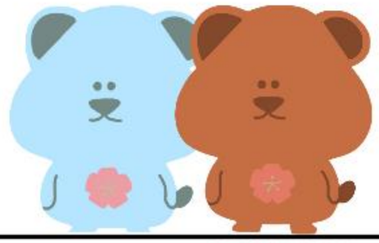
논문 스터디 1주차

Week17 구미진, 안서연, 최예은

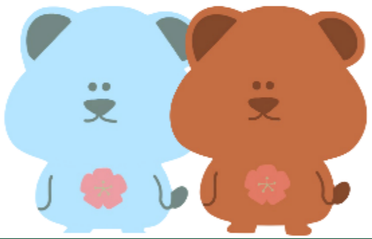
EWHA EURON

Index



EWHA EURON

Image-to-Image Translation with Conditional Adversarial Networks

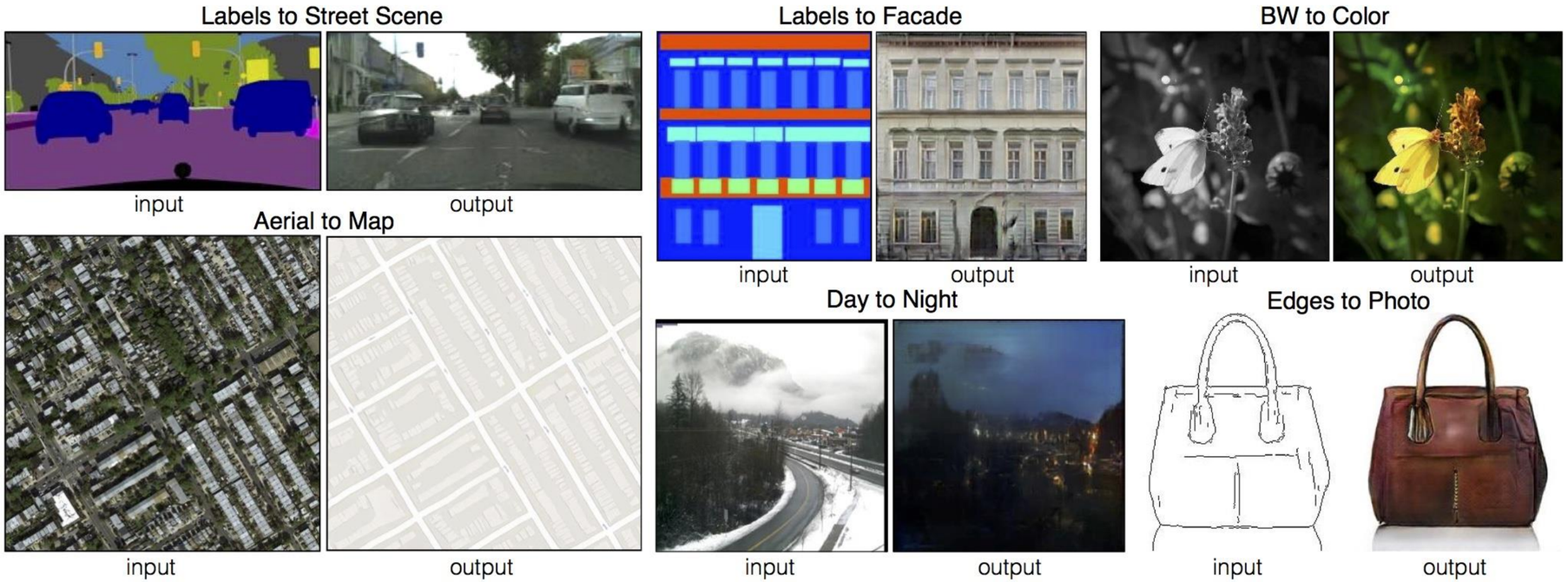


EWHA EURON

01 Image translation

Image-to-image translation with Conditional adversarial networks

- 결국 pixel로부터 pixel을 prediction하는 문제(pix2pix)
- CNN이 아닌 GAN
- Image-to-image translation에 적합한 conditional GAN(CGAN)



02 Related work

Structured losses for image modeling

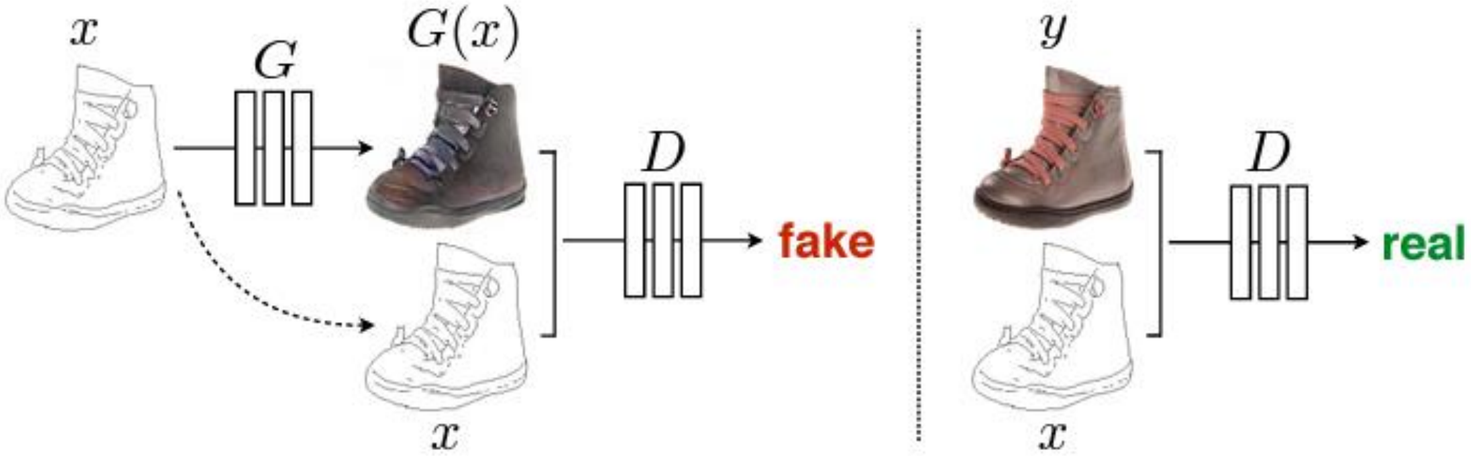
- 기존의 image-to-image translation은 픽셀 단위의 classification 혹은 regression으로 문제 접근
- cGAN은 structured loss를 사용

Conditional GANs

- 이전 연구들과 달리 generator로 “U-Net” 기반의 구조를 사용
- Discriminator로 convolutional “PatchGAN” 사용

02 Related work

Conditional GANs (CGAN)



- G는 condition x를 입력으로 받아 fake 이미지를 생성한다
- D 또한 condition x를 받아 판별한다

03 Method

일반적인 GAN

- 랜덤 노이즈 벡터 z로부터 이미지 y를 출력

$$G : z \rightarrow y$$

cGAN

- 관찰한 이미지 x와 랜덤 노이즈 벡터 z로부터 이미지 y를 출력

$$G : \{x, z\} \rightarrow y$$

- Discriminator D는 Generator G가 생성한 이미지를 fake로 구별할 수 있도록 학습함
- Generator G는 Discriminator D가 실제 이미지인지 아닌지를 구분하지 못하도록 real에 가까운 fake 이미지를 만들어내도록 학습함

03 Method

- cGAN의 손실 함수

$$\mathcal{L}_{cGAN}(G,D) = \mathbb{E}_{x,y}[\log D(x,y)] + \mathbb{E}_{x,z}[\log(1 - D(x, G(x,z))],$$

- G는 손실 함수를 최소화하는 방향으로, D는 최대화하는 방향으로 학습시킨다

$$\mathcal{L}_{GAN}(G,D) = \mathbb{E}_{y}[\log D(y)] + \mathbb{E}_{x,z}[\log(1 - D(G(x,z))].$$

$$\mathcal{L}_{L1}(G) = \mathbb{E}_{x,y,z}[\|y - G(x,z)\|_1].$$

- 최적의 생성자 G*

$$G^* = \arg\min_G \max_D \mathcal{L}_{cGAN}(G,D) + \lambda \mathcal{L}_{L1}(G).$$

03 Method

- cGAN의 손실 함수

$$\mathcal{L}_{cGAN}(G,D) = \mathbb{E}_{x,y}[\log D(x,y)] + \mathbb{E}_{x,z}[\log(1 - D(x, G(x,z))],$$

- G는 손실 함수를 최소화하는 방향으로, D는 최대화하는 방향으로 학습시킨다

$$\mathcal{L}_{GAN}(G,D) = \mathbb{E}_{y}[\log D(y)] + \mathbb{E}_{x,z}[\log(1 - D(G(x,z))].$$

$$\mathcal{L}_{L1}(G) = \mathbb{E}_{x,y,z}[\|y - G(x,z)\|_1].$$

- 최적의 생성자 G*

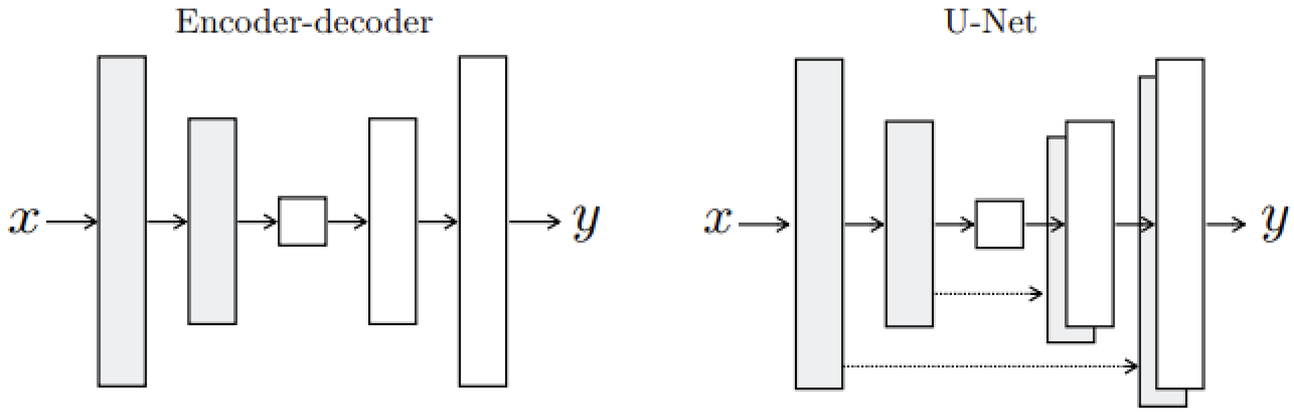
$$G^* = \arg\min_G \max_D \mathcal{L}_{cGAN}(G,D) + \lambda \mathcal{L}_{L1}(G).$$

"Instead, for our final models, we provide noise only in the form of dropout, applied on several layers of our generator at both training and test time"

Network architectures

Generator

- U-Net
- Encoder-decoder 구조에 skip connection이 추가된 U-Net 사용
- Encoder와 decoder가 대칭적으로 연결



03 Method

Discriminator

- PatchGAN
- L1 loss를 사용할 경우 blurry하지만 low-frequency 성분들을 잘 검출해냄
- L1 loss를 사용하면서도 discriminator가 high-frequency structure를 모델링할 수 있도록 하기 위해, local image patch를 사용
- 이를 위해 patchGAN이라는 discriminator 구조 설계
- 전체 이미지를 보는 것이 아닌 NxN patch 단위로 prediction
- N이 작더라도 high quality result를 만들어내고, 적은 파라미터 수와 빨리 실행된다는 장점이 있음

04 Experiments

Evaluation metrics

- 전통적인 방법으로 mean-squared error를 측정하는 방법이 있으나, structure를 측정하지 못한다는 단점이 있음
- 대신 두 가지 방법을 사용하여 평가에 이용했는데,

1. map generation, image colorization, aerial photo generation 문제와 같은 "real vs fake"의 조사
2. 생성된 cityscape가 충분히 실제적인지는 제공되는 인식 시스템을 이용하여 객체를 인식할 수 있는지 측정

04 Experiments

Loss	Per-pixel acc.	Per-class acc.	Class IOU
L1	0.42	0.15	0.11
GAN	0.22	0.05	0.01
cGAN	0.57	0.22	0.16
L1+GAN	0.64	0.20	0.15
L1+cGAN	**0.66**	**0.23**	**0.17**
Ground truth	0.80	0.26	0.21

Table 1: FCN-scores for different losses, evaluated on Cityscapes labels↔photos.

04 Experiments

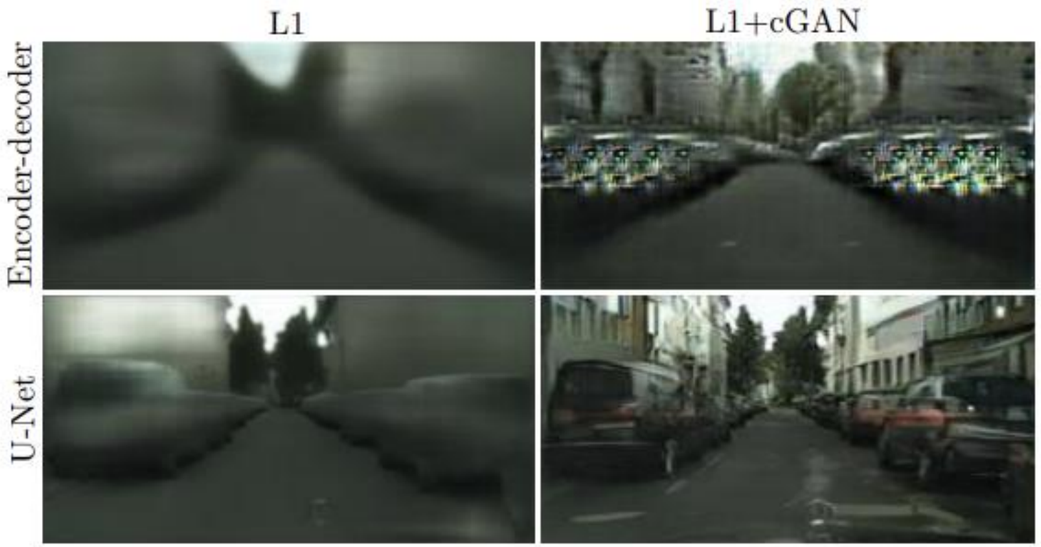


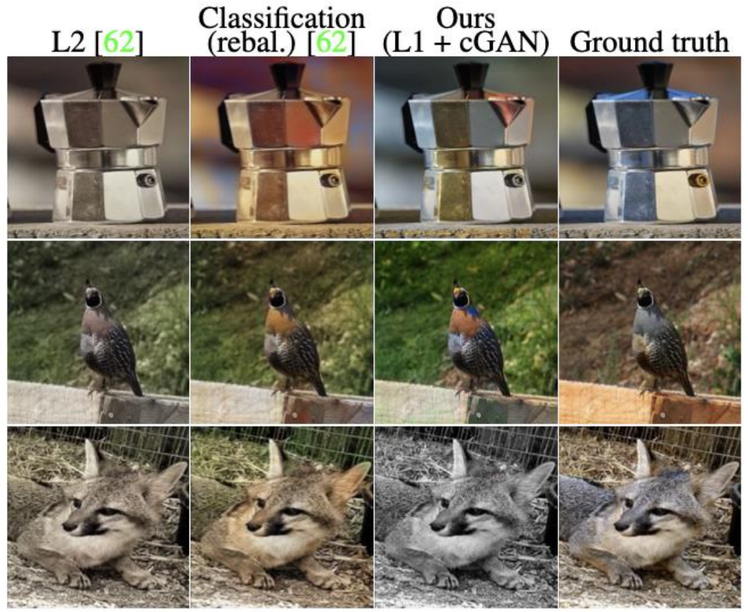
Figure 5: Adding skip connections to an encoder-decoder to create a "U-Net" results in much higher quality results.

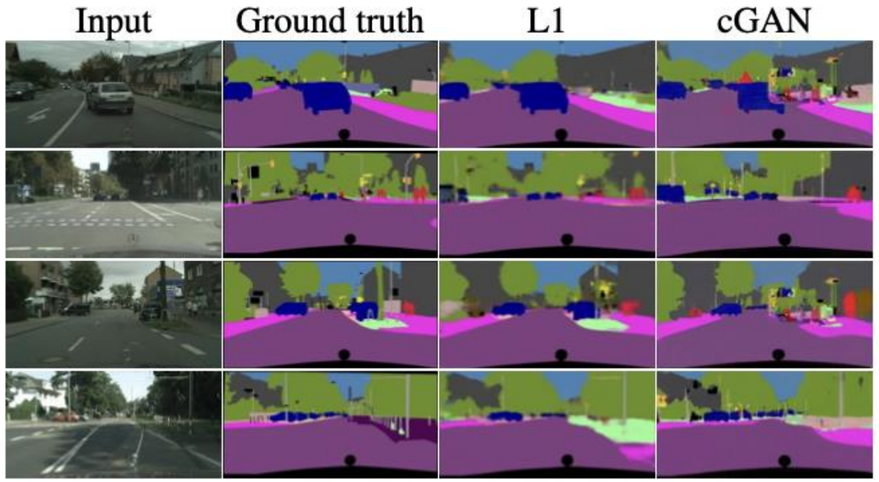
Loss	Per-pixel acc.	Per-class acc.	Class IOU
Encoder-decoder (L1)	0.35	0.12	0.08
Encoder-decoder (L1+cGAN)	0.29	0.09	0.05
U-net (L1)	0.48	0.18	0.13
U-net (L1+cGAN)	**0.55**	**0.20**	**0.14**

04 Experiments

Discriminator receptive field	Per-pixel acc.	Per-class acc.	Class IOU
1×1	0.39	0.15	0.10
16×16	0.65	0.21	**0.17**
70×70	**0.66**	**0.23**	**0.17**
286×286	0.42	0.16	0.11

04 Experiments





04 Experiments

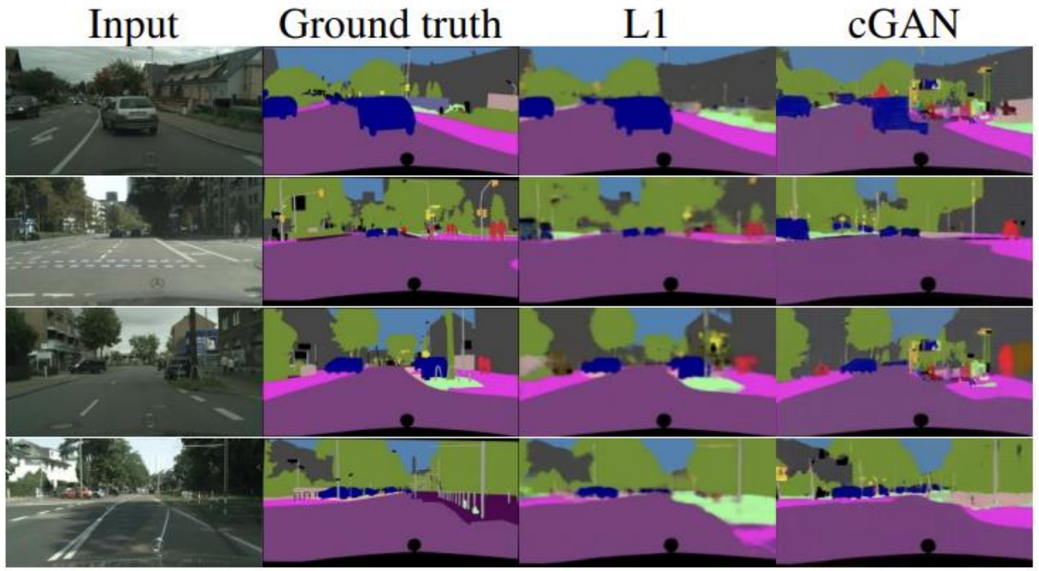
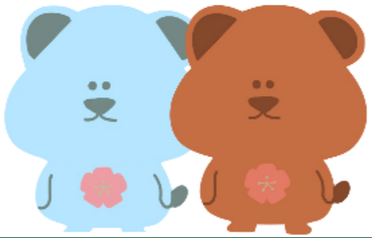


Figure 10: Applying a conditional GAN to semantic segmentation. The cGAN produces sharp images that look at glance like the ground truth, but in fact include many small, hallucinated objects.

Loss	Per-pixel acc.	Per-class acc.	Class IOU
L1	**0.86**	**0.42**	**0.35**
cGAN	0.74	0.28	0.22
L1+cGAN	0.83	0.36	0.29

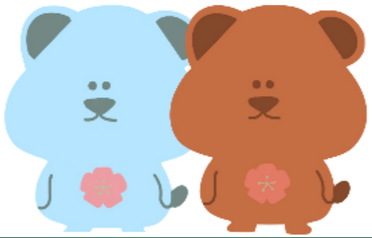
Table 6: Performance of photo→labels on cityscapes.

Bringing old photos back to life



EWHA EURON

Introduction



EWHA EURON

Introduction

Bringing Old photos Back to Life

1) Background

Image Restoration(이미지 복원)

- Quality restoration- 화질 복원
- Resolution restoration -해상도 복원
- Color restoration(colorization)-색상 복원

- 기존의 restoration 작업은 synthetic image(인위적으로 만들어낸 이미지)를 활용 Supervised Learning으로 수행

- 실제의 old image는 synthetic image와는 차이가 존재함
  ->Supervised Learning은 실제 old image에 대한 일반화된 모델로 적합x(generalization issue)

- 실제 old image는 복합적이고 다양한 degradation이 존재(mixed degradation issue)

Introduction

Bringing Old photos Back to Life

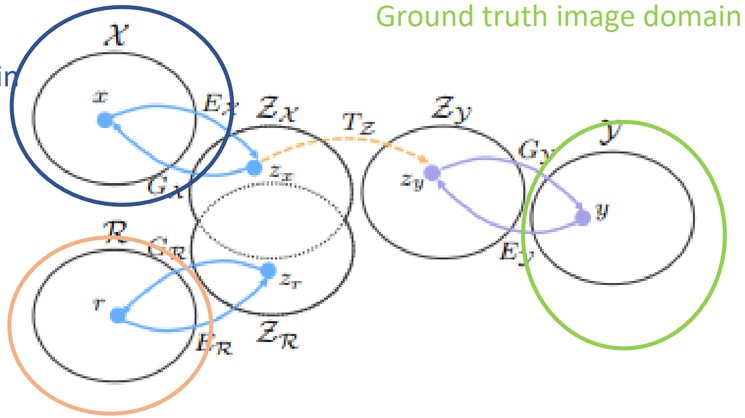
Synthetic image image domain

↓

Align

↑

Real old image domain

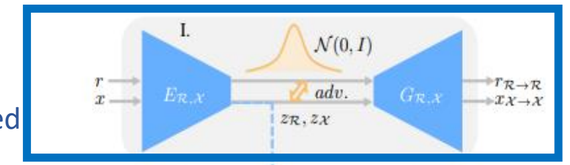


<Triplet Domain>

2) 논문에서 제시하는 해결방안

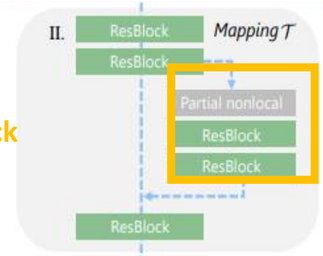
1) Generalizaion Issue 에 대한 방안
- **Triplet domain** translation network with 2 VAE 제시
  : real image domain 과 synthetic image domain gap을 줄이면서 하나로 병합해서 Ground truth image domain(깨끗한 이미지)으로 연결하는 network 만들기

VAE 1: real image domain + Synthetic image domain aligned

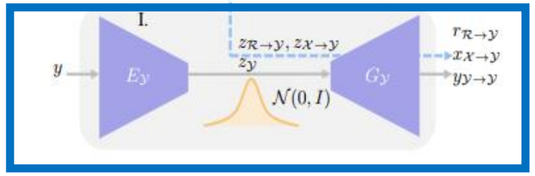


2) Mixed degradation Issue에 대한 방안
- Global branch with partial nonlocal block 제시

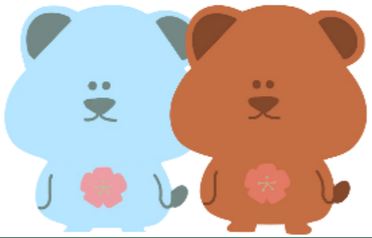
Global branch with partial non-local Block



VAE 2: Ground Truth image domain



Related Work



Related Work

Single degradation image restoration : learning-based method

- Unstructured degradation
  ex. Noise,blurriness,color fading,저해상도
  :딥러닝 기반 denoising,super-resolution,deblurring 연구

- Structured degradation (더 어려움)
  ex. Holes,scratches,spots
  : image inpainting (사진의 일부가 손상되었을때 복원해서 채워놓는 기술)
  https://wandb.ai/authors/enriching-words-with-subwords/reports/-Image-Inpainting---VmlIdzo0NzU5Njg

Mixed degradation image restoration

- 연구가 많이 진행되지 않음
- Syntheis data 기반 Supervised learning 연구가 대부분이라서 성능 안좋음
- unstructrued degradation만 해결하려함
- Deep learning 기반 연구도 있으나 해당 논문의 방법이 성능 및 효율성이 더 좋다고 함

Related Work

Old photo restoration

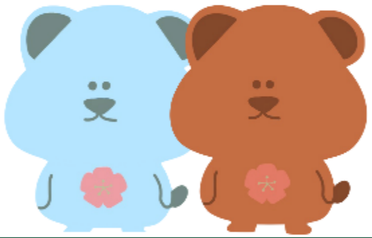
- 전형적인 mixed degradation problem

- 기존의 연구는 inpainting 에만 집중함, unstructured degradation을 복원하지 않아
  복원 후에 사진이 오래되보이는 경향이 있음

Bringing Old Photos Back to Life

“우리는 실제 오래된 사진에 적용할 수 있는 generalized restoration model을
제시하고,이는 unstructured degradation & structured degradation 모두 해결할수 있을것이다! "

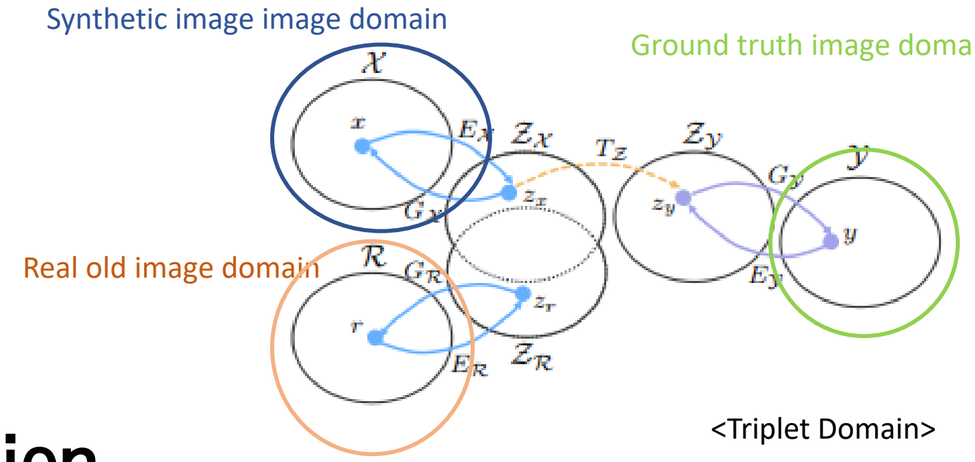
Method



EWHA EURON

Method

1. Generalizaion Issue 에 대한 해결방안



<Triplet Domain>

1.1 Restoration via latent space translation

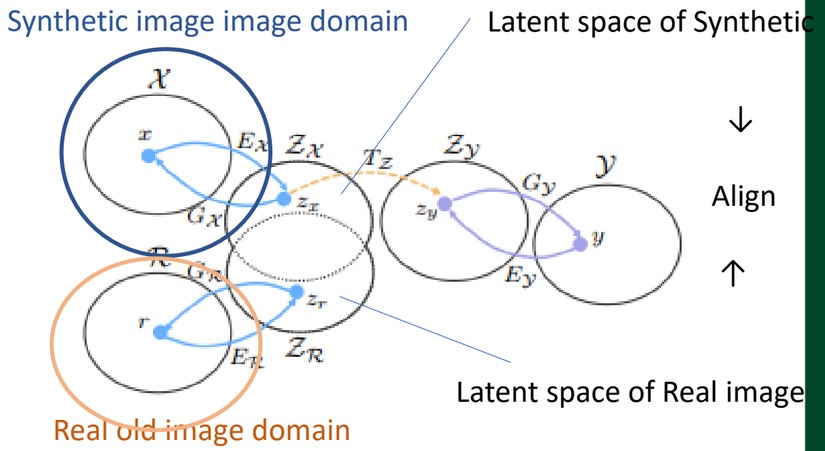
- 3 domain(real old image,synthetic,ground truth) 정의하고, latent space에 mapping
- Real old image와 synthetic은 모두 corrupted 되어 공통된 특징이 있을수 있음
  -> 공통된 부분을 중심으로 두 latent space를 align함 (ZR ≈ ZX)
- 공식: rR→Y = GY ○ TZ ○ ER(r)

ER : R → ZR, **EX** : X 7→ ZX , **EY** : Y 7→ ZY (latent space로 바꾸기)
ZR: real old image의 latent space, **ZX**: latent space of synthetic image **ZY**: latent space of ground truth
TZ : ZX → ZY (mapping) , **GY** : ZY→ Y

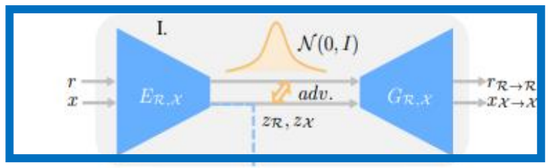
Method

1.2 Domain alignment in the VAE latent space

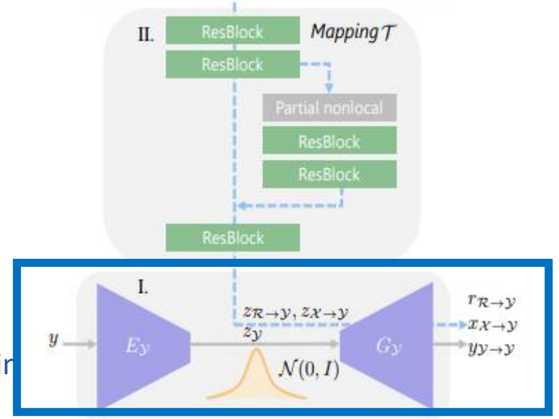
- VAE 1을 활용하여 Real image와 Synthetic image domain이 동일한 latent space로 encode함
- Adversarial discriminator를 학습하면서 두 도메인간 gap 줄어듬
- 결과적으로 하나의 compact latent space가 생김
- 이것을 Ground truth latent space와 mapping시켜 restoration을 할 수 있는 network 형성



VAE 1: real image domain +
Synthetic image domain aligned



$$\mathcal{L}_{\text{VAE}_1}(r) = \underbrace{\text{KL}(E_{\mathcal{R},\mathcal{X}}(z_r|r)||\mathcal{N}(0,I))}_{\text{Gaussian 분포를 따르지 않는 latent code 제거}} + \underbrace{\alpha\mathbb{E}_{z_r\sim E_{\mathcal{R},\mathcal{X}}(z_r|r)}\left[\|G_{\mathcal{R},\mathcal{X}}(r_{\mathcal{R}\to\mathcal{R}}|z_r) - r\|_1\right]}_{\text{Latent code가 주요 정보를 추출}} + \underbrace{\mathcal{L}_{\text{VAE}_1,\text{GAN}}(r)}_{\text{VAE의 over-smooth 문제 해결}}$$



VAE 2: Ground Truth image domain

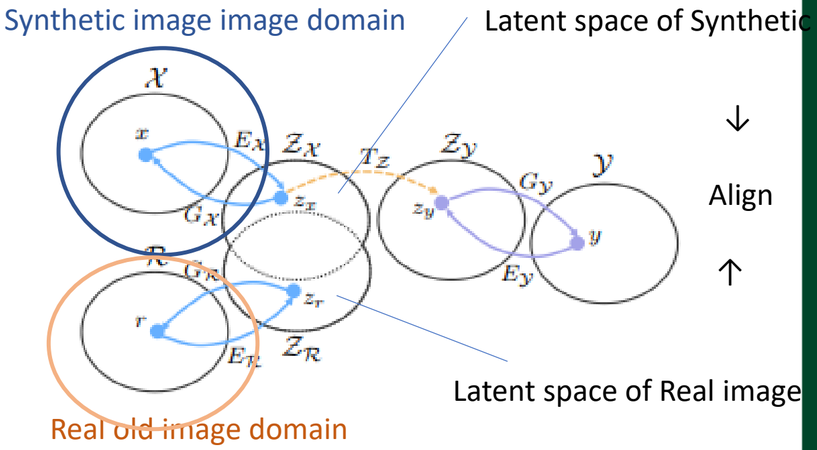
Method

1.2 Domain alignment in the VAE latent space

- 추가적으로 latent space에서 두 domain간 gap을 줄이기 위해 adversarial network 사용
  * Discriminator DR,X : ZR과 ZX의 차이를 확인
    loss function:

$$\mathcal{L}^{\text{latent}}_{\text{VAE}_1,\text{GAN}}(r,x) = \mathbb{E}_{x\sim\mathcal{X}}[D_{\mathcal{R},\mathcal{X}}(E_{\mathcal{R},\mathcal{X}}(x))^2] + \mathbb{E}_{r\sim\mathcal{R}}[(1-D_{\mathcal{R},\mathcal{X}}(E_{\mathcal{R},\mathcal{X}}(r)))^2].$$

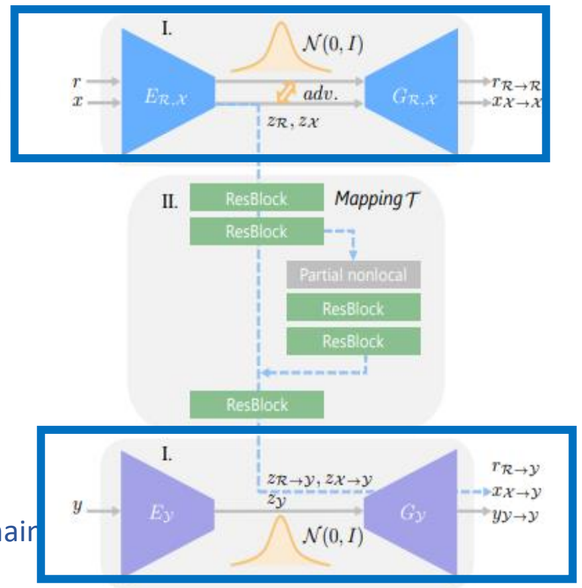
  * Encoder ER,X: discrinator을 fool하면서 R,X의 R(real image),X(synthetic)이 동일한 latent space로 mapping되도록 함



➡ 종합한 공식

$$\min_{E_{\mathcal{R},\mathcal{X}},G_{\mathcal{R},\mathcal{X}}} \max_{D_{\mathcal{R},\mathcal{X}}} \mathcal{L}_{\text{VAE}_1}(r) + \mathcal{L}_{\text{VAE}_1}(x) + \mathcal{L}^{\text{latent}}_{\text{VAE}_1,\text{GAN}}(r,x).$$

VAE 1: real image domain + Synthetic image domain aligned



VAE 2: Ground Truth image domain

Method

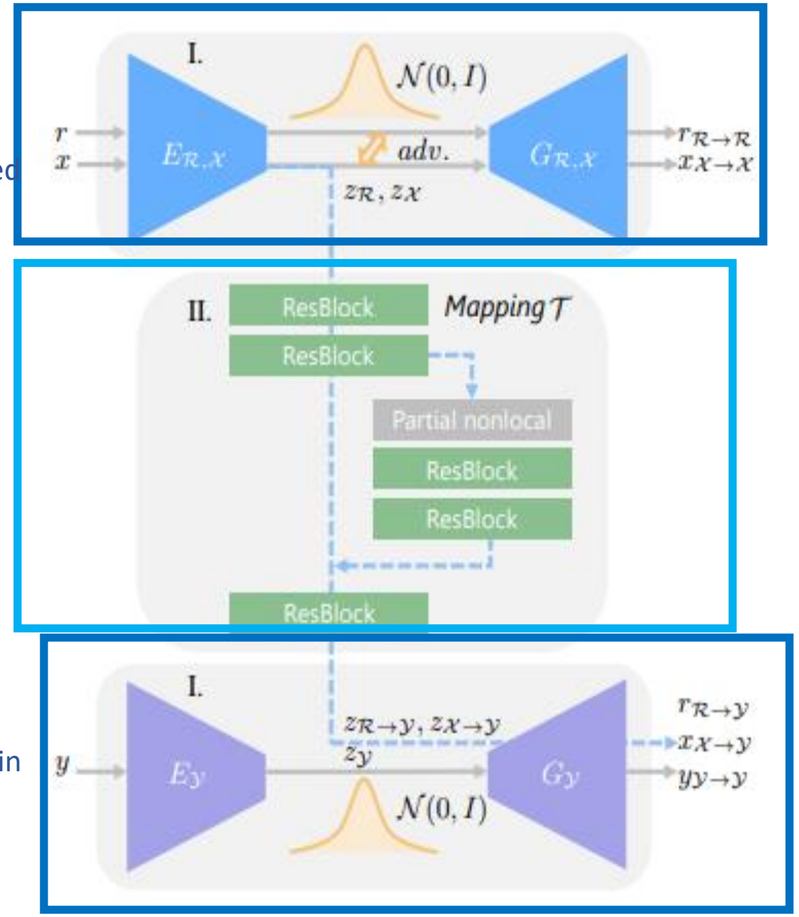
1.3 Restoration through latent mapping

- Fixed 2 VAE, mapping network T
- R and X 가 동일한 same latent space로 aligned 되었기때문에
  ZX to ZY mapping을 수행하면 R을
  복원하는 효과를 가짐

VAE 1: real image domain +
Synthetic image domain aligned

Mapping

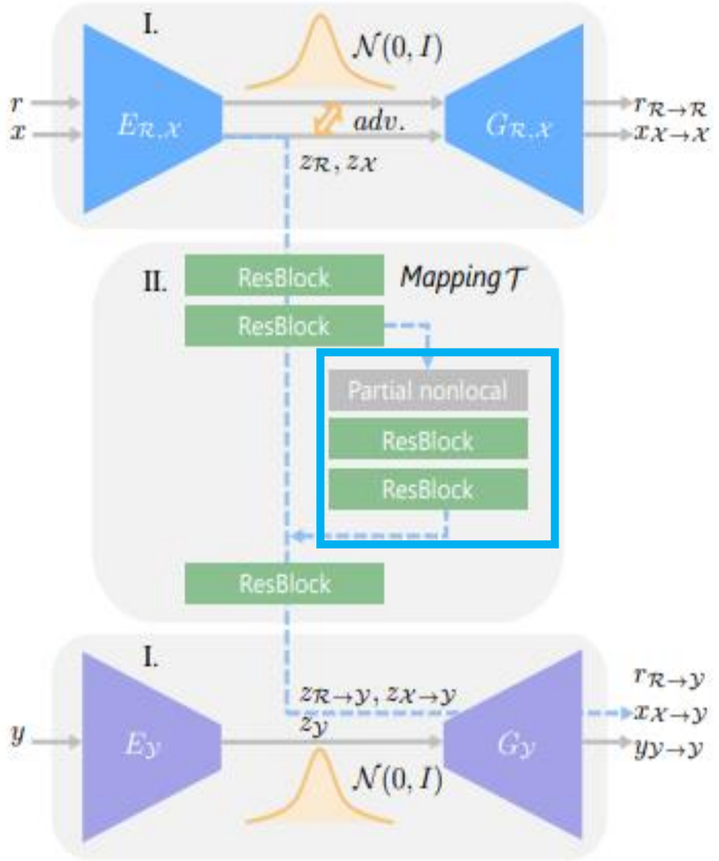
VAE 2: Ground Truth image domain



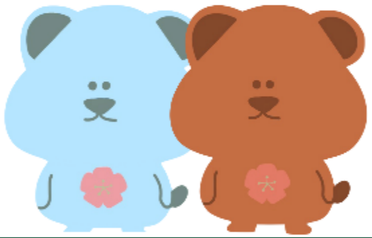
Method

2. Multiple degradation Issue에 대한 해결방안

- Residual Block으로만 mapping하면
- Partial nonlocal block이 있는 global branch를 추가해서 global context 속 여러 degradation을 파악할 수 있도록 함



Experiment



EWHA EURON

Implementaion

1) Training Dataset

: 5,718 old photos + Synthetic damaged Image(from Pascal VOC dataset)

2) Scratch detection

: U-Net (architecture for semantic segmentation)

3) Training details

: Adam solver with $\beta 1$ =0.5 and $\beta 2$ = 0.999, learning rate 0.0002 for the first 100 epochs with linear decay to zero

Comparisons

Quantitative comparison – DIV2K dataset

Method	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓
Input	12.92	0.49	0.59	306.80
Attention [42]	**24.12**	**0.70**	0.33	208.11
DIP [43]	22.59	0.57	0.54	194.55
Pix2pix [55]	22.18	0.62	**0.23**	**135.14**
Sequential [56, 57]	22.71	0.60	0.49	191.98
Ours w/o PN	23.14	0.68	0.26	143.62
Ours w/ PN	**23.33**	**0.69**	**0.25**	**134.35**

Table 1: **Quantitative results on the DIV2K dataset.** Upward arrows indicate that a higher score denotes a good image quality. We highlight the best two scores for each measure. In the table, PN stands for partial nonlocal block.

2nd place **PSNR/SSIM.**
2nd place **LPIPS** (Pix2pix 1st)
But FID better than pix2pix:
slight quantitative advantage.

PSNR(peak signal-to-noise-ration) , SSIM(Structural similarity index) : 복원된 ouput과 ground truth 간의 차이 계산에 쓰임. PSNR과 SSIM 높을수록 품질이 좋음
LPIPS(Learned perceptual image patch similarity): Perceptual similarity 계산에 쓰임. FID와 LPIPS를 사용하여 생성된 이미지의 품질과 다양성을 평가.값이 낮을수록 생성된 이미지가 ground truth와 유사
FID(Fechet Inception Distance) : 실제 이미지와 생성된 이미지 간의 확률 분포 차이을 계산. 값이 낮을수록 실제이미지와 확률분포가 유사

Comparisons

Qualitative comparison

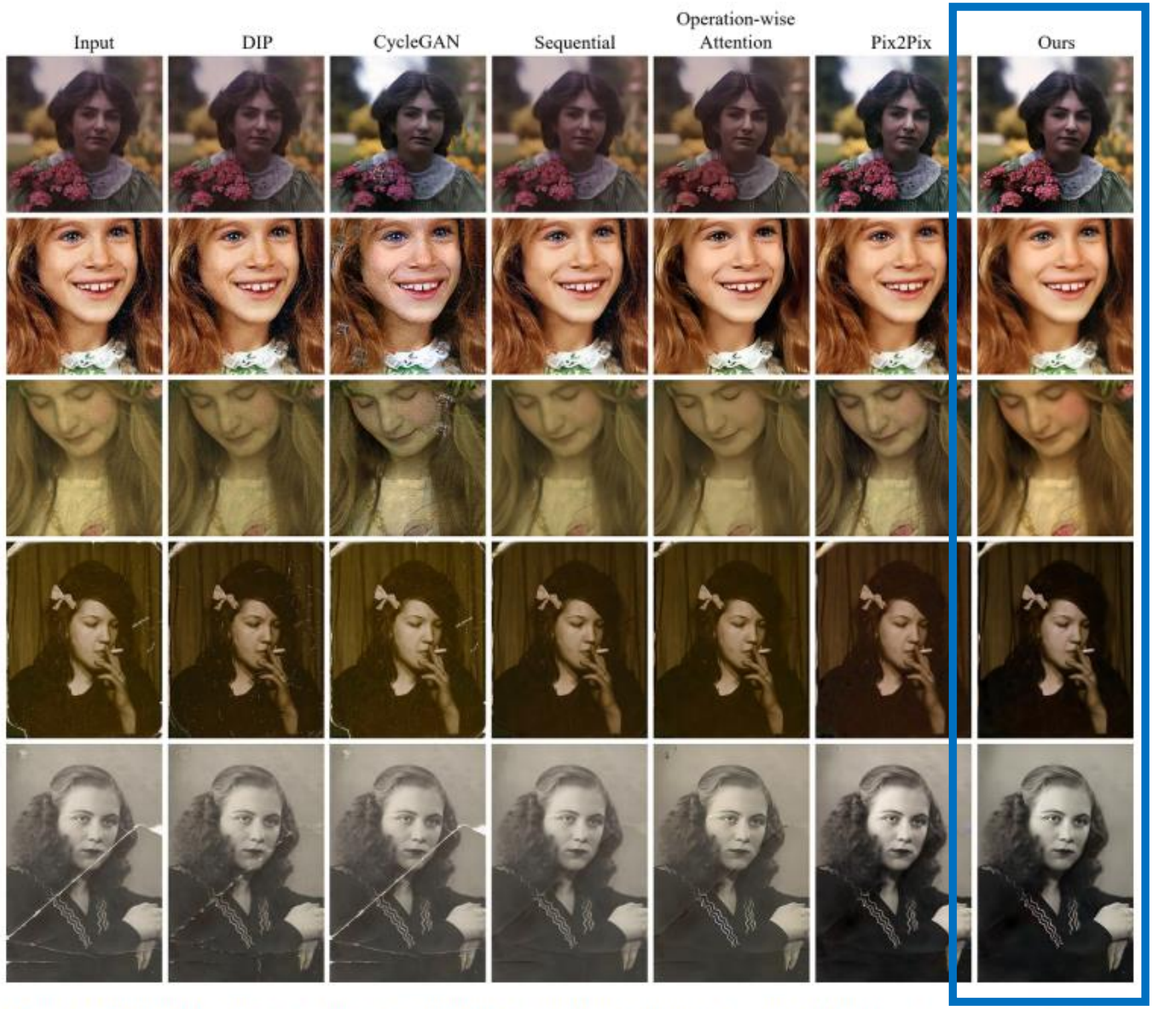


Figure 5: **Qualitative comparison against state-of-the-art methods.** It shows that our method can restore both unstructured and structured degradation and our recovered results are significantly better than other methods.

User Studv

Method	Top 1	Top 2	Top 3	Top 4	Top 5
DIP [43]	2.75	6.99	12.92	32.63	69.70
CycleGAN [44]	3.39	8.26	15.68	24.79	52.12
Sequential [56, 57]	3.60	20.97	51.48	83.47	93.64
Attention [42]	11.22	28.18	56.99	75.85	89.19
Pix2Pix [55]	14.19	54.24	72.25	86.86	96.61
Ours	**64.83**	**81.35**	**90.68**	**96.40**	**98.72**

Table 2: **User studv results.** The percentage (%) of user selection is shown.

CycleGan:모든 스크래치 제거x
operation-wise attention method and the sequential operations: Sepia issue, color fading 해결x
Pix2pix: film noise, structured defects 잔재

Our method
- gives clean, sharp images with the scratches plausibly filled
- enhance the photo color appropriately

EWHA EURON

Ablation Study

Ablation study: machine learning system에서 일부 building blocks을 제거해서 전체 성능에 미치는 효과를 연구

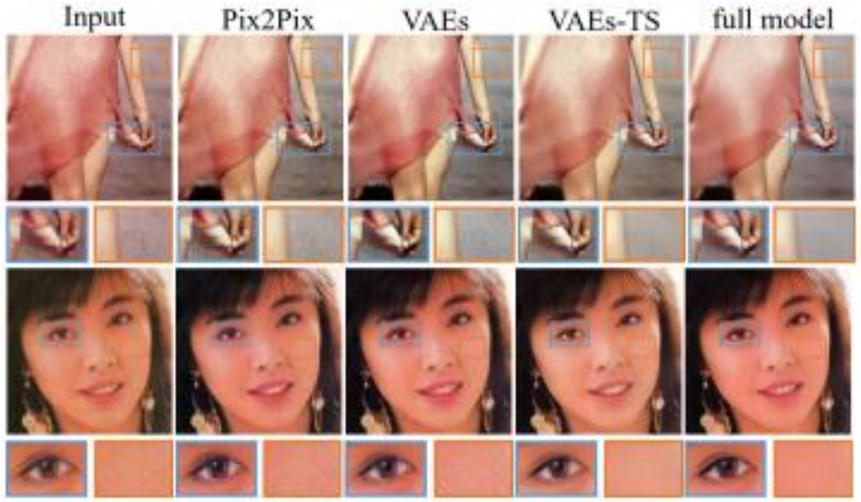


Figure 6: **Ablation study for two-stage VAE translation.**

Method	Pix2Pix	VAEs	VAEs-TS	full model
Wasserstein ↓	1.837	1.048	0.765	**0.581**
BRISQUE ↓	25.549	23.949	23.396	**23.016**

Table 3: **Ablation study of latent translation with VAEs.**

Figure 8: **Ablation study of partial nonlocal block.** Partial nonlocal does not touch the non-hole regions.

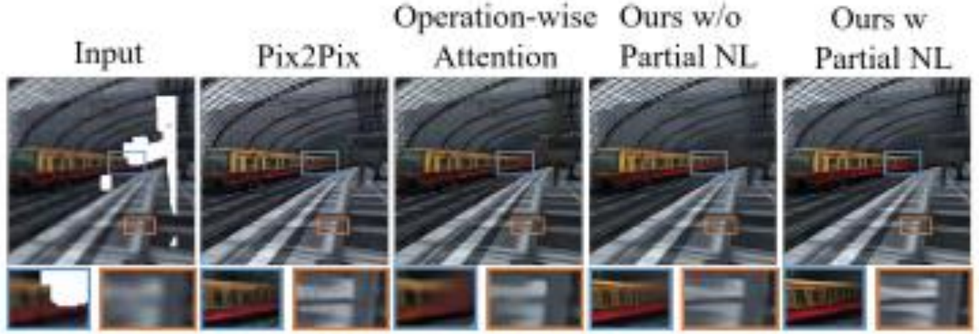


Figure 7: **Ablation study of partial nonlocal block.** Partial nonlocal better inpaints the structured defects.

Discussion and Conclusion

EWHA EURON

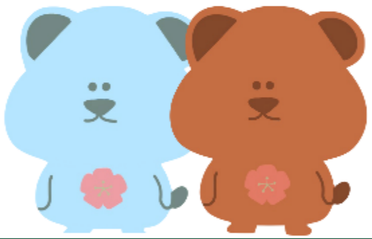
Discussion and Conclusion

Figure 9: **Limitation.** Our method cannot handle complex shading artifacts.

1) 한계: shading 관련 데이터셋 부족으로 compex shading을 해결하지 못함

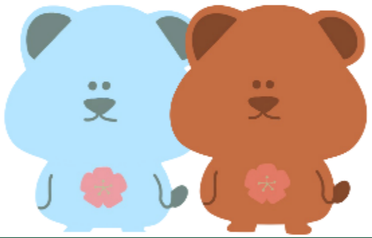
2) 의의:
- Triplet domain translation network를 사용하여 mixed degradation 해결
- 기존 방법들보다 Generalization issue 해결
- Scratch가 전체적으로 일관성 있게 복원됨

Denoising Diffusion Probabilistic Models



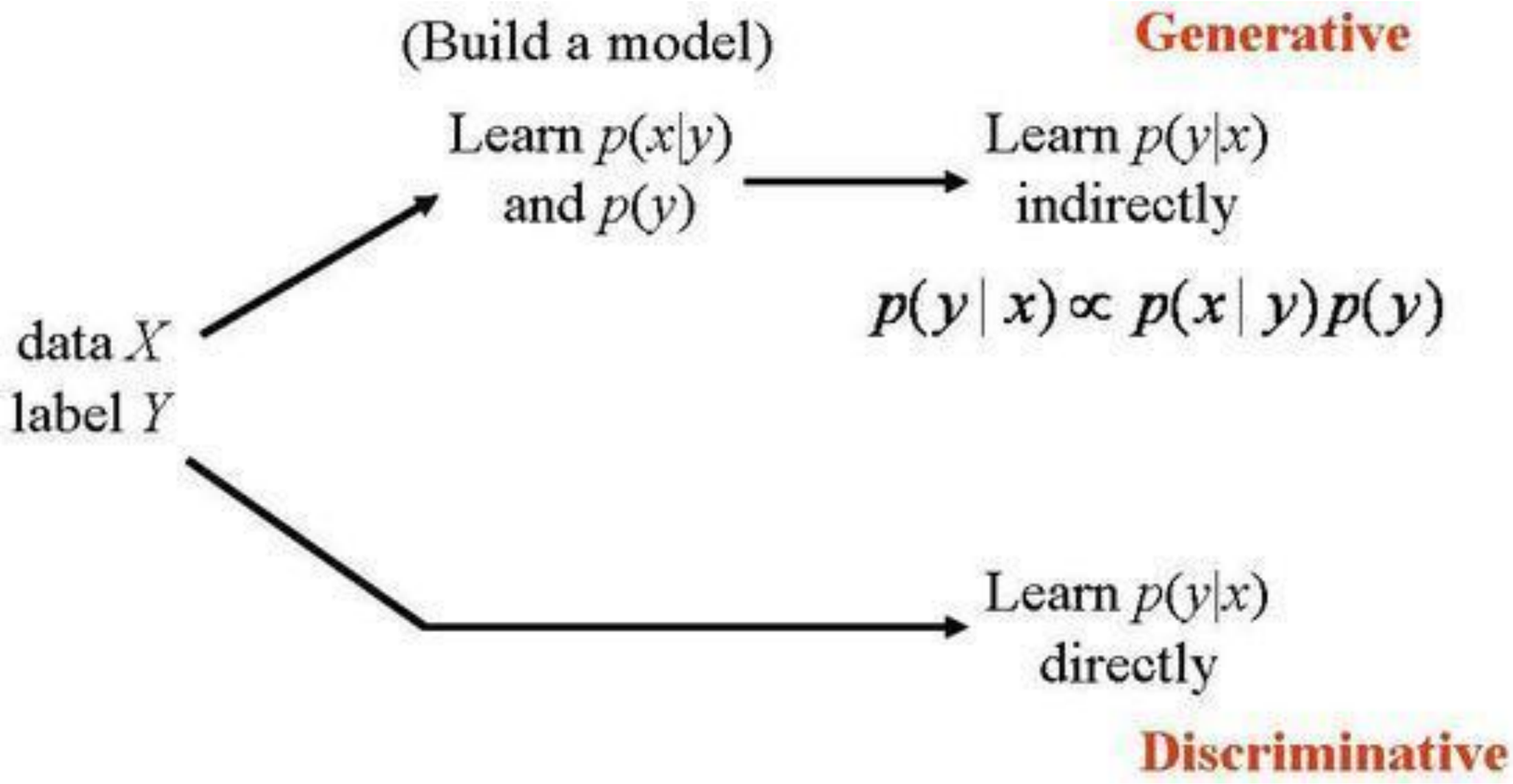
EWHA EURON

Generative model



EWHA EURON

Generative model



Discriminate model – class의 차이에 주목하여 바로바로 어떤 class에 들어가야 할지 결정해 주는 모델

Generative model – 각 class의 분포에 주목하여 어떤 분포에 들어갈 가능성이 가장 많은지 결정해 주는 모델

Generative model

Generative Model

: 학습한 data의 distribution을 따르는 새로운 data를 만들어내는 모델

Auto-regressive models(AR)

순서를 가지는 변수(variable)들의 조건부 확률(conditional probability)의 곱으로 데이터의 likelihood를 계산하는 모델.
자기 자신을 입력으로 하여 자기 자신을 예측

Variational Auto encoders(VAEs)

잠재변수(Latent variable)기반의 generative model로 데이터 x와 latent variable z의 결합확률분포(joint distribution)를 구해서 x에 대해서 주변화(marginalize)하는 모델

Energy Based Models(EBMs)

에너지함수(Energy function)를 이용해서 distribution을 estimate

Generative Adversarial Networks(GANs)

Discriminator와 Generator를 서로 adversarial 방향으로 학습시켜서 데이터를 생성하는 모델

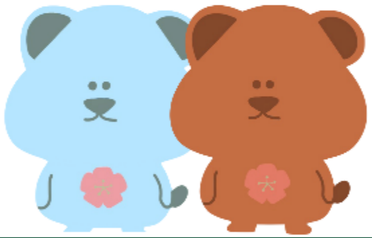
Normalizing Flows

Simple한 base 분포 p(z)에서 복잡한 데이터 분포 p(x)로 가는 역사상(invertible mapping)함수를 이용해 distribution을 estimate하는 모델

Diffusion

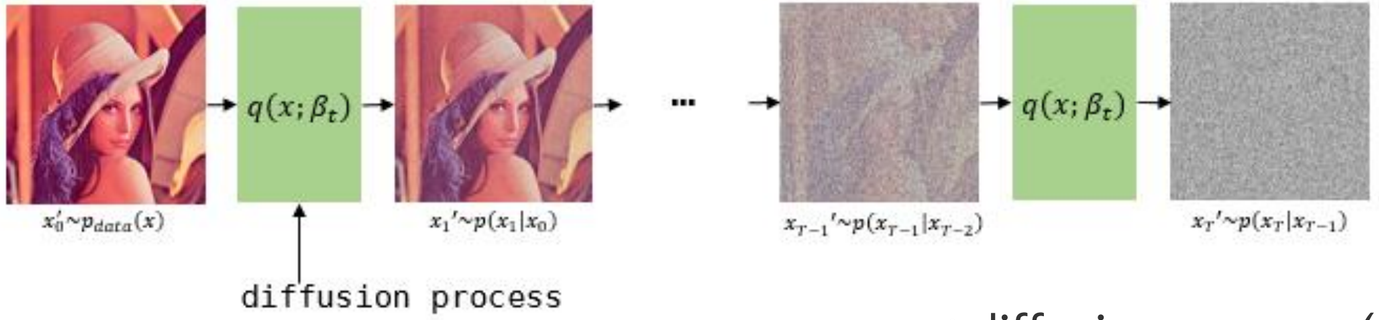
데이터 x에서 점점 noise를 추가해서 noise data로 만들고, noise data에서 데이터 x로 돌아오는 과정을 학습해 distribution을 estimate하는 모델

Abstract



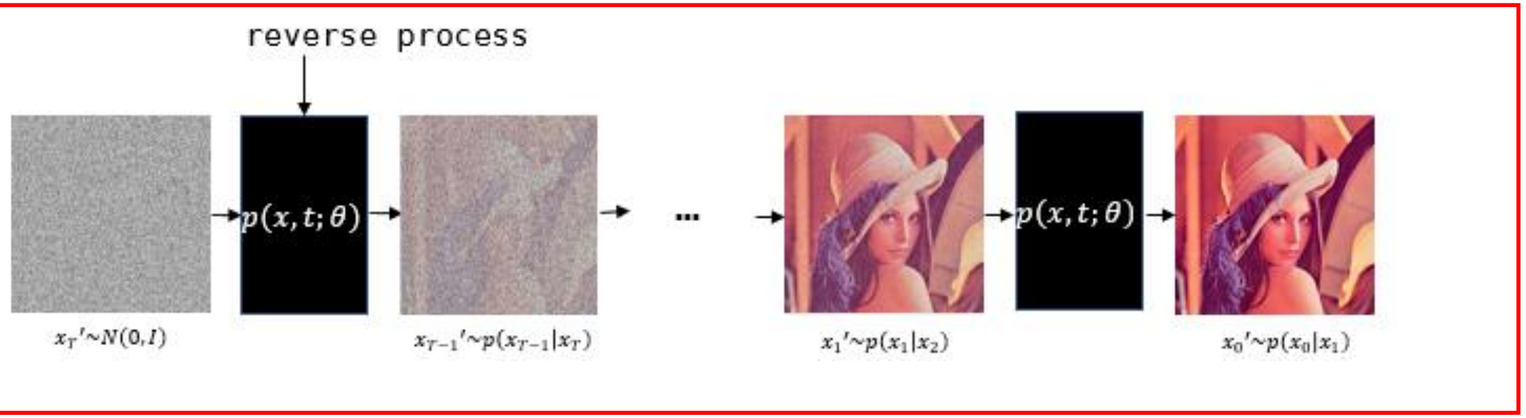
EWHA EURON

Abstract



q : diffusion process (노이즈 추가 과정)

p : reverse process (노이즈를 걷어내는 과정)



noising process의 역과정을 수학적으로 나타내서 역과정을 학습하는 방법이 DDPM

Abstract

*Markov process – 과거 상태들(s1, s2,...,st−1)과 현재 상태(st)가 주어졌을 때, 미래 상태(st+1)는 과거 상태와는 독립적으로 현재 상태에 의해서만 결정된다는 것

Abstract

VAE와 다른점

1. Encoder가 없으며 DDPM은 노이즈를 조금씩 입히는 fixed된 forward process를 가짐.
2. 그 forward distribution이 반드시 gaussian distribution을 따른다. (VAE는 여기에 또 loss function이 필요함)
3. VAE의 decoder는 각 layer마다 개별적인 파라미터를 가지지만, DDPM에서는 모든 time step t에 대해 같은 모델이 사용된다.
4. DDPM에서는 latent variable들의 dimension이 data dimension과 같다.

DDPM은 VAE의 generation process를 T개의 쉬운 markov process로 쪼갠 것과 같음.

DDPM

EWHA EURON

1) Process

- **Forward Process / Diffusion Process :** 점진적으로 gaussian noise를 추가하는 것

$$q(x_t|x_{t-1}) \coloneqq N\left(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t\mathbf{I}\right)$$

x_(t-1)이 주어졌을 때 x_t가 어떻게 나올 것 이냐

$$q(x_1, \dots, x_T|x_0) \coloneqq \prod_{t=1}^{T} q(x_t|x_{t-1})$$

- **Backward Process / Reverse Process :** 미세한 Gaussian noise를 걷어내는 과정

(exact reverse distribution 인 $q(x_{t-1}|x_t)$를 알 수 없기 때문에 $p_\theta x_{t-1}|x_t) \coloneqq N\left(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)\right)$ 로 구함

$$p_\theta(\mathbf{x}_{0:T}) \coloneqq p(\mathbf{x}_T) \prod_{t=1}^{T} p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$$

2) Loss function

- Variational inference를 사용

negative log likelihood로 최소화

$$
\begin{aligned}
&\mathbb{E}_{x_{1:T}\sim q(x_{1:T}|x_0)}[-\log p_\theta(x_0)] \\
&\leq \mathbb{E}_{x_{1:T}\sim q(x_{1:T}|x_0)}\left[-\log p_\theta(\boldsymbol{x_T}) - \sum_{t=1}^{T}\log\frac{p_\theta(x_{t-1}|x_t)}{q(x_t|x_{t-1})}\right] \\
&= \mathbb{E}_{x_{1:T}\sim q(x_{1:T}|x_0)}\left[-\log p_\theta(\boldsymbol{x_T}) - \sum_{t=2}^{T}\log\frac{p_\theta(x_{t-1}|x_t)}{q(x_t|x_{t-1})} - \log\frac{p_\theta(x_0|x_1)}{q(x_1|x_0)}\right] \\
&= \mathbb{E}_{x_{1:T}\sim q(x_{1:T}|x_0)}\left[-\log p_\theta(\boldsymbol{x_T}) - \sum_{t=2}^{T}\log\frac{p_\theta(x_{t-1}|x_t)}{q(x_{t-1}|x_t,x_0)}\cdot\frac{q(x_{t-1}|x_0)}{q(x_t|x_0)} - \log\frac{p_\theta(x_0|x_1)}{q(x_1|x_0)}\right] \quad \because * \\
&= \mathbb{E}_{x_{1:T}\sim q(x_{1:T}|x_0)}\left[-\log p_\theta(\boldsymbol{x_T}) - \sum_{t=2}^{T}\log\frac{p_\theta(x_{t-1}|x_t)}{q(x_{t-1}|x_t,x_0)} - \sum_{t=2}^{T}\log\frac{q(x_{t-1}|x_0)}{q(x_t|x_0)} - \log\frac{p_\theta(x_0|x_1)}{q(x_1|x_0)}\right] \\
&= \mathbb{E}_{x_{1:T}\sim q(x_{1:T}|x_0)}\left[-\log p_\theta(\boldsymbol{x_T}) - \sum_{t=2}^{T}\log\frac{p_\theta(x_{t-1}|x_t)}{q(x_{t-1}|x_t,x_0)} - \log\frac{q(x_1|x_0)}{q(x_T|x_0)} - \log\frac{p_\theta(x_0|x_1)}{q(x_1|x_0)}\right] \\
&= \mathbb{E}_{x_{1:T}\sim q(x_{1:T}|x_0)}\left[\underbrace{-\log\frac{p_\theta(\boldsymbol{x_T})}{q(x_T|x_0)}}_{①} \underbrace{-\sum_{t=2}^{T}\log\frac{p_\theta(x_{t-1}|x_t)}{q(x_{t-1}|x_t,x_0)}}_{②} \underbrace{-\log p_\theta(x_0|x_1)}_{③}\right]
\end{aligned}
$$

①은 VAE의 KL divergence와 비슷한 term
②는 reverse process와 diffusion process의 분포를 매칭시키는(KL divergence를 낮추는) loss
③은 reverse process의 마지막 과정으로, VAE의 reconstruction loss에 대응되는 term

2) Loss function — 기존 diffusion model에서 발전한 점

$$minimize\ \mathbb{E}_q\left[log\frac{q(x_{1:T}|x_0)}{p_\theta(x_{0:T})}\right]$$

$$= \mathbb{E}_q[\underbrace{KL(q(x_T|x_0)||p(x_T))}_{①\ \ L_T} + \underbrace{\sum_{t>1} KL(q(x_{t-1}|x_t,x_0)||p_\theta(x_{t-1}|x_t))}_{②\ L_{t-1}} - \underbrace{\log p_\theta(x_0|x_1)}_{③\ L_0}]$$

① L_T

DDPM의 forward process가 input을 gaussian noise로 만드는 fixed process이기 때문에 L_T는 항상 0에 가까운 상수이므로 학습 과정에서 무시

② L_{t-1}

$$q(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_{t-1};\tilde{\boldsymbol{\mu}}_t(\mathbf{x}_t,\mathbf{x}_0),\tilde{\beta}_t\mathbf{I}), \tag{6}$$

$$\text{where}\quad \underbrace{\tilde{\boldsymbol{\mu}}_t(\mathbf{x}_t,\mathbf{x}_0) := \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\mathbf{x}_0 + \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\mathbf{x}_t}_{Mean}\quad \text{and}\quad \underbrace{\tilde{\beta}_t := \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t}_{variance} \tag{7}$$

2) Loss function — 기존 diffusion model에서 발전한 점

② L_{t-1}

$$L_{t-1} = \mathbb{E}_q \left[\frac{1}{2\sigma^2} \|\tilde{\mu}_t(x_t, x_0) - \mu_\theta(x_t, t)\|^2 \right] + C$$

$$x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t}\boldsymbol{\epsilon}$$

$$L_{t-1} - C = \mathbb{E}_q \left[k \left\| \boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta\left(\sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t}\boldsymbol{\epsilon}, t\right) \right\|^2 \right]$$

$$L_{\text{simple}} = \mathbb{E}_q \left[\left\| \boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta\left(\sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t}\boldsymbol{\epsilon}, t\right) \right\|^2 \right]$$

$\boldsymbol{\epsilon}_\theta(x_t, t)$: x_t와 t가 주어지면 해당 이미지의 noise가 무엇인지 예측하는 network

$\boldsymbol{\epsilon}_\theta(x_t, t)$이 잘 학습되게 되면 노이즈로 부터 순차적으로 노이즈를 제거해가며 선명한 이미지를 얻음

Algorithm 1 Training

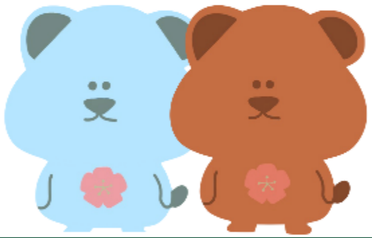
1: **repeat**
2: $\mathbf{x}_0 \sim q(\mathbf{x}_0)$
3: $t \sim \text{Uniform}(\{1, \dots, T\})$
4: $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
5: Take gradient descent step on
$\nabla_\theta \left\| \boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t}\boldsymbol{\epsilon}, t) \right\|^2$
6: **until** converged

Algorithm 2 Sampling

1: $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
2: **for** $t = T, \dots, 1$ **do**
3: $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ if $t > 1$, else $\mathbf{z} = \mathbf{0}$
4: $\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t) \right) + \sigma_t \mathbf{z}$
5: **end for**
6: **return** $\mathbf{x}_0$

EWHA EURON

Conclusion



EWHA EURON

Results

Table 1: CIFAR10 results. NLL measured in bits/dim.

Model	IS	FID	NLL Test (Train)
Conditional			
EBM [11]	8.30	37.9	
JEM [17]	8.76	38.4	
BigGAN [3]	9.22	14.73	
StyleGAN2 + ADA (v1) [29]	**10.06**	**2.67**	
Unconditional			
Diffusion (original) [53]			≤ 5.40
Gated PixelCNN [59]	4.60	65.93	3.03 (2.90)
Sparse Transformer [7]			**2.80**
PixelIQN [43]	5.29	49.46	
EBM [11]	6.78	38.2	
NCSNv2 [56]		31.75	
NCSN [55]	8.87 ± 0.12	25.32	
SNGAN [39]	8.22 ± 0.05	21.7	
SNGAN-DDLS [4]	9.09 ± 0.10	15.42	
StyleGAN2 + ADA (v1) [29]	$\mathbf{9.74} \pm 0.05$	3.26	
Ours (L, fixed isotropic $\mathbf{\Sigma}$)	7.67 ± 0.13	13.51	≤ 3.70 (3.69)
Ours (L_{simple})	9.46 ± 0.11	**3.17**	≤ 3.75 (3.72)

Table 2: Unconditional CIFAR10 reverse process parameterization and training objective ablation. Blank entries were unstable to train and generated poor samples with out-of-range scores.

Objective	IS	FID
$\tilde{\mu}$ prediction (baseline)		
L, learned diagonal $\mathbf{\Sigma}$	7.28 ± 0.10	23.69
L, fixed isotropic $\mathbf{\Sigma}$	8.06 ± 0.09	13.22
$\|\tilde{\boldsymbol{\mu}} - \tilde{\boldsymbol{\mu}}_\theta\|^2$	–	–
ϵ prediction (ours)		
L, learned diagonal $\mathbf{\Sigma}$	–	–
L, fixed isotropic $\mathbf{\Sigma}$	7.67 ± 0.13	13.51
$\|\tilde{\boldsymbol{\epsilon}} - \boldsymbol{\epsilon}_\theta\|^2$ (L_{simple})	$\mathbf{9.46 \pm 0.11}$	**3.17**

- **IS (inception score)**: 생성된 image로부터 classification을 할 때 얼마나 특정 class로의 추정을 잘 하는지에 대한 score. classification 성능이 좋으면서 전체 class를 고르게 생성해낼수록 IS score가 높다.
- **FID (Frechet Inception Distance)**: 실제 데이터를 참고하여 (정확히는 데이터 분포를 참고) 평균, 공분산을 비교하며 낮을수록 좋다.

Conclusion

"a corgi wearing a bow tie and a birthday hat"

"a fire in the background"

"only one cloud in the sky today"

GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models

a photo of a cat → an anime drawing of a super saiyan cat, artstation

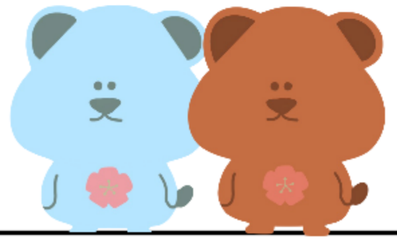
a photo of a victorian house → a photo of a modern house

a photo of an adult lion → a photo of lion cub

a photo of a landscape in winter → a photo of a landscape in fall

Hierarchical Text-Conditional Image Generation with CLIP Latents

THANK YOU



EWHA EURON