Paper Reading

Inversion-based Style Transfer with Diffusion Models

2023/9/10

Contents

Contents

• Seminal work of NST

Gatys L A, Ecker A S, Bethge M. Image style transfer using convolutional neural networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 2414-2423.

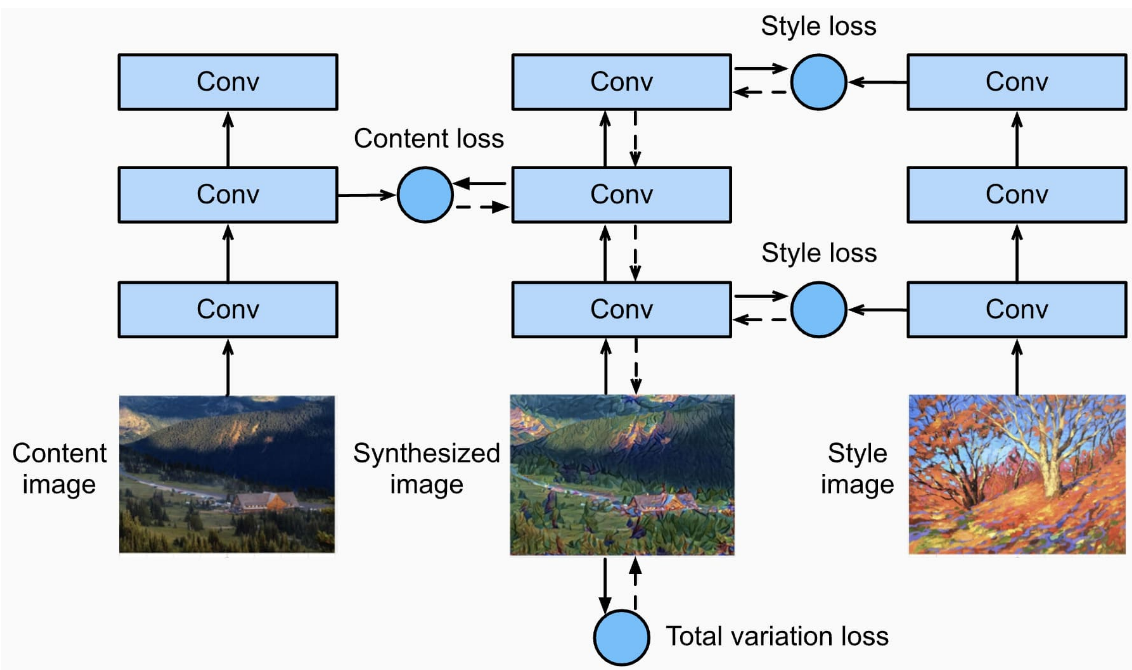
Landscape (content) + Scream (style)

Optimize a generated image that resembles a style image in style while preserving content of a content image.

Background

• Seminal work of NST

Gatys L A, Ecker A S, Bethge M. Image style transfer using convolutional neural networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 2414-2423.



Neural Style Transfer Architecture

$$\mathcal{L}_{\text{content}}(\vec{p},\vec{x},l) = \frac{1}{2}\sum_{i,j}\left(F_{ij}^{l} - P_{ij}^{l}\right)^2$$

$$G_{ij}^{l} = \sum_{k} F_{ik}^{l} F_{jk}^{l}.$$

$$E_l = \frac{1}{4N_l^2 M_l^2}\sum_{i,j}\left(G_{ij}^{l} - A_{ij}^{l}\right)^2$$

$$\mathcal{L}_{\text{style}}(\vec{a},\vec{x}) = \sum_{l=0}^{L} w_l E_l,$$

$$\mathcal{L}_{total} = \alpha\mathcal{L}_{content} + \beta\mathcal{L}_{style}$$

Background

- NST Applications

Art head portrait generation

A. Selim, M. Elgharib, and L. Doyle, "Painting style transfer for head portraits using convolutional neural networks," ACM Transactions on Graphics (ToG), vol. 35, no. 4, p. 129, 2016.

- NST Applications

Logo decoration

Content Style Decorated Content Style Decorated

G. Atarsaikhan, B. K. Iwana, and S. Uchida, "Contained neural style transfer for decorated logo generation," in 2018 13th IAPR Interna- tional Workshop on Document Analysis Systems (DAS). IEEE, 2018, pp. 317–322.

• NST Applications

Fashion design

S. Jiang and Y. Fu, "Fashion style generator." in IJCAI, 2017, pp. 3721–3727.

Why the Gram matrix can represent image style?

• Demystifying NST

Gram损失等效于二阶多项式核的MMD度量

$$\begin{aligned}
\mathcal{L}_{style}^{l} &= \frac{1}{4N_l^2M_l^2}\sum_{i=1}^{N_l}\sum_{j=1}^{N_l}(\sum_{k=1}^{M_l}F_{ik}^{l}F_{jk}^{l} - \sum_{k=1}^{M_l}S_{ik}^{l}S_{jk}^{l})^2 \\
&= \frac{1}{4N_l^2M_l^2}\sum_{i=1}^{N_l}\sum_{j=1}^{N_l}\Big((\sum_{k=1}^{M_l}F_{ik}^{l}F_{jk}^{l})^2 + (\sum_{k=1}^{M_l}S_{ik}^{l}S_{jk}^{l})^2 - 2(\sum_{k=1}^{M_l}F_{ik}^{l}F_{jk}^{l})(\sum_{k=1}^{M_l}S_{ik}^{l}S_{jk}^{l})\Big) \\
&= \frac{1}{4N_l^2M_l^2}\sum_{i=1}^{N_l}\sum_{j=1}^{N_l}\sum_{k_1=1}^{M_l}\sum_{k_2=1}^{M_l}(F_{ik_1}^{l}F_{jk_1}^{l}F_{ik_2}^{l}F_{jk_2}^{l} + S_{ik_1}^{l}S_{jk_1}^{l}S_{ik_2}^{l}S_{jk_2}^{l} - 2F_{ik_1}^{l}F_{jk_1}^{l}S_{ik_2}^{l}S_{jk_2}^{l}) \\
&= \frac{1}{4N_l^2M_l^2}\sum_{k_1=1}^{M_l}\sum_{k_2=1}^{M_l}\sum_{i=1}^{N_l}\sum_{j=1}^{N_l}(F_{ik_1}^{l}F_{jk_1}^{l}F_{ik_2}^{l}F_{jk_2}^{l} + S_{ik_1}^{l}S_{jk_1}^{l}S_{ik_2}^{l}S_{jk_2}^{l} - 2F_{ik_1}^{l}F_{jk_1}^{l}S_{ik_2}^{l}S_{jk_2}^{l}) \\
&= \frac{1}{4N_l^2M_l^2}\sum_{k_1=1}^{M_l}\sum_{k_2=1}^{M_l}\Big((\sum_{i=1}^{N_l}F_{ik_1}^{l}F_{ik_2}^{l})^2 + (\sum_{i=1}^{N_l}S_{ik_1}^{l}S_{ik_2}^{l})^2 - 2(\sum_{i=1}^{N_l}F_{ik_1}^{l}S_{ik_2}^{l})^2\Big) \\
&= \frac{1}{4N_l^2M_l^2}\sum_{k_1=1}^{M_l}\sum_{k_2=1}^{M_l}\Big(({\mathbf{f}_{\cdot k_1}^{l}}^{T}\mathbf{f}_{\cdot k_2}^{l})^2 + ({\mathbf{s}_{\cdot k_1}^{l}}^{T}\mathbf{s}_{\cdot k_2}^{l})^2 - 2({\mathbf{f}_{\cdot k_1}^{l}}^{T}\mathbf{s}_{\cdot k_2}^{l})^2\Big),
\end{aligned}$$

Li Y, Wang N, Liu J, et al. Demystifying neural style transfer[C]//Proceedings of the 26th International Joint Conference on Artificial Intelligence. 2017: 2230-2236.

Demystifying NST

Gram损失等效于二阶多项式核的MMD度量

$$\mathcal{L}^l_{style} = \frac{1}{4N_l^2 M_l^2} \sum_{k_1=1}^{M_l} \sum_{k_2=1}^{M_l} \Big(k(\mathbf{f}^l_{\cdot k_1}, \mathbf{f}^l_{\cdot k_2}) + k(\mathbf{s}^l_{\cdot k_1}, \mathbf{s}^l_{\cdot k_2}) - 2k(\mathbf{f}^l_{\cdot k_1}, \mathbf{s}^l_{\cdot k_2}) \Big) = \frac{1}{4N_l^2} \mathrm{MMD}^2[\mathcal{F}^l, \mathcal{S}^l],$$

(1) Linear kernel: $k(\mathbf{x}, \mathbf{y}) = \mathbf{x}^T \mathbf{y}$;

(2) Polynomial kernel: $k(\mathbf{x}, \mathbf{y}) = (\mathbf{x}^T \mathbf{y} + c)^d$;

(3) Gaussian kernel: $k(\mathbf{x}, \mathbf{y}) = \exp\big(-\frac{\|\mathbf{x}-\mathbf{y}\|_2^2}{2\sigma^2}\big)$.

Li Y, Wang N, Liu J, et al. Demystifying neural style transfer[C]//Proceedings of the 26th International Joint Conference on Artificial Intelligence. 2017: 2230-2236.

More Style Losses

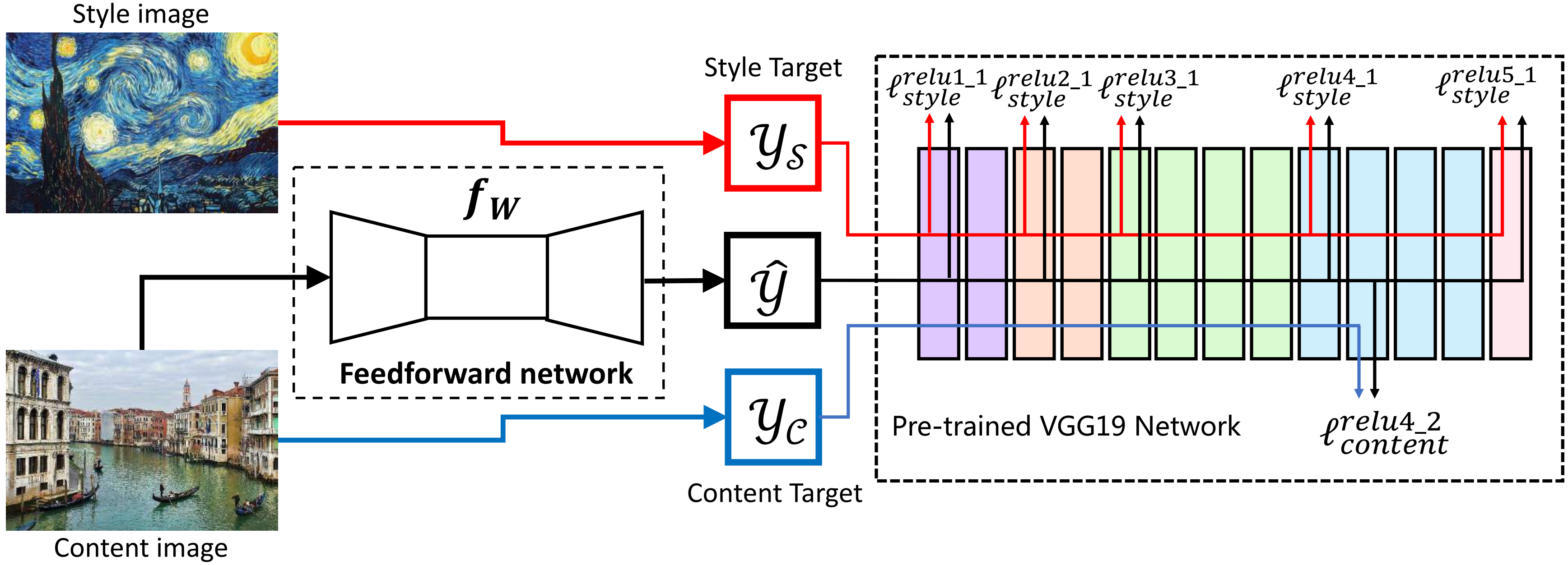
Style Loss	Paper	Publisher
Gram Loss	Image style transfer using convolutional neural networks	CVPR 2016
MMD Loss	Demystifying neural style transfer	IJCAI 2017
Mean-Variance Loss	Arbitrary style transfer in real-time with adaptive instance normalization	ICCV 2017
Histogram Loss	Stable and controllable neural texture synthesis and style transfer using histogram losses	Arxiv 2017
EMD Loss	Style Transfer by Relaxed Optimal Transport and Self-Similarity	CVPR 2019

Typical style losses wided used in style transfer literature

Contents

• Real-Time NST

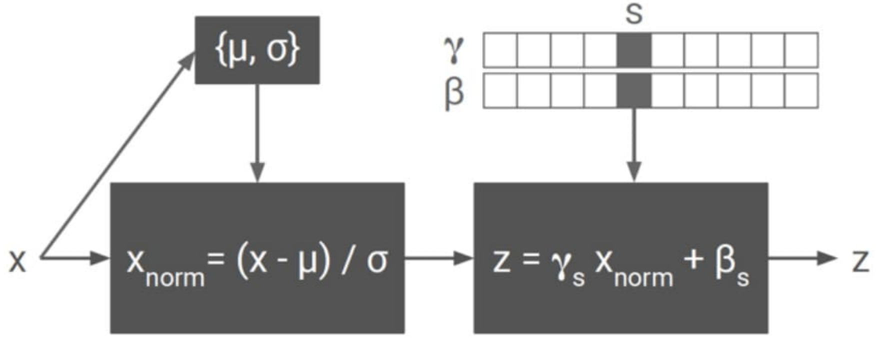
J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real- time style transfer and super-resolution," in *European conference on computer vision*. Springer, 2016, pp. 694–711.

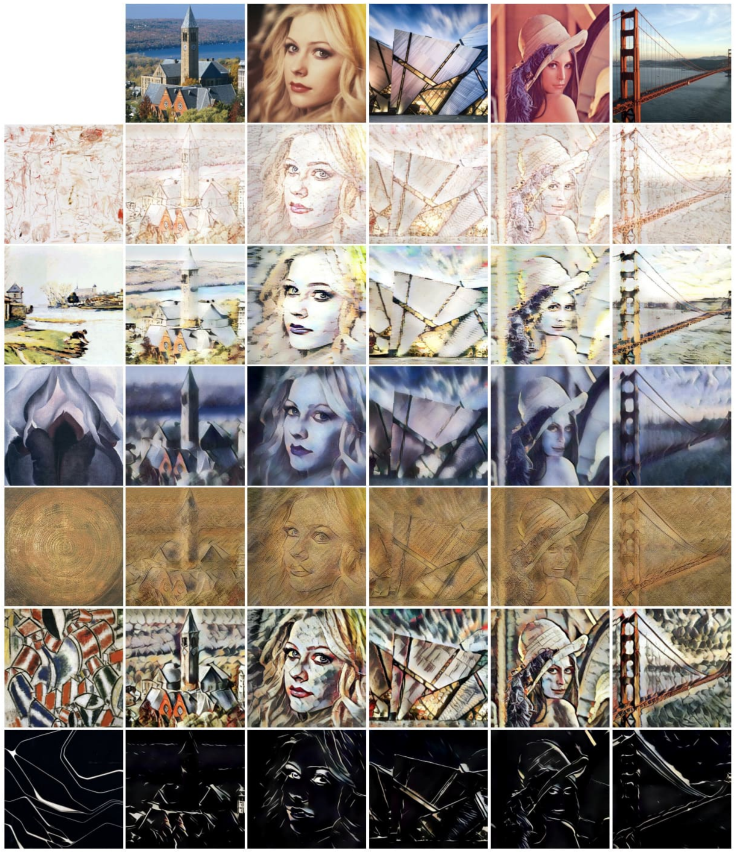


Extending on-line optimized based algorithm to real-time generation.

• Multi-Style NST

① Model style with Instance Normalization affine transformation parameters



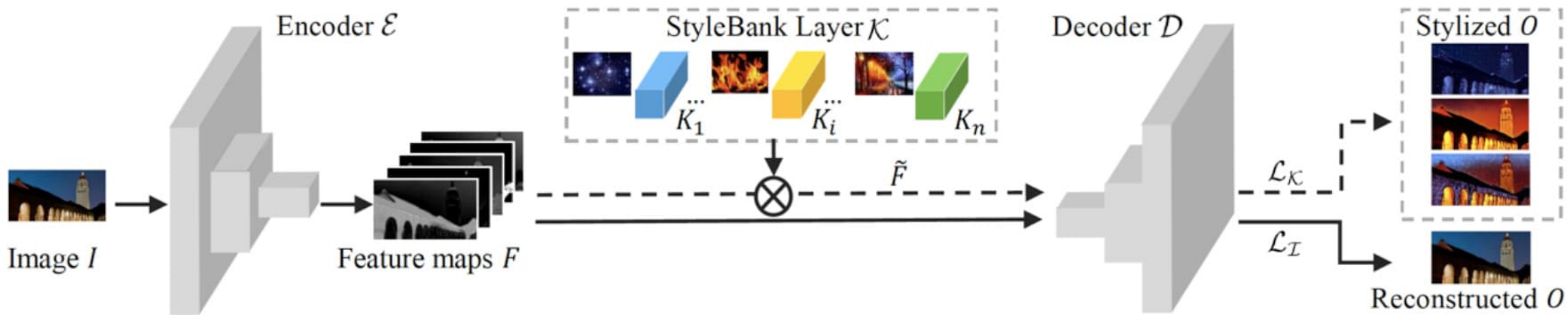


V. Dumoulin, J. Shlens, and M. Kudlur, "A learned representation for artistic style," *arXiv preprint arXiv:1610.07629*, 2016.

Related Work

• Multi-Style NST

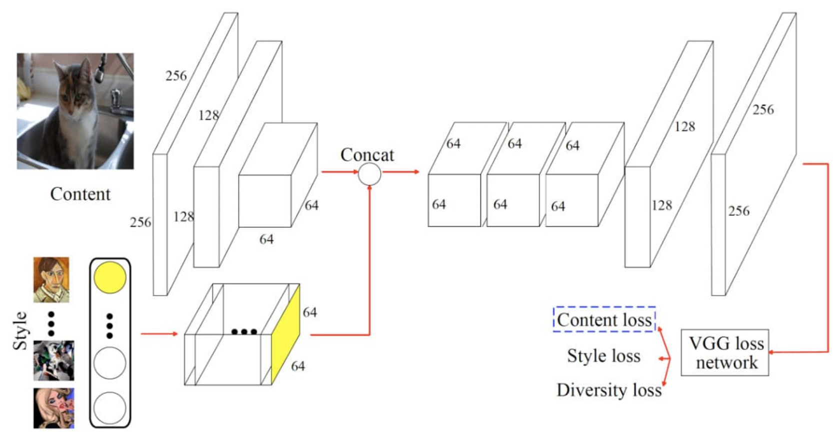
② Use specific convolutional kernel to render specific style



D. Chen, L. Yuan, J. Liao, N. Yu, and G. Hua, "Stylebank: An explicit representation for neural image style transfer," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1897–1906.

③ Use specific conditional signal to indicate specific style

Y. Li, C. Fang, J. Yang, Z. Wang, X. Lu, and M.-H. Yang, "Diversified texture synthesis with feed-forward networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3920–3928.

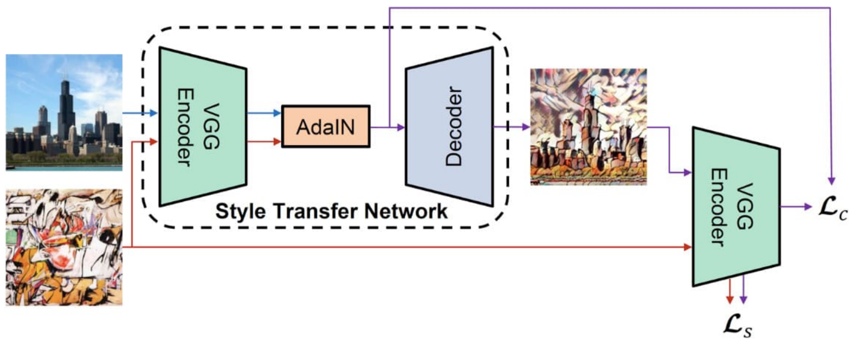


Related Work

Seminar 2023/9/10

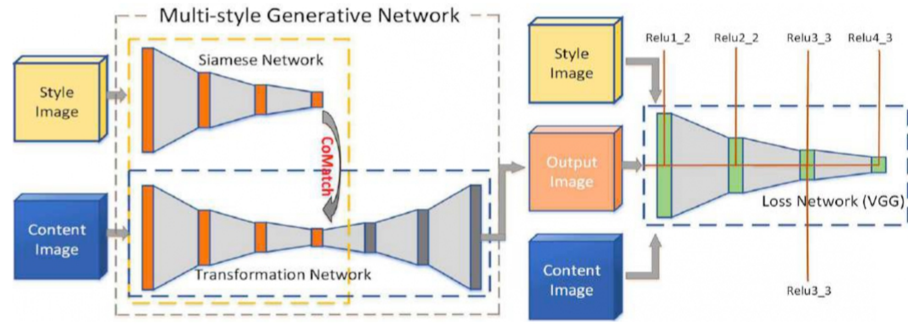
• Arbitrary-Style NST

(1) AdaIN



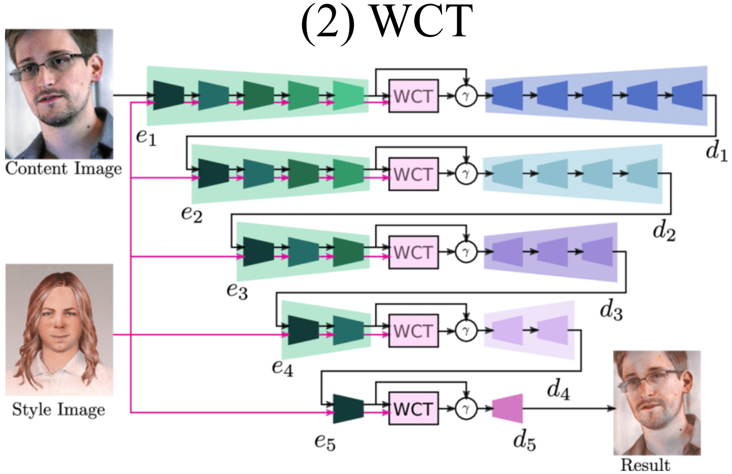
X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 1501–1510.

(3) MSG-Net



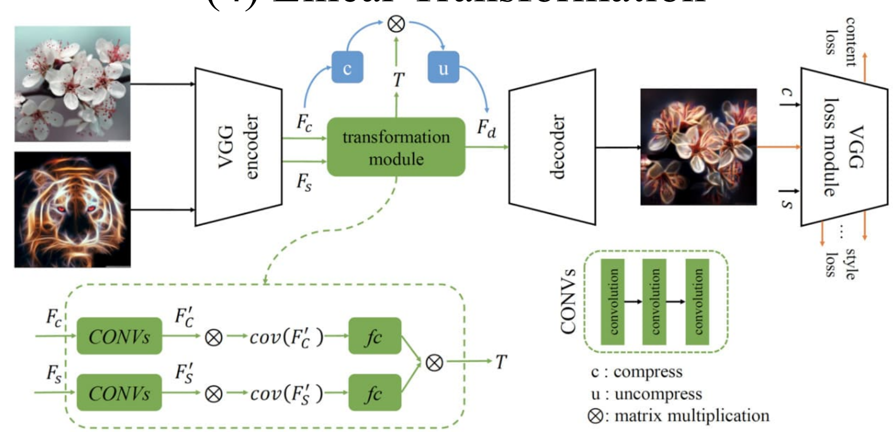
H. Zhang and K. Dana, "Multi-style generative network for real-time transfer," in Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 0–0.

(2) WCT



Y. Li, C. Fang, J. Yang, Z. Wang, X. Lu, and M.-H. Yang, "Universal style transfer via feature transforms," in *Advances in neural information processing systems*, 2017, pp. 386–396.

(4) Linear Transformation

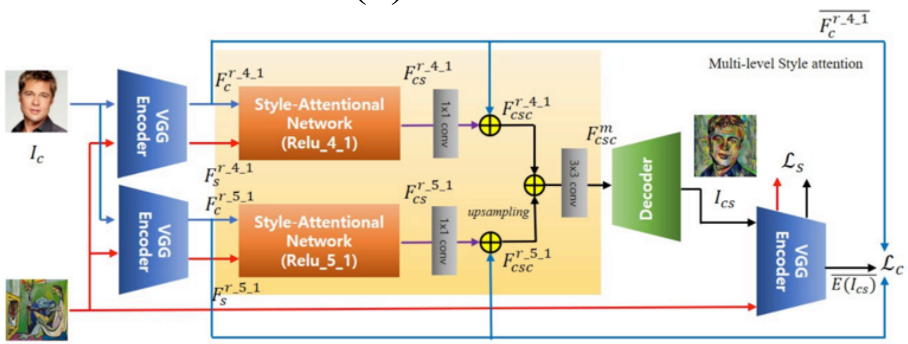


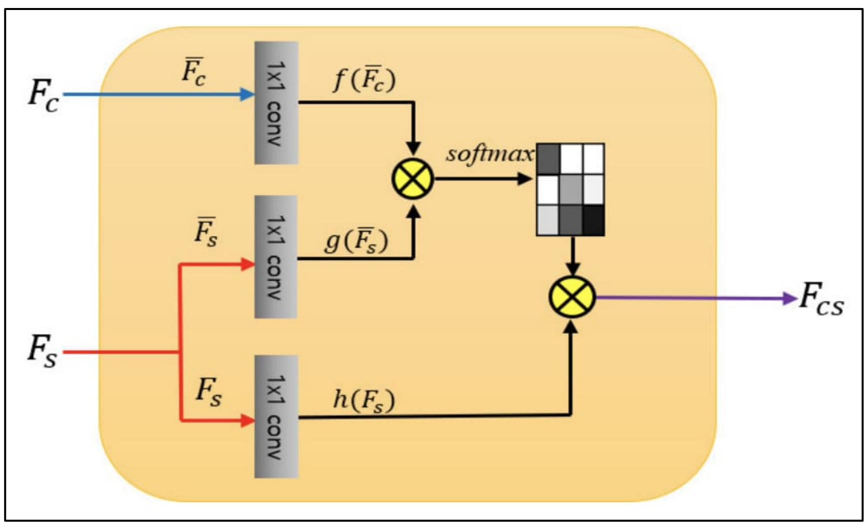
X. Li, S. Liu, J. Kautz, and M.-H. Yang, "Learning linear transforma- tions for fast arbitrary style transfer," arXiv preprint arXiv:1808.04537, 2018.

Related Work

- ## Arbitrary-Style NST

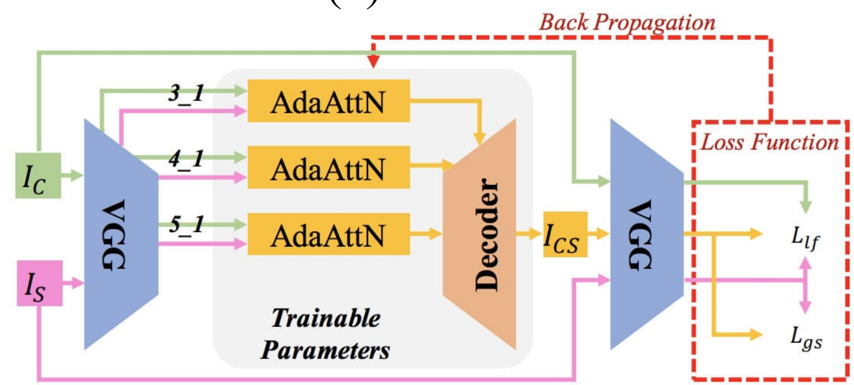
(5) SANet

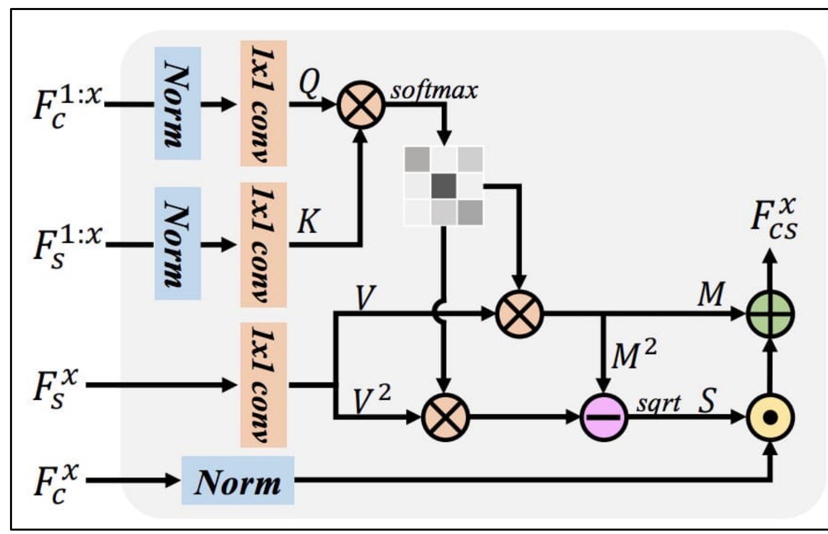




Park D Y, Lee K H. Arbitrary style transfer with style-attentional networks[C]//proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 5880-5888.

(6) Adaattn

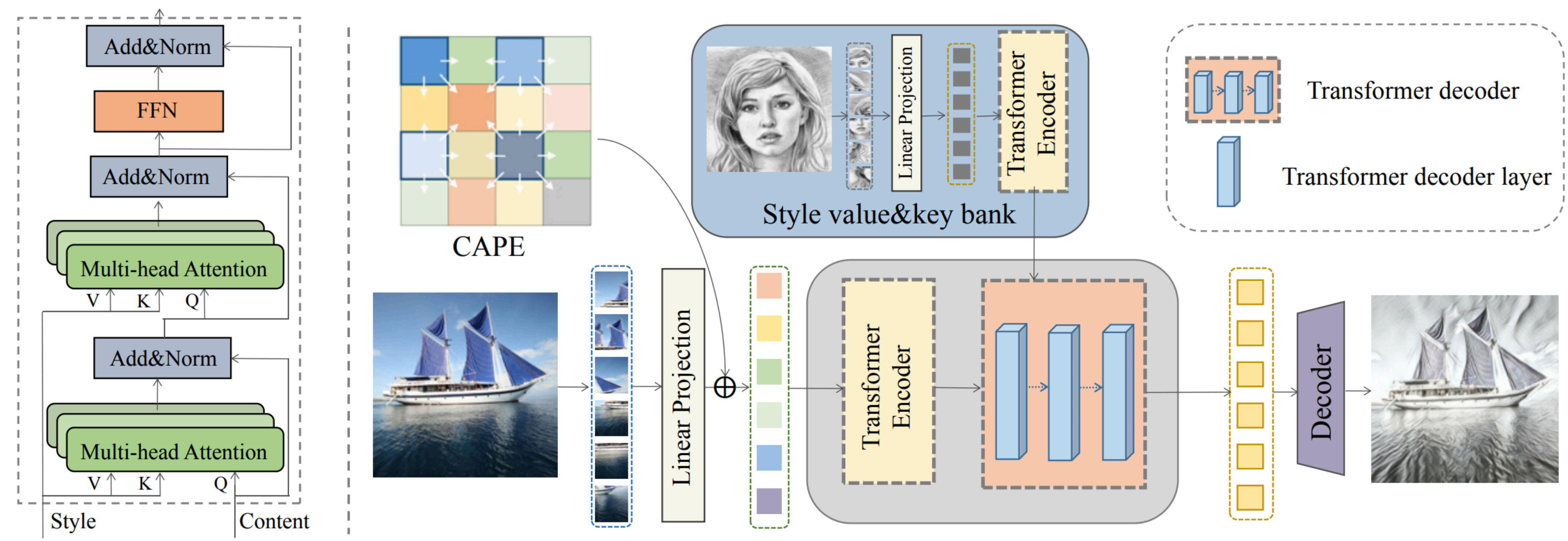




Liu S, Lin T, He D, et al. Adaattn: Revisit attention mechanism in arbitrary neural style transfer[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 6649-6658.

- Arbitrary-Style NST

(7) StyTr2



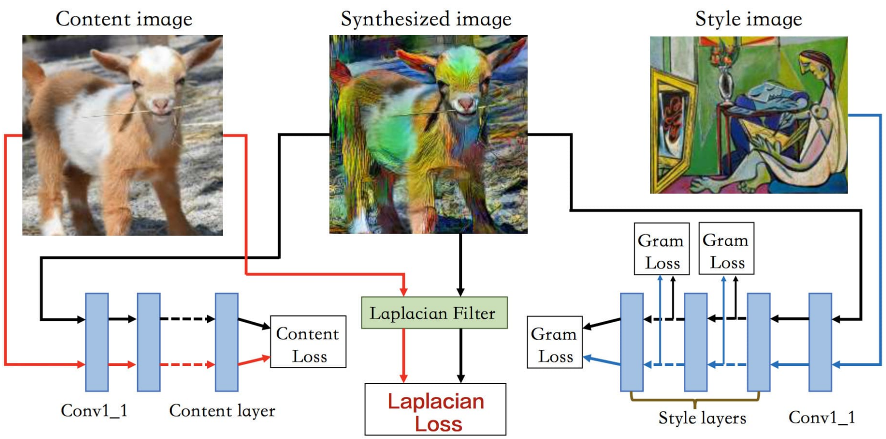
(a) Transformer decoder layer (b) StyTr² Network

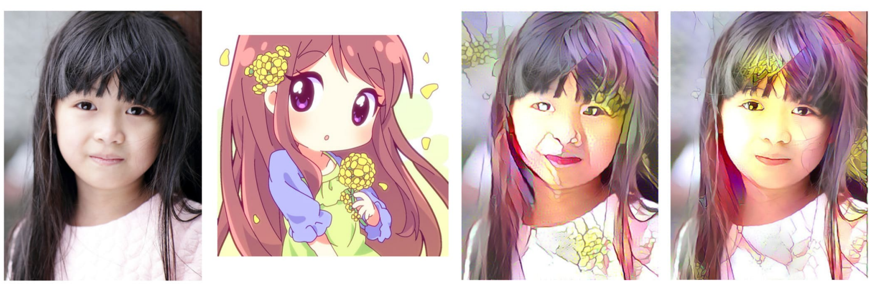
Deng Y, Tang F, Dong W, et al. Stytr2: Image style transfer with transformers[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 11326-11336.

Related Work

• NST Quality Improvement

(1) Remove noisy elements

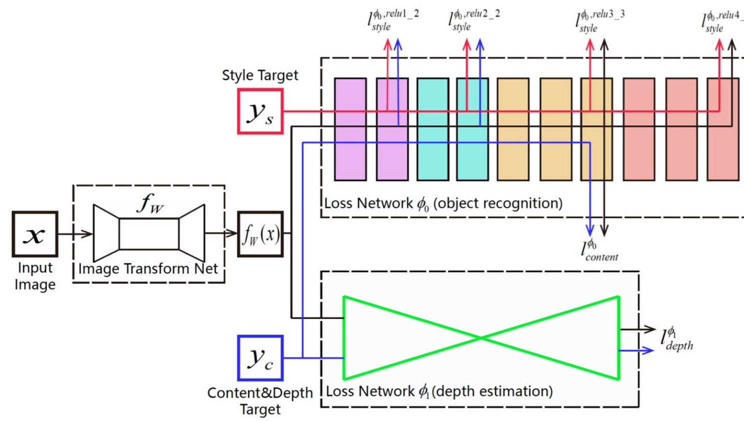


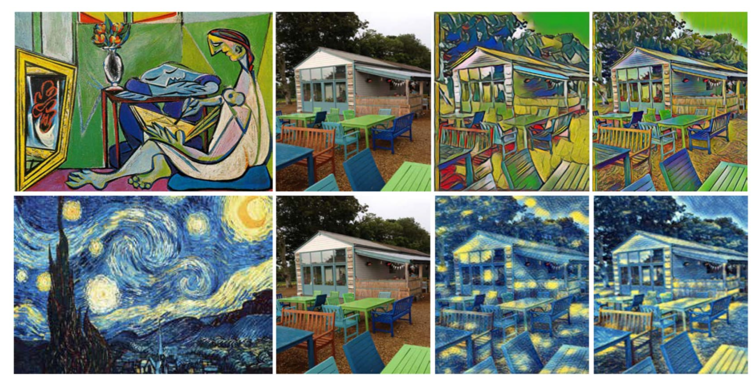


(a) Content image (b) Style image (c) Gatys-style output (d) Lapstyle output

S. Li, X. Xu, L. Nie, and T.-S. Chua, "Laplacian-steered neural style transfer," in Proceedings of the 25th ACM international conference on Multimedia. ACM, 2017, pp. 1716–1724.

(2) Promote structure consistency





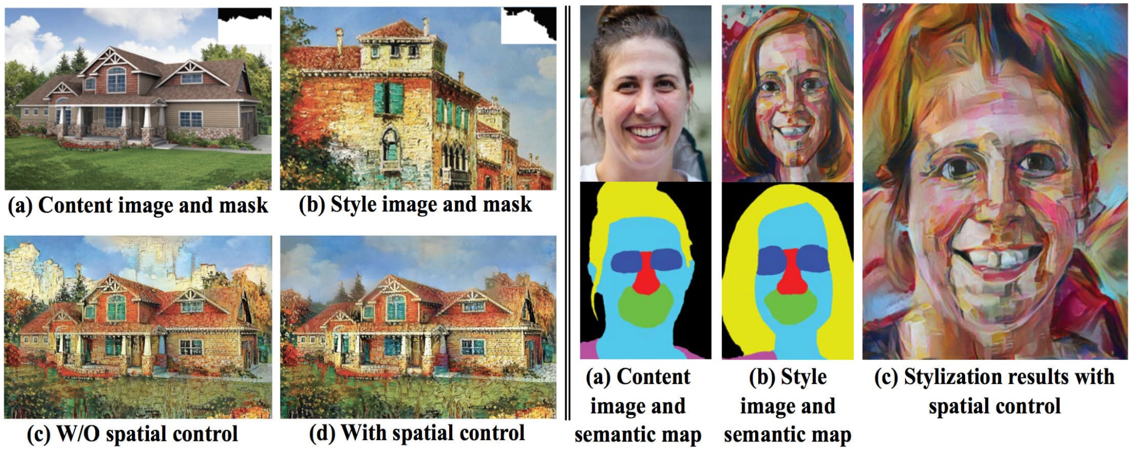
Style image Content image Fast NST Depth-aware NST

X.-C. Liu, M.-M. Cheng, Y.-K. Lai, and P. L. Rosin, "Depth-aware neural style transfer," in Proceedings of the Symposium on Non- Photorealistic Animation and Rendering, 2017, pp. 1–10.

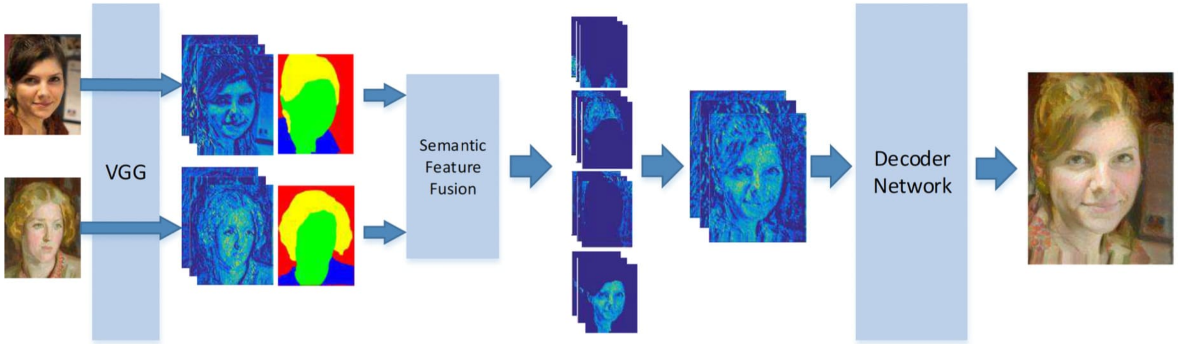
Related Work

• NST Quality Improvement

(3) Semantic Style Transfer



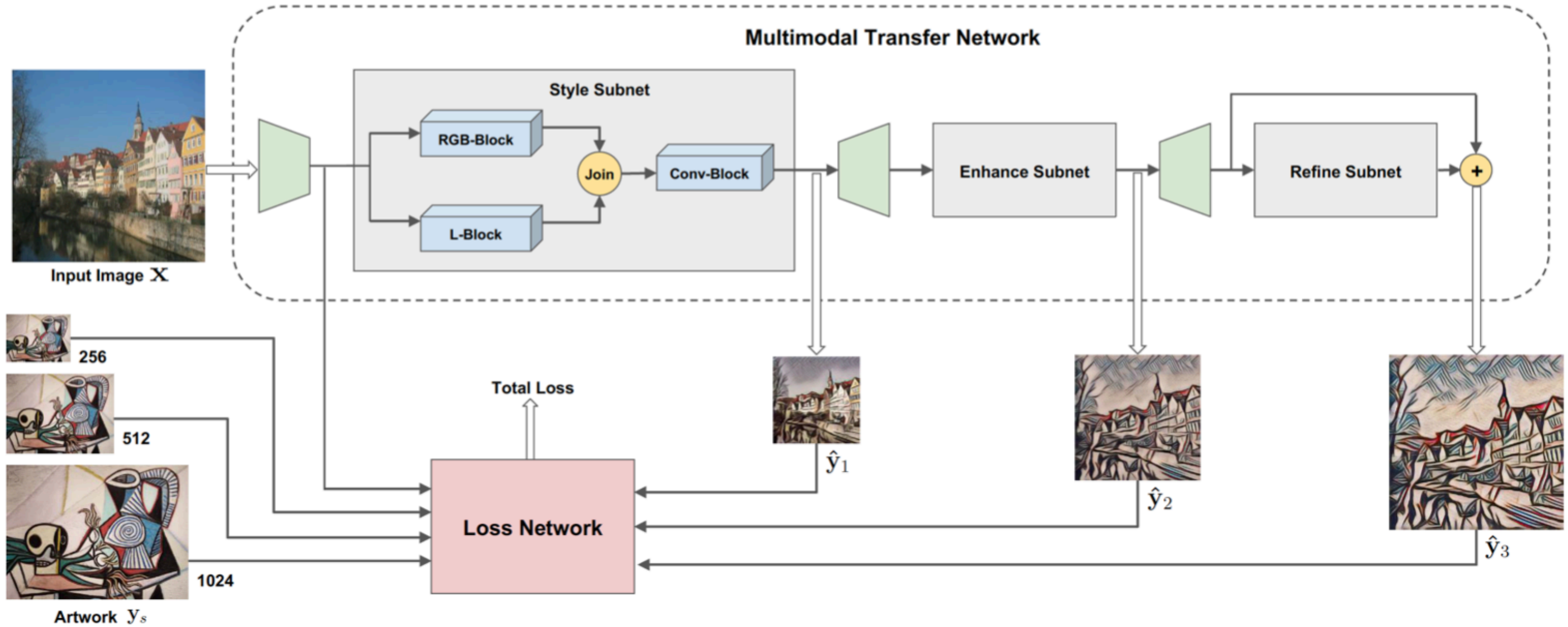
L.A.Gatys,A.S.Ecker,M.Bethge,A.Hertzmann,andE.Shechtman, "Controlling perceptual factors in neural style transfer," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 3985–3993.



M. Lu, H. Zhao, A. Yao, F. Xu, Y. Chen, and L. Zhan, "Decoder network over lightweight reconstructed feature for fast semantic style transfer," in Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2469–2477.

- NST Quality Improvement

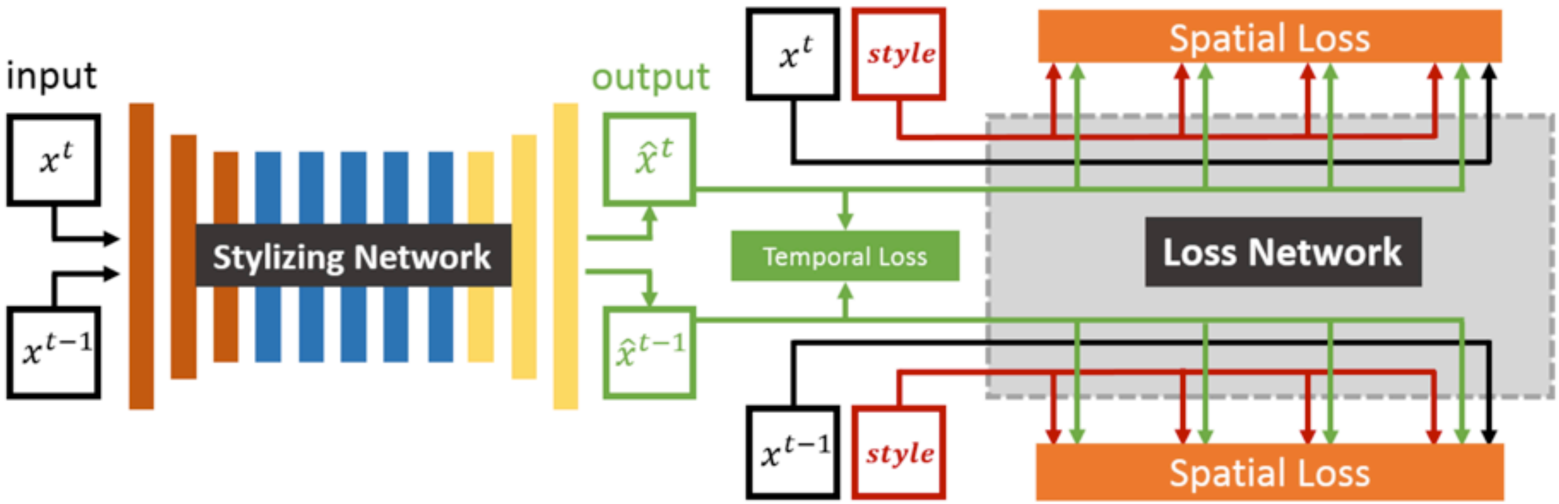
(4) Promote style saliency for large image



X. Wang, G. Oxholm, D. Zhang, and Y.-F. Wang, "Multimodal transfer: A hierarchical deep convolutional neural network for fast artistic style transfer," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 5239–5247.

- NST beyond image

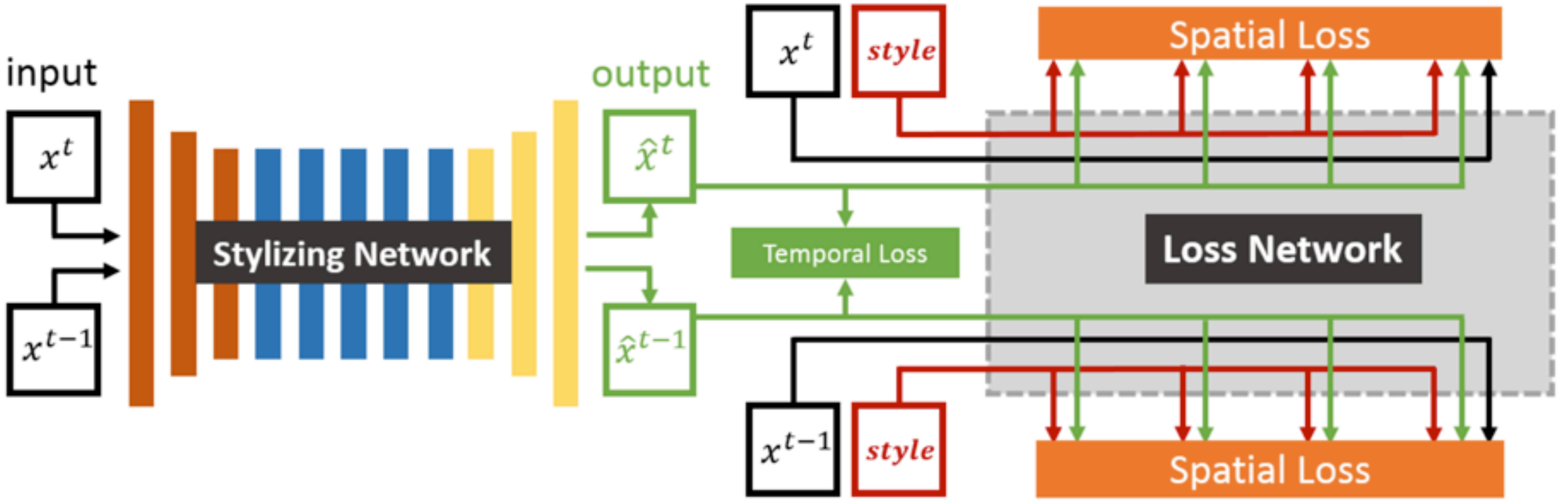
Video Style Transfer



Huang H, Wang H, Luo W, et al. Real-time neural style transfer for videos[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 783-791.

$$\mathcal{L}_{hybrid} = \underbrace{\sum_{i\in\{t,t-1\}} \mathcal{L}_{spatial}(\mathbf{x}^i, \hat{\mathbf{x}}^i, \mathbf{s})}_{\text{spatial loss}} + \underbrace{\lambda\mathcal{L}_{temporal}(\hat{\mathbf{x}}^t, \hat{\mathbf{x}}^{t-1})}_{\text{temporal loss}},$$

- NST beyond image

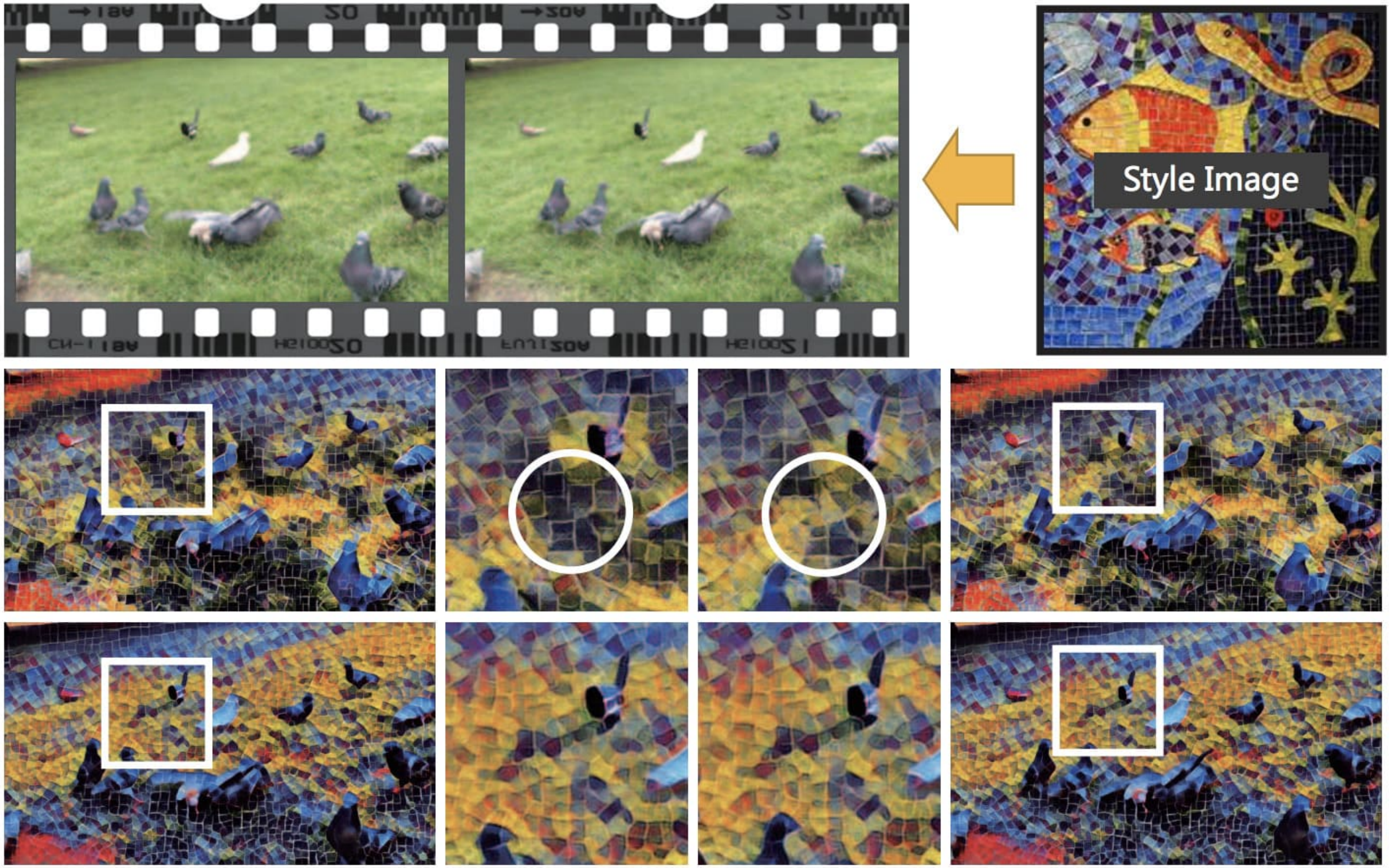


Huang H, Wang H, Luo W, et al. Real-time neural style transfer for videos[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 783-791.

$$\mathcal{L}_{spatial}(\mathbf{x}^t, \hat{\mathbf{x}}^t, \mathbf{s}) = \underbrace{\alpha \sum_l \mathcal{L}^l_{content}(\mathbf{x}^t, \hat{\mathbf{x}}^t)}_{\text{content loss}} + \underbrace{\beta \sum_l \mathcal{L}^l_{style}(\mathbf{s}, \hat{\mathbf{x}}^t)}_{\text{style loss}}$$

$$\mathcal{L}_{temporal}(\hat{\mathbf{x}}^t, \hat{\mathbf{x}}^{t-1}) = \frac{1}{D}\sum_{k=1}^{D} \mathbf{c}_k \left(\hat{\mathbf{x}}^t_k - f(\hat{\mathbf{x}}^{t-1}_k)\right)^2,$$

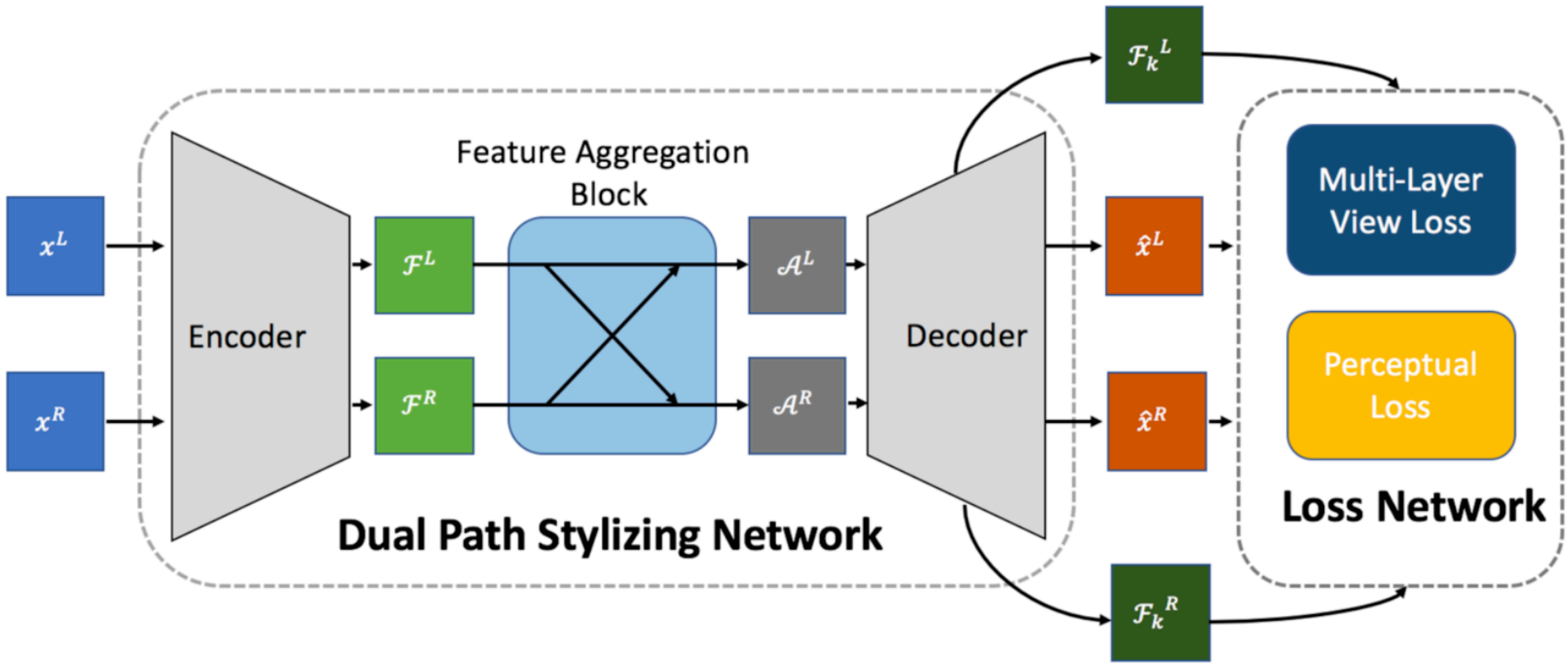
- NST beyond image



NST without / with temporal consistency

- NST beyond image

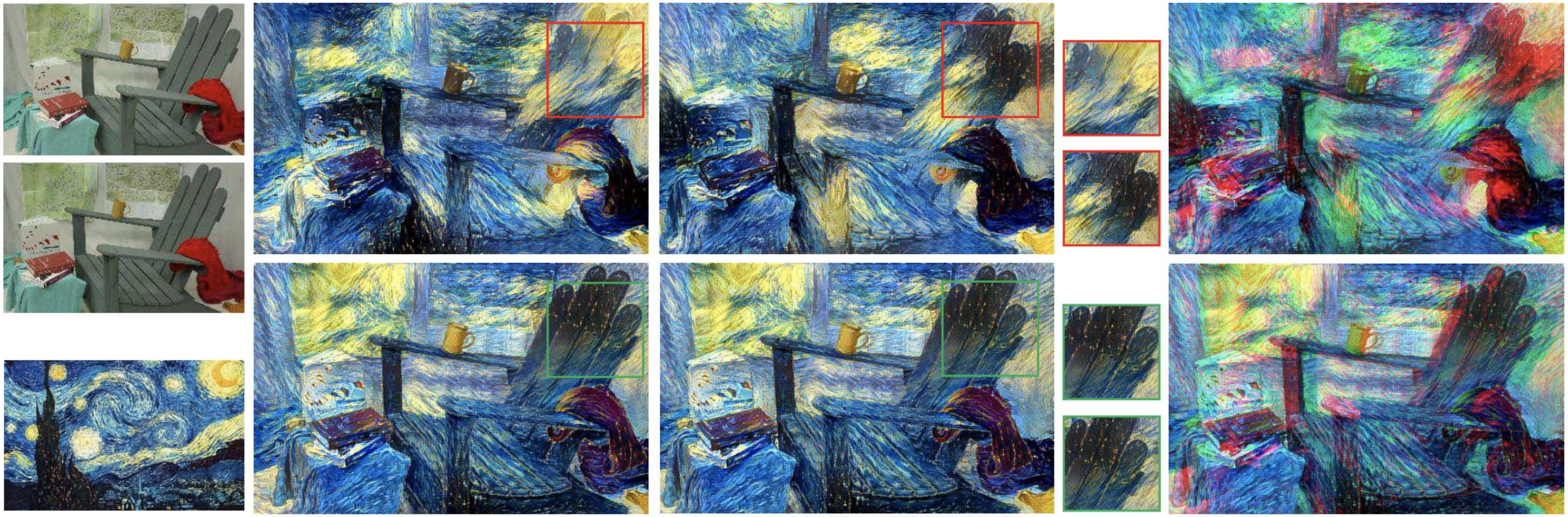
Stereoscopic Neural Style Transfer



Chen D, Yuan L, Liao J, et al. Stereoscopic neural style transfer[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 6654-6663.

• NST beyond image

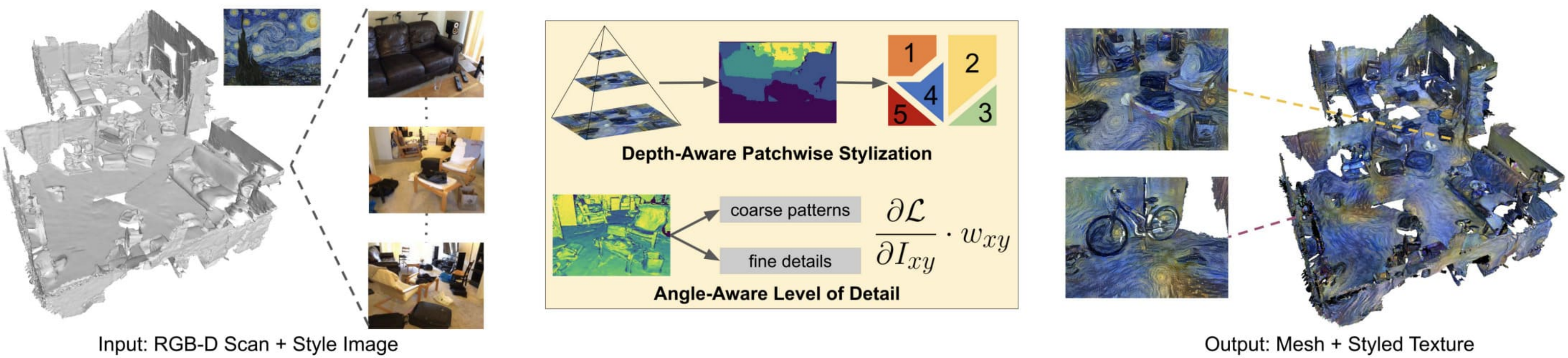
Stereoscopic Neural Style Transfer



Chen D, Yuan L, Liao J, et al. Stereoscopic neural style transfer[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 6654-6663.

- NST beyond image

Style Transfer for 3D reconstruction

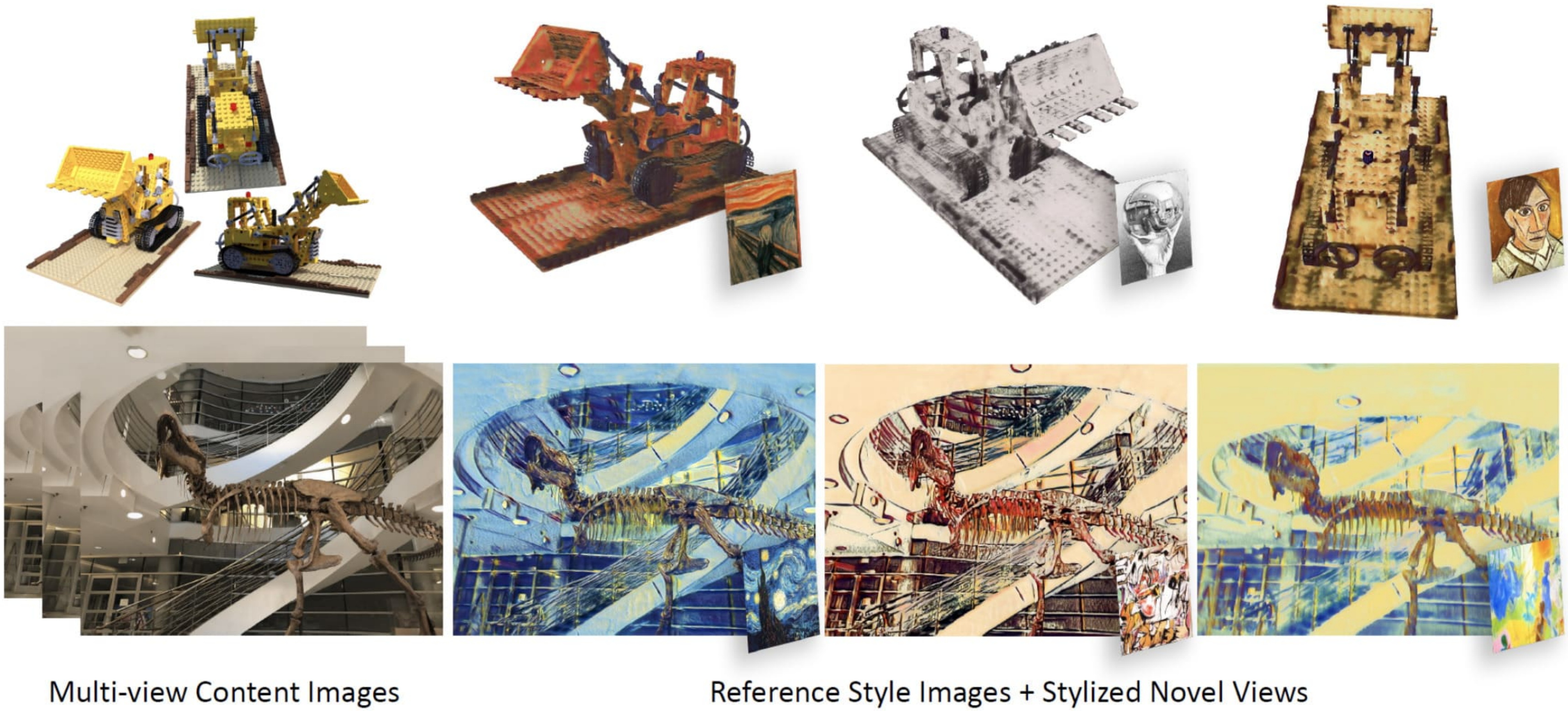


Höllein L, Johnson J, Nießner M. Stylemesh: Style transfer for indoor 3d scene reconstructions[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 6198-6208.

Related Work

- NST beyond image

Style Transfer for 3D reconstruction



Liu K, Zhan F, Chen Y, et al. StyleRF: Zero-shot 3D Style Transfer of Neural Radiance Fields[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 8338-8348.

- Patch-Matching Based NST

CNNMRF

C. Li and M. Wand, "Combining markov random fields and convo- lutional neural networks for image synthesis," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2479–2486.

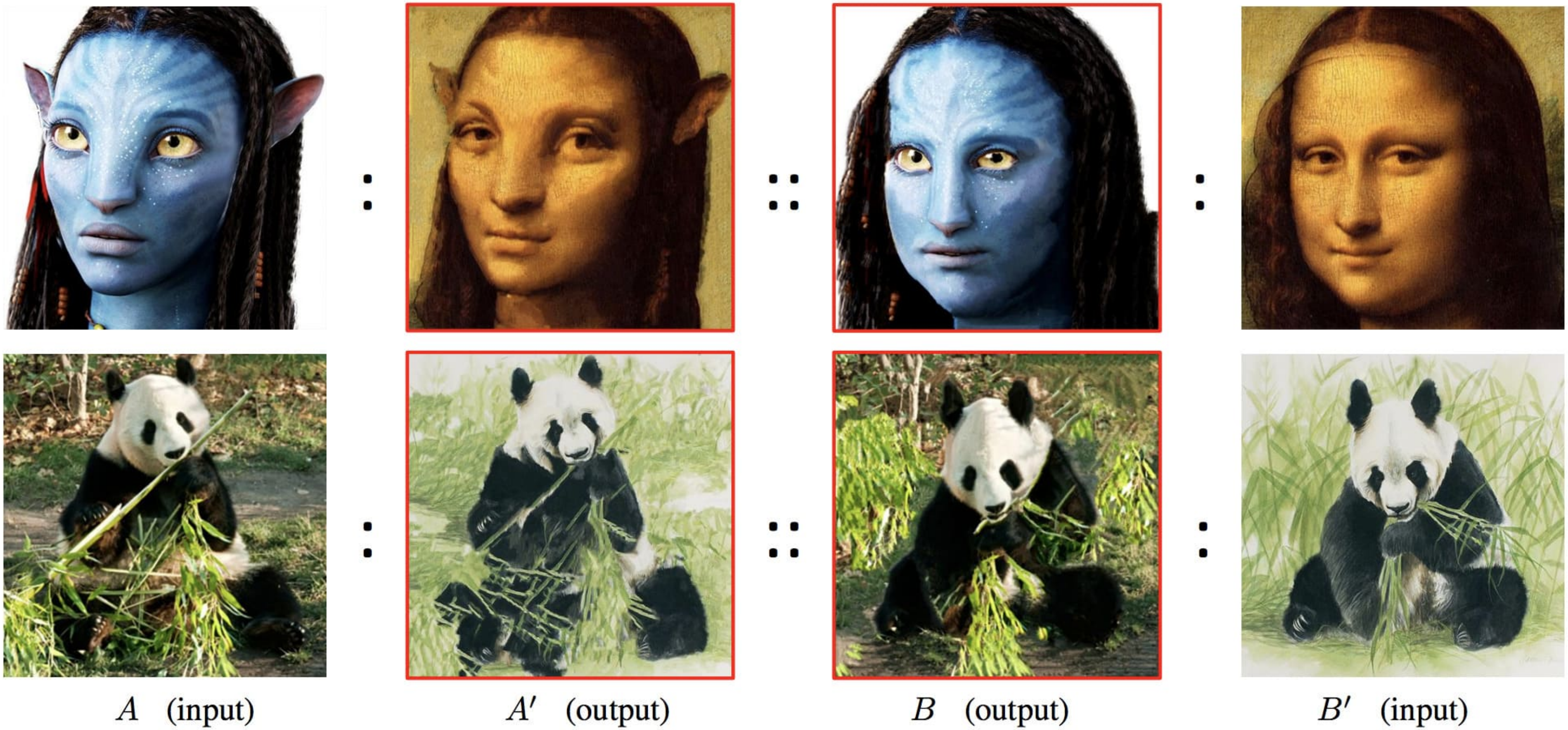
$$E_s(\Phi(\hat{y}), \Phi(y_s)) = \sum_{i=1}^{m} ||\Psi_i(\Phi(\hat{y})) - \Psi_{NN(i)}(\Phi(y_s))||^2$$

$$NN(i) := \underset{j=1,\ldots,m_s}{\arg\max} \frac{\Psi_i(\Phi(\hat{y})) \cdot \Psi_j(\Phi(y_s))}{||\Psi_i(\Phi(\hat{y}))|| \cdot ||\Psi_j(\Phi(y_s))||}$$

Related Work

- Patch-Matching Based NST

Deep Image Analogy



Liao J, Yao Y, Yuan L, et al. Visual attribute transfer through deep image analogy[J]. ACM Transactions on Graphics, 2017, 36(4): 1-15.

• Patch-Matching Based NST

Deep Image Analogy

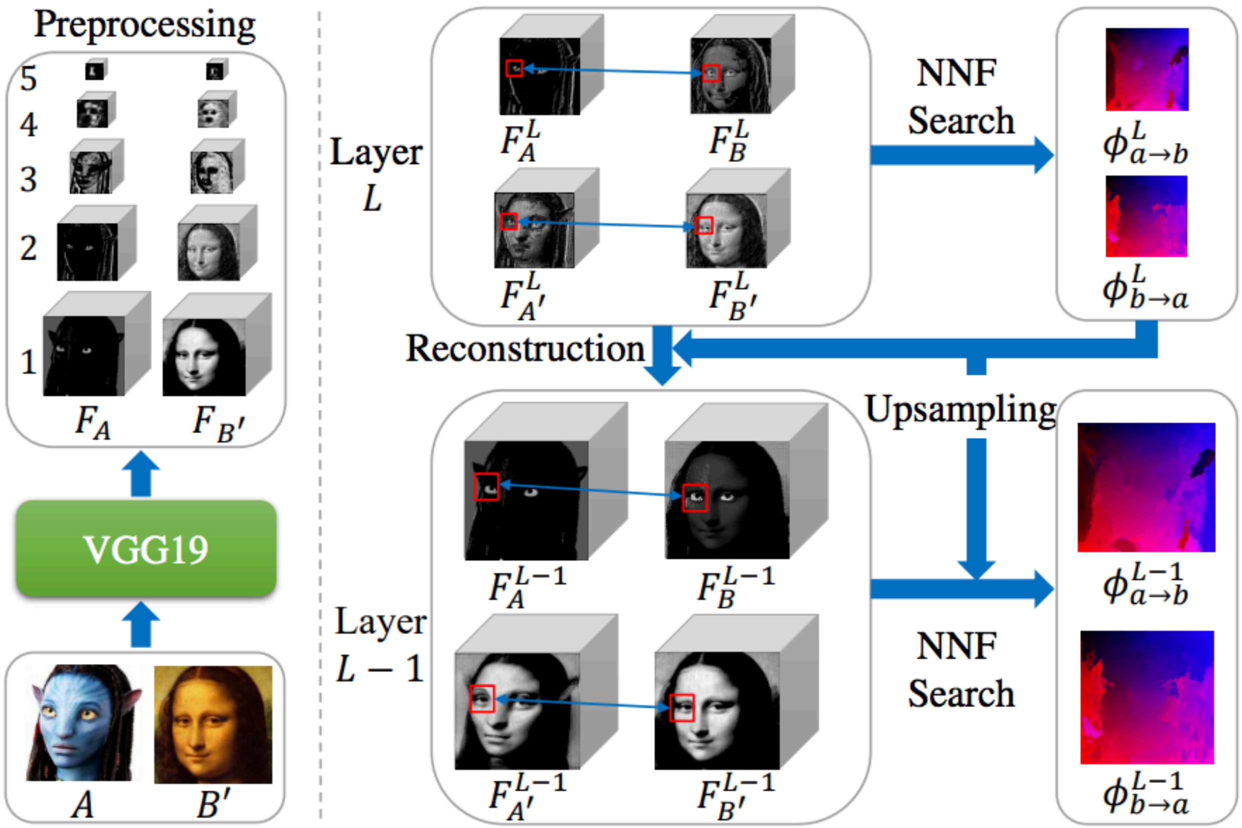
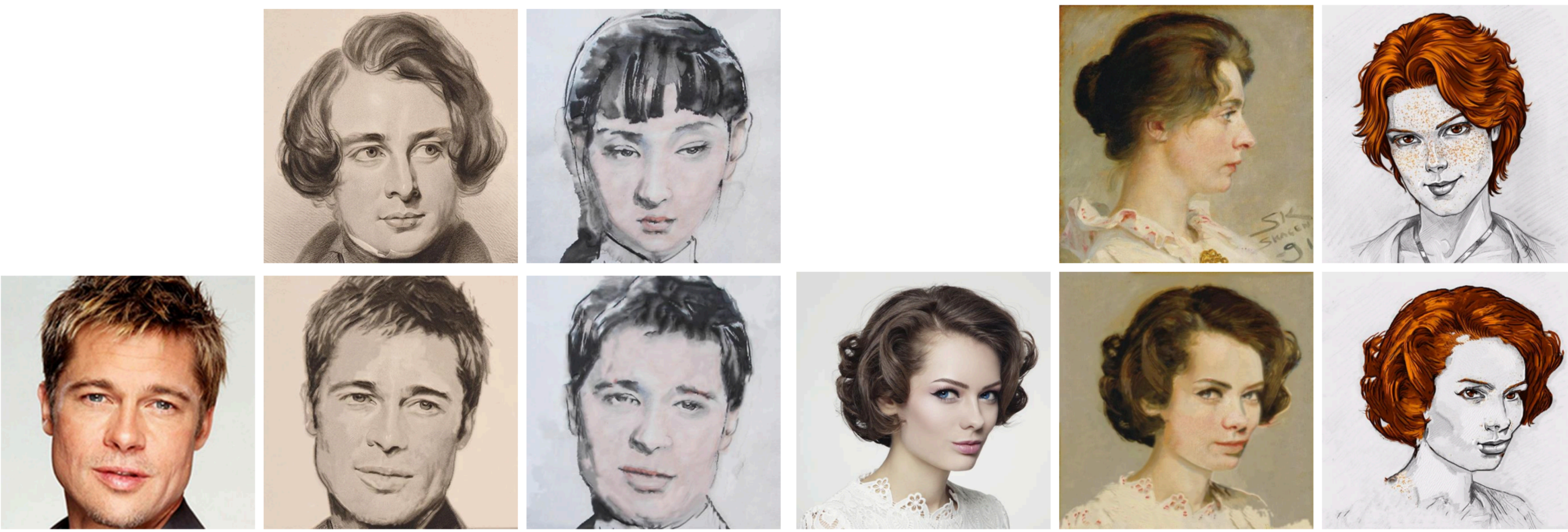


Figure 4: *System pipeline.*

Related Work

- Patch-Matching Based NST

Experiment results



Related Work

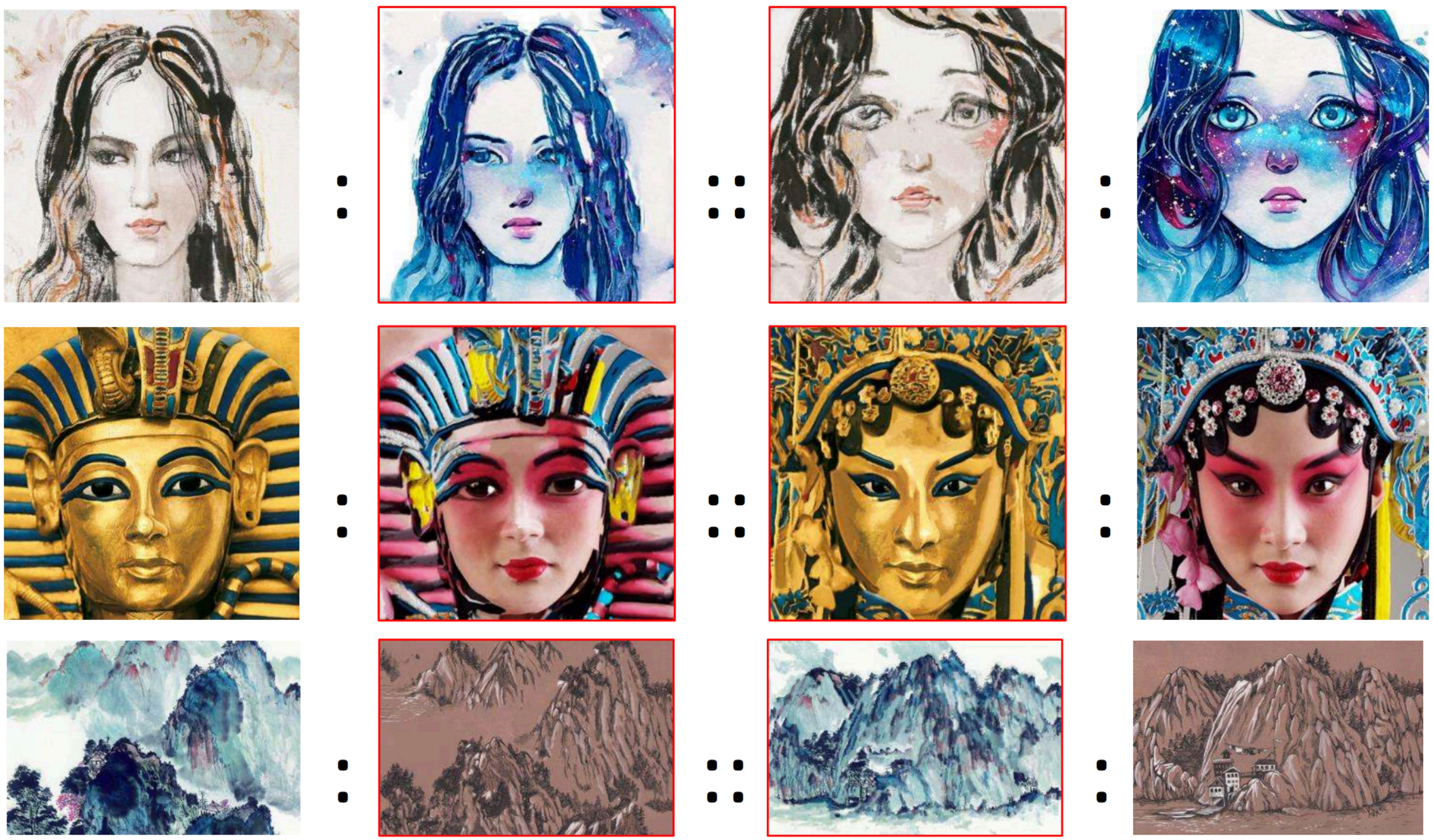
Seminar 2023/9/10

- Patch-Matching Based NST

Experiment results

- Patch-Matching Based NST

Experiment results

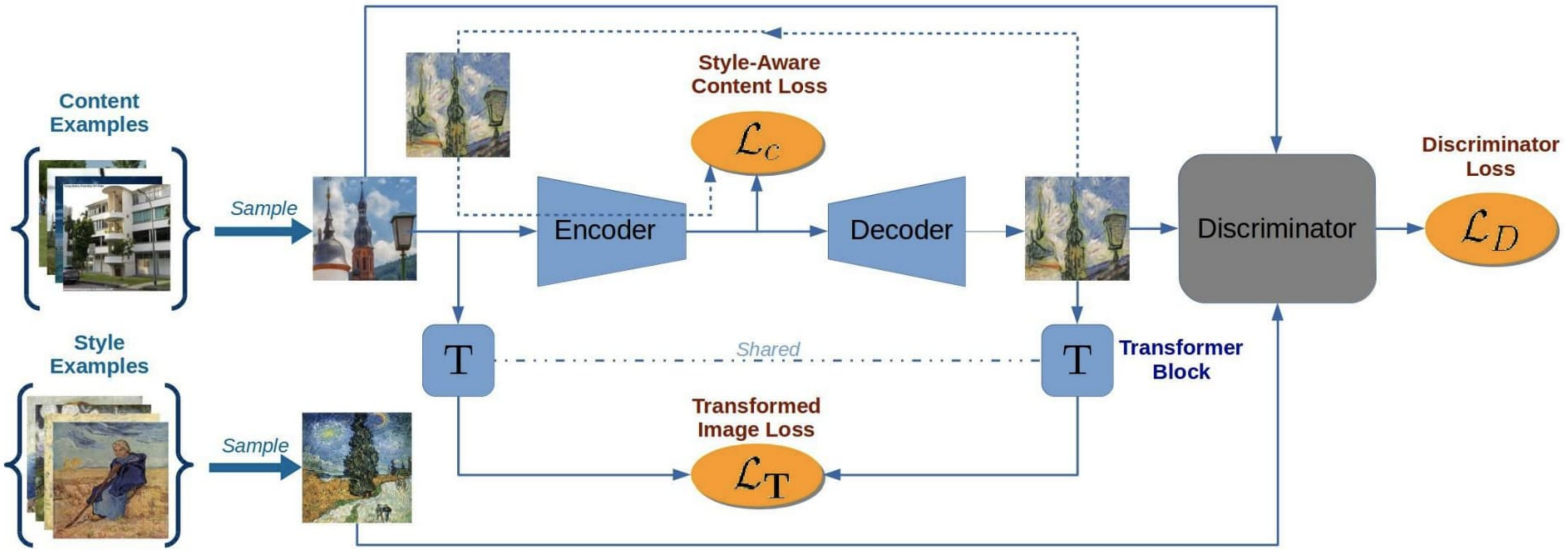


A (input) A' (output) B (output) B' (input)

Related Work

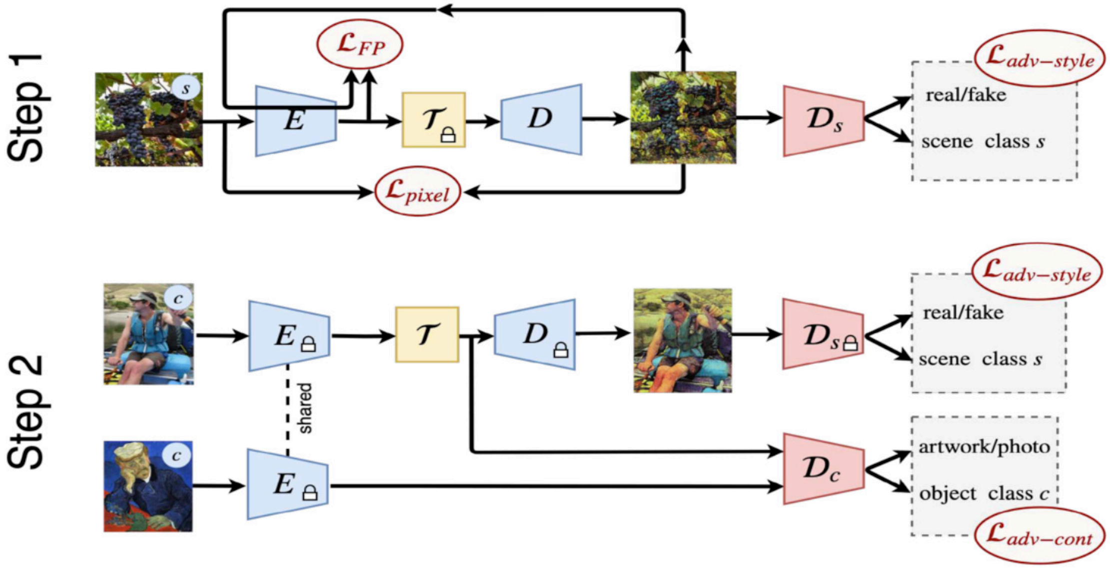
• GAN based NST

Sanakoyeu A, Kotovenko D, Lang S, et al. A style-aware content loss for real-time hd style transfer[C]//proceedings of the European conference on computer vision (ECCV). 2018: 698-714.



Generated Monet Art painting patches

- GAN based NST



D. Kotovenko, A. Sanakoyeu, P. Ma, S. Lang, and B. Ommer, "A content transformation block for image style transfer," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 10 032–10 041.

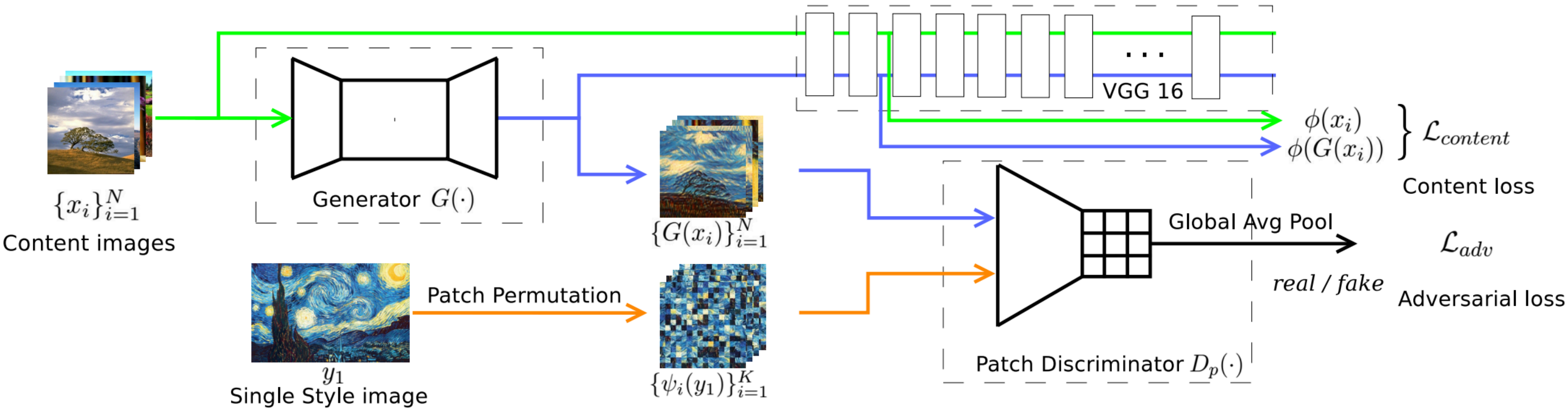
梵高风格

毕加索风格

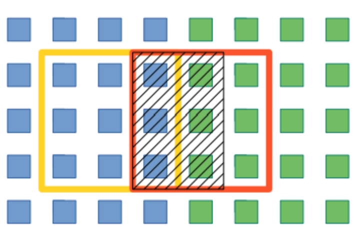
基尔希纳风格

Related Work

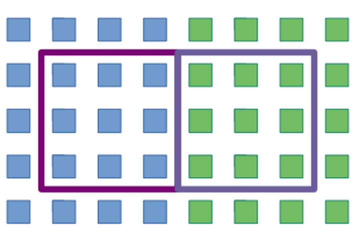
• GAN based NST



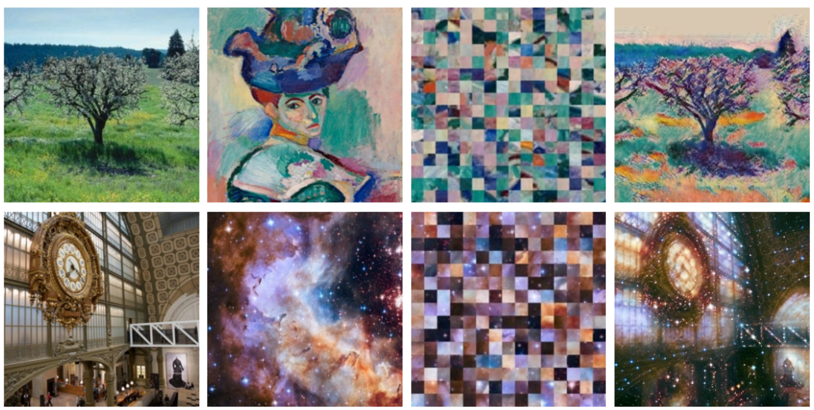
Zheng Z, Liu J. P2-GAN: efficient style transfer using single style image[J]. arXiv preprint arXiv:2001.07466, 2020.



(a) stride = 2, kernel_size = 3



(b) stride = kernel_size = 3

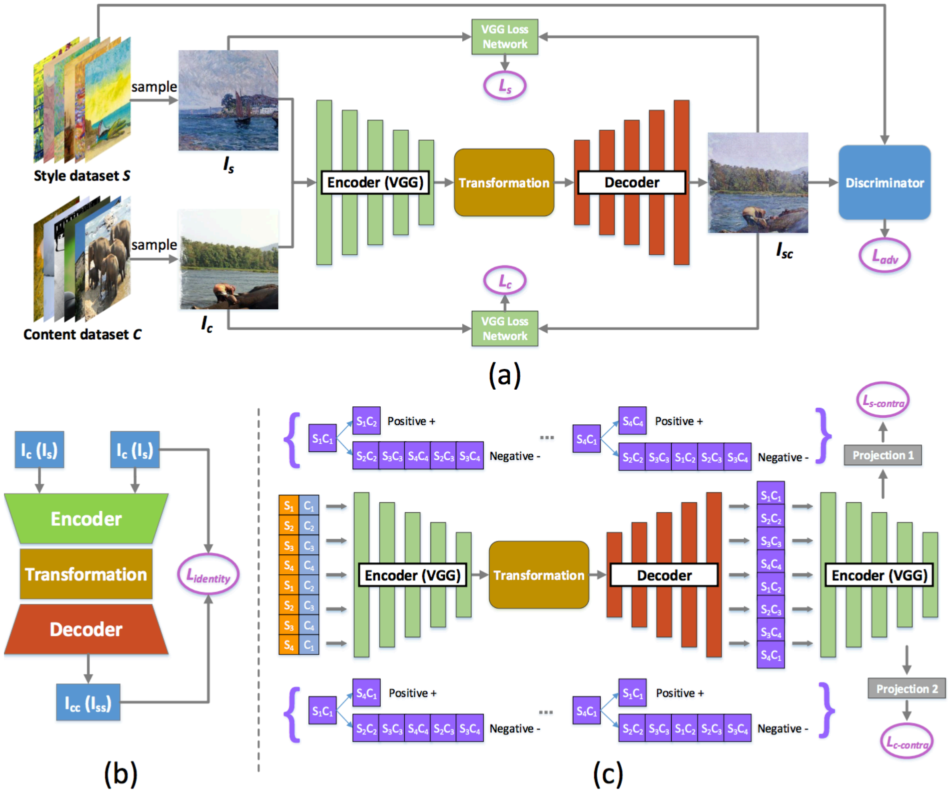


(a) Content images (b) Style images (c) Patch Permutation (d) Our results

- CL based NST

Artistic Style Transfer with Internal-external Learning and Contrastive Learning

Chen H, Wang Z, Zhang H, et al. Artistic style transfer with internal-external learning and contrastive learning[J]. Advances in Neural Information Processing Systems, 2021, 34: 26561-26573.



$$\mathcal{L}_{c-contra} := -log(\frac{exp(l_c(s_ic_j)^T l_c(s_yc_j)/\tau)}{exp(l_c(s_ic_j)^T l_c(s_yc_j)/\tau) + \sum exp(l_c(s_ic_j)^T l_c(s_mc_n)/\tau)})$$

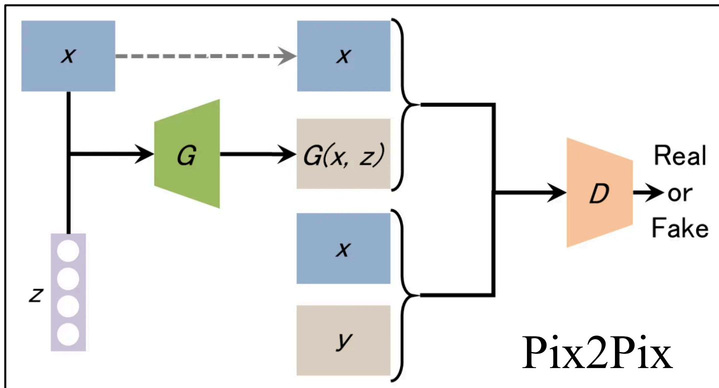
$$\mathcal{L}_{s-contra} := -log(\frac{exp(l_s(s_ic_j)^T l_s(s_ic_x)/\tau)}{exp(l_s(s_ic_j)^T l_s(s_ic_x)/\tau) + \sum exp(l_s(s_ic_j)^T l_s(s_mc_n)/\tau)})$$

Related Work

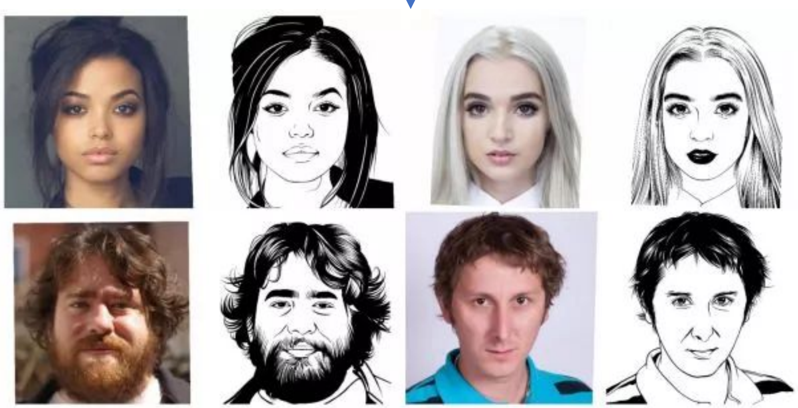
Seminar 2023/9/10

• Generalized NST

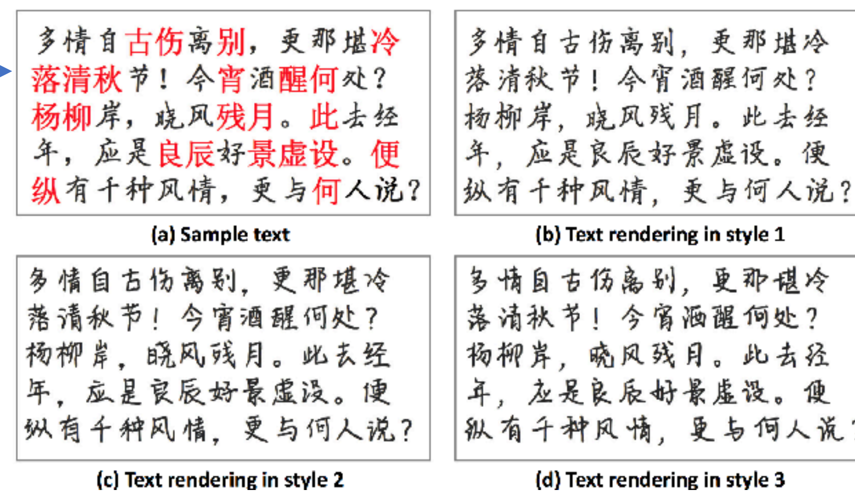
Isola P, Zhu J Y, Zhou T, et al. Image-to-image translation with conditional adversarial networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 1125-1134.



Qi X, Sun M, Wang W, et al. Face sketch synthesis via semantic-driven generative adversarial network[C]//2021 IEEE International Joint Conference on Biometrics (IJCB). IEEE, 2021: 1-8.



Yi R, Liu Y J, Lai Y K, et al. Apdrawinggan: Generating artistic portrait drawings from face photos with hierarchical gans[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 10743-10752.



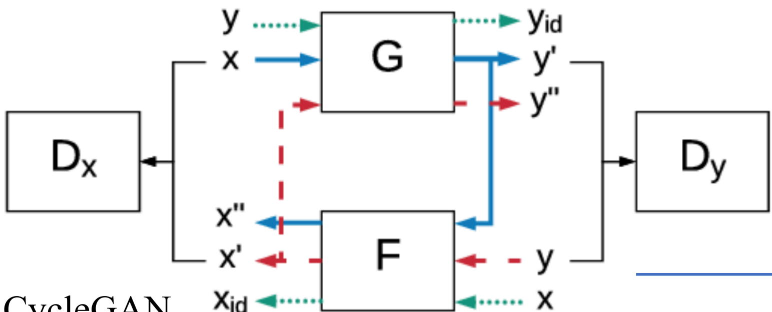
Jiang Y, Lian Z, Tang Y, et al. DCFont: an end-to-end deep Chinese font generation system[M]//SIGGRAPH Asia 2017 Technical Briefs. 2017: 1-4.

Related Work

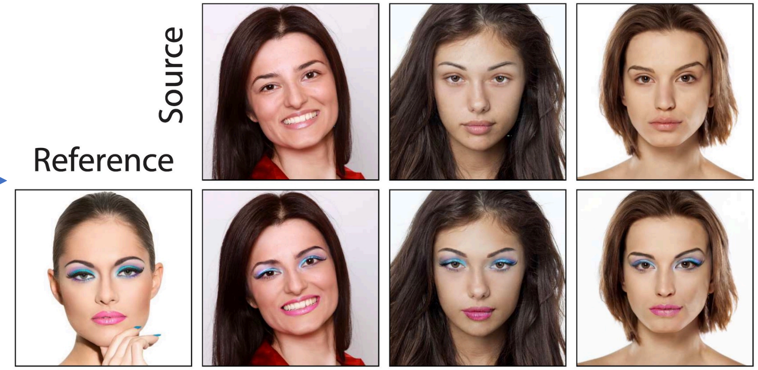
Seminar 2023/9/10

• Generalized NST

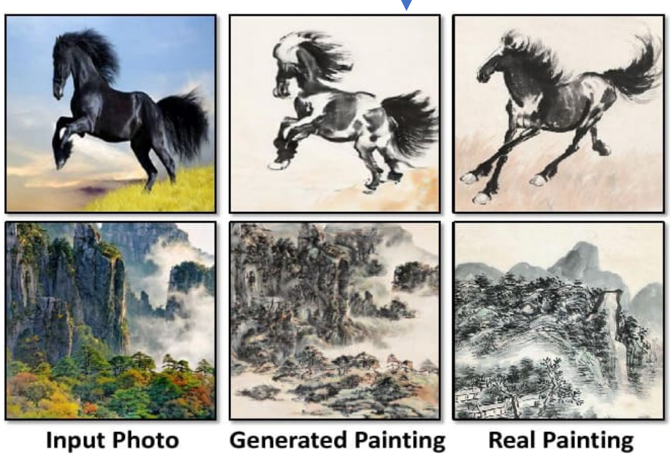
Zhu J Y, Park T, Isola P, et al. Unpaired image-to-image translation using cycle-consistent adversarial networks[C]//Proceedings of the IEEE international conference on computer vision. 2017: 2223-2232.



CycleGAN



Chang H, Lu J, Yu F, et al. Pairedcyclegan: Asymmetric style transfer for applying and removing makeup[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 40-48.



He B, Gao F, Ma D, et al. Chipgan: A generative adversarial network for chinese ink wash painting style transfer[C]//Proceedings of the 26th ACM international conference on Multimedia. 2018: 1172-1180.

Gao X, Zhang Y, Tian Y. Learning to Incorporate Texture Saliency Adaptive Attention to Image Cartoonization[C]//International Conference on Machine Learning. PMLR, 2022: 7183-7207.

Contents

Authors

First author:

Yuxin Zhang 2020 PhD at

Institute of Automation, Chinese Academy of Sciences

- Domain Enhanced Arbitrary Image Style Transfer via Contrastive Learning（SIGGRAPH 2022）
- Inversion-Based Style Transfer with Diffusion Models (CVPR 2023)
- A Unified Arbitrary Style Transfer Framework via Adaptive Contrastive Learning (ToG 2023)

Authors

Corresponding author:

Weiming Dong, Professor, Institute of Automation, Chinese Academy of Sciences

2016-11至今， 中国科学院自动化研究所**模式识别国家重点实验室**，研究员

2010-11~2016-10，中国科学院自动化研究所**模式识别国家重点实验室**，副研究员

2009-11~2010-10，中国科学院自动化研究所**模式识别国家重点实验室**，助理研究员

2007-10~2009-10，中国科学院自动化研究所**中欧信息、自动化与应用数学联合实验室**，博士后

2004-04~2007-06，法国国立信息与自动化研究院（INRIA）/法国**亨利▪庞加莱南锡第一大学**，博士

2001-09~2004-01，**清华大学**计算机科学与技术系，工学硕士

1997-09~2001-07，**清华大学**计算机科学与技术系，工学学士

Authors

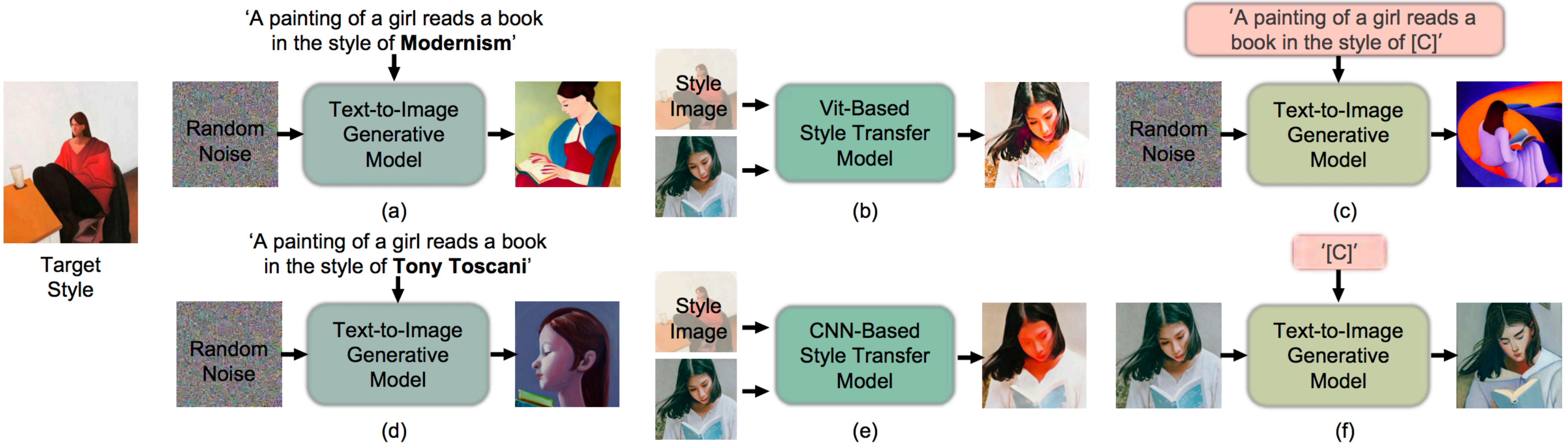
Corresponding author:

Weiming Dong, Professor, Institute of Automation, Chinese Academy of Sciences

Contents

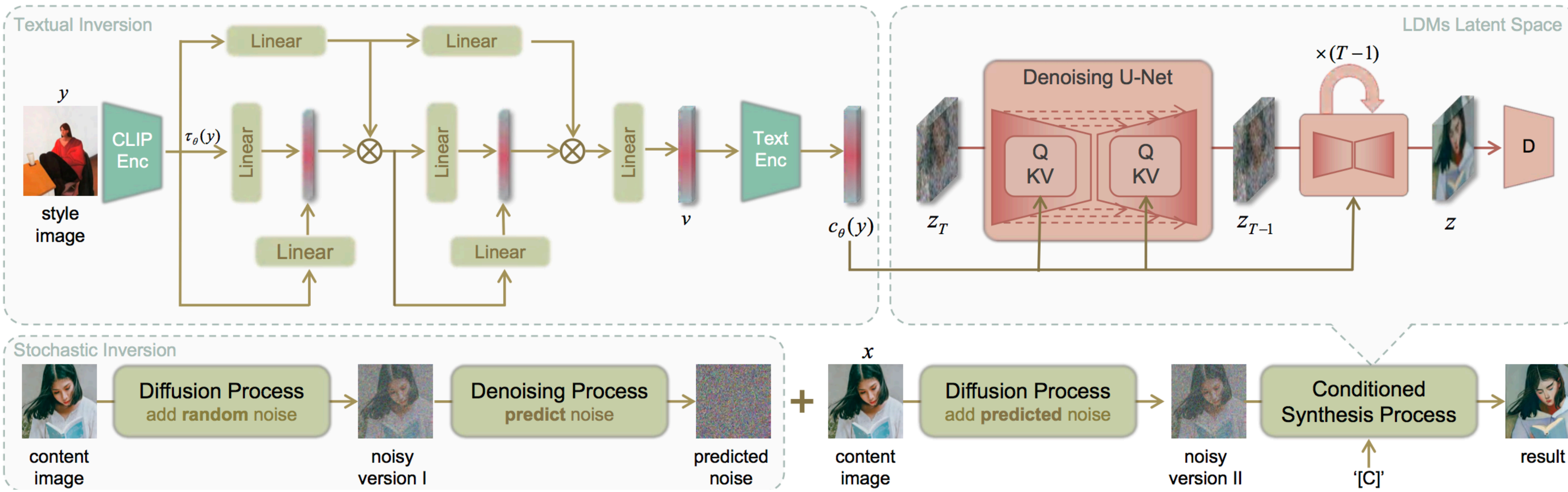
Paper Reading

Seminar 2023/9/10



If a photo speaks 1000 words, then every painting tells a story.

Model architecture

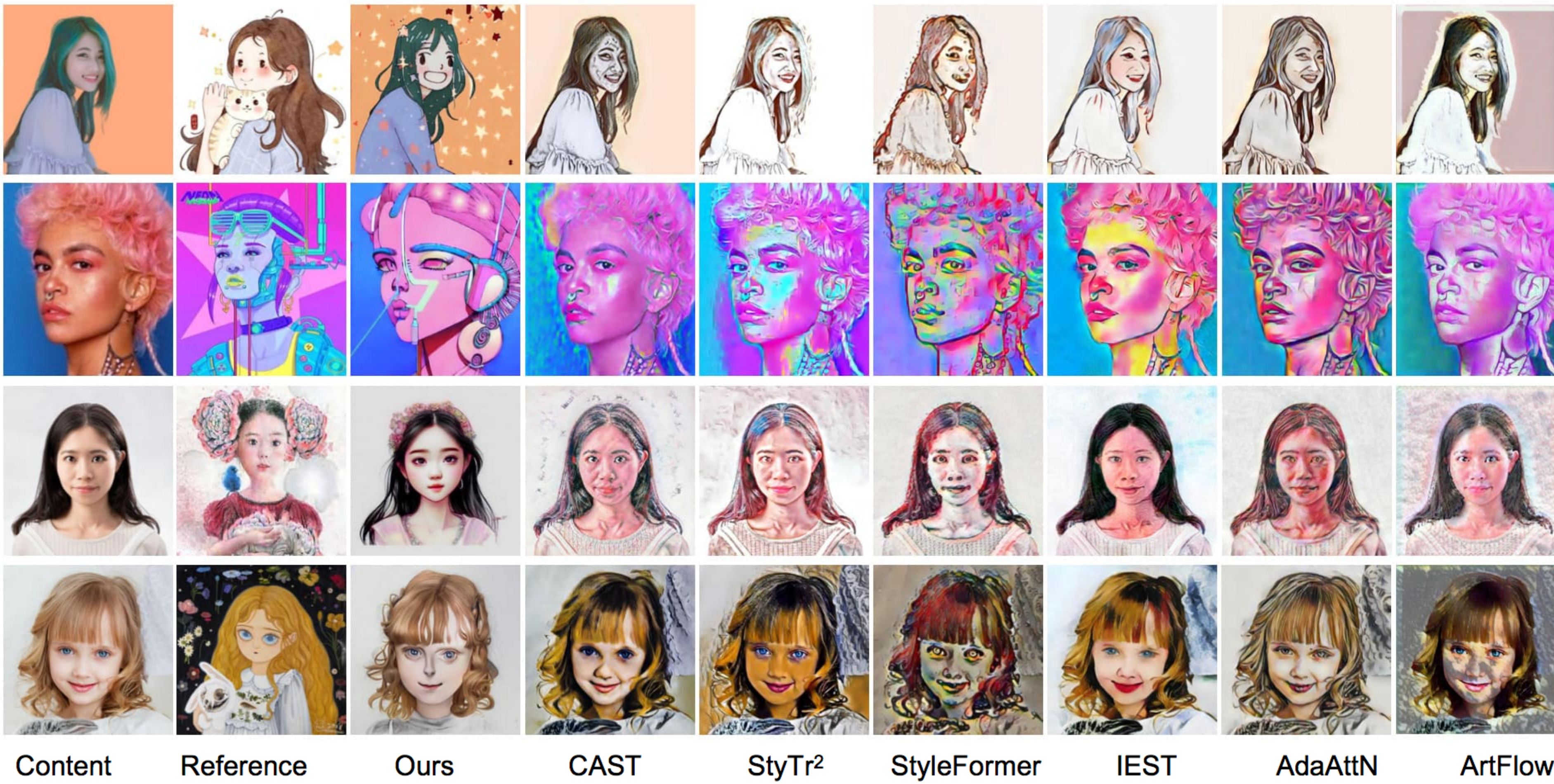


$$\hat{v} = \arg\min_{v} \mathbb{E}_{z,x,y,t} \left[\left\| \epsilon - \epsilon_\theta\left(z_t, t, \mathrm{MultiAtt}(\tau_\theta(y)\right)\right\|_2^2 \right]$$

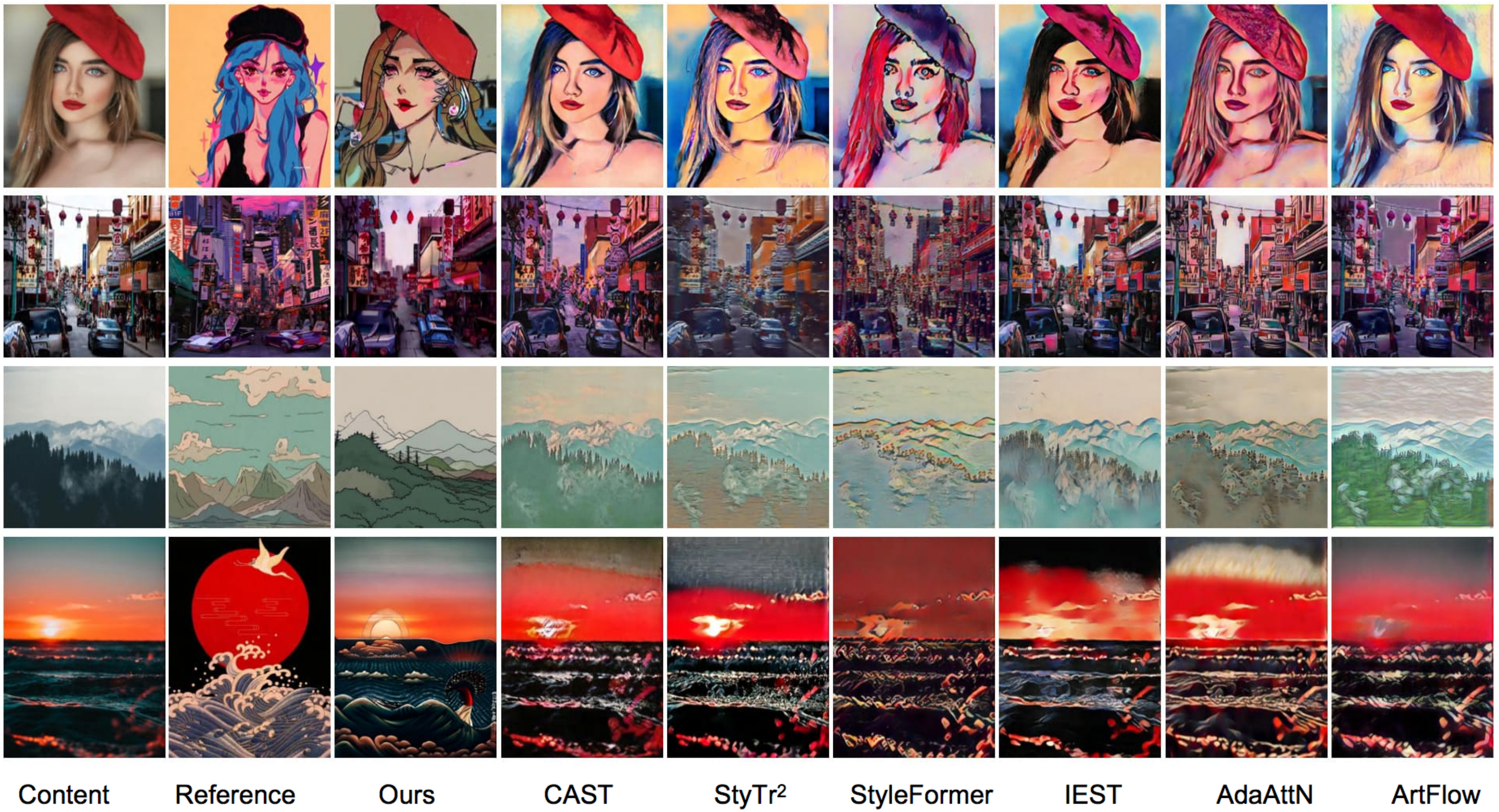
CLIP text embedding process

```python
class CLIPTextEmbedder(nn.Module):

    def __init__(self, version: str = "openai/clip-vit-large-patch14", device="cuda:0", max_length: int = 77):

        super().__init__()

        self.tokenizer = CLIPTokenizer.from_pretrained(version)

        self.transformer = CLIPTextModel.from_pretrained(version).eval()

        self.device = device
        self.max_length = max_length

    def forward(self, prompts: List[str]):

        batch_encoding = self.tokenizer(prompts, truncation=True, max_length=self.max_length, return_length=True,
                                        return_overflowing_tokens=False, padding="max_length", return_tensors="pt")

        tokens = batch_encoding["input_ids"].to(self.device)

        return self.transformer(input_ids=tokens).last_hidden_state
```

Experiment results



Experiment results



Experiment results

text-to-image synthesis with placeholder

Experiment results

Reference | Content | ‘A comic of a woman’ | ‘A cartoon of a woman’ | ‘A caricature of a woman’ | ‘A paintingof a woman’ | Ours

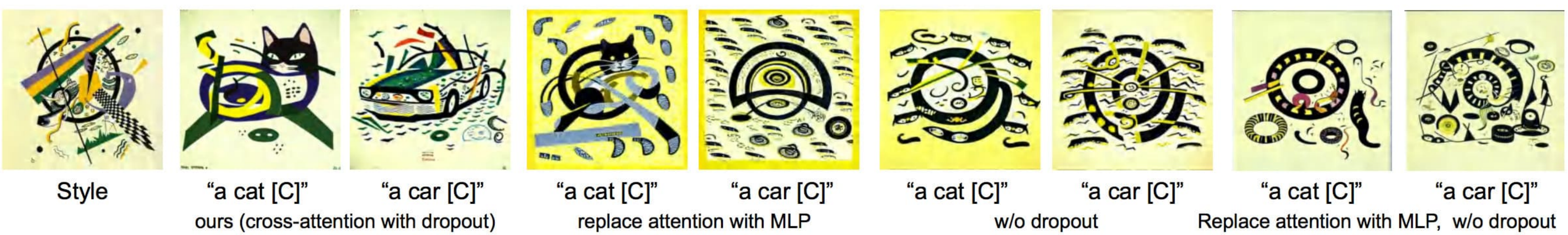
Comparison with SDM conditioned on human caption

	CAST [59]	StyTr^2 [7]	StyleFormer [55]	IEST [3]	AdaAttN [34]	ArtFlow [1]	TexIn [15]	
							img2img	txt2img
Preference↑	0.368	0.310	0.218	0.161	0.310	0.276	0.379	0.121
Ours	**0.632**	**0.690**	**0.782**	**0.839**	**0.690**	**0.724**	**0.621**	**0.879**

User Study

Experiment results

(a) The ablation study of our stochastic inversion and the impact of hyper parameter strength



(b) The ablation study of cross-attention and dropout.

Ablation Study

The End

Thanks for your attention.