CLIPstyler: Image Style Transfer with a Single Text Condition (CVPR 2022)

Team Member: Zekai Zhang, Chuanruo Ning, Rundong Luo

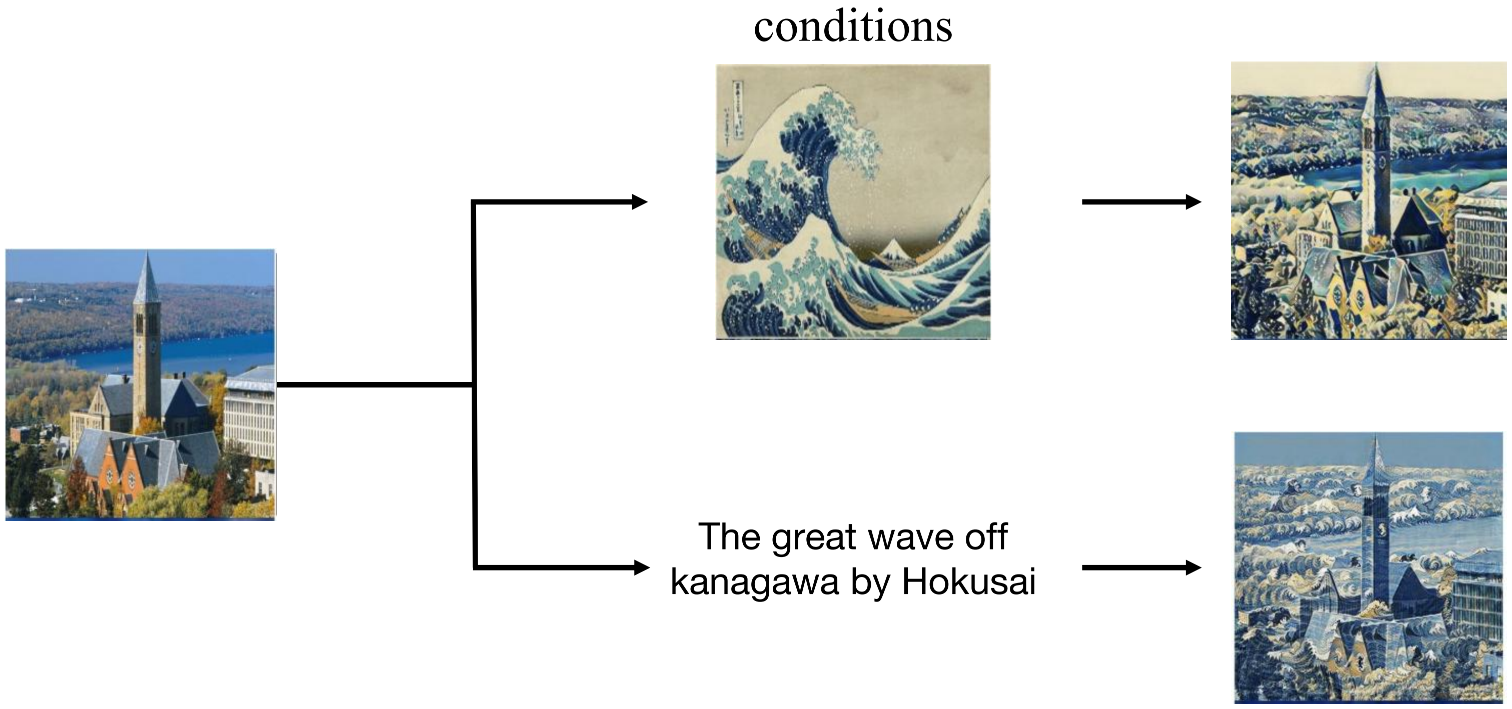
Contents

Contents

Background

Image Style Transfer: Transfer a content image to a given target style.

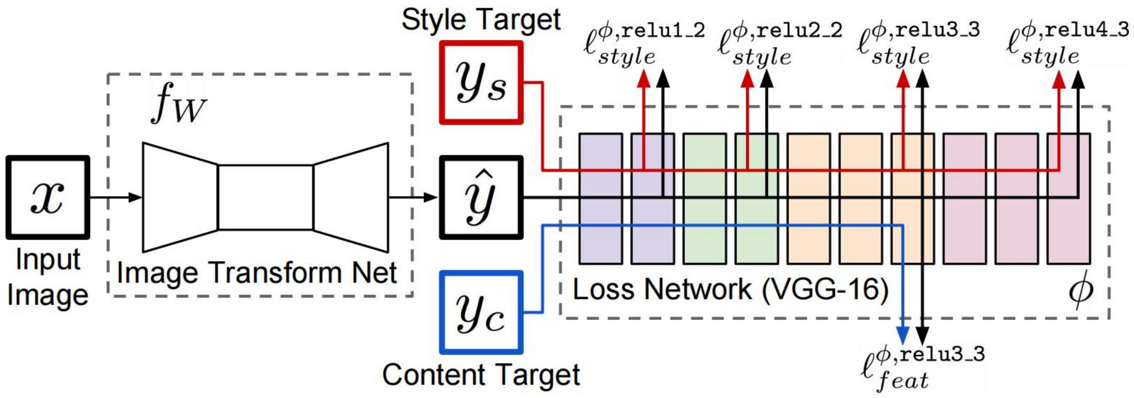
- Transfer under image conditions (a pure computer vision task)
- Transfer under text conditions (a tough multi-model task)



Background

Image-conditioned Style Transfer

- How to extract and preserve content?
  - Use pre-trained convolutional networks to extract content
  - Align feature activations to preserve content
- How to extract and apply style?
  - Use the Gram Matrix between extracted feature activations to represent style
  - Minimize the Gram Matrix difference to align style



Background

Text-conditioned Style Rransfer

- How to extract and preserve content?
  - Use pre-trained convolutional networks to extract content
  - Align feature activations to preserve content
- How to extract and apply style?
  - An intuitive idea: use **CLIP** to extract the style of both the content image and the given text, then minimize their difference.

$$L_{global} = D_{CLIP}(f(I_{\text{content}}), t_{\text{sty}}),$$

where f is a generative model.

Contents

- Background and Related Works
- Method
- Experiments
- Extensions
- Timeline

Method

CLIP loss: aligns the CLIP-space direction between the text-image pairs

- Intuitive version: cosine distance in the CLIP space (not necessarily works well)

$$L_{global} = D_{CLIP}(f(I_c), t_{sty})$$

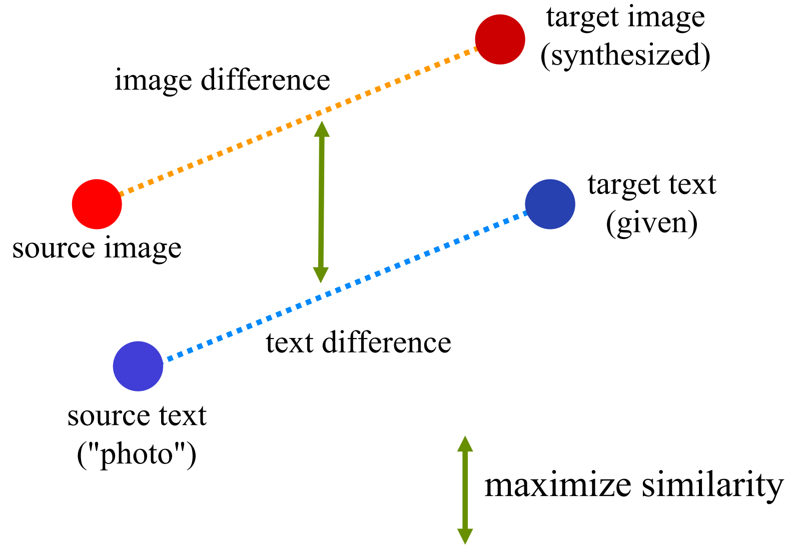
- Advanced version: directional CLIP loss

$$\Delta I = E_I(f(I_c)) - E_I(I_c)$$

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

$$L_{dir} = 1 - \frac{\Delta I \cdot \Delta T}{|\Delta I||\Delta T|}$$

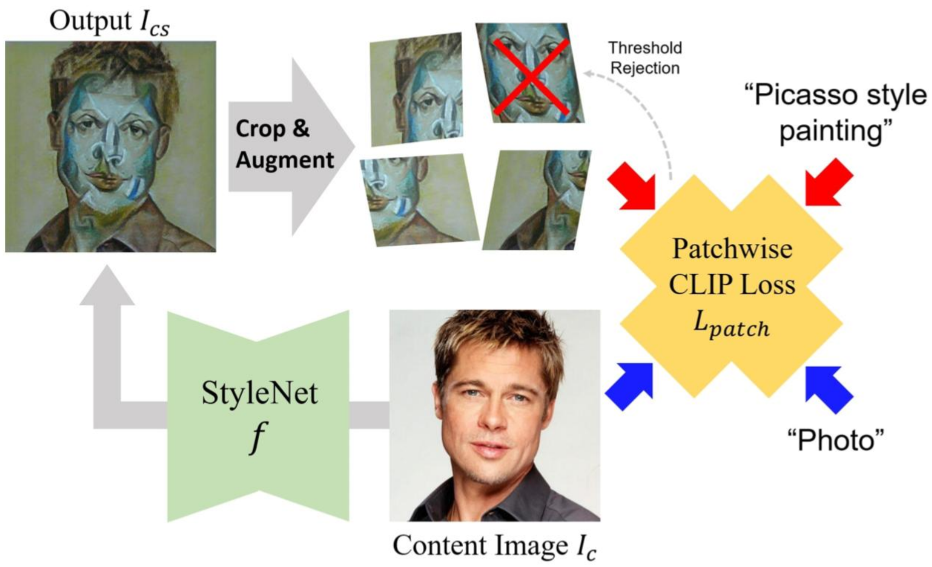
Idea: align the difference between source image-text pair and the difference between target image-text pair in the embedding space. For natural image, source text is set as "photo".



Method

Patch-wise CLIP loss: compute the CLIP loss between image patches

- Prevent over-stylization



Content loss

- Preserve contents

Total variance loss

- Avoid abnormal neighboring pixels

Contents

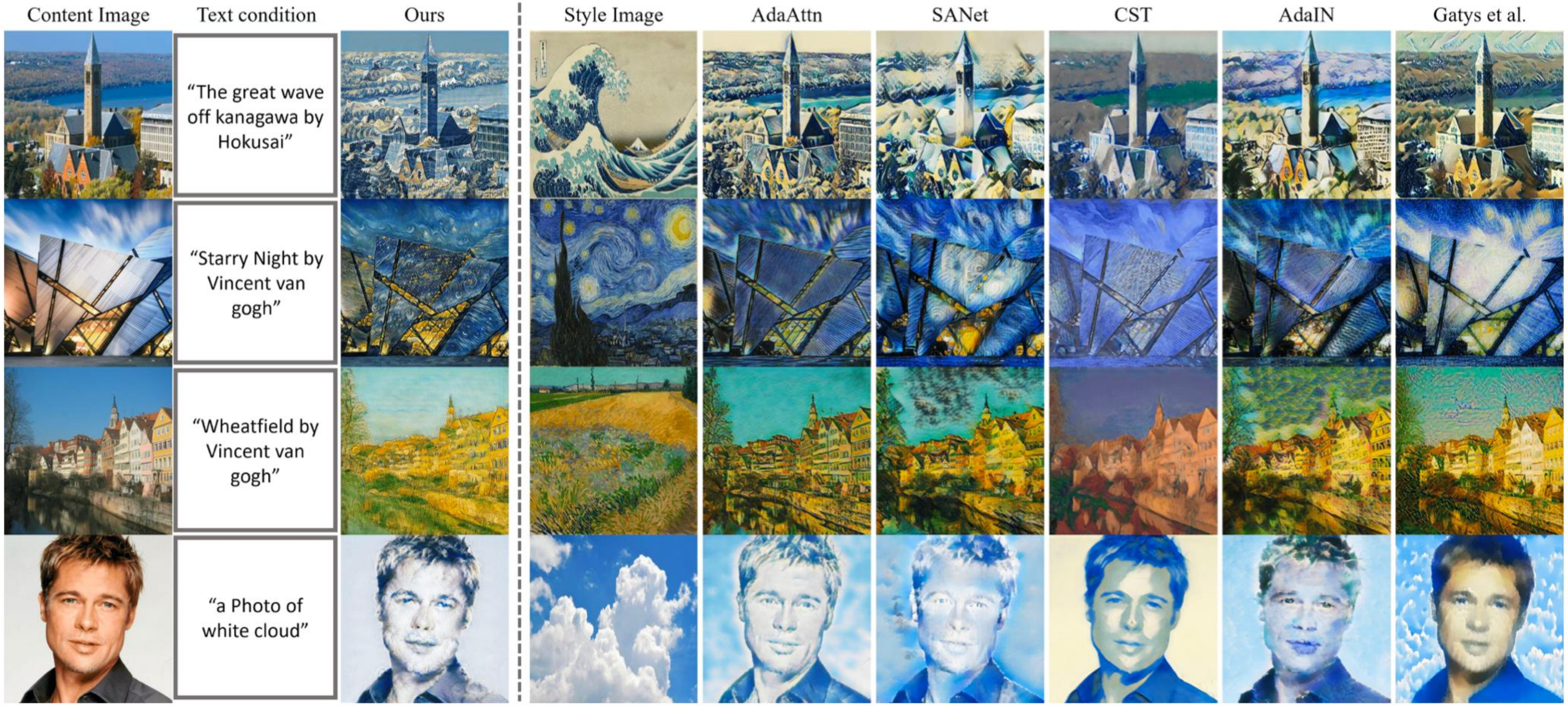
Experiments

- Qualitative evaluations

Content Image | “Mosaic” | “A sketch with crayon” | “White wool” | “Green crystal” | “A graffiti style painting”

Experiments

- Comparison with image-conditioned methods



Experiments

input

target text: sketch with black pencil

input

target text: a photo of night

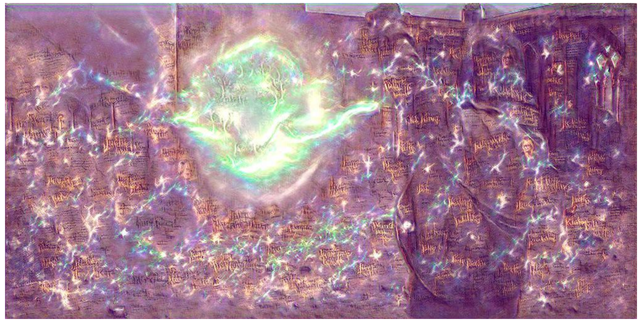
target text: night

Contents

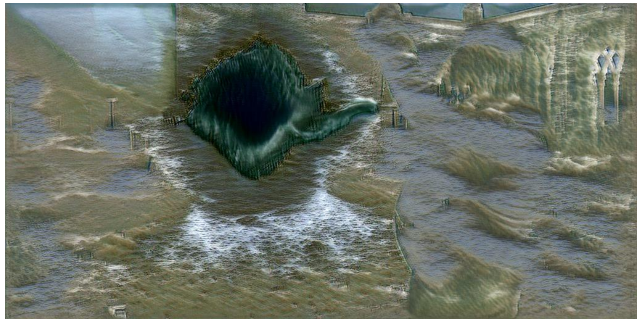
Extensions

Failure on complicated scenes

source image



source text: photo
target text: magical world

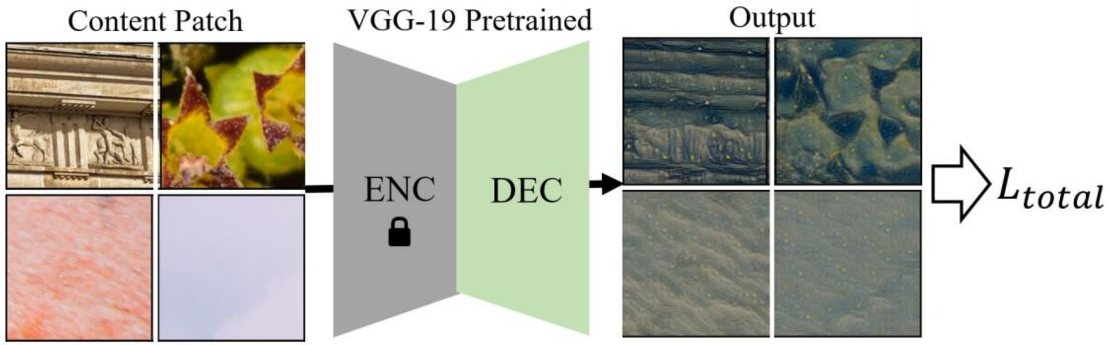


source text: fire burst
target text: tidal surge

Over-stylization: we only want part of image to be stylized.
Improvement: Detect region of interest (that best matches the source text), then apply stylization. Only compute stylization loss on interested regions.

Extentions

Fast Style Transfer: train once, and then fast inference



Video demo: Harry Potter versus Voldemort

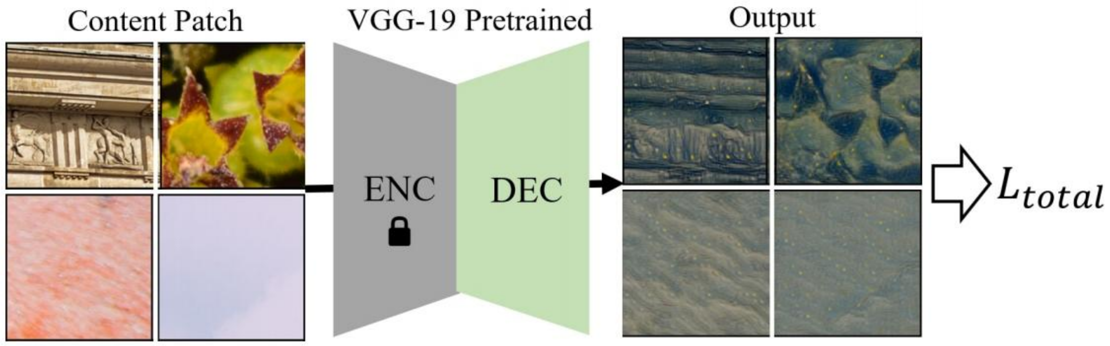
Target text: sketch with black pencils

Original audio clipped from: *Harry Potter and the Deathly Hallows: Part 2*

Harry Potter and the Deathly Hallows: Part 2
with text "Sketch with black pencil"

Extentions

Fast Style Transfer: train once, and then fast inference



Video demo: Harry Potter versus Voldemort

Target text: sketch with black pencils

Original audio clipped from: *Harry Potter and the Deathly Hallows: Part 2*

Improvement: For video translation, temporal consistency loss could be introduced.

Contents

Timeline

Finished

- Data preprocessing, debugging, etc.
- All experiments mentioned in the original paper
- Future work proposals

Todo

- Put the proposals into practice
- Overcome the faliure cases
- Apply temporal loss on videos

Thanks for listening!