961

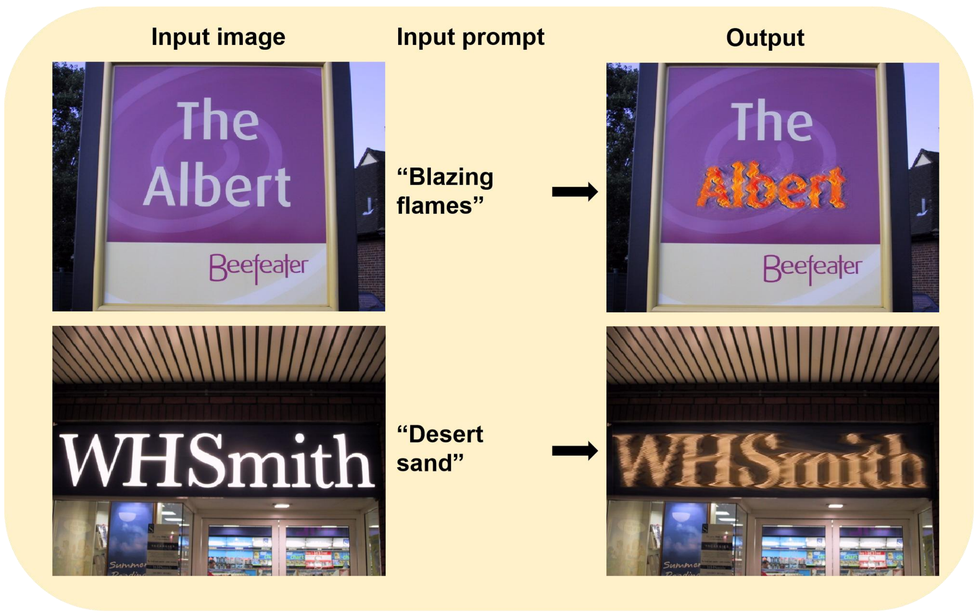
Font Style Translation in Scene Text Images with CLIPstyler

Honghui Yuan, Keiji Yanai
The University of Electro-Communication, Tokyo, JAPAN

Abstract

We proposed a new framework named FontCLIPstyler to realize **scene text style transformation**.

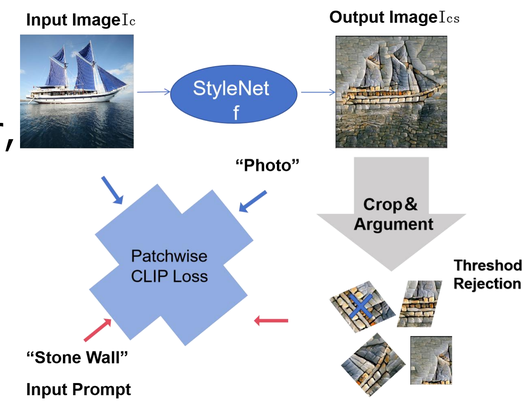
Our method could freely change the style of text areas in scene images **using prompts** based on CLIPStyler[1].



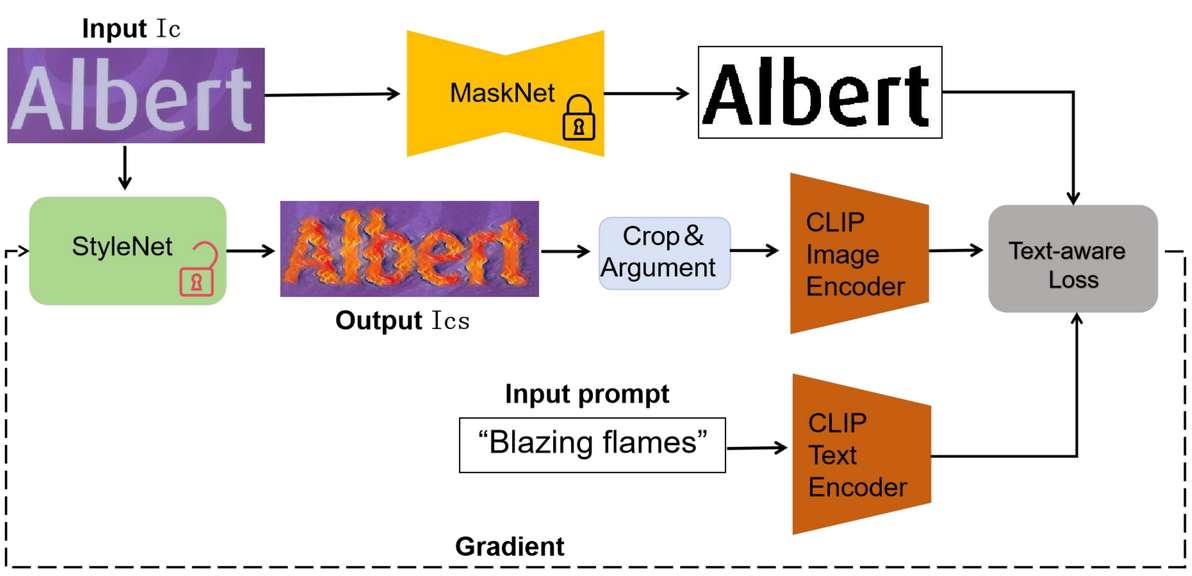
Methodology

Previous Work: ClipStyler[1]

- For general image style transfer, and **not applicable to text images**.
- **Unable to transform styles to specific areas** in the image.



Proposed Method



1. Network for extracting text masks in images(**MaskNet**)
   - Pre-trained with 2000 images using Unet
2. Transfer semantic style of prompts(**StyleNet**)
   - Optimized with loss function Text-aware Loss
3. Only transform the style of the text area without changing the background(**Text-aware Loss**)

Text-aware Loss

$$L_{ta} = \lambda_d L_{distance} + \lambda_p L_{patch} + \lambda_r L_{recon}$$

$L_{distance}$ Allows style transformation within a limited region based on distance transformation of the input image.

L_{patch} Transfer semantic style of prompts to text area

L_{recon} Background reconstruction

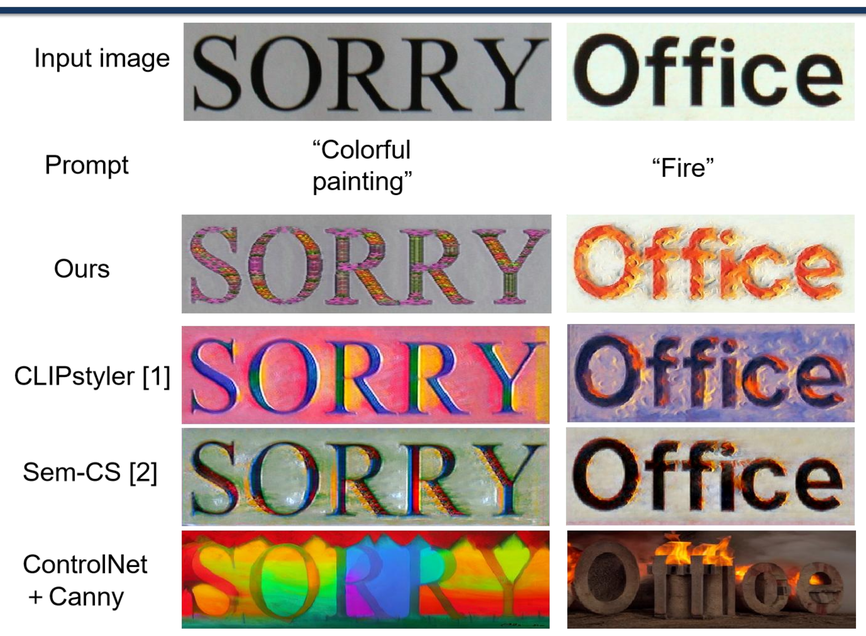
[1] Kwon, G., Ye, J.C.: Clipstyler: Image style transfer with a single text condition. In: CVPR. 2022
[2] Kamra, C.G., Mastan, I.D., Gupta, D.: Sem-cs: Semantic clipstyler for text-based image style transfer. In: ICIP. 2023

Results

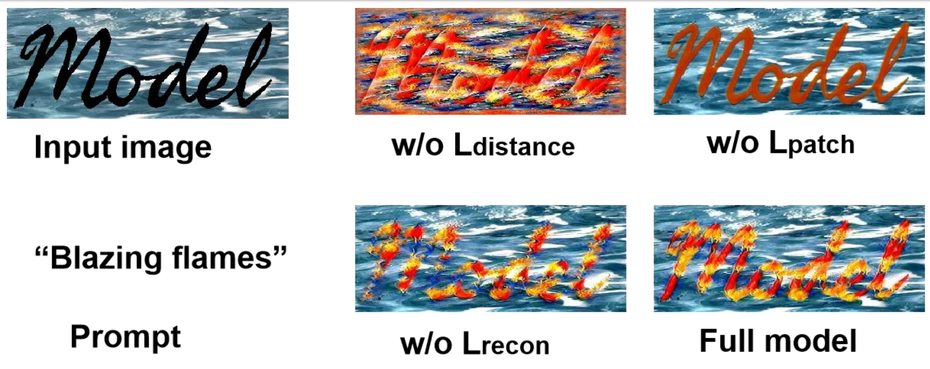
1. Qualitative evaluation

Input scene images, Input prompts and Generated results

2. Comparison with existing methods



3. Ablation Studies



4. Quantitative evaluation

- NIMA、DISTS : Evaluating image quality from human perspective
- LPIPS、FID : Similarity between images
- CLIP score : Consistency of images and prompts

Ablation Results

	DISTS↓	NIMA↑	LPIPS↓	FID↓	CLIP SCORE↑
w/o Ldistance	0.4997	4.5776	0.6082	910.50	0.2428
w/o Lpatch	**0.3132**	4.4650	**0.5489**	**318.72**	0.2149
w/o Lrecon	0.4196	4.7785	0.5950	713.20	0.2536
Full model	0.4075	**4.9550**	0.5846	485.00	**0.2583**

The effectiveness of each component was demonstrated

Comparison with existing methods

	DISTS↓	NIMA↑	LPIPS↓	FID↓	CLIP SCORE↑
CLIPstyler	0.3901	4.6776	0.7171	**372.40**	0.1848
Sem-CS	0.3838	4.7322	0.6984	460.79	0.2065
Ours	**0.3324**	**4.8632**	**0.6667**	445.55	**0.2101**

Natural stylized scene text images could be generated with high consistency from input prompts.