Forward and Reverse Processes**

While both DDPM and DDIM can be applied to both forward and reverse processes, we conduct a comparative study to highlight their distinctions in generated images. Figure 4-17 illustrates that results from the forward DDIM exhibit superior performance in preserving content compared to DDPM. Notably, in the second row, the tooth and lips appear unchanged in the output images generated by DDIM & DDPM, while the texture and shape with other ways are altered in the images produced by other ways such as DDPM & DDPM, DDPM & DDIM, and DDIM & DDIM. Consequently, we have chosen to default to use DDIM for the forward process and DDPM for the reverse process based on these observed differences.

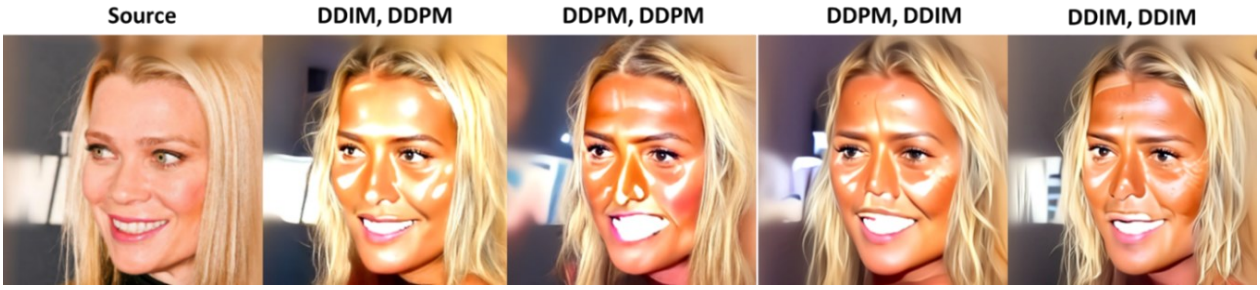

Figure 4-17 Ablation study results on diffusion processes. From the second column to the right, the combinations of methods (forward, reverse) are (DDIM, DDPM), (DDPM, DDPM), (DDPM, DDIM), and (DDIM, DDIM).

4.4.6 **Diffusion Process Time Control**

Given that the diffusion process is typically time-consuming, two widely adopted techniques to address this are respacing and skipping time steps [6],[55]. The last time step *T* is respaced into *T'*. Subsequently, the diffusion model is forwarded to time *t0* < *T'*, followed by the reversal of the diffusion process from x_{t_0} to *T'*. The choice of *T'* and *t0* has significant impacts on both image quality and time consumption. Figure 4-18 (a) and (b) demonstrate that image quality, concerning style transformation, improves as the respacing time step *T'* increases. However, the growth rate of improvement decreases, and the difference becomes imperceptible even as the sampling time continues to increase. Simultaneously, the CLIP score increases as t0 increases, as shown in Figure 4-18 (c). On the other hand, content information is not fully preserved at time steps *t0 = 10 or 15*, as depicted in Figure 4-18 (a). Hence, we have set (*T', t0*) as (50, 25) for our baseline configuration.

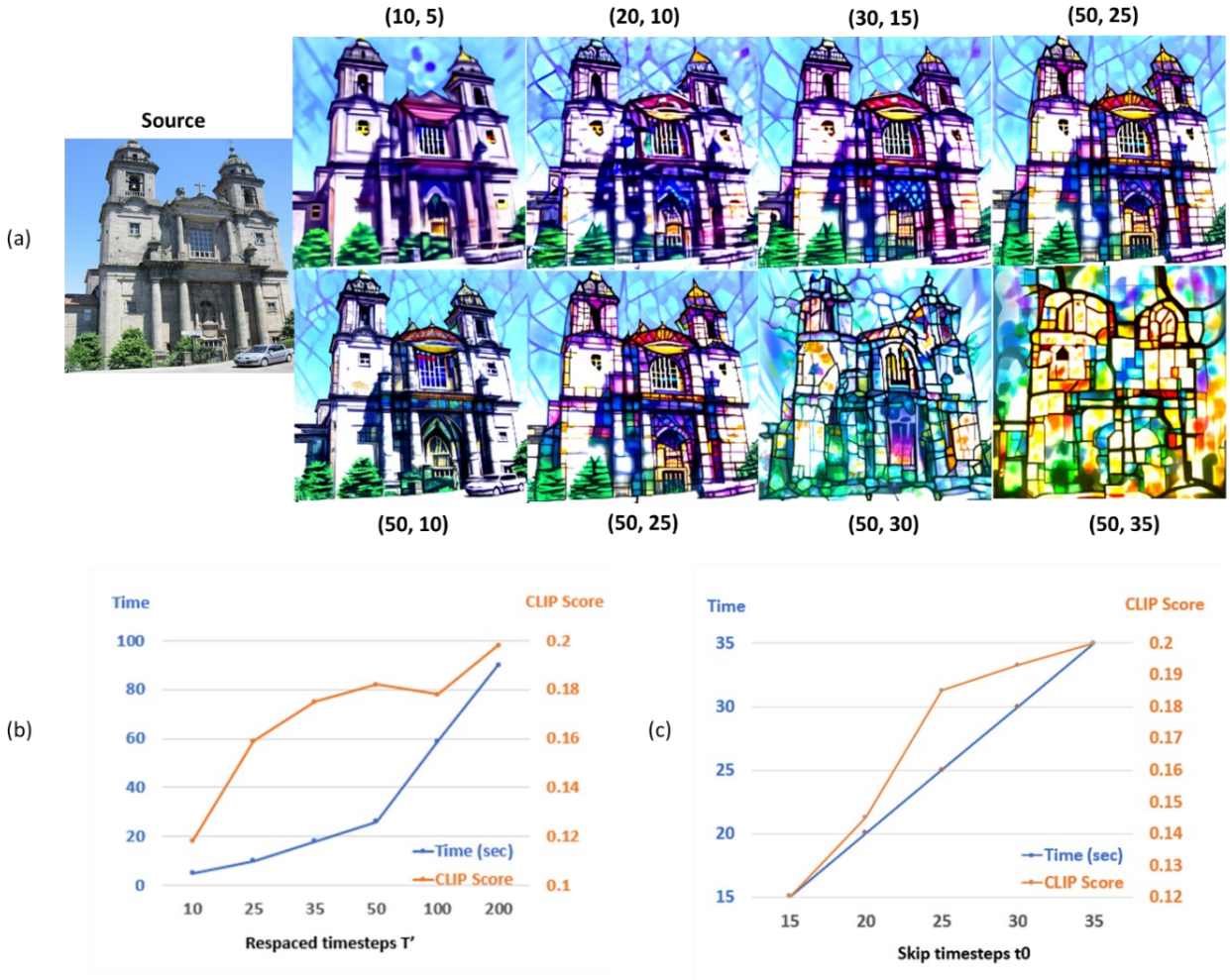

Figure 4-18 The impact of respaced timesteps *T'* and skip timesteps *t0*. (a) showcase images sampled with different combinations of (*T'*, *t0*). The first row highlights the variation in *T'* when *t0* is at its half, while the second row depicts the differences when *T'* is fixed at 50 and various *t0* values are used. Furthermore, (b) and (c) present graphical representations of the relationship between sampling time and CLIP score for the first and second rows of (a), respectively. These graphs provide insights into how changes in sampling time, influenced by the selected values of T' and t0, correlate with the resulting CLIP scores for the generated images

4.4.7 **Role of Augmentation**

We employ patch-based CLIP losses to provide style guidance, utilizing denoised images. In our approach, N patches are randomly cropped and subjected to augmentation through a perspective function and random affine transformations. To assess the significance of augmentation, we conducted an ablation study on both augmentation and the number of patches. The results as illustrated in Figure 4-19, reveal that images generated without augmentation fail to sufficiently transform to a desirable requested style standard. Additionally, when N is less than 32, the transformation of the dog is inadequate in aligning with the target prompt "Painting by Cezanne". Based on these findings, we opted for $N = 32$ with augmentations, a choice that not only enhances the quality of results but also leads to a noteworthy reduction in inference time to 24 seconds.

Figure 4-19 Ablation study results on augmentation and the number on and the number of patches *N*. The target prompt is "Painting by Cezanne", e.g., our default setting W/, *N*=32

4.5 More Sampling Results

4.5.1 **Novelty of our proposed model compared to ZeCon model**

To demonstrate the superiority of our proposed model over ZeCon, we further conduct sampling results using various original images and text prompts. As depicted in Figure 4-20, ZeCon struggles to preserve facial features accurately. Additionally, Figure 4-21 shows that ZeCon fails to maintain the shape of objects and introduces noise into the stylized images. However, our proposed method works well.

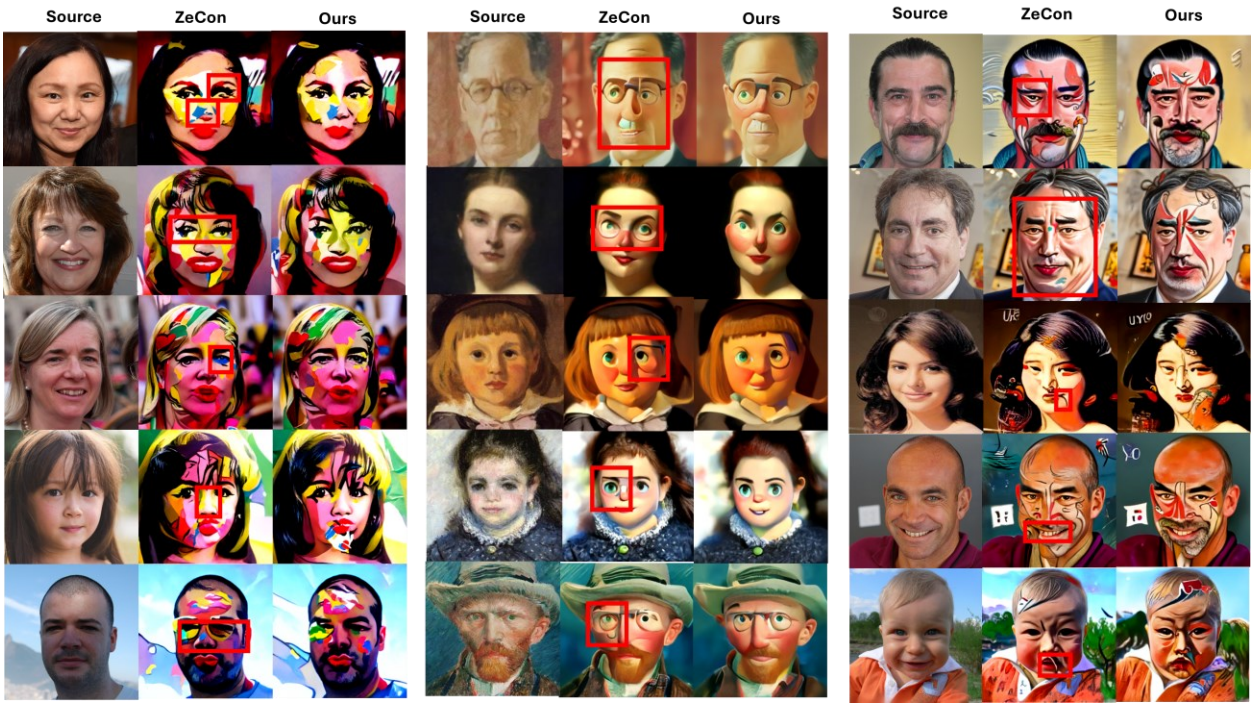

Figure 4-20 More results on our proposed model and ZeCon. The target prompt is "Pop art, Pixar, Ukiyo-E". ZeCon model has problem to preserve face part like eye, nose or lips.

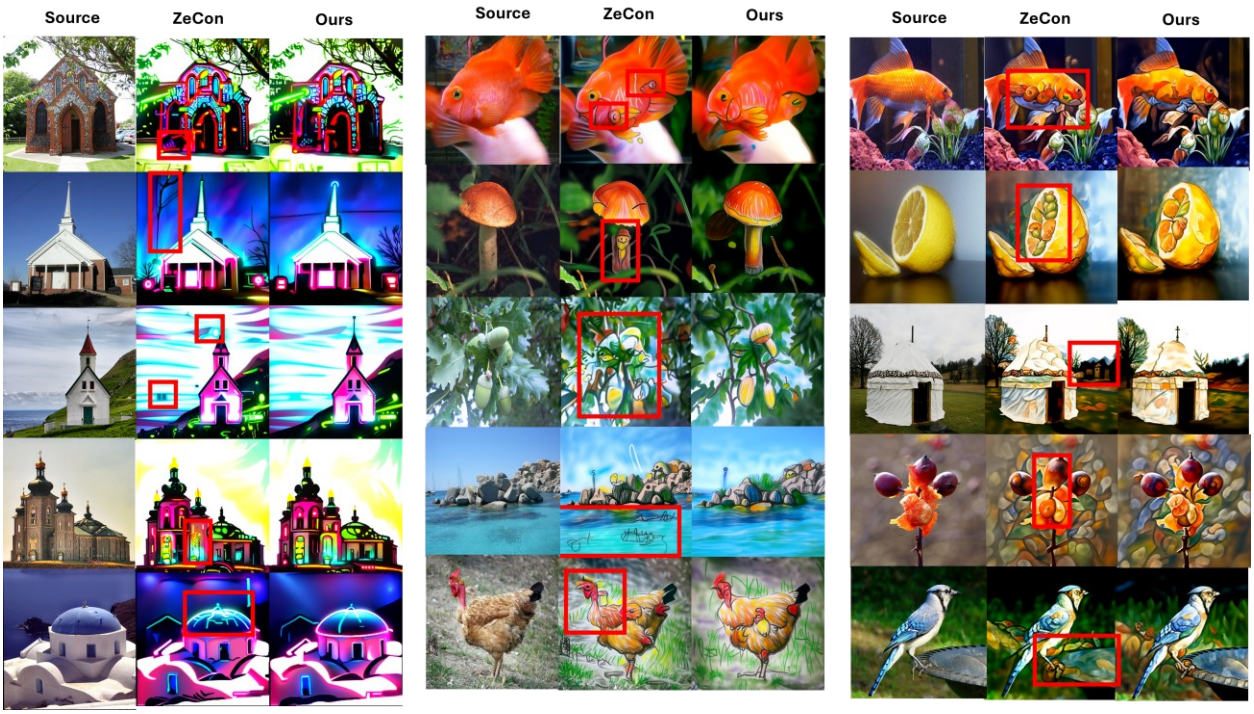

Figure 4-21 More results on our proposed model and ZeCon. The target prompt is "Neon Light, Sketch of Crayon, Portrait by Cezanne". ZeCon model has problem to preserve object part like building or object shape.

4.5.2 My own image result

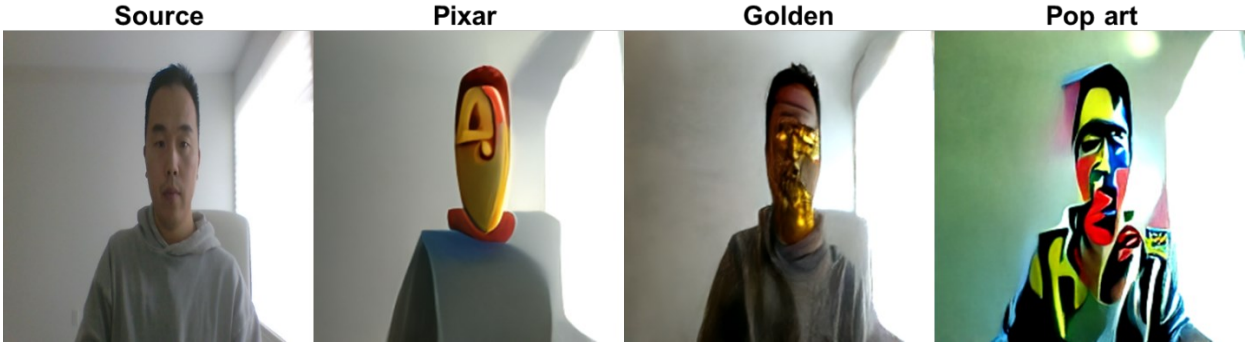

Figure 4-22 More results on my own image. The target prompt is "Pixar, Golden, Pop art". Pretrained model has no information on my own image so that reconstruction is not worked well.

When applying image style transfer with my own image, all resulting images are failed to preserve the original face as shown in Figure 4-22. This is because the faces are not reconstructed well under reverse process, and the pretrained model does not include my image data. Consequently, the stylized images are failed to maintain the original my own content features.

4.5.3 **Failure Case**

Figure 4-23 Failure case. The target prompt is "Staine Glass, Portrait by Cezanne, Pixar". Some of images are not preserving content's part well. Especially posed face is not working well.

When stylizing certain images, some of the stylized results fail to preserve the original content features as illustrated in Figure 4-23. This issue arises because our model's hyperparameters for controlling gradient loss are highly sensitive to variations in tuning values. Consequently, some images require adjustments to the hyperparameters to better preserve the content parts and to control stylization.

4.6 Summary

Our work showcases the outcomes derived from our research efforts, emphasizing the superior performance of our method in comparison to other GAN-based or other diffusion-based approaches. Our comprehensive evaluation encompasses diverse metrics and user preference considerations to provide our model's superior capabilities with content aware stylized images. A key metric in our analysis is patch wise contrastive learning with attention map guidance, which

keep the similarity between the generated image and the source image. Remarkably, our method consistently achieves higher CLIP scores, signifying a closer resemblance to the original face when compared to alternative GAN-based methods. This underscores the efficacy of our approach in accurately maintaining facial features and stylization during the diffusion reverse process. Moreover, we conduct an evaluation of the inference time for our model, measuring the speed at which our method can generate stylized images. Our approach exhibits competitive performance in terms of inference time, demonstrating efficiency and suitability for real-time applications comparison with other models. In addition to quantitative metrics, our assessment such as user study extended to the quality of images by enhanced details and realistic representations. These findings collectively highlight the robustness and versatility of our approach in style transfer tasks. Furthermore, our evaluation extends to assessing the overall realism of the stylized images. Leveraging subjective assessments and visual comparisons, our method consistently keeps content aware images. This underscores the capability of our approach to capture intricate details, leading to more believable and realistic representations.

By outperforming other methods across various metrics such as user study, inference time, CLIP score, and face ID similarity, our results unequivocally showcase the effectiveness and superiority of our proposed approach. These findings not only validate the advancements made but also emphasize the valuable contributions of our research in the realm of image style transfer. The demonstrated potential for practical applications in computer vision, graphics, and related domains further underscores the significance of our work in pushing the boundaries of this field.

Chapter 5 Conclusion and Future works

5.1 Conclusion

We propose a novel method for text-driven image style transfer with the aid of the diffusion model. Our approach involves QS-attn module, which achieves preserving contents of an image while transferring style to a target image. The QS-attn module selects content-related keys and queries by measuring feature significance by assessing the distribution entropy of an QS-Attn matrix formed with keys and queries from the encoder. We utilize contrastive loss with selected positive parts, negative parts, anchors to preserve content of source image.

The outcomes reveal an improved preservation of content, establishing superior effectiveness when compared to other diffusion models or GAN models. Additionally, extensive experiments have been conducted to optimize our model for the generation of high-quality stylized images.

We can summarize our contribution as follows:

- Diverse and Content-Preserving Image Generation: Our proposed model, equipped with the QS-attn module, excels in generating a wide range of synthetic images while effectively preserving the content from the source images. The editing capacity includes modifying facial attributes and even changing identities on animals, showcasing its versatility and applicability in image manipulation tasks.
- No Additional Training or Fine-Tuning Required: Unlike other diffusion models, our approach achieves high-quality images without the need for any additional training or fine-tuning, simplifying the model deployment process.

- Rapid Zero-Shot Image Generation Capability: Our model supports zero-shot image generation, enabling rapid interference for user-based tasks, providing a faster solution compared to other methods that may require more elaborate procedures.

In summary, our work has made significant contributions to the field of style transfer by introducing a novel methodology and demonstrating its effectiveness through extensive experimentation. These achievements not only mark advancements in the current state of style transfer but also lay the groundwork for future developments.

5.2 Future Works

Despite the notable success achieved in our work, there remain areas for further exploration and improvement. There are some limitations using only text prompting such as when similarity between prompting text and image is far from, stylized image is not working well and sometimes reveals prompted text on the stylized images.

Thus, a potential avenue for future research can involve leveraging multi-modality, i.e. image-, speech- and text-driven, to enhance the applicability in various situations. Exploring novel diffusion models and incorporating advanced attention map methods can be another focus, aimed at further enhancing the generation of high-quality and content-aware stylized images.

References

[1] Xia, Weihao, et al. "Gan inversion: A survey." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45.3 (2022): 3121-3138.

[2] Radford, Alec, et al. "Learning transferable visual models from natural language supervision." International conference on machine learning. PMLR, 2021.

[3] Tov, Omer, et al. "Designing an encoder for stylegan image manipulation." ACM Transactions on Graphics (TOG) 40.4 (2021): 1-14.

[4] Rombach, Robin, et al. "High-resolution image synthesis with latent diffusion models." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022.

[5] Kingma, Diederik P., and Max Welling. "An introduction to variational autoencoders." Foundations and Trends® in Machine Learning 12.4 (2019): 307-392.

[6] Kim, Gwanghyun, Taesung Kwon, and Jong Chul Ye. "Diffusionclip: Text-guided diffusion models for robust image manipulation." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.

[7] Kwon, Gihyun, and Ye, Jong Chul. "Diffusion-based image translation using disentangled style and content representation." arXiv preprint arXiv:2209.15264 (2022).

[8] Dosovitskiy, Alexey, et al. "An image is worth 16x16 words: Transformers for image recognition at scale." arXiv preprint arXiv:2010.11929 (2020).

[9] Mathilde Caron, Hugo Touvron, Ishan Misra, Herve Jegou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 9650–9660, 2021.

[10] Yang, Serin, Hyunmin Hwang, and Jong Chul Ye. "Zero-shot contrastive loss for text-guided diffusion image style transfer." arXiv preprint arXiv:2303.08622 (2023).

[11] Park, Taesung, et al. "Contrastive learning for unpaired image-to-image translation." Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16. Springer International Publishing, 2020.

[12] Hu, Xueqi, et al. "Qs-attn: Query-selected attention for contrastive learning in i2i translation." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.

[13] Abiodun, Oludare Isaac, et al. "State-of-the-art in artificial neural network applications: A survey." Heliyon 4.11 (2018).

[14] Aggarwal, Charu C. "Neural networks and deep learning." Springer 10.978 (2018): 3.

[15] u, Jiuxiang, et al. "Recent advances in convolutional neural networks." Pattern recognition 77 (2018): 354-377.

[16] Bebis, George, and Michael Georgiopoulos. "Feed-forward neural networks." Ieee Potentials 13.4 (1994): 27-31.

[17] Agarap, Abien Fred. "Deep learning using rectified linear units (relu)." arXiv preprint arXiv:1803.08375 (2018).

[18] Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation." Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. Springer International Publishing, 2015.

[19] Balestriero, Randall, et al. "A cookbook of self-supervised learning." arXiv preprint arXiv:2304.12210 (2023).

[20] Michelucci, Umberto. "An introduction to autoencoders." arXiv preprint arXiv:2201.03898 (2022).

[21] Chen, Ting, et al. "A simple framework for contrastive learning of visual representations." International conference on machine learning. PMLR, 2020.

[22] Harshvardhan, G. M., et al. "A comprehensive survey and analysis of generative models in machine learning." Computer Science Review 38 (2020): 100285.

[23] Jaiswal, Ashish, et al. "A survey on contrastive self-supervised learning." Technologies 9.1 (2020): 2.

[24] Gatys, Leon A., Alexander S. Ecker, and Matthias Bethge. "A neural algorithm of artistic style." arXiv preprint arXiv:1508.06576 (2015).

[25] Goodfellow, Ian, et al. "Generative adversarial nets." Advances in neural information processing systems 27 (2014).

[26] Karras, Tero, Samuli Laine, and Timo Aila. "A style-based generator architecture for generative adversarial networks." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019.

[27] Isola, Phillip, et al. "Image-to-image translation with conditional adversarial networks." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.

[28] Zhu, Jun-Yan, et al. "Unpaired image-to-image translation using cycle-consistent adversarial networks." Proceedings of the IEEE international conference on computer vision. 2017.

[29] Patashnik, Or, et al. "StyleCLIP: Text-driven manipulation of stylegan imagery." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021.

[30] Gal, Rinon, et al. "StyleGAN-NADA: CLIP-guided domain adaptation of image generators." ACM Transactions on Graphics (TOG) 41.4 (2022): 1-13.

[31] Crowson, Katherine, et al. "VQGAN-CLIP: Open domain image generation and editing with natural language guidance." European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022.

[32] Yu, Jiahui, et al. "Vector-quantized image modeling with improved VQGAN." arXiv preprint arXiv:2110.04627 (2021).

[33] Kwon, Gihyun, and Jong Chul Ye. "CLIPStyler: Image style transfer with a single text condition." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.

[34] Ching, Wai-Ki, and Michael K. Ng. "Markov chains." Models, algorithms and applications (2006).

[35] Song, Yang, et al. "Score-based generative modeling through stochastic differential equations." arXiv preprint arXiv:2011.13456 (2020).

[36] Ho, Jonathan, Ajay Jain, and Pieter Abbeel. "Denoising diffusion probabilistic models." Advances in neural information processing systems 33 (2020): 6840-6851.

[37] Song, Jiaming, Chenlin Meng, and Stefano Ermon. "Denoising diffusion implicit models." arXiv preprint arXiv:2010.02502 (2020).

[38] Voynov, Andrey, Kfir Aberman, and Daniel Cohen-Or. "Sketch-guided text-to-image diffusion models." ACM SIGGRAPH 2023 Conference Proceedings. 2023.

[39] Nichol, Alex, et al. "Glide: Towards photorealistic image generation and editing with text-guided diffusion models." arXiv preprint arXiv:2112.10741 (2021).

[40] Wei, Rongting, Chunxiao Fan, and Yuexin Wu. "Diffusion-Adapter: Text Guided Image Manipulation with Frozen Diffusion Models." International Conference on Artificial Neural Networks. Cham: Springer Nature Switzerland, 2023.

[41] Zhang, Chenshuang, et al. "Text-to-image diffusion model in generative ai: A survey." arXiv preprint arXiv:2303.07909 (2023).

[42] Chefer, Hila, et al. "Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models." ACM Transactions on Graphics (TOG) 42.4 (2023): 1-10.

[43] Kawar, Bahjat, et al. "Imagic: Text-based real image editing with diffusion models." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.

[44] An, J.; Joe, I. Attention Map-Guided Visual Explanations for Deep Neural Networks. Appl. Sci. 2022, 12, 3846.

[45] Hessel, Jack, et al. "Clipscore: A reference-free evaluation metric for image captioning." arXiv preprint arXiv:2104.08718 (2021).

[46] Yang, Qingsong, et al. "Low-dose CT image denoising using a generative adversarial network with Wasserstein distance and perceptual loss." IEEE transactions on medical imaging 37.6 (2018): 1348-1357.

[47] Janocha, Katarzyna, and Wojciech Marian Czarnecki. "On loss functions for deep neural networks in classification." arXiv preprint arXiv:1702.05659 (2017).

[48] 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. In International Conference on Machine Learning, pages 8748–8763. PMLR, 2021. 1, 2, 4.

[49] Liu, Ziwei, et al. "Large-scale celebfaces attributes (celeba) dataset." Retrieved August 15.2018 (2018): 11.

[50] Choi, Yunjey, et al. "Stargan v2: Diverse image synthesis for multiple domains." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020.

[51] Yu, Fisher, et al. "Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop." arXiv preprint arXiv:1506.03365 (2015).

[52] Deng, Jia, et al. "Imagenet: A large-scale hierarchical image database." 2009 IEEE conference on computer vision and pattern recognition. Ieee, 2009.

[53] Mohammad, Saif, and Svetlana Kiritchenko. "Wikiart emotions: An annotated dataset of emotions evoked by art." Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018). 2018.

[54] Deng, Jiankang, et al. "Arcface: Additive angular margin loss for deep face recognition." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019.

[55] Hyungjin Chung, Byeongsu Sim, and Jong Chul Ye. Come-closer-diffuse-faster: Accelerating conditional diffusion models for inverse problems through stochastic contraction. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 12413–12422, 2022. 1, 11