2021 ASIAN CONFERENCE ON INNOVATION IN TECHNOLOGY (ASIANCON 2021)

One-shot style transfer using Wasserstein Autoencoder

Paper ID: 1014

Hidemoto Nakada, Hideki Asoh

AIRC, AIST

Background

- Image Style Transfer
  - Input：Content Image and Style Image
  - Output：Content rendered with the specified style
- Issue
  - The original method takes too long to render the image
- This work
  - Instant rendering in any style

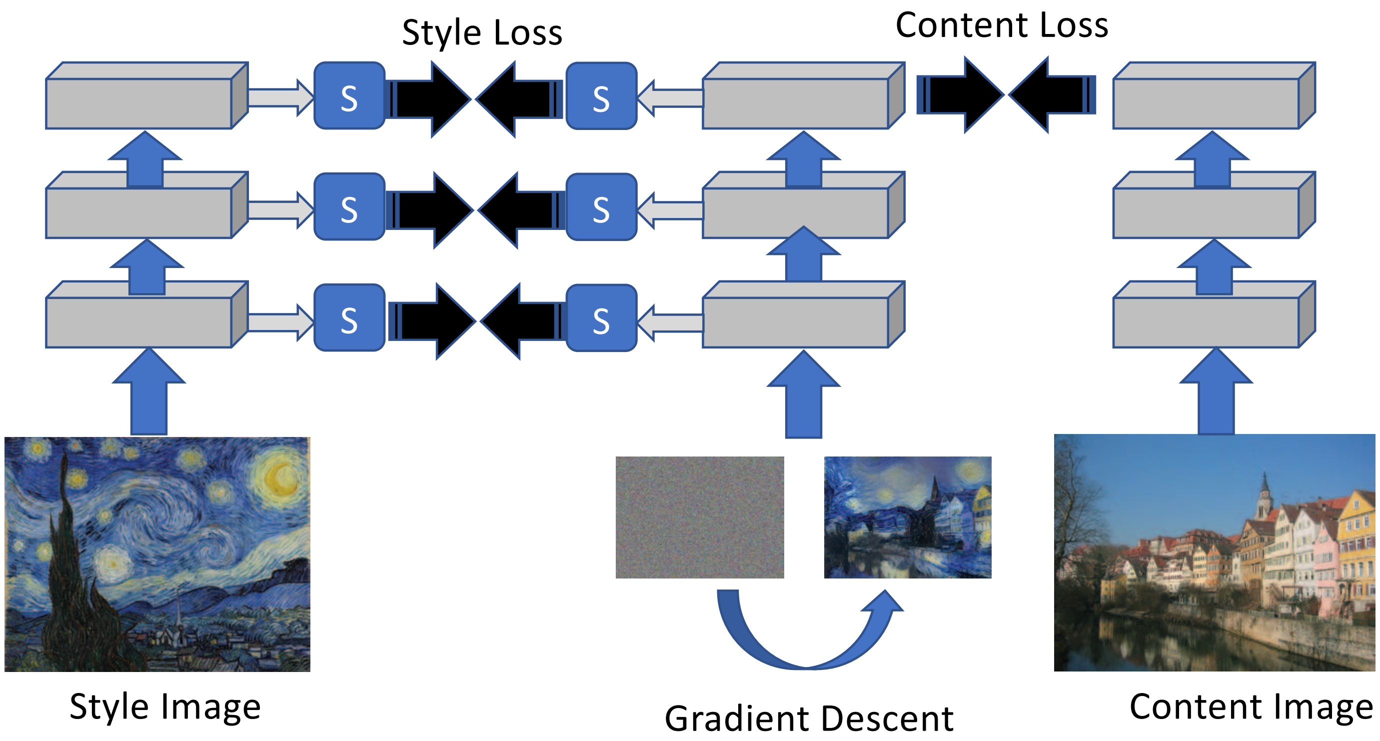
Style Image

Content Image

Images are from [Gatys, et.al '16]

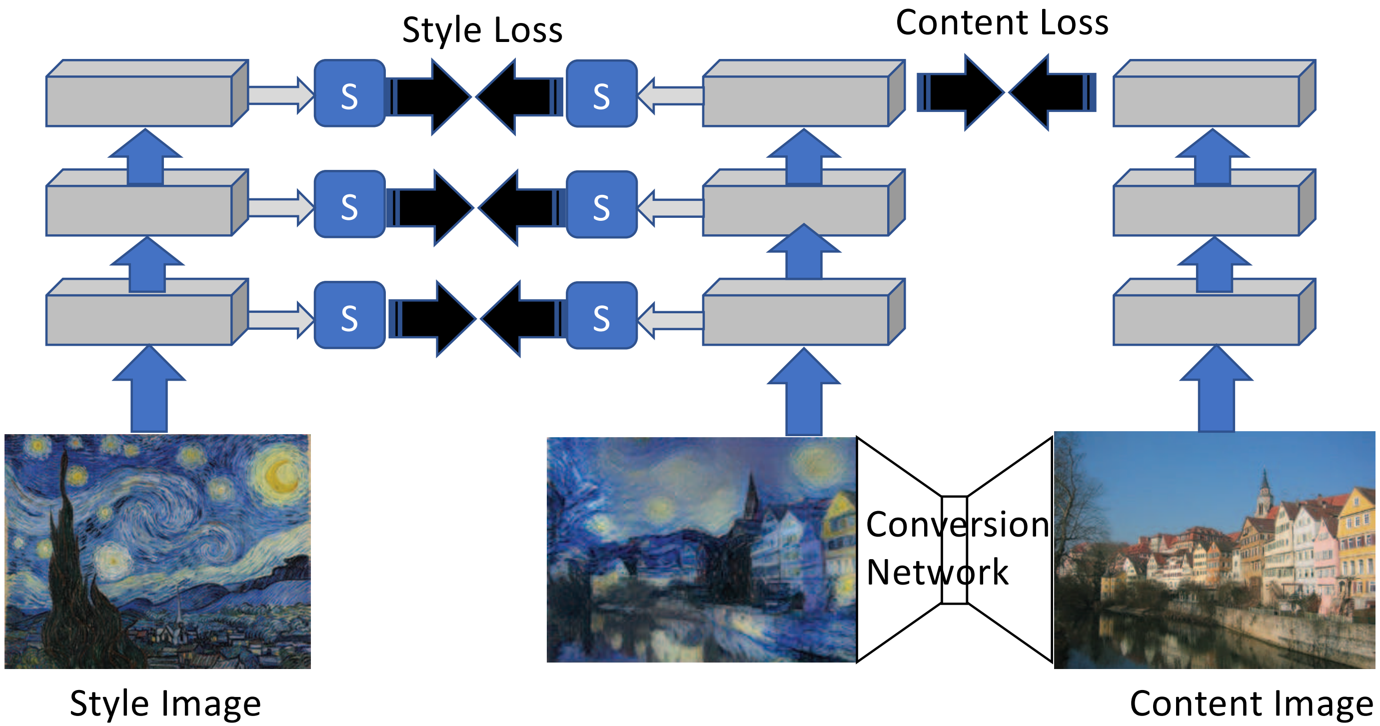
The original method [Gatys' et.al, '16]

- Gradually refines image with SGD
  - Takes long time to render



Conversion Network [Johnson '17]

- Train the conversion network, instead of each image
  - Once the network has been trained, the conversion is instant
  - The conversion network is style dependent
    - We have to train a network for each style.



Motivation and Goal

- Existing methods takes too long time
  - Gatys, et al.: Optimize the image itself
    - For each style and each content.

  - Johnson, et al.: Optimize style converters for each style
    - For each style, can reuse the converter network.
    - For any new style, need to train the converter.

→ **Enable instant style transfer for any style and content**

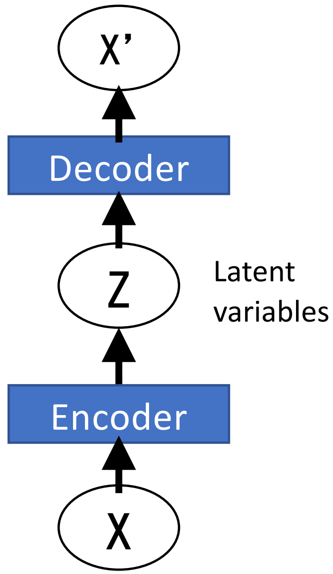
Wasserstein Autoencoder

- A method for Representation Learning

content

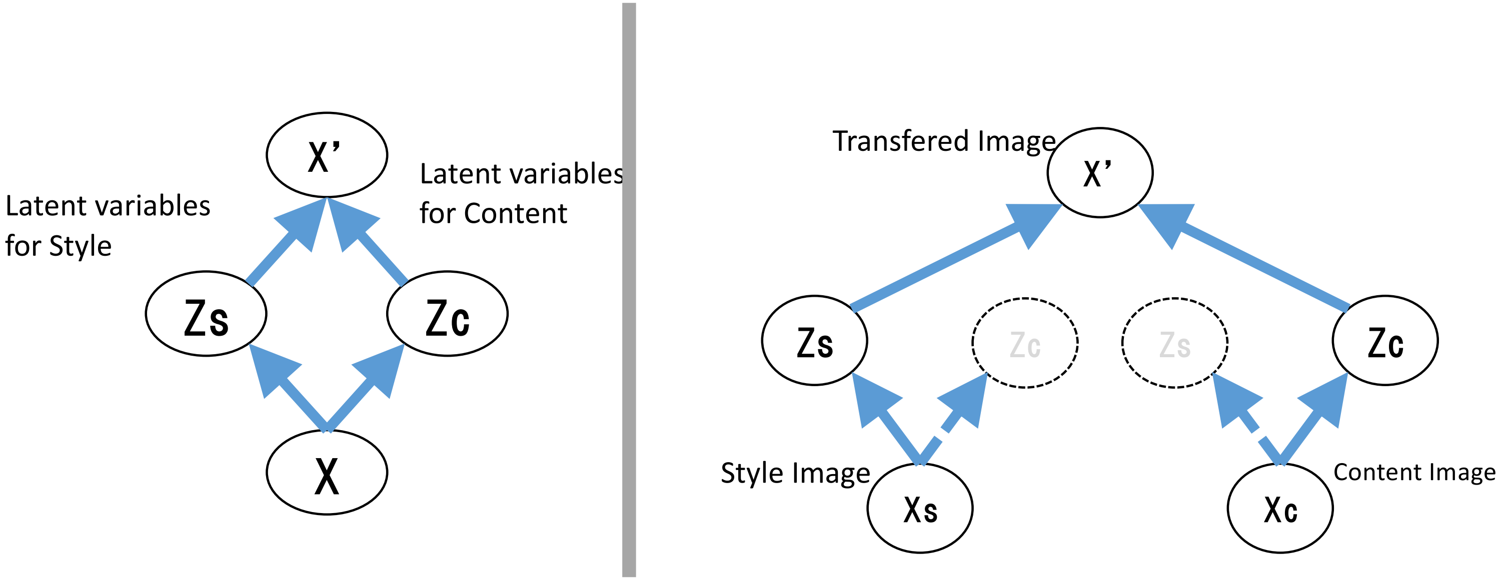
- Reconstruct X' from Latent Representation Z
  - Train the encoder and the decoder simultaneously

- Difference from 'Classical Autoencoder'
  - Train the networks so that the latent variables will have normal distribution with $\mu=0$, $\sigma=1$
  - Enforce the latent variables have 'meaningful' distribution
  - WAE uses Wasserstein distance
    - C.f: VAE uses KL divergence



Proposed Method

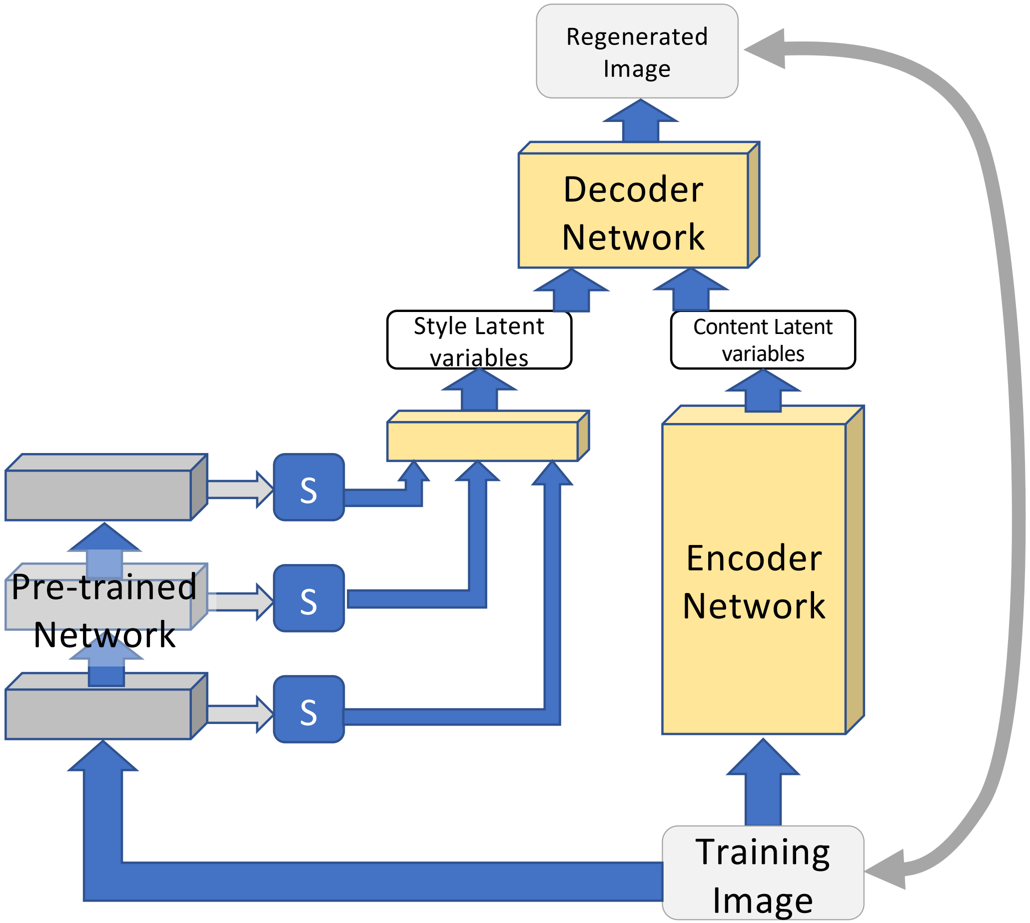
- WAE enables to represent images with latent variables
- If we can disentangle latent variables for style and content, we can render any image with any style.



Proposed Network

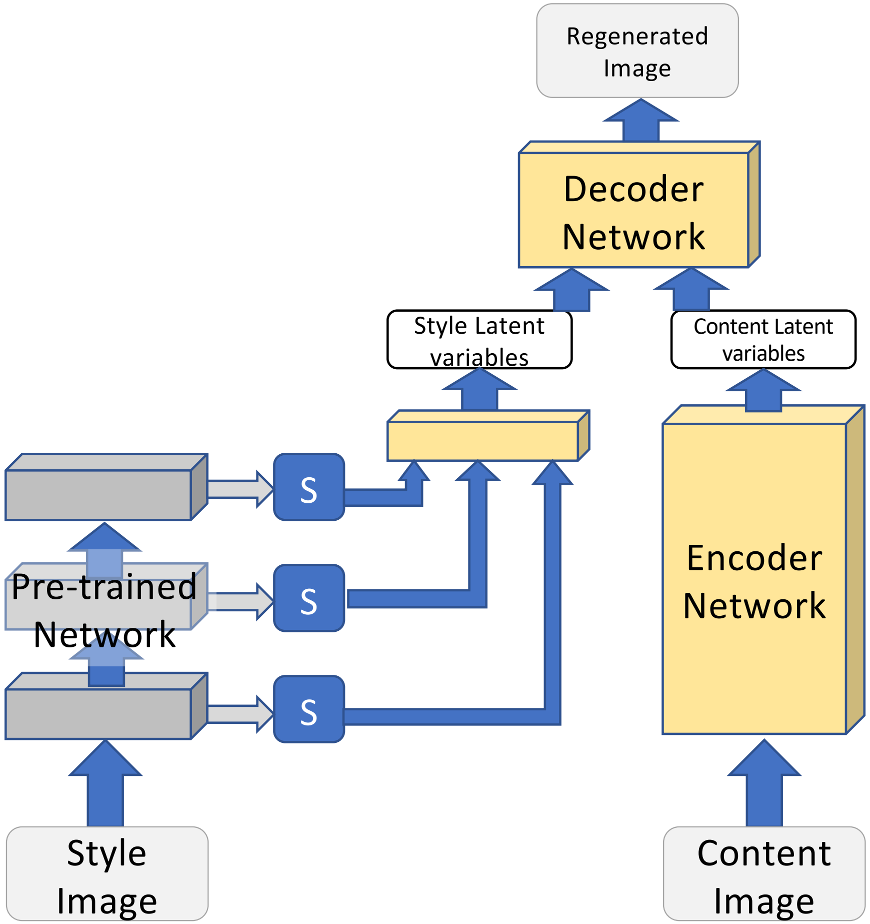
- **Training Time:**
  Input one image and minimize the Wasserstein loss and the image loss (difference between input and output)

- Style matrix is calculated with pretrained VGG network



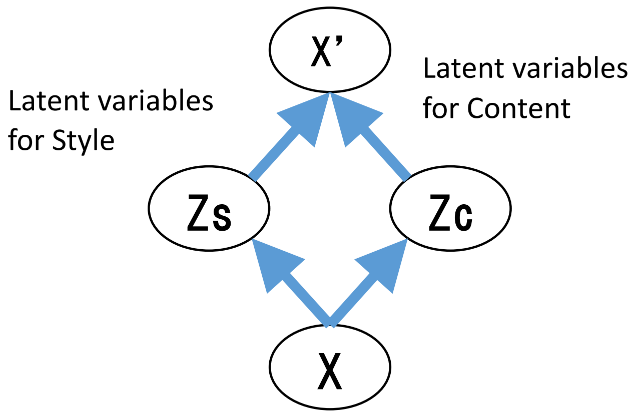
Proposed Network

- **Transfer:**
  Input two images, style and content, concatenate the latent representation, and decode it using the decoder network.



Disentanglement with regularization

- Latent variable for content will contain some style information
  - For better disentanglement, we have to 'squeeze out' the style information from the content latent variable



- Introduce regularization to latent variables
  - Enforce the variance of variables close to 1.
  - Effectively, minimize the number of variables that are actually used.

$$\frac{\lambda}{N}\sum_{n=1}^{N}\sum_{i=1}^{d_z}\|\log(\sigma_i^2(x_n))\|^1$$

Experimental Settings

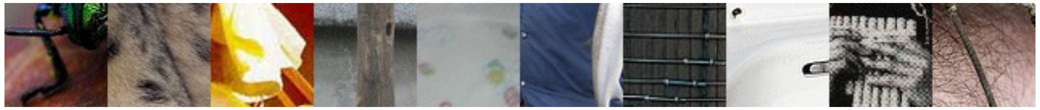
- Dataset Diversity
  - CelebA only
  - CelebA + Anime-Face + Imagenet
- Control the content-variable contribution
  - Changing # content latent variable
  - 512,256,128,64,32
- Latent variable Regularization
  - No Regularization
  - Style only, Content Only
  - Both

Dataset

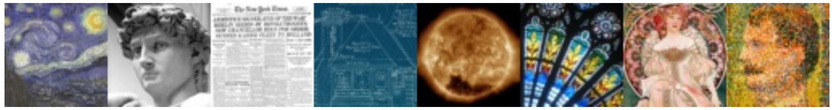
- CelebA
  - Cropped centering the face: 193,800

- Anime-Face-Dataset
  - Resized: 14,490

- ImageNet
  - Center Cropped: 196,371



- Style Images



Reconstruction

Trained /w Celeb A only Trained /w Celeb A + Anime + ImageNet

Reconstruction

Trained /w Celeb A only

Original

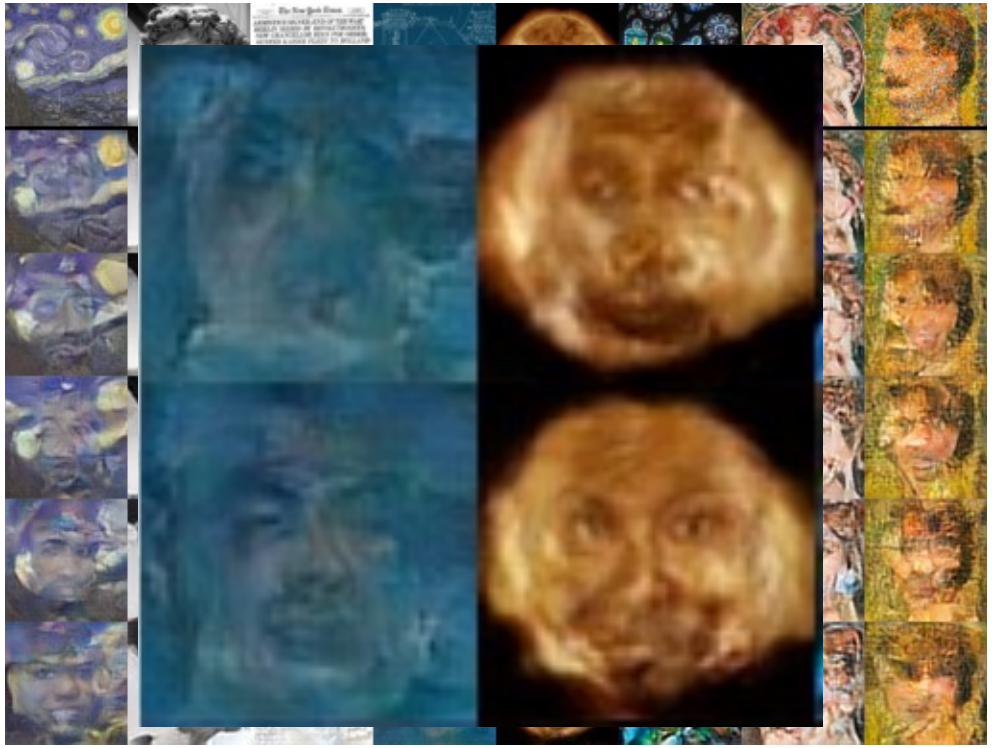
#Zc=512

#Zc=256

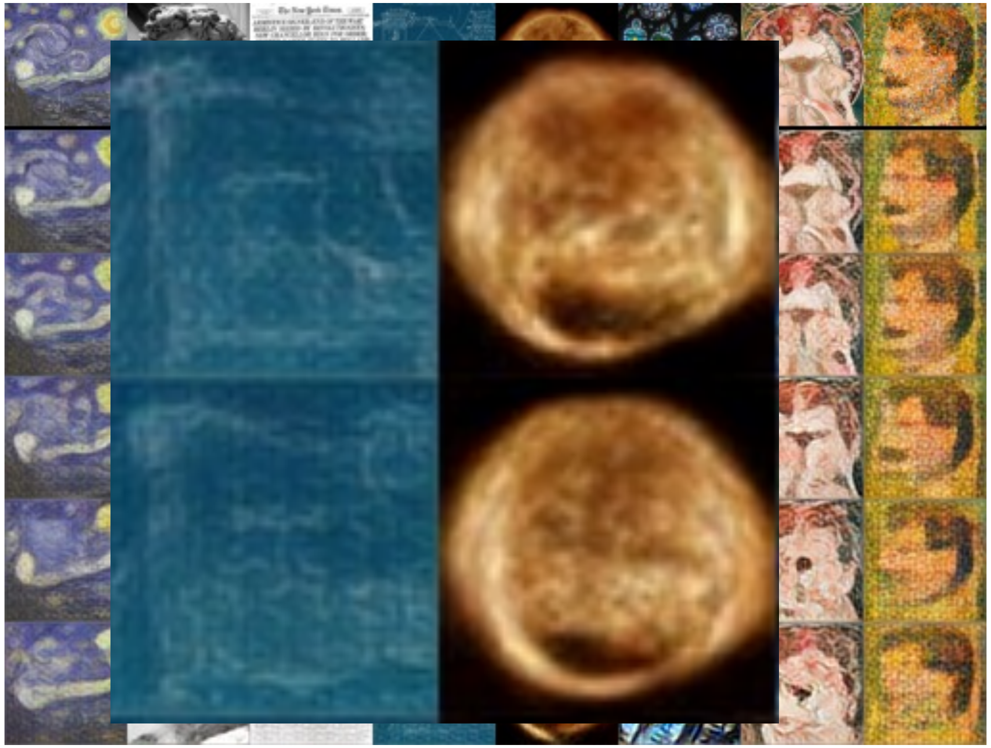
#Zc=128

#Zc=64

#Zc=32



Trained /w Celeb A + Anime + ImageNet



Face Artifact

Style Transferred

Trained /w Celeb A only

Content

Style

#Zc=512

#Zc=256

#Zc=128

#Zc=64

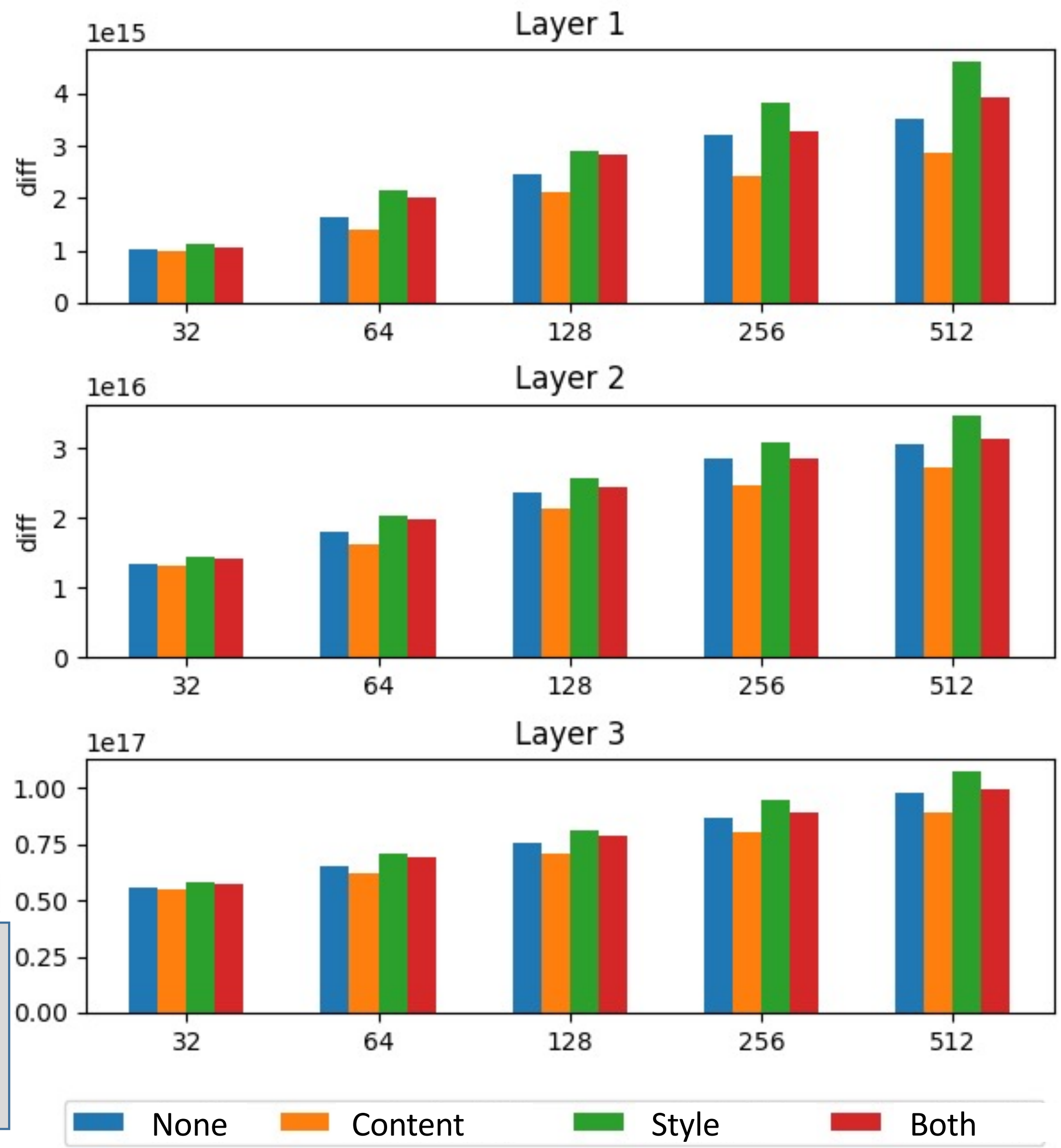
#Zc=32

Trained /w Celeb A + Anime + ImageNet

Contribution of Regularization

- Difference between style matrix of the style image and generated image
  - Smaller the better disentanglement.
- Regularization on
  - None
  - Content
  - Style
  - Content and Style Both

‘Content only’ shows the best disentanglement



Contribution of Regularization

Content
Style
#Zc=512
#Zc=256
#Zc=128
#Zc=64
#Zc=32

NoSP | Content Only | Style Only | Both | NoSP | Content Only | Style Only | Both | NoSP | Content Only | Style Only | Both

Conclusion

- Summary
  - Proposed WAE based Style transfer
  - Instant Style transfer with arbitrary style and content
  - Disentanglement with regularization
- Future Work
  - Strong Disentanglement
  - Improve Image Decode Network
    - GAN
    - PixelCNN

Acknowledgement

We would like to show our deep gratitude to **Tatsuhiko Inoue**, who helped us on implementation.

This paper is based on results obtained from a project commissioned by the New Energy and Industrial Technology Development Organization (NEDO).