TOMASZ OGIOŁDA & SZYMON PAJĄK

Neural Style Transfer

COMPUTATIONAL INTELLIGENCE

Based on the following papers

Xu, Y., Xia, M., Hu, K., Zhou, S., & Weng, L. (2025). **Style Transfer Review: Traditional Machine Learning to Deep Learning. Information,** 16(2), 157. https://doi.org/10.3390/info16020157

Zhang, S., Kang, H., Liu, Y., Mei, F., & Li, H. (2025). **HSI: A Holistic Style Injector for Arbitrary Style Transfer.** In Proceedings of the Computer Vision and Pattern Recognition Conference (pp. 23433-23442).

Yu, X., Tian, J., & Hu, Z. (2024). **An Analysis for Image-to-Image Translation and Style Transfer**. arXiv preprint arXiv:2408.06000.

Yeo, Yunha & Um, Daeho. (2025). **Can AI Recognize the Style of Art? Analyzing Aesthetics through the Lens of Style Transfer.** 10.48550/arXiv.2504.14272.

Zhu, Qiang & Lu, Kuan & Huo, Menghao & Li, Yuxiao. (2025). **Image-to-Image Translation with Diffusion Transformers and CLIP-Based Image Conditioning.** 10.48550/arXiv.2505.16001.

Style Transfer Tasks

SINGLE-STYLE TRANSFER

Uses a separate network trained for each style.

MULTI-STYLE TRANSFER

Uses one model for several styles.

ARBITRARY STYLE TRANSFER

Allows transfer to any style. Desired style image on input.

Style Transfer Architectures

CNN

Uses feature correlation (e.g., Gram Matrix) to define and transfer style.

GAN

Uses a Generator to create styled images and a Discriminator to judge realism/style match.

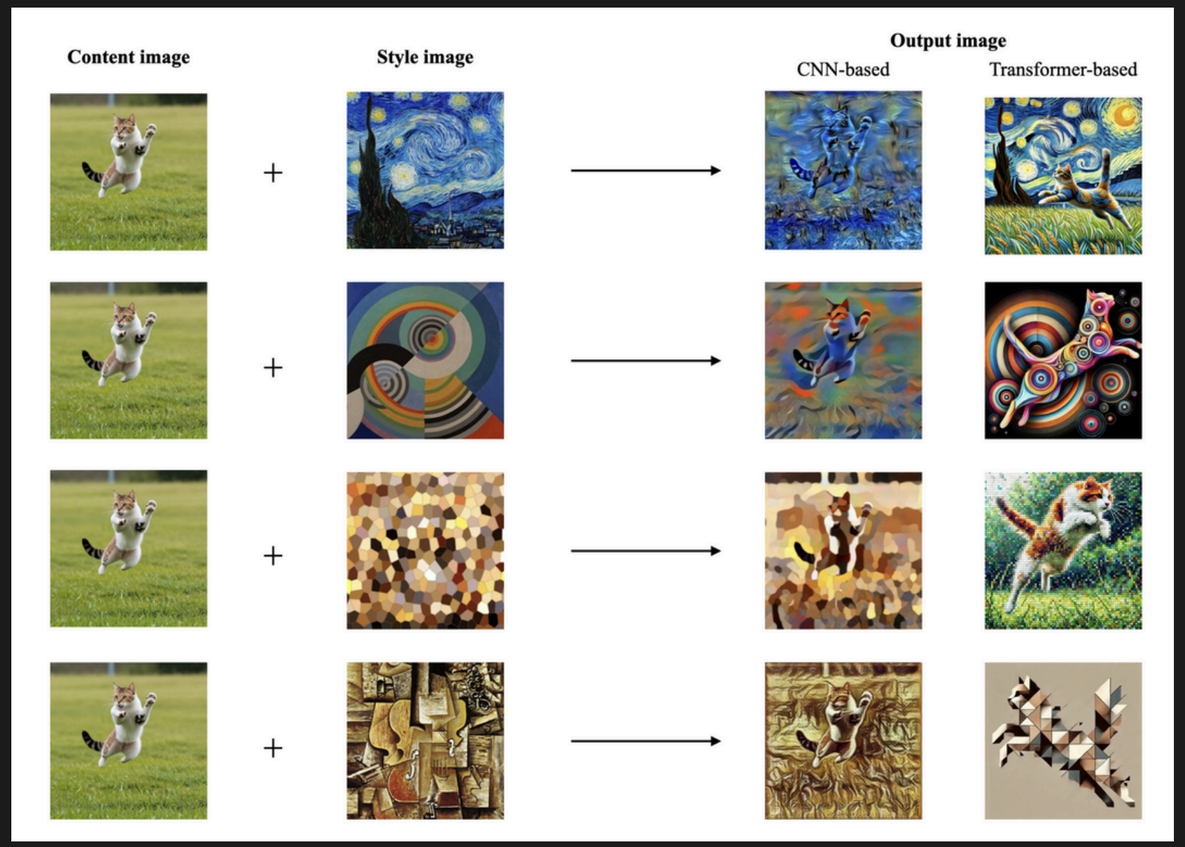
ATTENTION BASED

Uses attention mechanism to focus on relevant style features for better local transfer.

DIFFUSION BASED

Generates images by iteratively denoising a random noise input, guided by content and style.

Something about CNNs



- CNNs capture superficial visual elements like color and texture in images.
- The CNN mechanism achieves this by using a loss function that preserves the content image's boundaries and spatial arrangement.
- However, this local focus limits CNN-based models, making them largely incapable of comprehending deeper artistic concepts, or global cultural elements embedded in a style.
- The core mechanism is to minimize the total loss, which is the sum of a content loss function (preserving content) and a style loss function (using a Gram matrix to reflect style).

Something about Diffusion

- Diffusion models, particularly the Diffusion Transformer (DiT) architecture, offer a stable and scalable alternative for image-to-image style transfer.

- They operate in a latent space for efficiency and leverage the self-attention mechanism of transformers to capture global contextual relationships.

- **Semantic Conditioning**
  - Contrastive Language–Image Pre-training (CLIP) is a model trained to align text and image embeddings in a shared space, granting it a strong semantic understanding of visual content.
  - In the paper, the pre-trained CLIP image encoder is used to extract a semantic feature vector from the source image, serving as a powerful conditioning signal for the Diffusion Transformer without needing text prompts or class labels.

Evaluation Metrics

Used to quantify the quality, fidelity, and realism of images generated by style transfer or image-to-image translation models.

Content Preservation Measures

DOMAIN-INVARIANT PERCEPTUAL DISTANCE (DIPD)

- Measure content similarity between the generated image and the corresponding input image.

STRUCTURAL SIMILARITY INDEX MEASURE (SSIM)

- Calculates a score (between 0 and 1) by comparing the similarity of two images in terms of brightness, contrast, and structure.
- A higher score indicates greater similarity

LEARNED PERCEPTUAL IMAGE PATCH SIMILARITY (LPIPS)

- Assesses the perceptual similarity of images by comparing their deep features (extracted through a CNN like VGG)

Image Realism and Distribution Measures

FRÉCHET INCEPTION DISTANCE (FID)

- Measures the distributional difference between generated and real images in feature space using the Fréchet distance
- It compares the mean (μ) and covariance (Σ) of the distributions. A lower FID value indicates higher realism.

$$d^2 = |\mu_X - \mu_Y|^2 + \mathrm{tr}(\Sigma_X + \Sigma_Y - 2(\Sigma_X \Sigma_Y)^{1/2})$$

- μ_r, Σ_r = mean and covariance of **real** image features
- μ_g, Σ_g = mean and covariance of **generated** image features

KERNEL INCEPTION DISTANCE (KID)

- Similar to FID, KID is used to measure the authenticity of a batch of samples.

Holistic Style Injector (HSI)

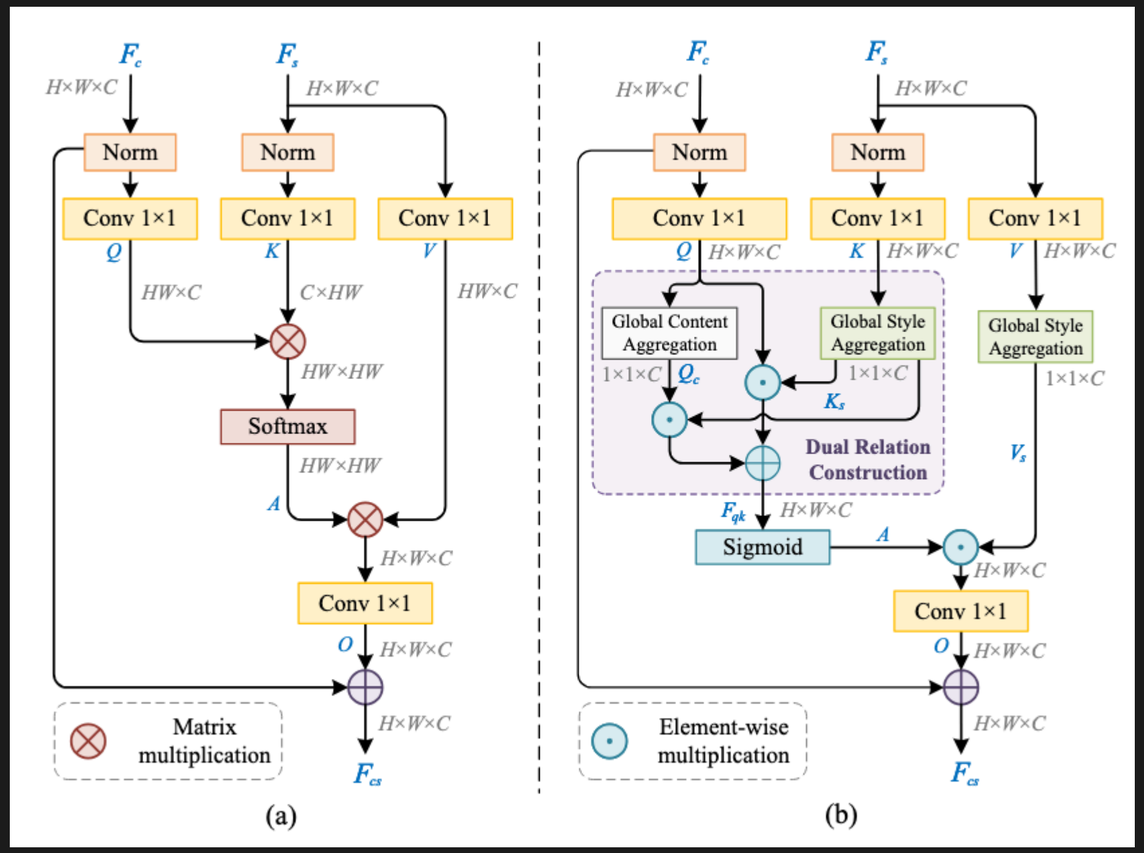


Image from paper by Zhang, S., Kang, H., Liu, Y., Mei, F., & Li, H. (2025). HSI: A Holistic Style Injector for Arbitrary Style Transfer.

The structure and feature encoding process comparison with self-attention (a) is standard self-atention, (b) is HSI

Attention-based style transfer model. A novel approach for style transfer.

Linear Computational Complexity

HSI achieves linear computational complexity (O(H×W×C)) by replacing the resource-intensive matrix multiplication of traditional attention with efficient element-wise multiplication.

Authors state

This crucial innovation eliminates the quadratic complexity that caused GPU memory overflow (OOM) in previous attention-based methods (like SANet and AdaAttN) when processing high-resolution images

Holistic Style Injector (HSI)

Global style aware
HSI avoids style bias and repetitive local distortions, it utilizes four channel-wise global statistical features—mean (μ), variance (σ2), skewness (γ1), and kurtosis (γ2)

Simultaneously constructs local-content-to-global-style and global-content-to-global-style relations. The weighting between these two relations is dynamically adjusted based on the semantic similarity of the content and style images

Project Decisions Left

STYLE TRANSFER TASK

Eiter single-style, multi-style or arbitrary style transfer.

TARGET STYLE(S)

Which styles our models will support.

EVALUATION METRICS

How to measure performance of our models?

SPECIFIC MODEL ARCHITECTURES

We'd like to compare 3 architectures, the specific model archs will be decided. Depends on the style transfer task. The ones we consider are:

GAN
- CycleGAN
- DE-GAN (Depth Extraction GAN)
- ESA-CycleGAN

Attention-based
- AdaAttN (Adaptive Attention Normalization)
- StyTr2

Diffusion-based
- Some of LDMs (Latent Diffusion Models) due to computational complexity
- InST (Inversion-based Style Transfer)

THANK YOU! 🎨🤖