clip2latent: Text driven sampling of a pre-trained StyleGAN using denoising diffusion and CLIP

Justin N. M. Pinkney and Chuan Li
Lambda, Inc. San Francisco, USA

Abstract

We introduce clip2latent a new method to efficiently create text-to-image models from a pretrained CLIP and StyleGAN.

clip2latent enables text-driven sampling with an existing generative model without any external data or fine-tuning.

We train a diffusion model conditioned on CLIP embeddings to sample latent vectors of a pre-trained StyleGAN. Leveraging the alignment between CLIP's image and text embeddings we can avoid the need for any text labelled data for training.

clip2latent allows us to generate high-resolution (1024x1024 pixels) images based on text prompts with fast sampling, high image quality, and low training compute and data requirements.

We also show that the use of the well studied StyleGAN architecture, without further fine-tuning, allows us to directly apply existing methods to control and modify the generated images adding a further layer of control to our text-to-image pipeline.

Data Generation

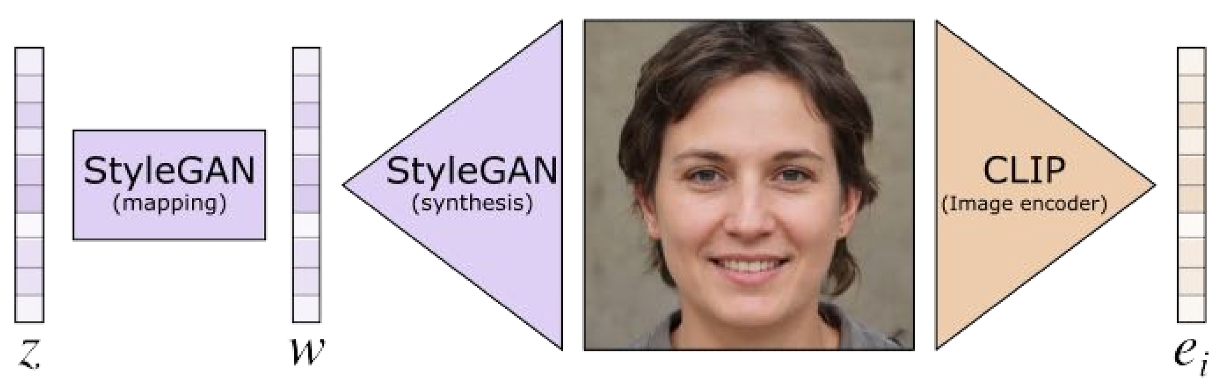
To generate our dataset we randomly sample StyleGAN latents and generate images from these, we then encode these images with CLIP. Giving us paired StyleGAN latent and CLIP image embedding training data.

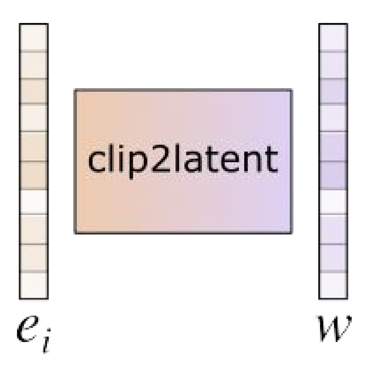
Training

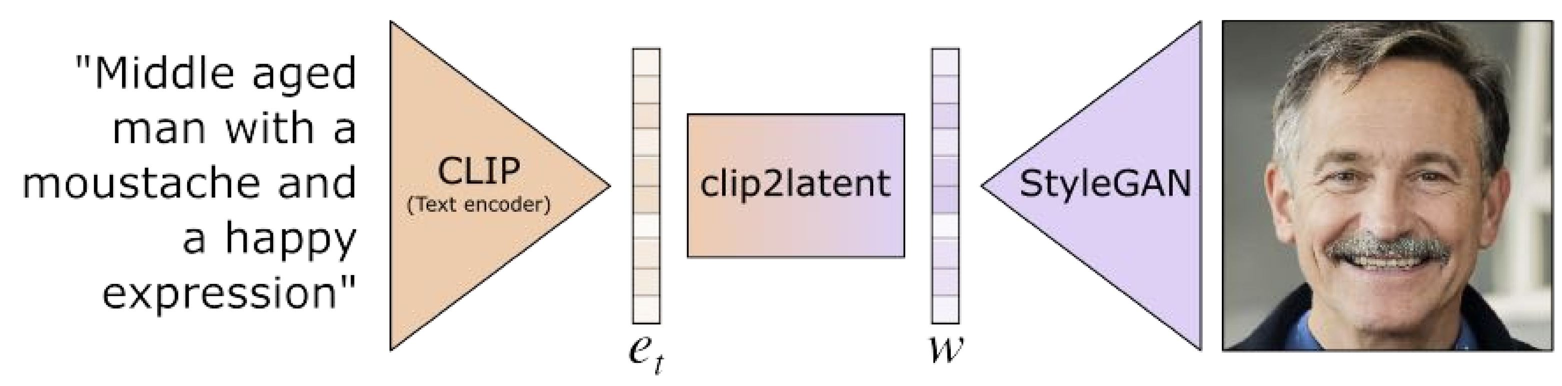
We train the clip2latent model using the same approach as the diffusion based prior from DALL-E 2. i.e. we train a denoising diffusion model to generate StyleGAN latent vectors conditioned on CLIP image embeddings. During training we add noise to the image embeddings to help the model generalise to text embeddings during inference.

Inference

Once we have trained clip2latent, we rely on the fact that CLIP can embed images and text into a shared latent space. At inference time we generate a CLIP embedding for a text description and use this as the conditioning to generate a StyleGAN latent vector, from which we can create a high-resolution image using the StyleGAN generator.







Results

Method	CLIP score	run-time (s)
clip2latent (ours)	0.316	11.760
clip2latent (ours) + timestep respacing	0.315	0.244
Direct Optimisation	0.321	19.191
LAFITE	0.278	0.045
DALLE-2	0.291*	—

Code and models available at:
`https://github.com/justinpinkney/clip2latent`

Try the model at Huggingface Spaces:
`huggingface.co/spaces/lambdalabs/clip2latent-demo`

Truncation

a British politician laughing happily | a Nigerian professor of economics | a person with very tight curly blonde hair

At high guidance scales our latent vectors can sometimes fall outside the typical domain of StyleGAN latent space, generating unnatural artefacts. We use the well-known technique of truncation to move the generated latent closer to the mean latent vector reducing artefacts.

Style mixing

a university graduate

We can add extra colour and lighting diversity to our generated images by performing Style Mixing. By mixing our generated latent vector with a randomly sampled latents for higher resolution layers we can create variations in colour, lighting and texture.

Editing

an arctic explorer, edited for: age, pose, smile

We can make use of the wealth of existing high-quality facial editing directions available for StyleGAN adding an extra level of control to our text-to-image pipeline.

Conclusion

Here we have used StyleGAN as our generator due to the wide range of pre-trained models available and its high-resolution, fast inference and state of the art performance in many domains. However, there is no reason a different GAN (e.g. BigGAN) or entirely different class of generative model (e.g. VAE) couldn't also be used.

We believe the application of diffusion models can allow the conditional sampling of previously unconditional models based on any image encoding, for example facial recognition/attribute networks or other classification models. We look forward to future applications of diffusion models as tools for arbitrarily mapping between latent spaces of pre-trained models.

Lambda