# Multiple Style Transfer via Variational AutoEncoder

Zhi-Song Liu, Vicky Kalogeiton and Marie-Paule Cani
LIX, École Polytechnique, CNRS, IP Paris

2021 IEEE ICIP 19-22 SEPTEMBER

Website Github

## Background

**Style transfer (single style transfer)**

Given a target and a reference image → synthesize an image with content from the target and style from the reference.


**Style transfer (Multiple style transfer)**

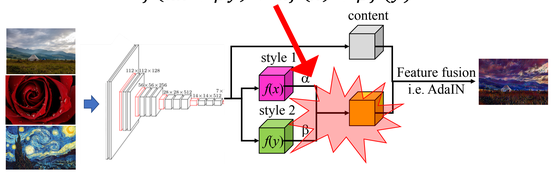
Fuse multiple style images into one content image → generate one stylized result.


**Challenge for multiple style transfer**

Most approaches resolve it as *linear weighted sum of deep features*.
Problem: deep features are nonlinear, not follow additivity or superposition principle.

Given style images $x$ and $y$, deep learning model as $f$,

$$f(\alpha x + \beta y) \neq \alpha f(x) + \beta f(y)$$

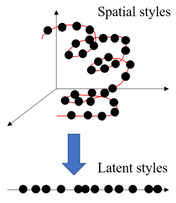



Can we find a projection process P that can transfer feature from nonlinear to linear correlation?

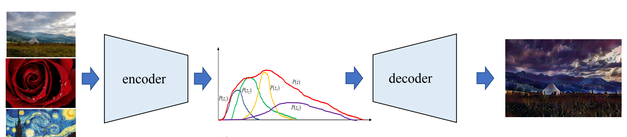
$$f(\alpha x + \beta y) \neq \alpha f(x) + \beta f(y)$$

→

$$f(\alpha x + \beta y) = \alpha P[f(x)] + \beta P[f(y)]$$



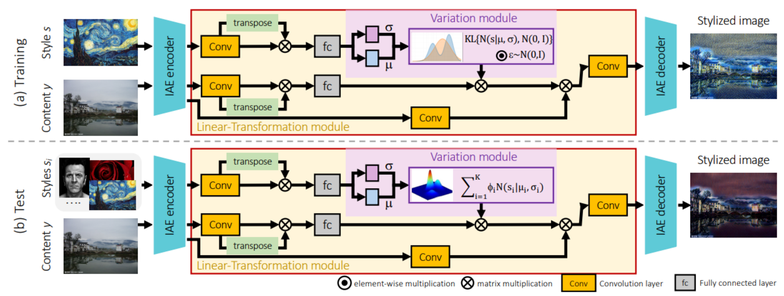

**Variational AutoEncoders:** compress the input information into a constrained multivariate latent distribution (encoding) to reconstruct it as accurately as possible (decoding)




$$P(\mathbf{X}) \approx \frac{1}{N}\sum_i P(\mathbf{X}|z_i = \mu(X) + \Sigma(X)^{0.5} * \varepsilon)$$

VAE projects images onto a multivariance gaussian model, where we can use linear computation to manipulate the features

## Proposed ST-VAE Method



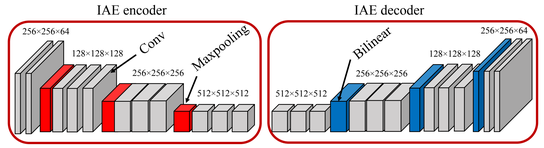

Three components:
1. (**IAE**): IAE encoder and IAE decoder
2. VAE-based Linear Transformation (**VLT**)
   - Linear-Transformation module
   - Variational module

**IAE**: symmetric encoder-decoder that extracts features for image reconstruction.

**Structure**:

Encoder follows VGG-19, keep all conv layers, discard the fc ones.

Decoder: symmetric to encoder, upsamples the feature abstraction to reconstruct the image




VAE-based Linear Transformation (VLT) is a combination of Linear Transformation module and Variational module

Given style feature $F_s$ , content feature $F_c$, target feature $F_d$:
**Goal:** learn a transformation matrix $T$ so that the covariance of the target feature $cov(F_d)$ is close to the covariance of the style feature $cov(F_s)$.

$$L_{style} = \frac{1}{NC}\|cov(F_d) - cov(F_s)\|^l$$

$$s.t.\ cov(F_s) = \bar{F}_s\bar{F}_s^T,\ cov(F_d) = \bar{F}_d\bar{F}_d^T,\ \ \bar{F}_d = T\bar{F}_c,\ \ \bar{F}_c = F_c - mean(F_c)$$

To obtain T, we have $T\bar{F}_c\bar{F}_c^T T^T = \bar{F}_s\bar{F}_s^T$ → $T = (E_s D E_s^T)\times\Sigma\times(E_c D E_c^T)^T$

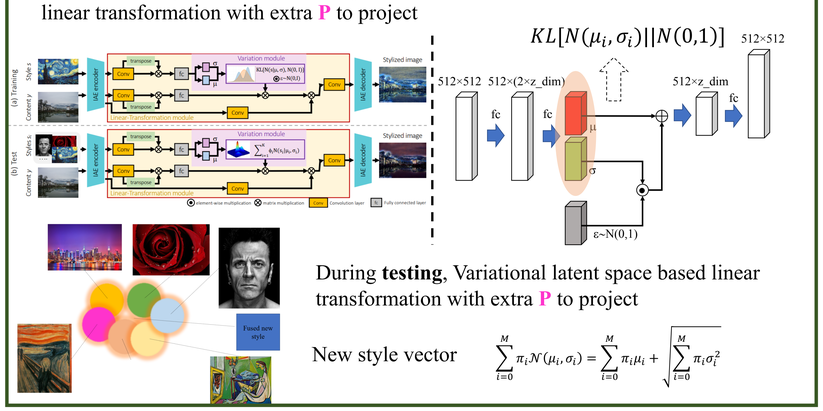
Σ is a covariance matrix between style and content features that can be learned by convolution as,

$$T = T_1(E_s D E_s^T)\times T_2(E_c D E_c^T)^T$$

**Variational module** is used to project the style covariance

$$F_d = F_c \times \boldsymbol{P}[T_1(\,cov(F_s))]\times T_2(cov(F_c))^T$$

During **training**, Variational latent space based linear transformation with extra **P** to project



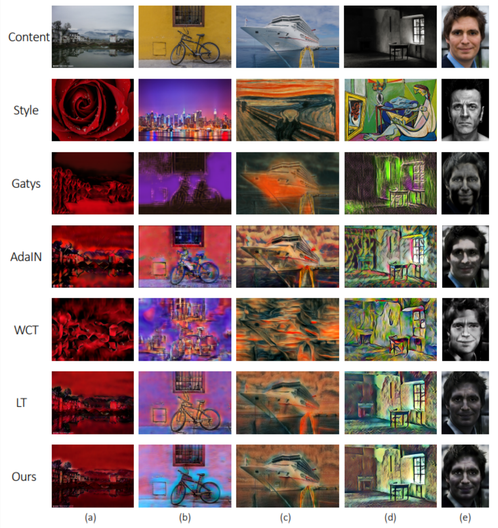

During **testing**, Variational latent space based linear transformation with extra **P** to project

New style vector $\sum_{i=0}^{M}\pi_i N(\mu_i,\sigma_i) = \sum_{i=0}^{M}\pi_i\mu_i + \sqrt{\sum_{i=0}^{M}\pi_i\sigma_i^2}$

## Experiments

We train Image AutoEncoder (**IAE**) on COCO [1].
For VLT, we use COCO as content and WikiArt [2] as style.

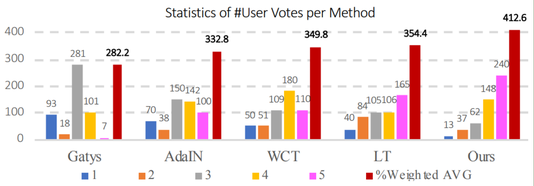
### Single style transfer




### Quantitative comparison

**User study:**

**Users:** 5 synthetic images in random order, to rate their quality from 1 to 5 (5 the highest).
**Results:** 400 images/method (10 content, 40 styles), randomly select 20. 500 votes from 20 users.




**Runtime**

| Image size | 256× 256 | 512× 512 | 1024× 1024 |
|---|---|---|---|
| Gatys [10] | 16.51 | 59.45 | 251.44 |
| AdaIN [3] | 0.019 | 0.071 | 0.288 |
| WCT [4] | 0.922 | 1.080 | 4.001 |
| LT [13] | 0.01 | 0.036 | 0.146 |
| Ours | 0.01 | 0.041 | 0.145 |

### Multiple style transfer