Image Style Transfer
CLIPstyler

CVPR

2022

CLIPstyler: Image Style Transfer with a Single Text Condition

Gihyun Kwon[1] Jong Chul Ye[1,2]
Dept. of Bio and Brain Engineering[1], Kim Jaechul Graduate School of AI[2], KAIST
{cyclomon, jong.ye}@kaist.ac.kr

Figure 1. Our style transfer results on various text conditions. Translated images have spatial structure of the content images with realistic textures corresponding to the text.

Abstract

Existing neural style transfer methods require reference style images to transfer texture information of style images to content images. However, in many practical situations, users may not have reference style images but still be interested in transferring styles by just imagining them. In order to deal with such applications, we propose a new framework that enables a style transfer 'without' a style image, but only with a text description of the desired style. Using ... only with a single text condition. Specifically, we propose a patch-wise text-image matching loss with multiview augmentations for realistic texture transfer. Extensive experimental results confirmed the successful image style transfer with realistic textures that reflect semantic query texts.

1. Introduction

Style transfer aims to transform a content image by transferring the semantic texture of a style image. The seminar

Kwon, G., & Ye, J. C. (2022). Clipstyler: Image style transfer with a single text condition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 18062-18071).

Image Style Transfer

CLIPstyler

Background & Goal

Previous research limitation

- Existing neural style transfer methods require reference style images
- Tried manipulating images with text conditions, but the performance of the embedding model is limited and the manipulation is restricted to specific content domains

Goal

- Propose a image style transfer method to deliver the semantic textures of text conditions using CLIP

Kwon, G., & Ye, J. C. (2022). Clipstyler: Image style transfer with a single text condition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 18062-18071).

Image Style Transfer

CLIPstyler

CLIP

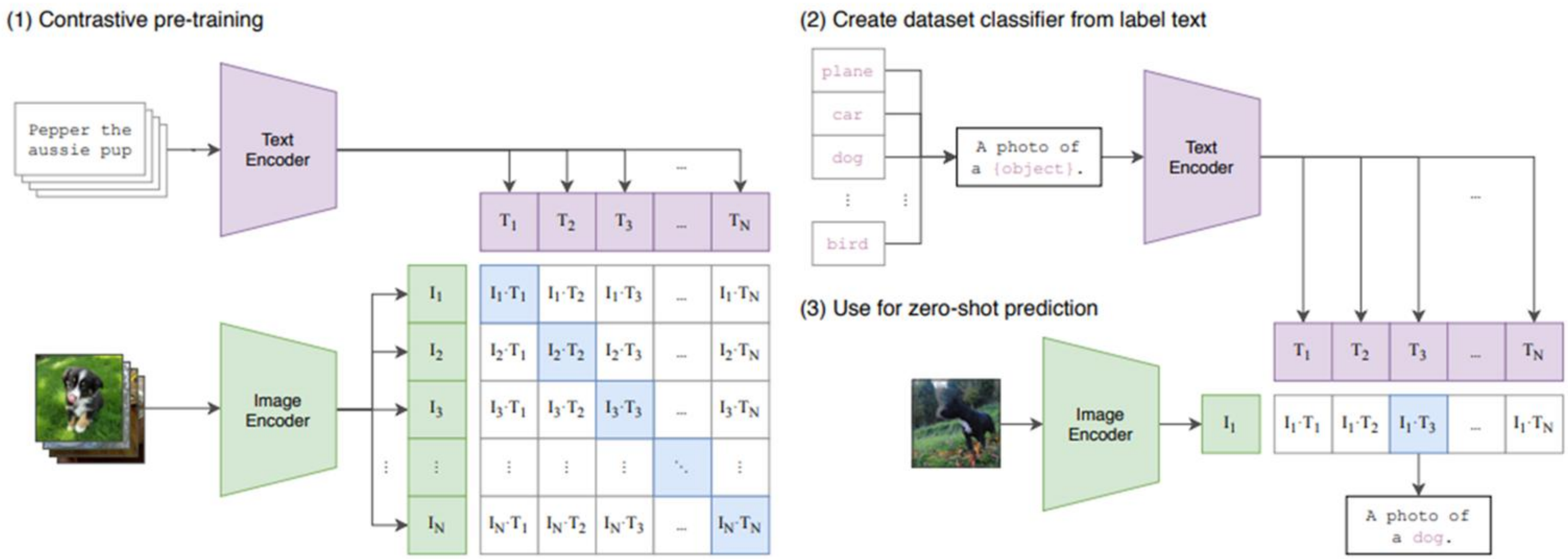


Fig 1. Overview of CLIP

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021, July). Learning transferable visual models from natural language supervision. In *International conference on machine learning* (pp. 8748-8763). PmLR.

Image Style Transfer

CLIPstyler

Network Architecture

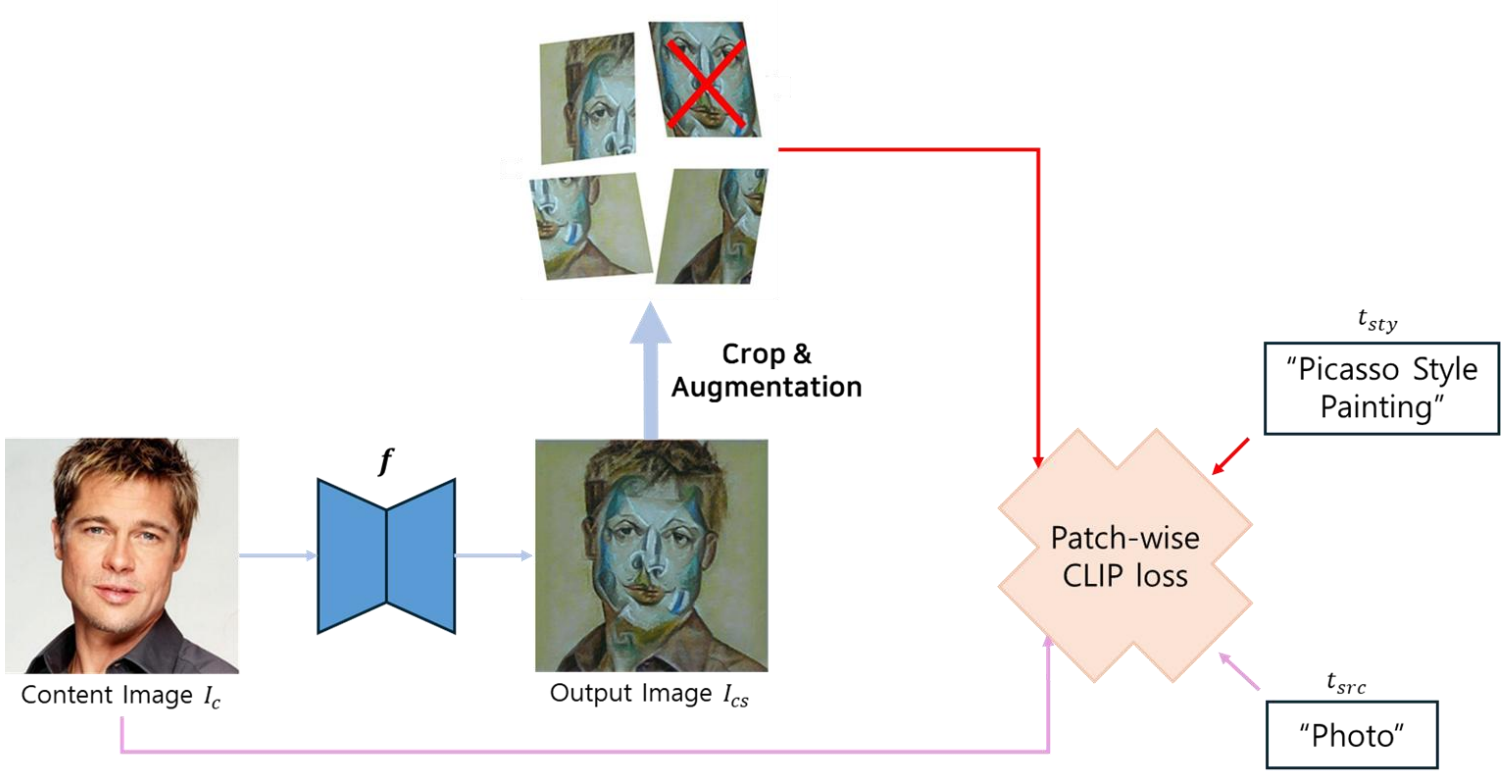


Fig 2. Overview of CLIPStyler

Kwon, G., & Ye, J. C. (2022). Clipstyler: Image style transfer with a single text condition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 18062-18071).

Image Style Transfer

CLIPstyler

Loss functions

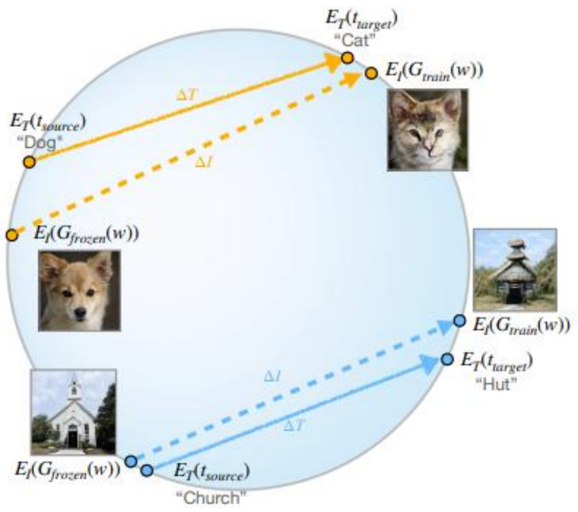
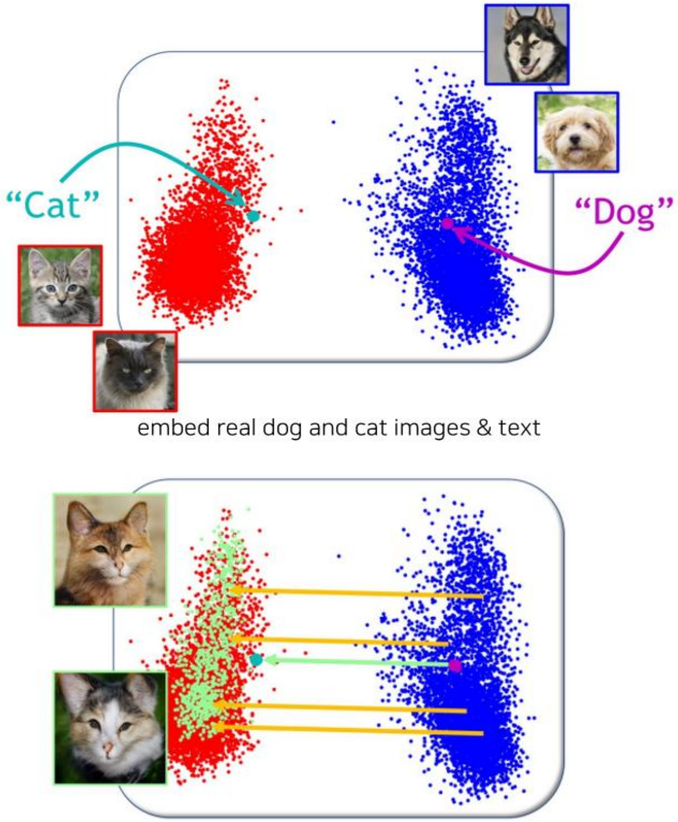


Fig 3. Overview of Directional CLIP loss



Global CLIP loss

- Employ the directional CLIP loss that aligns the CLIP-space direction between the source and output

$$\Delta T = E_T(t_{sty}) - E_T(t_{src}),$$
$$\Delta I = E_I(f(I_c)) - E_I(I_c),$$
$$L_{dir} = 1 - \frac{\Delta I \cdot \Delta T}{|\Delta I||\Delta T|},$$

Eq 1. Directional CLIP loss

Gal, R., Patashnik, O., Maron, H., Bermano, A. H., Chechik, G., & Cohen-Or, D. (2022). Stylegan-nada: Clip-guided domain adaptation of image generators. *ACM Transactions on Graphics (TOG)*, *41*(4), 1-13.

Image Style Transfer

CLIPstyler

Loss functions

PatchCLIP loss

- CLIPStyler's goal is to apply the semantic texture of t_{sty} → Global CLIP loss doesn't perfectly match to CLIPStyler
- Propose a PatchCLIP loss that is a method of calculating loss by using patches of an image

Augmentation

- Using Augmentations on each patch assist the network to represent more vivid and diverse textures
- Using Perspective Augmentation, all patches are guided to have the same semantic when viewed in multiple points

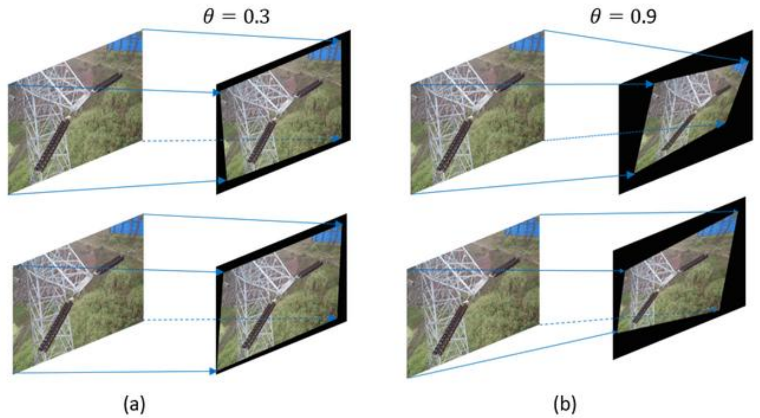


Fig 4. Example of Perspective Augmentation

Kwon, G., & Ye, J. C. (2022). Clipstyler: Image style transfer with a single text condition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 18062-18071).

Image Style Transfer

CLIPstyler

Loss functions

Threshold Rejection

- Due to stochastic randomness of path sampling and augmentations, method suffer from over-stylization
- Include regularization to reject the gradient optimization process for high-scored patches

Additional loss

- Content loss: To maintain the content information of input image → calculating the MSE between features of content and output images
- Total Variation loss: To alleviate the side artifacts from irregular pixels

$$\triangle T = E_T(t_{sty}) - E_T(t_{src}),$$

$$\triangle I = E_I(aug(\hat{I^i_{cs}})) - E_I(I_c)$$

$$l^i_{patch} = 1 - \frac{\Delta I \cdot \Delta T}{|\Delta I||\Delta T|},$$

$$L_{patch} = \frac{1}{N}\sum_i^N R(l^i_{patch}, \tau)$$

$$\text{where} \quad R(s,\tau) = \begin{cases} 0, & \text{if} \quad s \le \tau \\ s, & \text{otherwise} \end{cases}$$

Eq 2. Patch-wise CLIP loss with Threshold rejection

$$\mathcal{L}_{\text{content}} = \frac{1}{N}\sum_{i=1}^N (F_l(I_{cs})_i - F_l(I_c)_i)^2$$

Eq 3. Content Loss

$$\mathcal{L}_{TV} = \sum_{i,j}((I(i+1,j) - I(i,j))^2 + (I(i,j+1) - I(i,j))^2)$$

Eq 4. Total Variation Loss

Kwon, G., & Ye, J. C. (2022). Clipstyler: Image style transfer with a single text condition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 18062-18071).

Image Style Transfer

CLIPstyler

Experiments

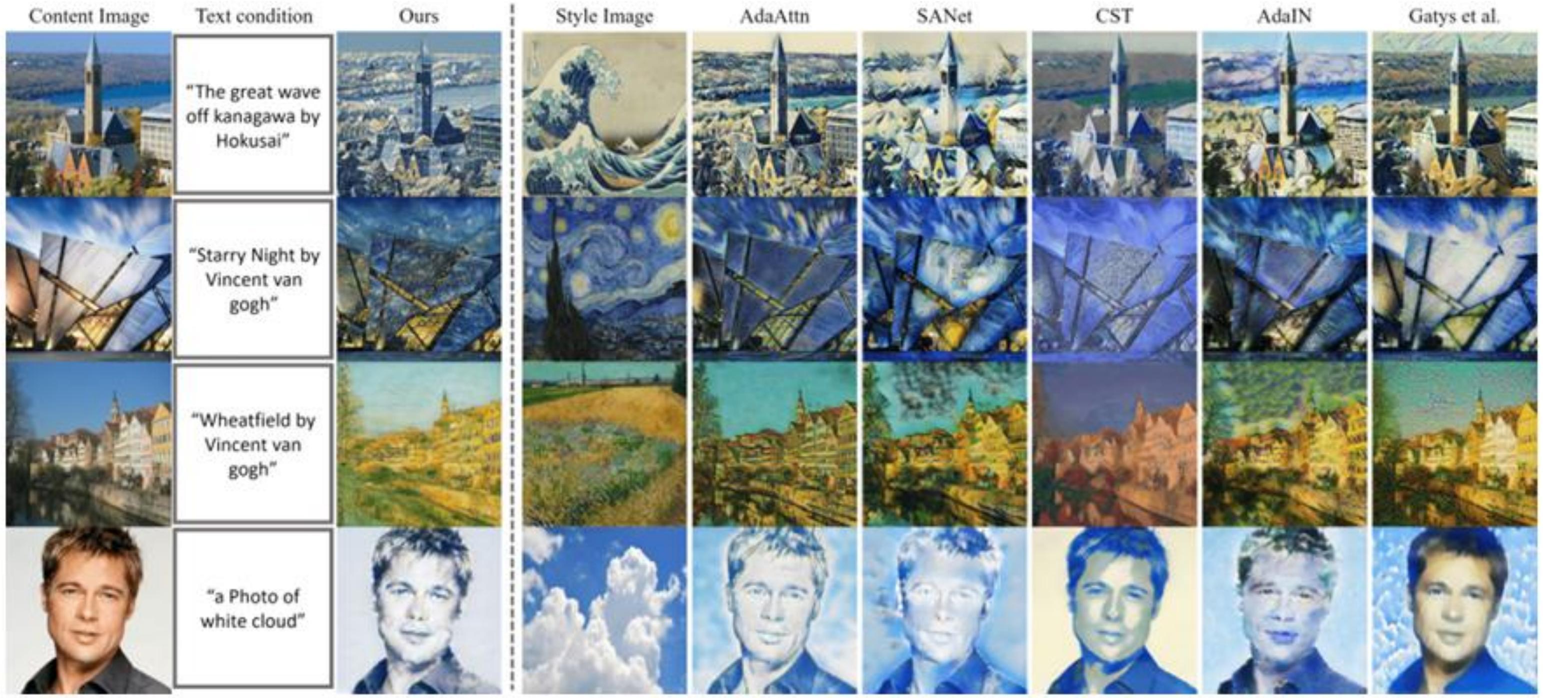


Fig 5. Comparison results with baseline style transfer models

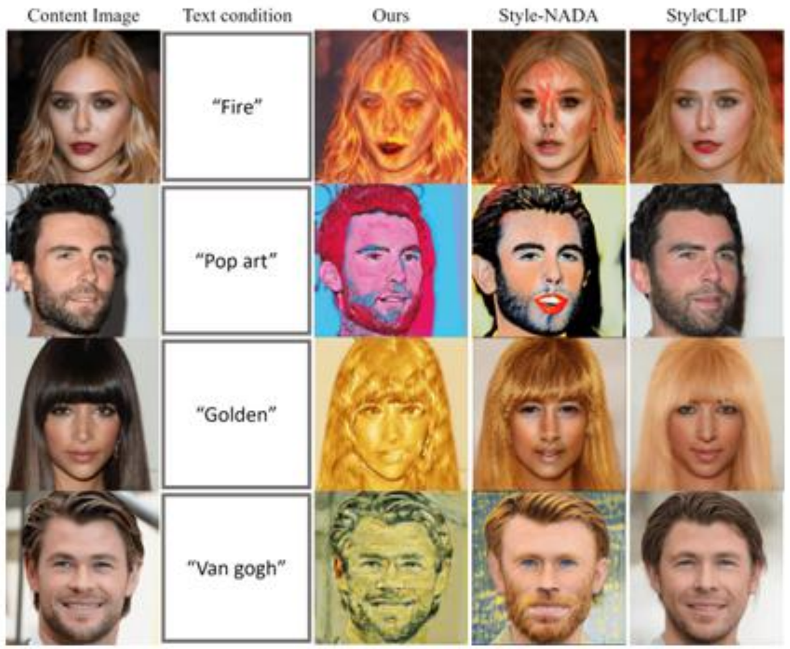


Fig 6. Comparison results with other text-guide d manipulation models

Kwon, G., & Ye, J. C. (2022). Clipstyler: Image style transfer with a single text condition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 18062-18071).

Image Style Transfer

CLIPstyler

Experiments

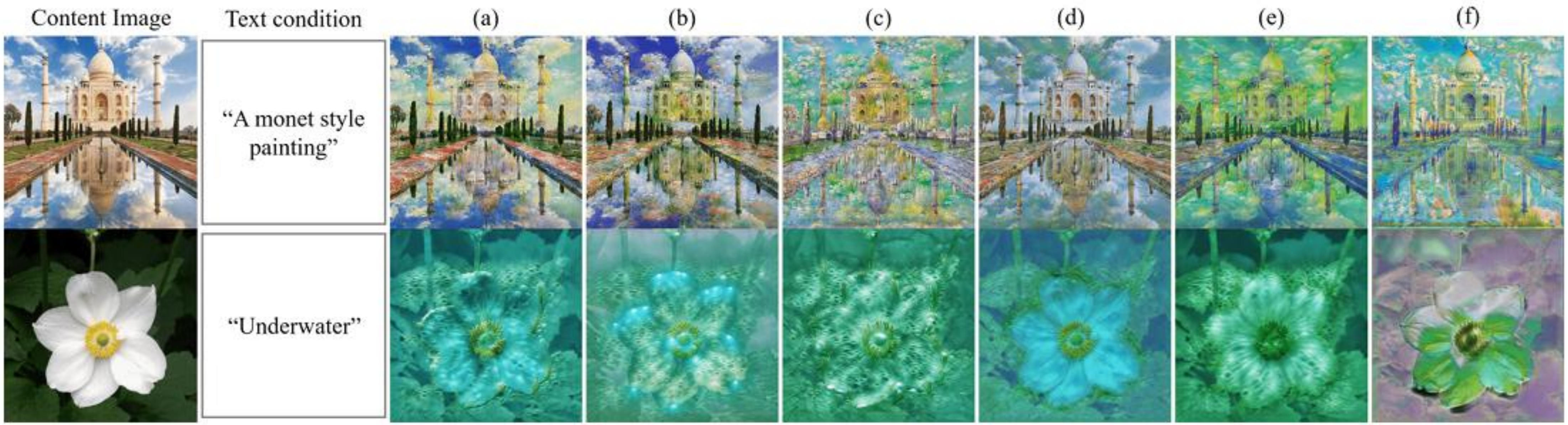


Fig 7. Ablation study results

Kwon, G., & Ye, J. C. (2022). Clipstyler: Image style transfer with a single text condition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 18062-18071).

Image Style Transfer

CLIPstyler

Implications & Limitations

Implications

- Proposed a novel image style transfer framework to transfer the semantic texture information only using text c ondition

Limitations

- It is difficult to maintain consistency because the generated style may be different even when the same text pr ompt is entered

Kwon, G., & Ye, J. C. (2022). Clipstyler: Image style transfer with a single text condition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 18062-18071).