Adding Conditional Control to Text-to-Image Diffusion Models

arXiv 2302

Lvmin Zhang*, Maneesh Agrawala*
*Stanford University

马逸扬
2023/03/26

Content

- Authors
- Background
- Method
- Experiments

Content

- Authors
- Background
- Method
- Experiments

Background

First, we briefly review Diffusion models.

Diffusion models are a kind of generative models.

Generative models: to build a projection between a known distribution (*e.g.* Gaussian distribution) to an unknown distribution (*e.g.* natural image distribution).

Background

Diffusion models consist of two processes: *forward* and *reverse*.

- *forward* process:

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}) \tag{1}$$

- *reverse* process:

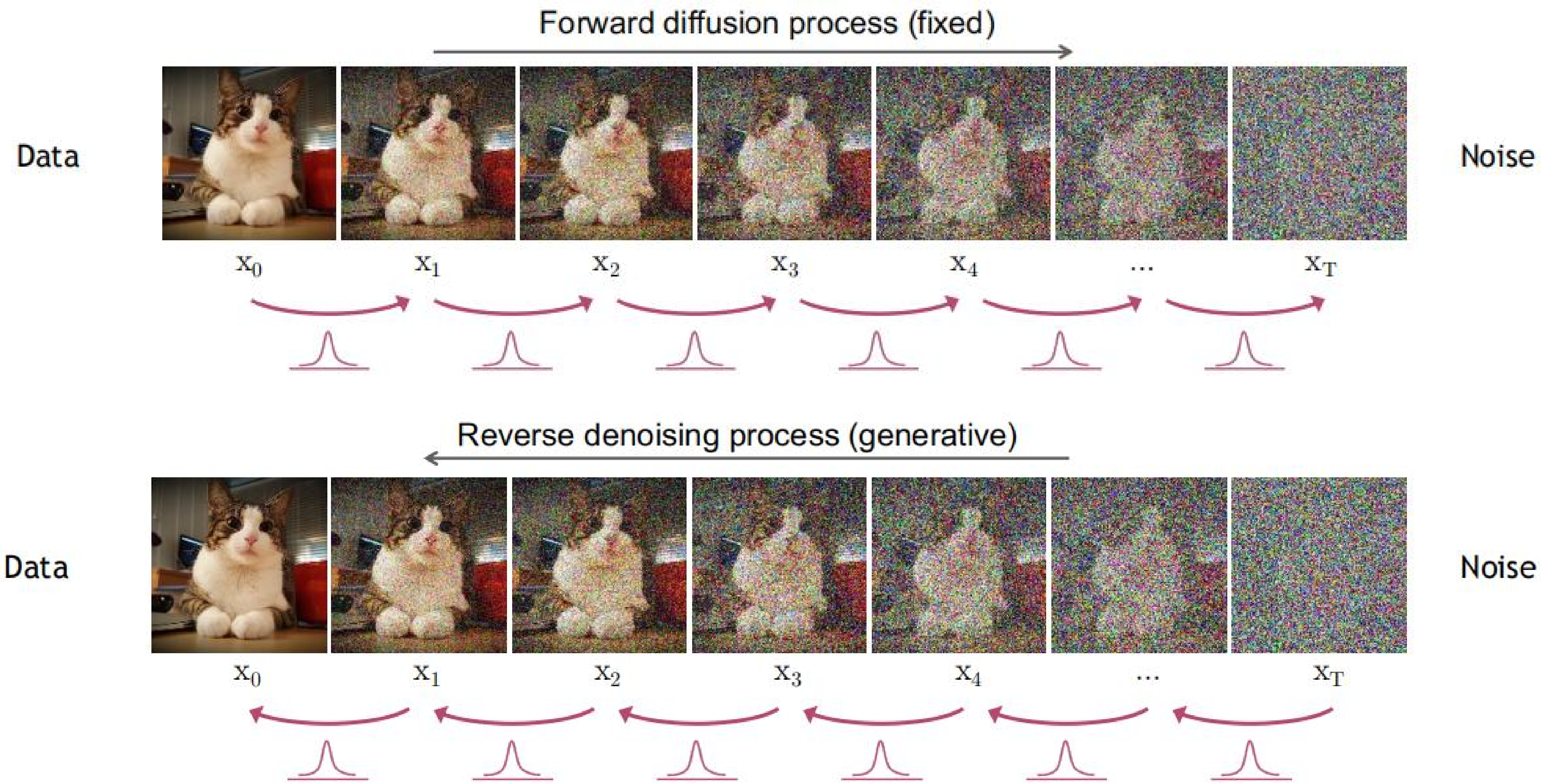
$$q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) = \mathcal{N}(\mathbf{x}_{t-1}; \tilde{\boldsymbol{\mu}}(\mathbf{x}_t, \mathbf{x}_0), \tilde{\beta}_t\mathbf{I}) \tag{2}$$

where

$$\begin{aligned} \tilde{\beta}_t &= 1/(\frac{\alpha_t}{\beta_t} + \frac{1}{1-\bar{\alpha}_{t-1}}) = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t} \cdot \beta_t \\ \tilde{\boldsymbol{\mu}}_t(\mathbf{x}_t, \mathbf{x}_0) &= (\frac{\sqrt{\alpha_t}}{\beta_t}\mathbf{x}_t + \frac{\sqrt{\bar{\alpha}_t}}{1-\bar{\alpha}_t}\mathbf{x}_0)/(\frac{\alpha_t}{\beta_t} + \frac{1}{1-\bar{\alpha}_{t-1}}) = \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\mathbf{x}_t + \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\mathbf{x}_0 \end{aligned} \tag{3}$$

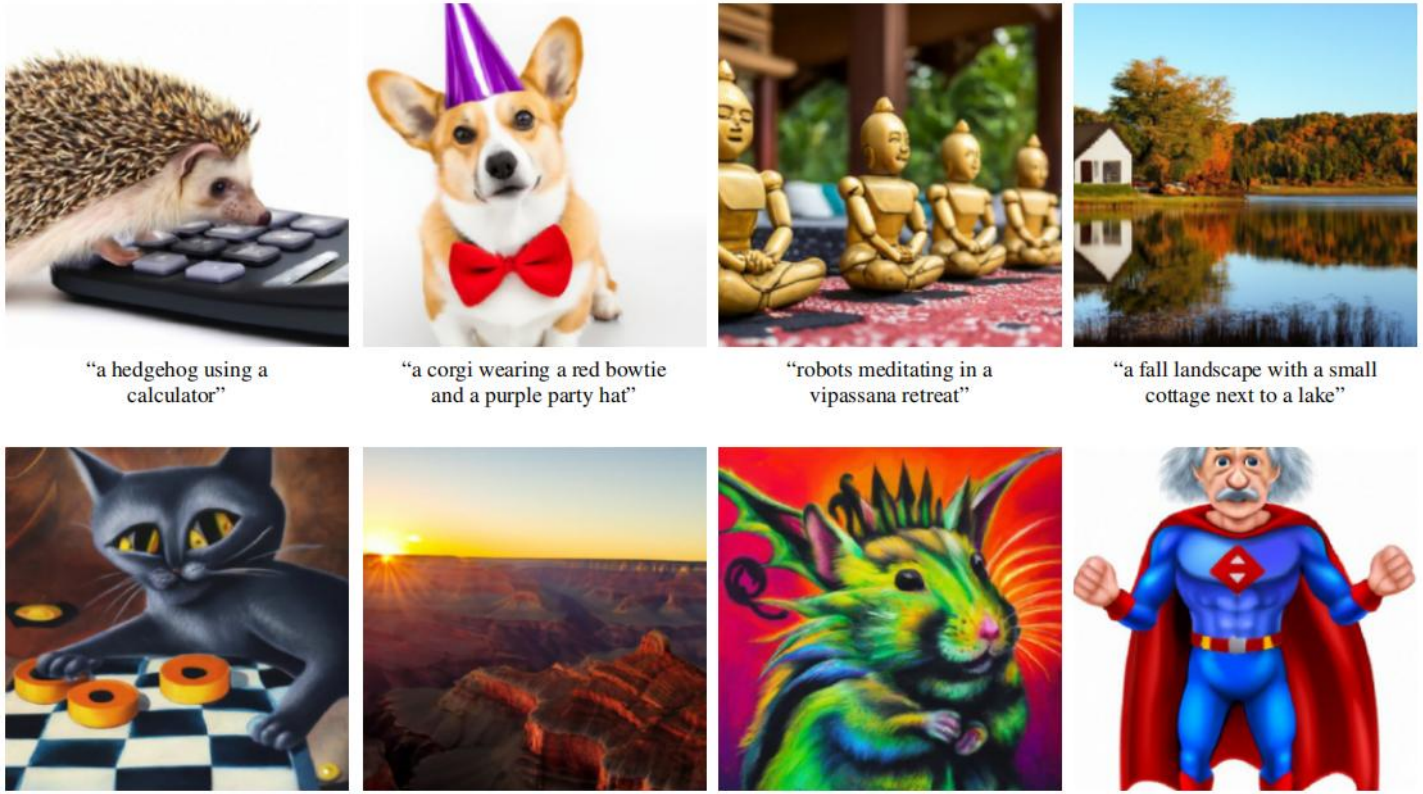
Background

Diffusion models consist of two processes: *forward* and *reverse*.



Background

Text-to-image diffusion models have achieved great results.



"a hedgehog using a calculator"

"a corgi wearing a red bowtie and a purple party hat"

"robots meditating in a vipassana retreat"

"a fall landscape with a small cottage next to a lake"

"a surrealist dream-like oil painting by salvador dalí of a cat playing checkers"

"a professional photo of a sunset behind the grand canyon"

"a high-quality oil painting of a psychedelic hamster dragon"

"an illustration of albert einstein wearing a superhero costume"

GLIDE (arXiv 2112)

Background

Text-to-image diffusion models have achieved great results.

an espresso machine that makes coffee from human souls, artstation

panda mad scientist mixing sparkling chemicals, artstation

a corgi's head depicted as an explosion of a nebula

DALLE-2 (arXiv 2204)

Background

Text-to-image diffusion models have achieved great results.

A chromeplated cat sculpture placed on a Persian rug.

Android Mascot made from bamboo.

Intricate origami of a fox and a unicorn in a snowy forest.

Imagen (NIPS 22’)

Background

However, text-to-image diffusion models still have several problems:

- Texts can be less detailed to control the process of generation.

(*e.g.* to generate a several persons with specific poses.)

- They often fail when generating subjects with strong priors.

(*e.g.* to generate human hands.)

How to provide additional control during generation?

Content

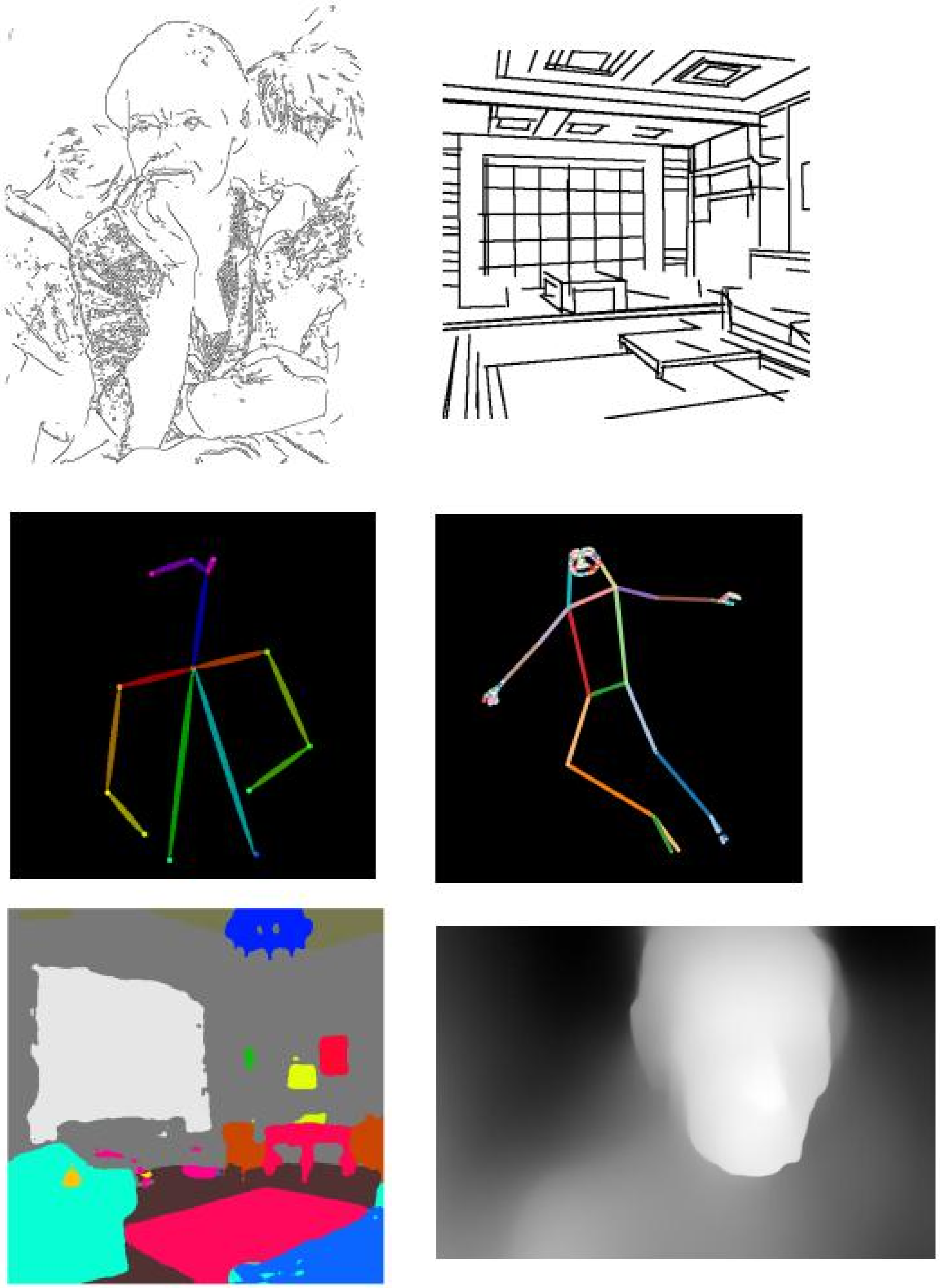
- Authors
- Background
- Method
- Experiments

Method

Additional controls contain:

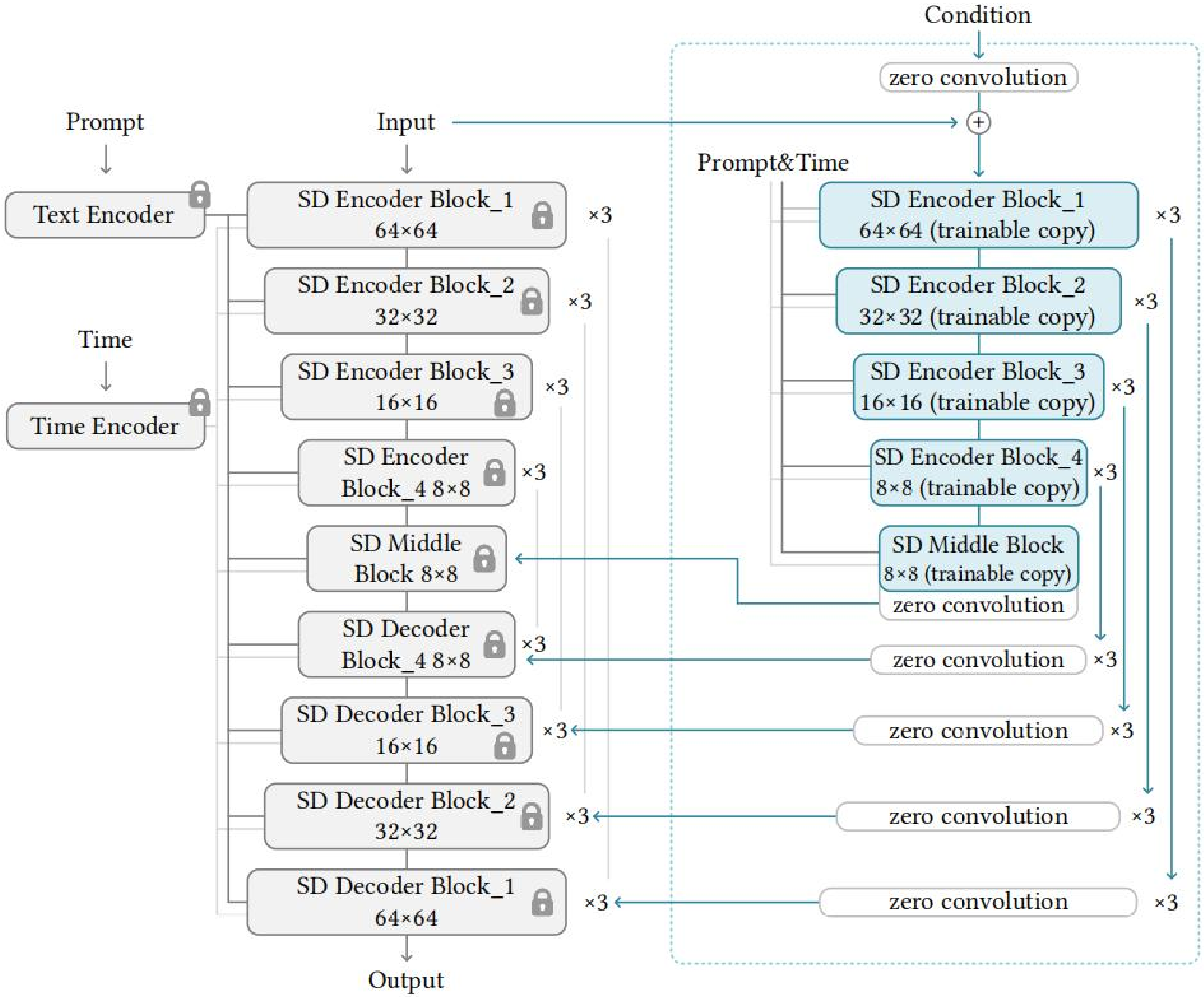
- Canny edge / Hough line.
- Skeleton.
- Segmentation maps.
- Depth maps.

Different from text embeddings, they all have same size to the generated images.



Method

The proposed framework.



Method

To leverage the pre-trained text-to-image models:

Use zero-convolution layers to ensure the outputs of the conditioning model are zero at first.

To leverage the character of size:

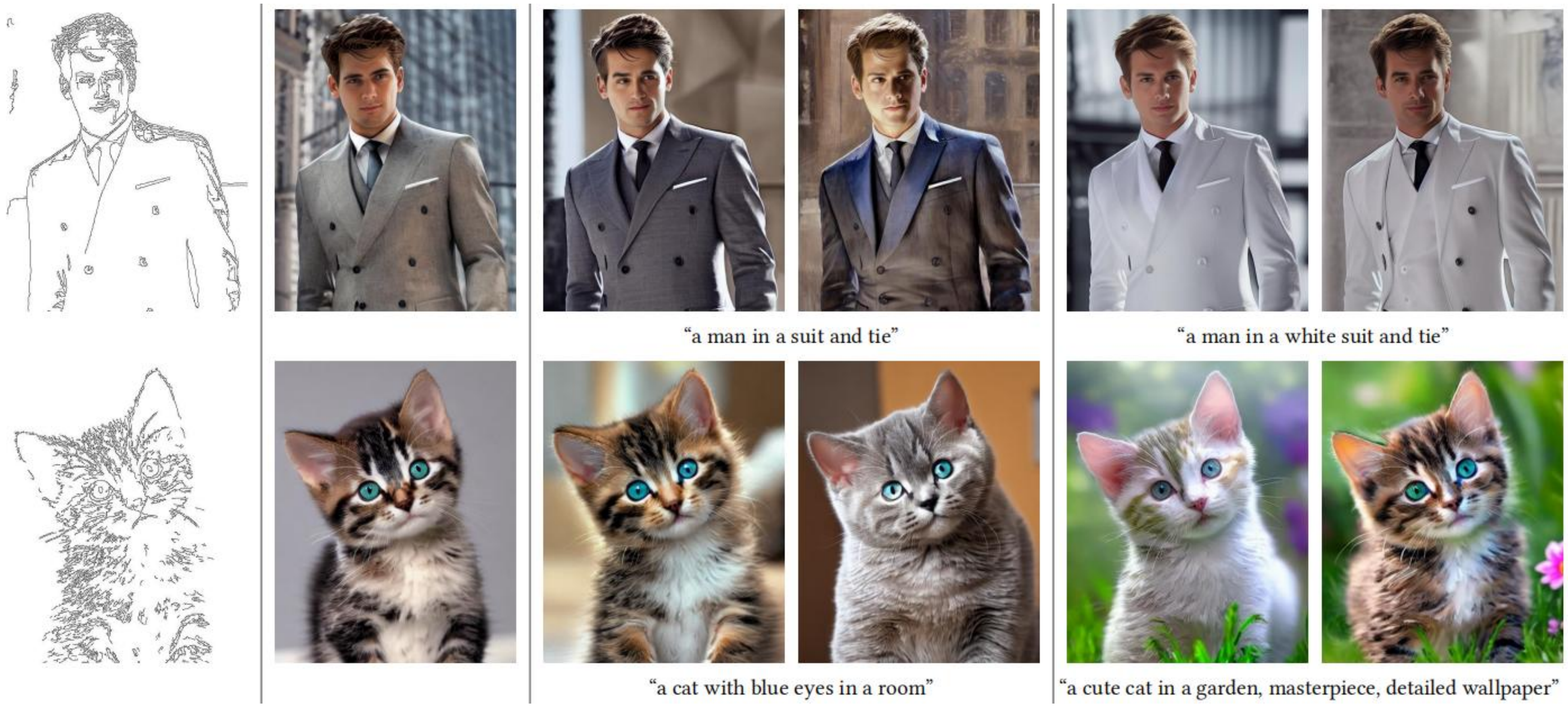
Use copied main block to extract features with the same size.

Content

- Authors
- Background
- Method
- Experiments

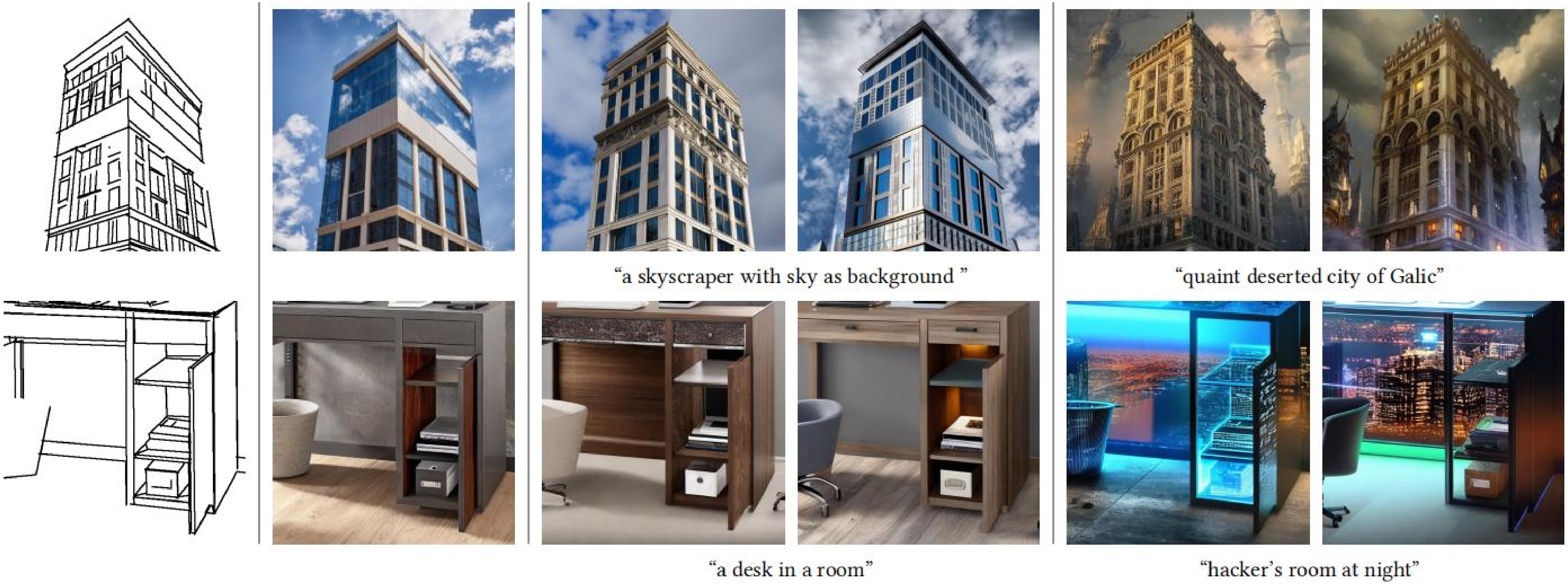
Experiments

Canny edge. 600 GPU hours. (A100 80G)



Experiments

Hough line. 150 GPU hours. (A100 80G)



Experiments

HED boundary. 300 GPU hours. (A100 80G)

Experiments

User sketching. 600+150 GPU hours. (A100 80G)

"a turtle in river"

"a masterpiece of cartoon-style turtle illustration"

"a cow with horns standing in a field"

"a robot ox on moon, UE5 rendering, ray tracing"

Experiments

Human pose (Openpifpaf). 400 GPU hours. (RTX 3090Ti)

Experiments

Human pose (Openpose). 300 GPU hours. (A100 80G)

Experiments

Semantic segmentation (COCO). 400 GPU hours. (RTX 3090Ti)

“fantastic artwork, fairy tail”

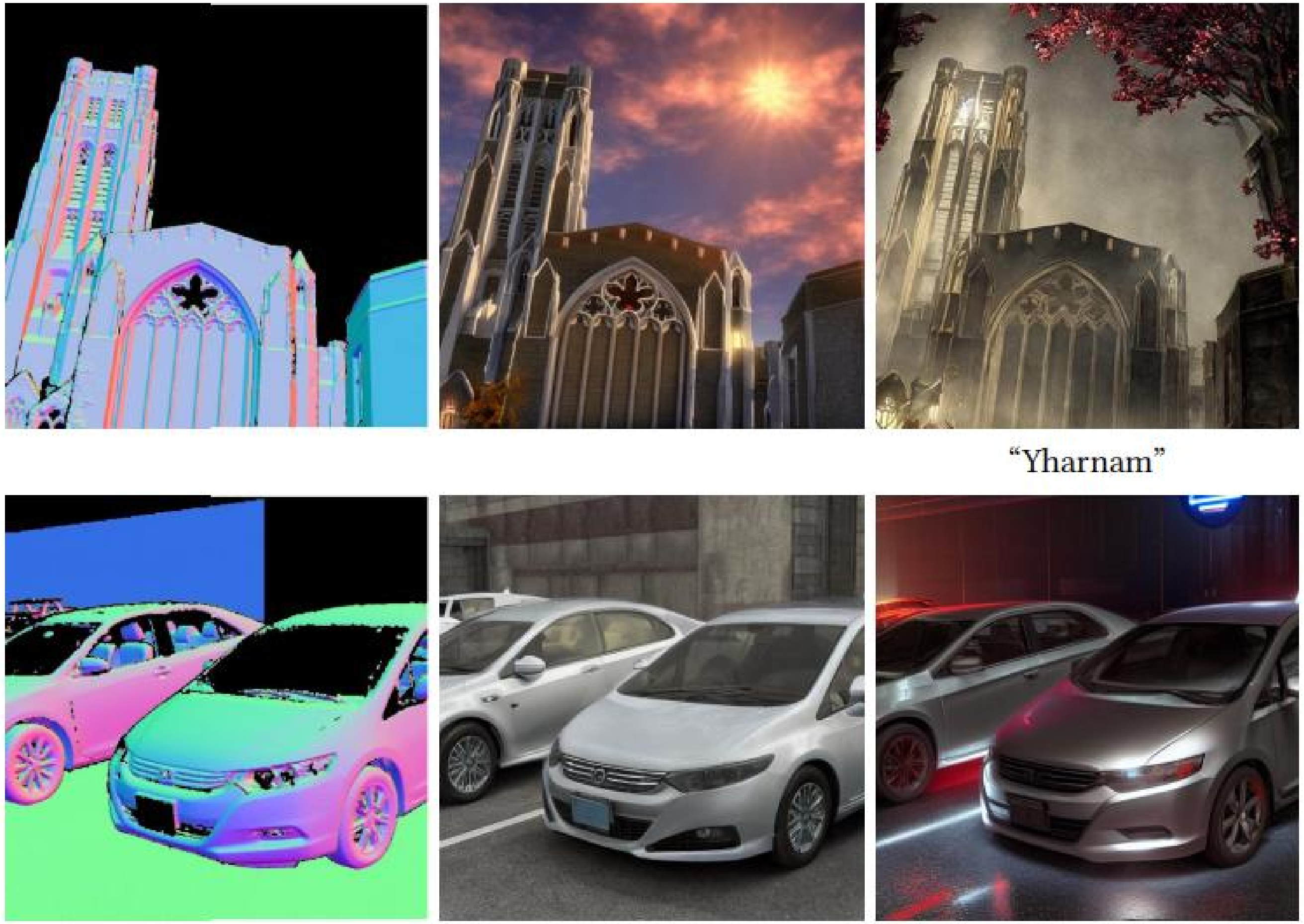
“cyberpunk, city at night”

Experiments

Semantic segmentation (ADE20k). 200 GPU hours. (A100 80G)

Experiments

Normal map. 100 GPU hours. (A100 80G)



"Yharnam"

"cars parked in a city night"

Experiments

Depth map. 500 GPU hours. (A100 80G)

Experiments

Cartoon line drawings. 300 GPU hours. (A100 80G)

Cartoon line drawing “1girl, masterpiece, best quality, ultra-detailed, illustration”

Experiments

Ablation study: comparison between ControlNet and Concatenate.

without ControlNet
(using Stability's "official" method to add the channels to input layer, same as their depth-to-image structure)

SD + ControlNet

Thanks for watching.

马逸扬

2023/03/26