Style Transfer using Stable Diffusion

Zuzanna Warchoł Filip Dąbkowski Sergiusz Pyskowacki

Gdańsk University of Technology

What is Style Transfer ?

Neural Style Transfer is a technique that takes two images—a content image and a style image—and blends them together in such a way that the resulting image combines the content of the first image with the style of the second.
For example, if our content image is a picture of a cat and our style image is a painting by Van Gogh, the output will be an image of the cat that looks as if it was painted by Van Gogh.

Diffusion Models

As the name suggests, diffusion models operate in a similar way to the physical phenomenon of diffusion. They are a class of score-based generative models, meaning that they learn the probability distribution of a dataset, enabling the generation of new samples from this distribution.

A simple way to visualize this is by imagining a drop of ink in a glass of water. Initially, the ink is concentrated, but over time, it spreads evenly throughout the water. Diffusion models apply this principle to images, gradually adding noise to training data until they become pure noise. Then, they learn how to reverse this process and reconstruct realistic images.

In the context of image-generation Diffusion Models, the data space consists of all training images of a given resolution, treated as data points. The underlying probability distribution, that the model seeks to learn, assigns high probability to realistic images and low probability to everything else. During the sampling process, a random point is drawn from the data space, initially representing random noise. Score-based models learn to predict a vector that, when added to this data point, increases its probability within the estimated distribution.

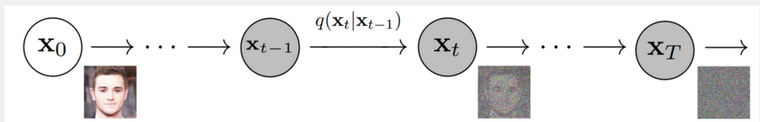
Key components:

1. Forward Diffusion Process
   Here we add Gaussian noise to the initial data in series of small steps moving it further from high-probability points and closer to the random noise during sampling.
2. Reverse Diffusion Process
   The model learns how to reconstruct the original data by reversing the forward distribution. Removing the noise increases its probability, bringing it closer to realistic images.

Mathematical approach

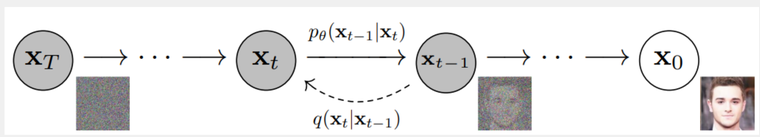
Forward process:

$$q(x_t|x_{t-1}) = N(x_t; \sqrt{\alpha_t}x_{t-1}, (1-\alpha_t)I) \quad (1)$$



Reverse process:

$$p_\theta(x_{t-1}|x_t) = N(x_{t-1}; \mu_\theta(x_t, t), \sigma_\theta(x_t, t)) \quad (2)$$

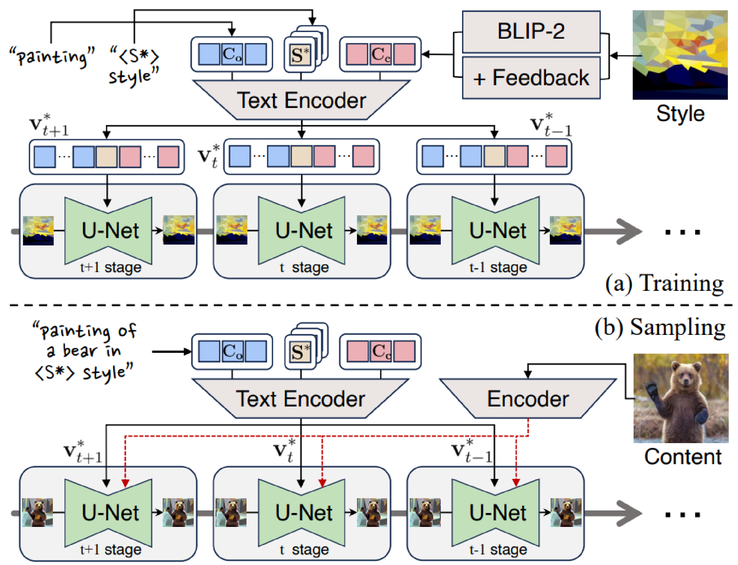


Methodology

Our project is based on the the *D*reamStyler paper, which builds upon Stable Diffusion and Textual Inversion for training while leveraging ControlNet to assist with style-transfer tasks.

- **Stable Diffusion**,is a groundbreaking Diffusion Model (DM) and an implementation of the Latent Diffusion Model (LDM), which operates in the latent space rather than directly in pixel space.
- **Contrastive Language-Image Pre-Training**(CLIP), is integrated into SD as its conditioning input for text prompts. CLIP itself was designed for multiple tasks, with its core purpose being to map the embedding of an image into the same space as the embedding of its textual description.
- **Textual Inversion**, is a technique for fine-tuning large text-guided image generation models to learn new concepts. Instead of fine-tuning the entire model, Textual Inversion freezes the diffusion model (DM) and trains a new token embedding vector.
- **ControlNet**, is a method for fine-tuning SD to accept additional conditioning inputs, such as depth maps, bounding boxes, and edge detections. ControlNet freezes Stable Diffusion's weights and trains only a small additional network, referred to as ControlNet.

Model Architecture



(a) DreamStyler constructs a training prompt with an opening text C_o, multi-stage style tokens S^*, and a context description C_c, which is captioned with BLIP-2 and human feedback. DreamStyler projects the training prompt into multi-stage textual embeddings $v^* = \{v_1^*, \ldots, v_T^*\}$, where T is the number of stages (a chunk of the denoising timestep). As a result, the denoising U-Net provides distinct textual information at each stage.
(b) DreamStyler prepares the textual embedding using a provided inference prompt. For style transfer, DreamStyler employs ControlNet to comprehend the context information from a content image.

Training and Fine-Tuning

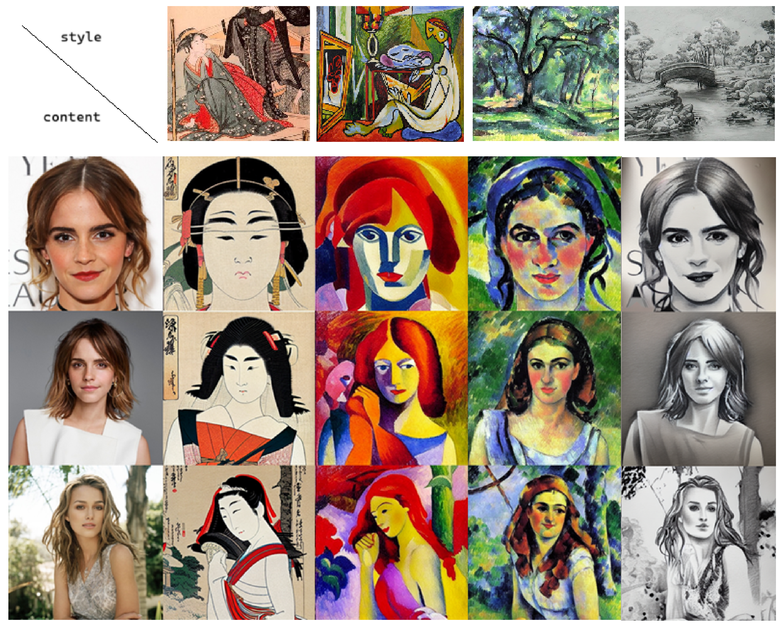
DreamStyler Model leverages Stable Diffusion (SD) for an image generation and Textual Inversion (TI) to learn an embedding vector representing the style of a given image. During training, we iteratively generate the style image using the prompt: *"Picture of C^* in the style of S^*"*

Where:

- S^* represents the learned style token and its corresponding embedding vector.
- C^* is a description of the training image, focusing on objects and composition.

Since Stable Diffusion weights remain frozen throughout training, this approach helps prevent overfitting and avoids distorting the pretrained SD model. However, the training process is highly dependent on the quality of the C^* prompt.

Style Transfer Exemple



References

Gatys et al., "A Neural Algorithm of Artistic Style," arXiv:1508.06576, 2015.
Ho et al., "Denoising Diffusion Probabilistic Models," NeurIPS, 2020.
GeeksforGeeks, "What are Diffusion Models?"
Ahn et al., "DreamStyler: Paint Your Style in Your Dream," arXiv:2309.06933, 2023.
Zhang et al., "Adding Conditional Control to Text-to-Image Diffusion Models," arXiv:2303.06131, 2023.
Gal et al., "Image Worth More Than a Thousand Words: Personalizing Text-to-Image Diffusion Models with Few-Shot Learning," arXiv:2203.06153, 2022.
Rombach et al., "High-Resolution Image Synthesis with Latent Diffusion Models," CVPR, 2022.