Query-Selected Global Attention for Text guided Image Style Transfer using Diffusion Model

Jungmin Hwang
School of EECS, Faculty of Engineering
University of Ottawa
Ottawa, ON, Canada
jhwan091@uottawa.ca

Won-Sook Lee
School of EECS, Faculty of Engineering
University of Ottawa
Ottawa, ON, Canada
wslee@uottawa.ca

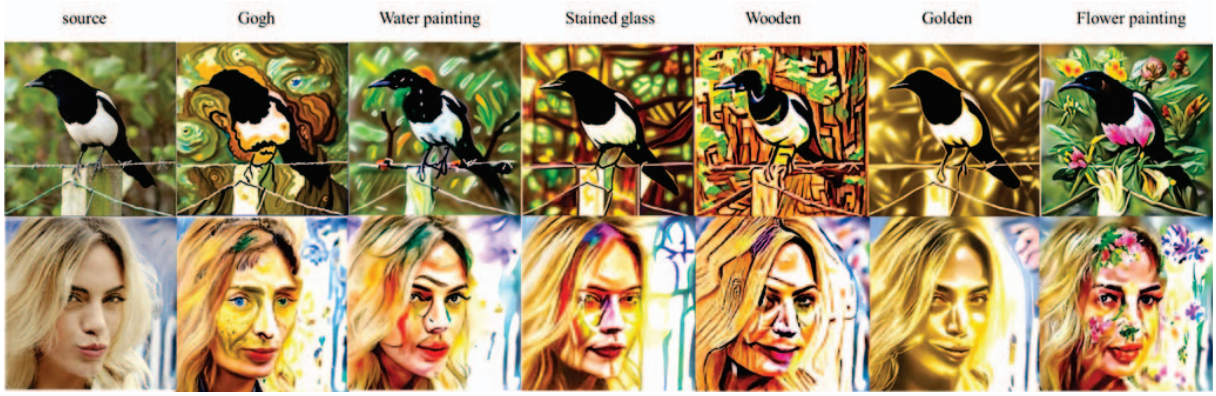

Fig. 1. This style transfer technique demonstrates remarkable outcomes when applied to various artistic styles. This method effectively retains the fundamental structure of the original images while seamlessly converting them into the desired aesthetic styles.

***Abstract*— Diffusion models have gained tremendous interest in image generation. Additionally, guided text methods for manipulating source images have shown successful progress. However, research on style transfer using diffusion models is still ongoing to address the trade-off between style transfer and content preservation. One representative solution to the issue is contrastive learning in a self-supervised manner, which is useful for extracting specific features from the same location on source and generated images for every pixel. However, there are instances where it is necessary to preserve certain areas, which contain more information from the source image compared to other areas in the image. Therefore, we propose anchoring the areas for preservation and intentionally selecting features at the anchor points through a query-selected global attention method. This enables our method to generate an image that preserves the content of the source while transferring the style without the need for additional fine-tuning or auxiliary network. Our diffusion model follows a simple architecture to enhance image quality and speed up inference time, in comparison to other diffusion methods. Our experimental results also demonstrate superior performance.**

Keywords*— *Diffusion, Style Transfer, Query Selection, Global Attention

I. Introduction

GAN inversion methods combined with CLIP have gained popularity for zero-shot image manipulation guided by text prompts for style transfer. Challenges include limited GAN inversion performance [1], especially in handling diverse types of images with novel poses, views, and details. SOTA encoder-based GAN inversion methods such as pix2pix [2], cycleGAN [3], and Contrastive Unpaired Image-to-Image Translation (CUT) [4] often fail to reconstruct images with unexpected attributes or details, leading to unintended changes in the input content for style transfer. Reconstruction issues are exacerbated in the case of images with high variance, such as those from datasets like LSUN-Church [5], ImageNet [6], etc. Fortunately, diffusion models, such as denoising diffusion probabilistic models, DDPM [7] and score-based generative models [8], have shown success in image generation and manipulation tasks, surpassing other generative models like VAEs [9], flows [10], auto-regressive models [11], and GANs [12]. Inspired by the success of diffusion models in various image tasks, researchers have explored their application in image-to-image style transfer. DiffusionCLIP [13] is proposed as a CLIP-guided robust image manipulation method using diffusion models. The method involves converting an input image to latent noises through forward diffusion. Latent noises can be inverted nearly perfectly to the original image using reverse diffusion if the score function remains the same. The key idea is to fine-tune the score function in the reverse diffusion process using a CLIP loss based on text prompts, allowing control over the attributes of the generated image. DiffusionCLIP addresses the limitations of existing GAN inversion methods, especially in handling diverse images, by combining diffusion models and CLIP for robust image manipulation guided by text prompts. The method demonstrates effectiveness in various scenarios and outperforms existing baselines in terms of accuracy and robustness. However, conditional diffusion models for image-to-image style transfer require paired datasets with matched source and target styles. Unconditional diffusion models have challenges in maintaining content due to the stochastic nature of the reverse sampling procedure. DiffusionCLIP leverages pretrained diffusion models and CLIP encoder for text-driven image style transfer but requires additional fine-tuning for the

979-8-3503-5409-6/24/$31.00 ©2024 IEEE
DOI 10.1109/CAI59869.2024.00207

desired style. DiffuseIT [14] uses disentangled style and content representation but faces a trade-off between transforming texture and maintaining content. In order to solve the problems, ZeCon loss [15] which is patch-wise contrastive loss between the input image and generated images is added to achieve zero-shot style transfer while preserving semantic content. Unlike DiffusionCLIP, ZeCon doesn't require additional training, achieving effective content preservation in a zero-shot manner. Additionally, it provides more accurate texture modification compared to DiffuseIT. However, the model based on overall structure of CUT selects randomly an anchor point from translated features and computing contrastive loss for mutual information maximization so that it can't choose ideal anchor selection and can have limited receptive field of anchor features. Our proposed solution involves inserting the QS-Attn module into diffusion model without introducing additional parameters. The module evaluates feature significance using features from encoder as queries and keys to calculate the attention matrix in the source domain. The distribution entropy of the attention matrix is computed as a metric to measure feature significance at different locations. Intuitive illustrations using a heat map of entropy values are provided based on input images. The paper aims to quantitatively measure the significance of each anchor feature. The entropy of each row in the attention matrix is calculated, and those with smaller values are retained to form the QS-Attn matrix. Thus, our primary contributions include the introduction of the QS-Attn module to address issues in anchor selection and limited receptive fields. Our paper proposes a method to quantitatively measure feature significance and selects relevant features for the contrastive loss, aiming for more accurate guided text to image translation with diffusion model.

II. RELATED WORKS

A. Diffusion model on DDPM and DDIM

Diffusion model operates by incrementally introducing Gaussian noise in a Markov chain forward process. Subsequently, a trained noise estimation model is employed to iteratively denoise and generate clean samples from the latent noise. DDPM directly samples x_t from x_0 by adding Gaussian noise with $\beta_t \in (0, 1)$ at time $t \in [1, ..., T]$,

$$x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t}\varepsilon \quad (1)$$

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:

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

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. DDIM [16] addresses this issue by adopting a sampling process that ensures the preservation of content:

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

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

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

B. Style Transfer using Diffusion models through CLIP

The diffusion model 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 a representative model, it is DiffusionCLIP. However, it 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 adopts a disentangled approach to style and content representation, drawing inspiration from the slicing Vision Transformer [17]. 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 necessitates an extra network for the computation of content losses. In order to solve this problem, Zero-shot Contrastive (ZeCon) loss is proposed 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 applied between the input image and the generated images as CUT. Unlike DiffusionCLIP, the method does not necessitate additional training as well as achieves more precise texture modification while upholding content integrity. However, CUT 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 from in the Horse to Zebra task. It's 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. Our aim is to purposefully select the anchor q and compute contrastive loss specifically on the salient features that encompass more domain-specific information.

III. MAIN CONTRIBUTIONS

A. Preliminaries on Ze-Contrastive loss function

DDPM's and DDIM's sampling equation follows as Eq (2, 3). ZeCon loss function is defined as follows as Eq (5).

$$\ell_{zecon}(\hat{x}_{0,t}, x_0) = \mathbb{E}_{X_0}[\sum_l \sum_s \ell(\hat{z}_l^s, z_l^s, z_l^{S/s})] \quad (5)$$

The loss function is based on CUT loss, which can get spatial feature from source image during training. With -the feature, the encoder within CUT model can preserve the content of

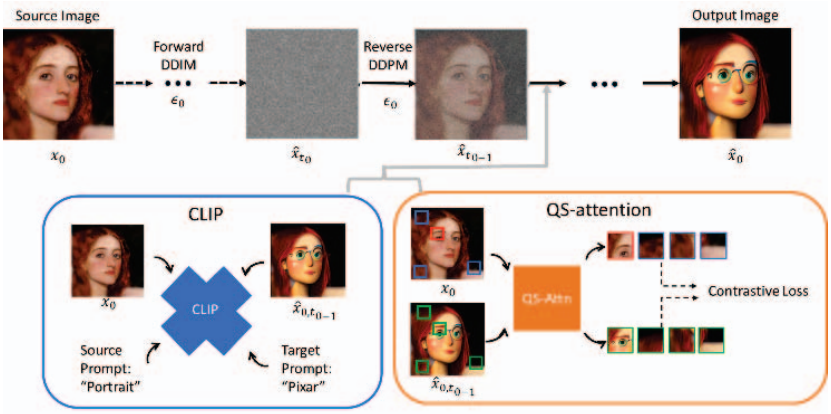

Fig. 2. Query-Selected Global Attention Diffusion model's reverse process conceptual structure, i.e. From Unseen domain to Pixar.

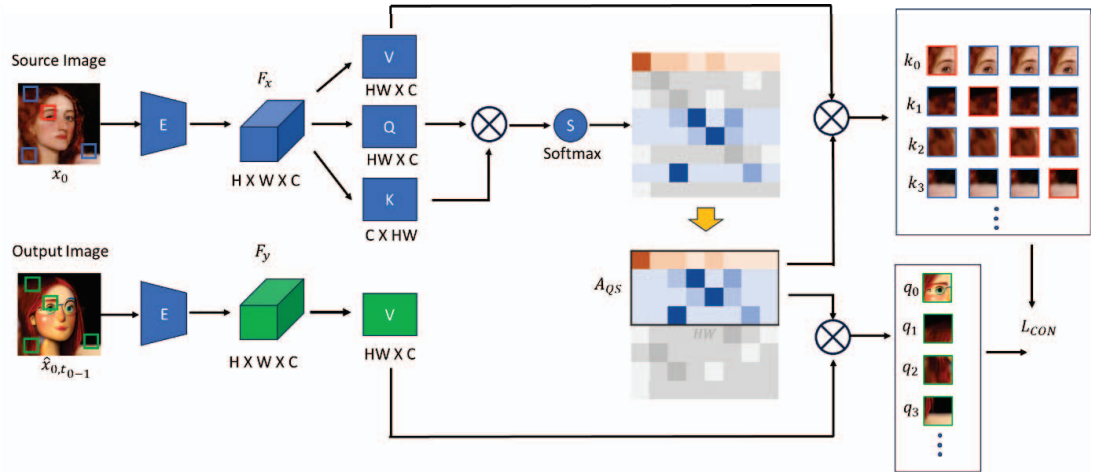

Fig. 3. The details structure of QS-Attn module. The encoder E extracts each feature from source image and Target image. First, Fx is reshaped and computed to derive the attention matrix. Each row in the matrix is sorted by its metric of the significance, and the selected N rows forming the A_{QS}. We 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.

source image through patch-wise contrastive loss, which maximizes the mutual information between positive pairs like pixels from the same location and minimizes the mutual information between negative pairs like pixels from different locations. In addition, U-Net [18] noise predictor has spatial information. Without additional training, it is possible to get spatial information as shown in Figure. In order to apply the loss, the pixels of feature maps are randomly selected and used to cross-entropy loss during every reverse timestep.

B. *QS-Attn for Contrastive Learning*

Rather than relying on a straightforward random strategy, we utilize an attention-based approach [19]. 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 elucidated in the subsequent two subsections. According to attention-based approach, some features do not reflect the domain characteristics due to random selection of patches. We propose three attention methods to select intentionally anchor q and then, compute the loss with more domain-specific patch information during diffusion reverse process.

C. *QS-Attn for Contrastive Learning*

Rather than relying on a straightforward random strategy, we utilize an attention-based approach [19]. 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 elucidated in the subsequent two subsections. According to attention-based approach, some features do not reflect the domain characteristics due to random selection of patches. We propose three attention methods to select intentionally anchor q and then, compute the loss with more domain-specific patch information during diffusion reverse process.

D. 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.6, To select all the significant queries, the rows of A_g are sorted by the entropy H_g in the ascending order, and the smallest N rows are selected as the QS-Attn matrix $A_{QS} \in R \times HW$. Note that A_{QS} is fully determined by the features.

$$H_g(i) = -\sum_{j=1}^{HW} A_g(i,j) log A_g(i,j) \quad (6)$$

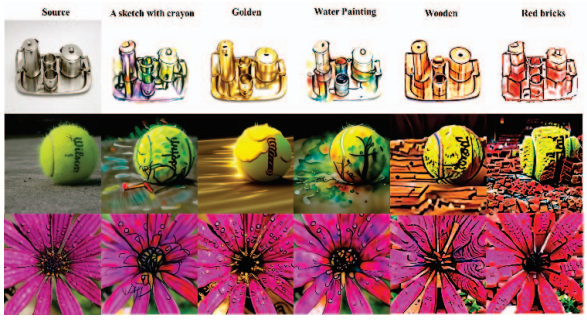

Fig. 4. Image Style Transfer through Query-Selected Global Attention Diffusion model's reverse process.

IV. Experiments

A. Implementation Details

We use pretrained models and use various dataset such as CelebHQ [20] datasets, ImageNet for style transfer. All images of human faces, dogs, bedrooms, and churches are 256 x 256 size. We also use images from the test-set on these datasets for experiments. Our model can use both types such as DDPM or DDPM of diffusion process. Through experiments, we adopt DDIM method as the forward process and then, adopts DDPM method as the reverse process. When T is 1000 as the total time step, we follow to reduce the step size (50, 25) for Image manipulation as other paper is proven. From this latent xt0, the stylized output image is sampled through diffusion processes. It is better to preserve spatial features from source image and to manipulate generated image with new style. In addition, we can save inference time by reducing the number of iterations.

B. Comprehensive studies

Our method can translate various style using text guiding method while maintaining contents of source images as Fig.4, i.e, A sketch with crayon, Golden, Water Painting, Wooden, Red bricks. 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 Fig. 5 and Table 1. The third row of Figure 5 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 are less colorful than others even if maintaining overall contents of images. Additionally, the diffusion model has to 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 shown in Fig. 5. 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 its identity. These results are confirmed with user study results presented in Table 1, 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, i.e, our method 30s, DiffusionCLIP 30 min (including fine-tune), DiffuseIT 48s.

Fig. 5. Comparison Our method with other diffusion model for Image Style Transfer.

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 1. Comparison Our method with other diffusion model for Image Style Transfer with User Preference.

C. Ablations studies

Our diffusion model can manipulate unseen domain in image style transfer. The qualitative results for image manipulation are shown in Fig. 6, where our method can translate pixar, zombie, even multiple manipulation while preserving contents of the images.

Fig. 6. Image Style Transfer for unseen domain and multiple manipulation

V. CONCLUSION

In this work, we propose a novel method for text-driven image style transfer with the aid of the diffusion model. Specifically, our approach involves query selection global attention, which achieve maintaining contents of images while translating style to target images. To further improve style quality of the translated image, we use image augmentation and self-supervised learning method to leverage denoising diffusion model to generate the high quality output. Extensive experiments demonstrate the superiority of our method for generating high-fidelity images than other diffusion models.

REFERENCES

[1] Wang, Tengfei, et al. "High-fidelity gan inversion for image attribute editing." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.

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

[3] 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.

[4] 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.

[5] 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).

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

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

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

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

[10] Rezende, Danilo, and Shakir Mohamed. "Variational inference with normalizing flows." International conference on machine learning. PMLR, 2015.

[11] Germain, Mathieu, et al. "Made: Masked autoencoder for distribution estimation." International conference on machine learning. PMLR, 2015.

[12] Goodfellow, Ian, et al. "Generative adversarial networks." Communications of the ACM 63.11 (2020): 139-144.

[13] 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.

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

[15] 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).DD

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

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

[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] 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.

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