Diffusion in Style

Martin Nicolas Everaert Marco Bocchio Sami Arpa Sabine Süsstrunk Radhakrishna Achanta

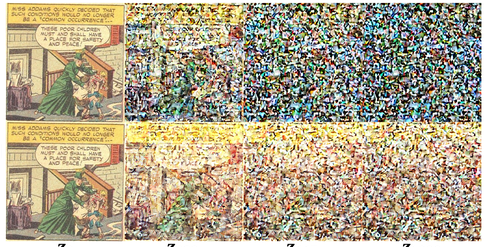
ICCV23 PARIS

Method

We perform **style adaptation** of Stable Diffusion by fine-tuning it with a **style-specific noise distribution** instead of the default $\mathcal{N}(\mathbf{0}_d, \mathbf{I}_{d\times d})$ [1,2].

Original diffusion $\mathcal{N}(\mathbf{0}_d, \mathbf{I}_{d\times d})$

Our style-adapted diffusion $\mathcal{N}(\boldsymbol{\mu}_{\text{style}}, \boldsymbol{\Sigma}_{\text{style}})$



$\mathbf{z}_0$ $\mathbf{z}_{333}$ $\mathbf{z}_{667}$ $\mathbf{z}_{1000}$

We compute the style-specific noise parameters $\boldsymbol{\mu}_{\text{style}}$ and $\boldsymbol{\Sigma}_{\text{style}}$ from **a small set of images of the desired style**.

Apart from the style-specific noise distribution $\mathcal{N}(\boldsymbol{\mu}_{\text{style}}, \boldsymbol{\Sigma}_{\text{style}})$, the fine-tuned model **can be used like Stable Diffusion**.

Intuition

The initial latent tensor $\hat{\boldsymbol{z}}_{1000}$ affects images composition and style, so adapting it to the style facilitates style adaptation.

Project website
https://ivrl.github.io/diffusion-in-style/

Acknowledgement:
This work is supported by Innosuisse grant 48552.1 IP-ICT.

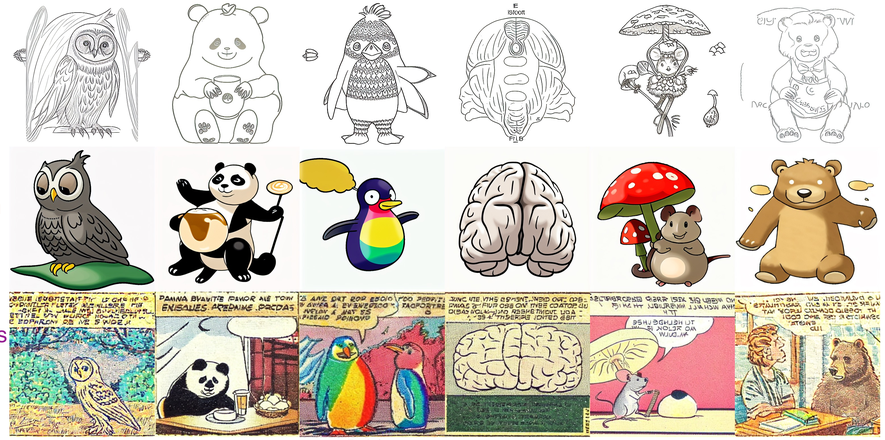
Results

We use our approach to fine-tune Stable Diffusion v1.5[1] to different styles, such as anime sketches, few-shot Pokemon images, and comics images.

Style 1, anime sketches[10]:

Style 2, few-shot Pokemon images[11]:

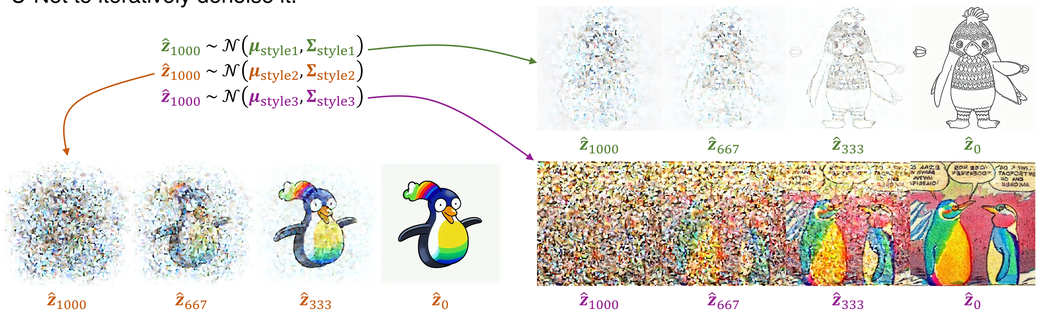
Style 3, comics images[12]:



A side view of an owl sitting in a field. | *A panda making latte art.* | *Rainbow coloured penguin.* | *A cross-section view of a brain.* | *A mouse using a mushroom as an umbrella.* | *A confused grizzly bear in calculus class.*

We sample the initial latent tensor $\hat{\boldsymbol{z}}_{1000}$ from the style-specific noise distribution and use the fine-tuned U-Net to iteratively denoise it.

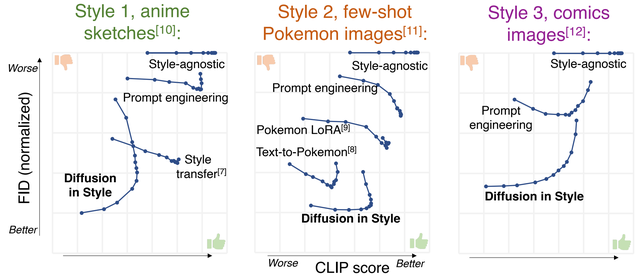
$\hat{\boldsymbol{z}}_{1000} \sim \mathcal{N}(\boldsymbol{\mu}_{\text{style1}}, \boldsymbol{\Sigma}_{\text{style1}})$
$\hat{\boldsymbol{z}}_{1000} \sim \mathcal{N}(\boldsymbol{\mu}_{\text{style2}}, \boldsymbol{\Sigma}_{\text{style2}})$
$\hat{\boldsymbol{z}}_{1000} \sim \mathcal{N}(\boldsymbol{\mu}_{\text{style3}}, \boldsymbol{\Sigma}_{\text{style3}})$



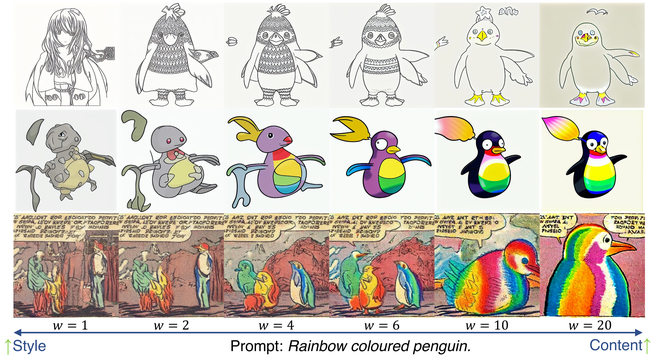
$\hat{\boldsymbol{z}}_{1000}$ $\hat{\boldsymbol{z}}_{667}$ $\hat{\boldsymbol{z}}_{333}$ $\hat{\boldsymbol{z}}_0$

Evaluation

Evaluating **CLIP**[3] **and FID**[4,5] **scores** on a range of guidance weights[6] w, **our method outperforms** prompt engineering, style transfer[7], and fine-tuning without noise distribution change[8,9].



The ⌋-shape of the curves indicates a **trade-off between style and content**.



References

[1] Rombach *et al.* CVPR 2022
[2] Ho *et al.* NeurIPS 2020
[3] Radford *et al.* PMLR 2021
[4] Heusel *et al.* NIPS 2017
[5] Wright *et al.* GCPR 2022
[6] Ho *et al.* NeurIPS Workshop 2021
[7] Chan *et al.* CVPR 2022
[8] Lambda Labs. Text-to-Pokemon model (2022), https://huggingface.co/lambdalabs/sd-pokemon-diffusers
[9] Paul. Pokemon LoRA model (2023), https://huggingface.co/sayakpaul/sd-model-finetuned-lora-t4
[10] Taebum, Anime Sketch Colorization dataset (2018), https://www.kaggle.com/datasets/ktaebum/anime-sketch-colorization-pair
[11] Liu *et al.* ICLR 2021, https://huggingface.co/datasets/huggan/few-shot-pokemon
[12] Simon & Kirby. 48 Famous Americans (1947), https://digitalcomicmuseum.com/index.php?did=24742