Interpolating between Images with Diffusion Models

AI 키워드	
날짜	@2024년 12월 31일
콘텐츠	논문
태그	유런

- 24 Jul 2023
- latent diffusion 모델을 활용하여 외형적으로 차이가 큰 두 이미지에 대해 보간을 수행하는 방법

https://arxiv.org/pdf/2307.12560

Introduction

다른 스타일의 이미지에 대해 높은 수준의 보간을 생성하고자 함

현실에서 많이 시도된 것이 아님 → 평가 방법 부족, 예술계에서의 창의적인 활용 기대할 수 잇음

사전학습된 latent diffusion model(LDM) 활용

파이프라인: 텍스트를 통한 조건 제어/노이즈 조절/생성된 이미지 중에서 선택 가능 등의 기능이 있으므로 현실에 도입 쉬움(하이퍼파라미터 튜닝 거의 없이)

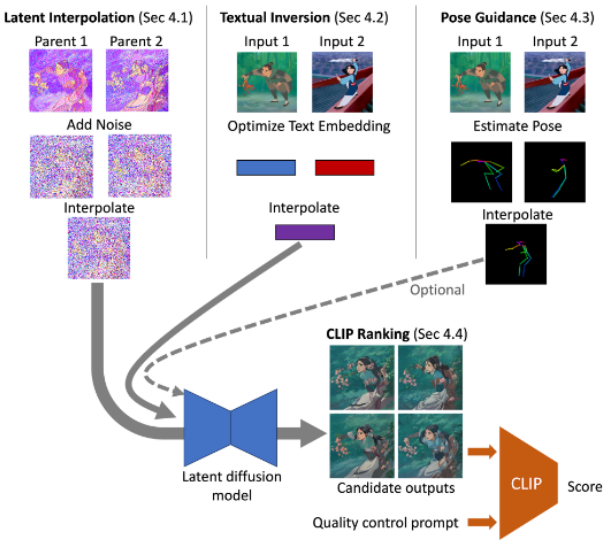


Figure 2: **Our pipeline.** To generate a new frame, we interpolate the noisy latent images of two existing frames (Section 4.1). Text prompts and (if applicable) poses are extracted from the original input images, and interpolated to provide to the denoiser as conditioning inputs (Section 4.2 and 4.3). This process can be repeated for different noise vectors to generate multiple candidates. The best candidate is selected by computing its CLIP similarity to a prompt describing desired characteristics (Section 4.4).

Preliminaries

$$\text{encoder } \mathcal{E} : x \mapsto z_0, \text{ decoder } \mathcal{D} : z_0 \mapsto \hat{x},$$

$$\text{denoising U-Net } \epsilon_\theta : (z_t; t, c_{\text{text}}, c_{\text{pose}}) \mapsto \hat{\epsilon}.$$

timestemp t

latent vector z_0

$$z_t = \alpha_t z_0 + \sigma_t \epsilon, \epsilon \sim \mathcal{N}(0, I)$$

- ϵ : denoising U-Net에 의해 추정됨
- α_t, σ_t: 파라미터

c_{text}: 타겟 이미지를 설명하는 텍스트

c_{pose}: 타겟 이미지 내 인간이 취할 포즈

Real Image Interpolation

Latent interpolation

기본 전략: 이미지 쌍을 반복적으로 보간하는 것

1. z_0에 공유하는 노이즈 추가

   이미지가 서로 가까울수록 더 작은 노이즈 활용 → 더 부드러운 보간

   타임스템프에 따라 parent 이미지를 새로 할당하고 노이즈를 각각 추가 → parent 이미지와는 더 가깝게, sibling 이미지와는 더 멀게 만듦

2. 보간
3. 디노이징 → 중간 이미지 생성

Interpolation type

- latent space와 text embedding에 대해 spherical linear interpolations (slerp)
- pose에 대해 선형 보간

Noise schedule

DDIM sampling

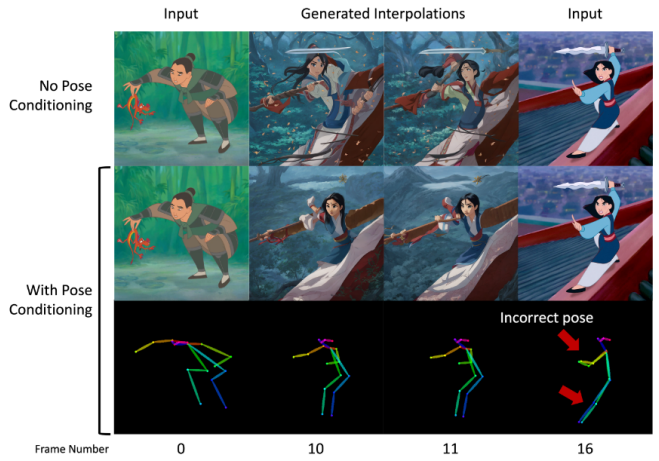
Textual inversion

전체 컨텐츠/스타일이 설명된 텍스트 프롬프트를 textual inversion(텍스트 반전 =positive/negative 프롬프트 사용)을 사용해 각 이미지에 특정하게 적용

LDM이 랜덤 노이즈 단계에서 디노이징 할 때 에러 최소화하도록 프롬프트 임베딩을 파인튜닝

$$\mathcal{L}(c_{\text{text}}) = \|\hat{\epsilon}_\theta(\alpha_t z_0 + \sigma_t \epsilon; t, c_{\text{text}}) - \epsilon\|$$

Pose guidance



포즈의 주체가 많이 다르다면 얼굴이나 관절이 여러개가 되는 식의 오류가 발생

1. OpenPose를 활용해 입력 이미지에서 포즈 추출
2. 모든 공유되는 keypoint 선형 보간 → 각 이미지의 중간이 되는 포즈 추출
3. 이 중간 포즈가 LDM에게 전달됨

2번에서 예측된 포즈가 잘못되어도 크게 벗어나지 않기 때문에 생성된 이미지 품질 유지 가능

CLIP ranking

각기 다른 랜덤시드로 여러 개의 후보 이미지 생성 후 CLIP으로 랭킹 매김

positive와 negative prompt, 생성된 이미지에 대해 CLIP similarity가 positive와 높고 negative와 낮은 이미지를 선택

유저가 직접 선택하거나 다른 프롬프트를 쓰는 방식으로 바꿀 수도 있음

Experiments

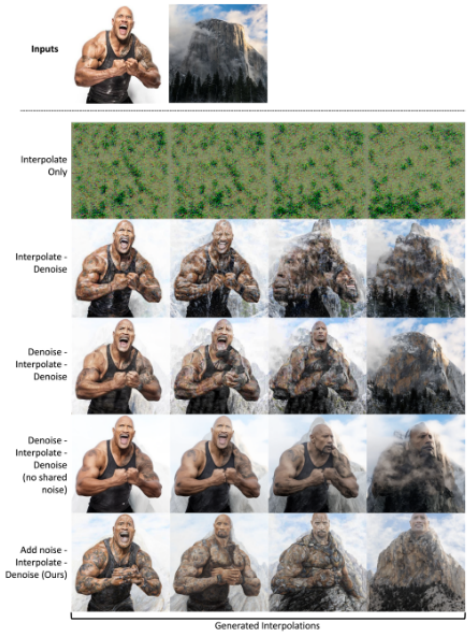
몇가지 베이스라인과 비교

1. interpolating without denoising(interpolate only)
2. interpolating between noisy versions of the input vectors (interpolate-denoise)

$$z_t^0 = \alpha_t z_{t-1}^0 + \beta_t \epsilon_t,$$
$$z_t^N = \alpha_t z_{t-1}^N + \beta_t \epsilon_t,$$

3. interpolating partially denoised versions of generated latents (denoise-interpolate-denoise)
4. denoise-interpolate-denoise with no shared noise added to the input latents.

정성적 비교



정량적 비교

Interpolation Scheme	FID	PPL
Interpolate only	436	56±8
Interpolate-denoise	179	172±32
Denoise-interpolate-denoise (DID)	169	144±26
DID w/o shared noise	199	133±22
Add noise-interpolate-denoise (ours)	214	193±27

Conclusion

- 다른 비디오/이미지 생성 모델과 결합될 수 있음
- 스타일과 레이아웃에 “큰” 차이가 있는 이미지 쌍 보간 어려움

ıre cases. Our approach is still limited in its ability to bridge large gaps in style, semantic: