Research Proposal

- Learning to Measure Generalized Domain Gap in the Wild

2022.12.13.

Advanced Computer Vision

Sohee Kim

KENTECH / Department of Energy Engineering / Institute for Energy AI

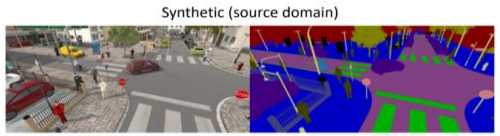
Visual Intelligence & Energy Wise Lab

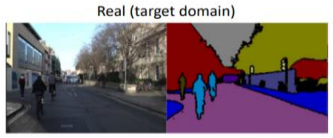
Contents

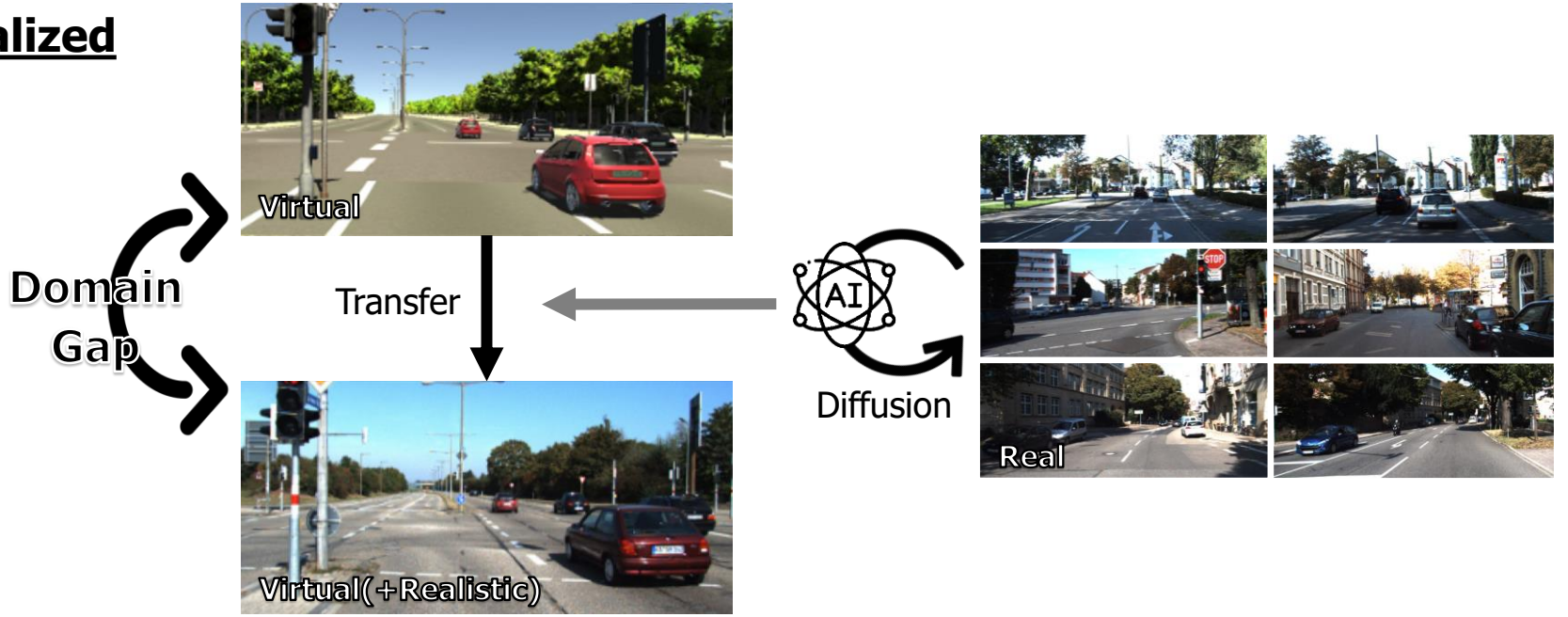
◆ Introduction

❖ Learning to Measure Generalized Domain Gap in the Wild

- **Motivation**
  - 전혀 다른 상황의 두 도메인 간의 거리를 측정할 방법?
    예) 가상의 이미지와 실제 이미지 사이







- Limitation: Previous methods are difficult to measure domain distance between few-shot samples and a target domain
- Necessity of a method to measure generalized domain gap in unpaired scenarios.
- Domain gap이 측정가능 하다면 다양한 task에 적용가능 – semantic segmentation, object detection 등

❖ Proposals

❑ Method to **measure generalized domain gap** between **unpaired image**

❑ Unpaired 상황에서 일반화된 도메인을 측정하는 방법론을 제안한다.

❑ 제안된 도메인 갭 측정방법으로 target task를 수행할 때 발생하는 도메인 차이에 의한 성능 드랍을 완화시키는 방법론을 제안한다.

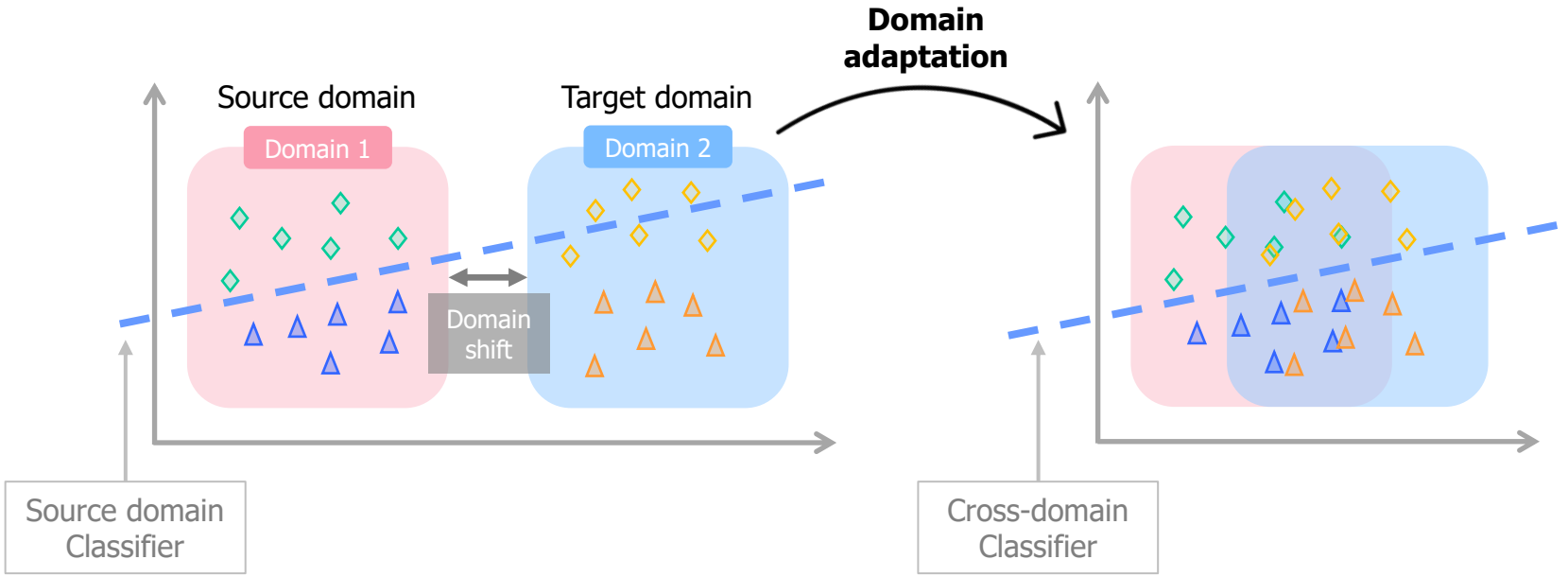
- 타겟 데이터에서 소량의 데이터로 학습할 때 사용할 수 있다.

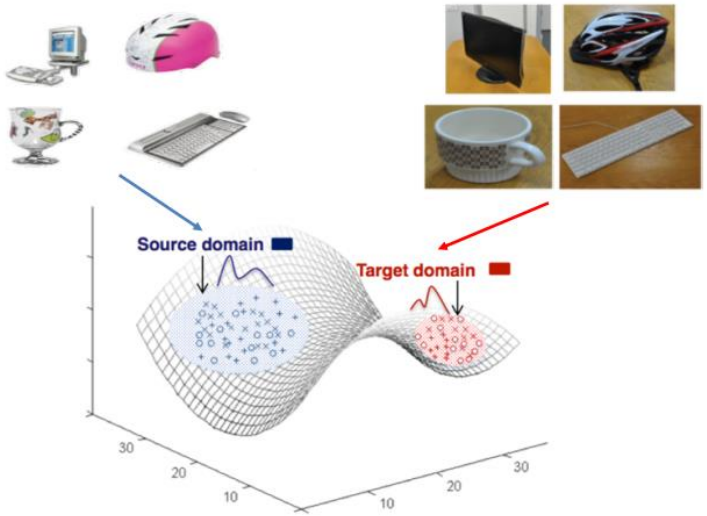
◆ Background – Domain Adaptation

❖ Domain adaptation (DA)

- Domain adaptation is the ability to apply an algorithm trained in one or more "source domains" to a different (but related) "target domain"
- 풍부한 label이 있는 데이터 (Source domain)에서 학습한 지식을 label이 있는 데이터가 부족한 target domain으로 transfer
- 서로 다른 distribution을 가진 **두 도메인**에 robust한 모델을 만드는 것을 목적으로 하는 분야

  ⇒ 데이터를 synthetic 환경에서 얻어서 실제 환경에서 동작 시키길 원하는 모든 문제에 적용가능

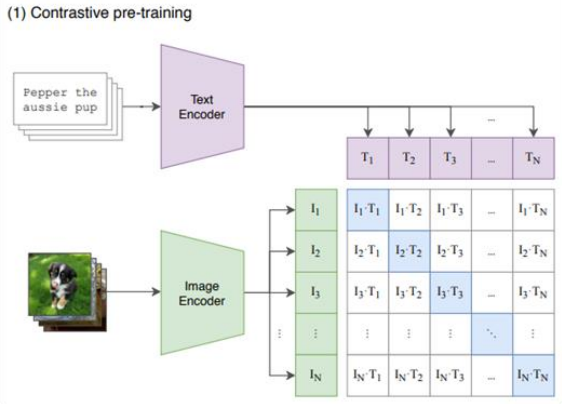




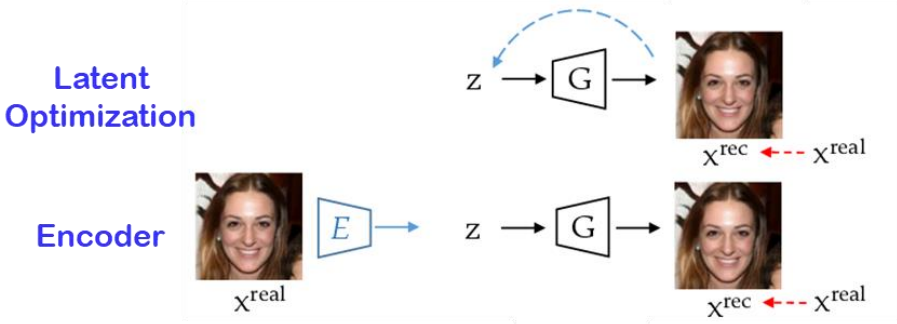
◆ Background – Style Transfer

❖ Style Transfer

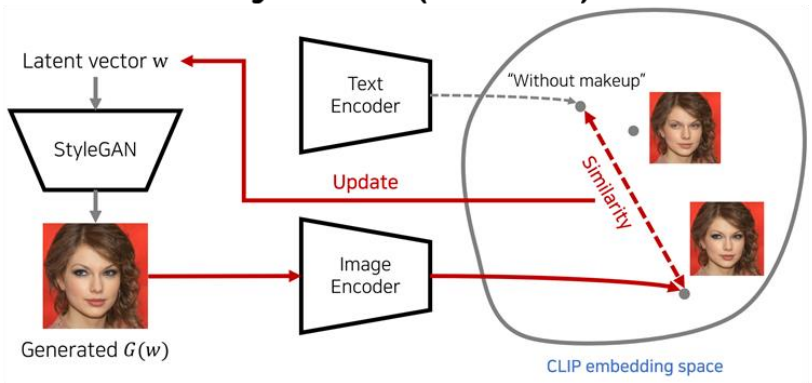
CLIP (ICML21)



GAN-Inversion

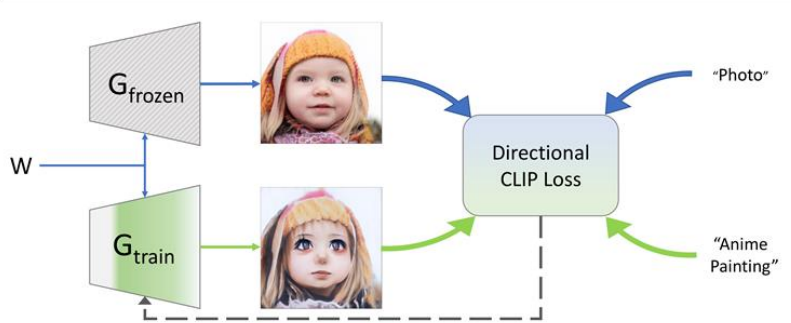


StyleCLIP (ICCV21)



$$\mathcal{L}_{\text{global}}(\boldsymbol{x}_{\text{gen}}, y_{\text{tar}}) = D_{\text{CLIP}}(\boldsymbol{x}_{\text{gen}}, y_{\text{tar}})$$

StyleGAN-NADA (arXiv21)



$$\mathcal{L}_{\text{direction}}(\boldsymbol{x}_{\text{gen}}, y_{\text{tar}}; \boldsymbol{x}_{\text{ref}}, y_{\text{ref}}) := 1 - \frac{\langle \Delta I, \Delta T \rangle}{\|\Delta I\| \|\Delta T\|}$$

where $\Delta T = E_T(y_{\text{tar}}) - E_T(y_{\text{ref}})$, $\Delta I = E_I(\boldsymbol{x}_{\text{gen}}) - E_I(\boldsymbol{x}_{\text{ref}})$

◆ Background – Style Transfer

❖ **CLIP** (Contrastive Learning-Image Pretraining) [1] (Open AI)

- 이미지와 텍스트를 같은 공간으로 보내서 (Multimodal) representation learning을 수행하는 모델
- **Image representation** = Image의 특성을 최대한 잘 설명하는 어떤 feature(representation)을 잘 뽑아 이를 다른 downstream task에 활용 ⇒ 다른 종류의 task로도 유연하게 zero-shot transfer 가능

한번도 보지 못한 datasets 에 대해 분류를 하는 작업

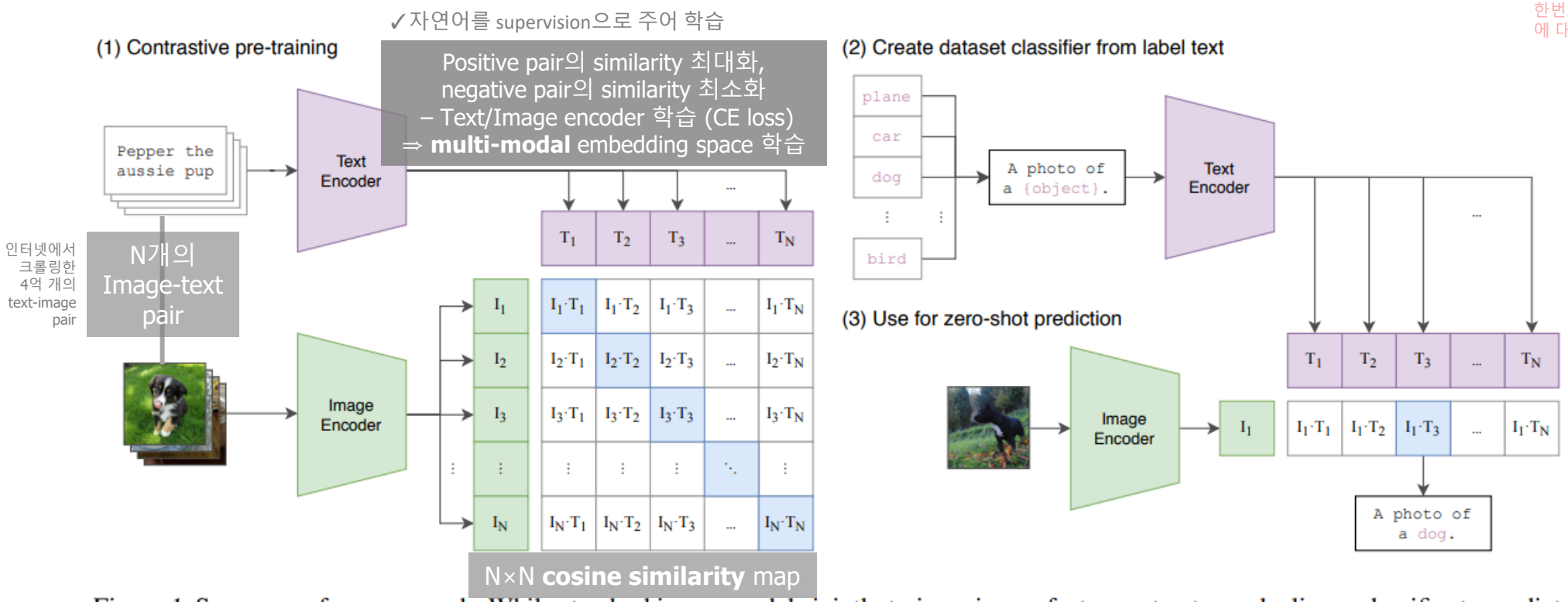


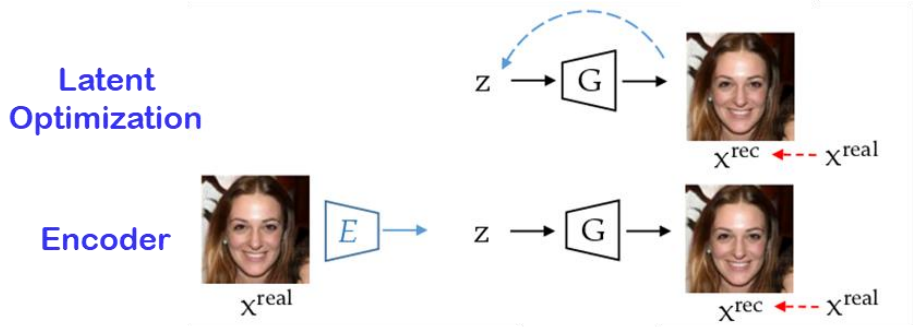
Figure 1. Summary of our approach. While standard image models jointly train an image feature extractor and a linear classifier to predict some label, CLIP jointly trains an image encoder and a text encoder to predict the correct pairings of a batch of (image, text) training examples. At test time the learned text encoder synthesizes a zero-shot linear classifier by embedding the names or descriptions of the target dataset's classes.

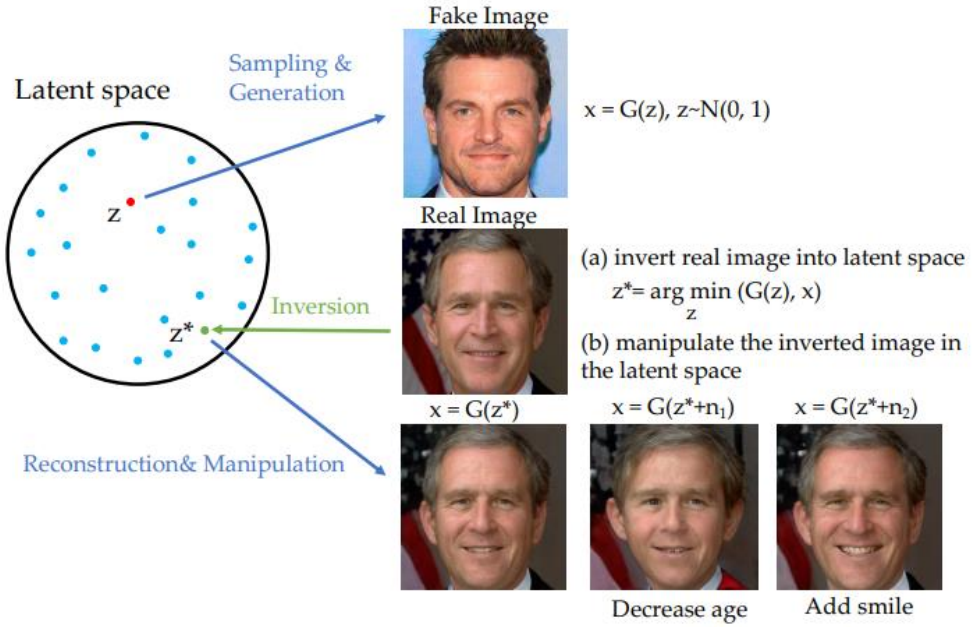
[1] Radford, Alec, et al. "Learning transferable visual models from natural language supervision." *International Conference on Machine Learning*. PMLR, 2021.

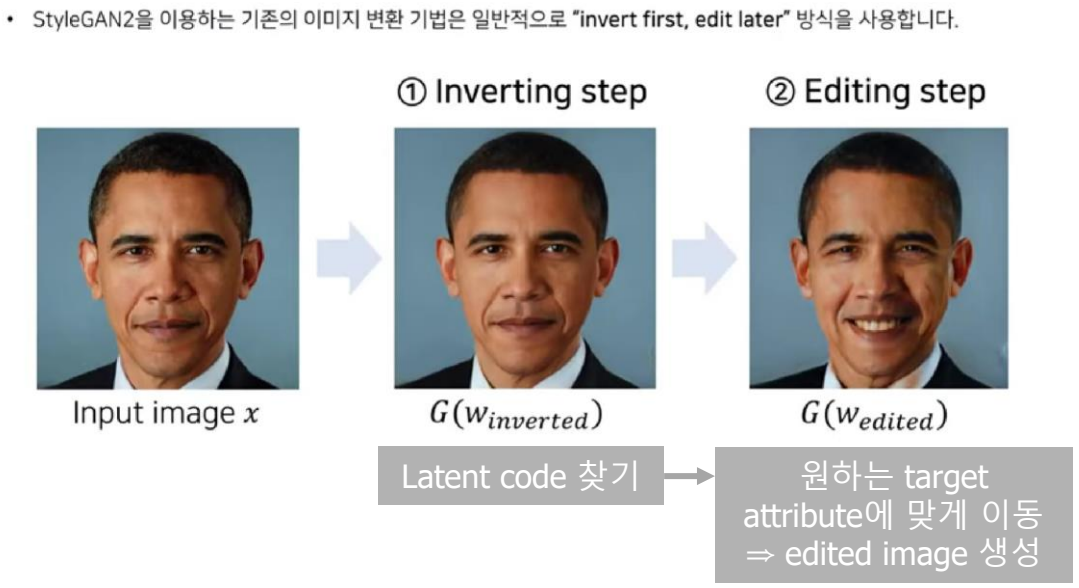
◆ Background – Style Transfer

❖ GAN Inversion

- 입력 이미지와 유사한 결과 이미지를 얻을 수 있도록 하는 **latent vector를 찾는 과정**
- Input image를 먼저 원하는 latent space상의 latent vector로 invert해준 뒤에(inverting) 해당 latent vector를 원하는 semantic 변형 방향의 특정 vector를 더해 operation된 latent vector를 만들어내고(editing) 다시 generator에 태워 editing된 이미지를 얻는 방식





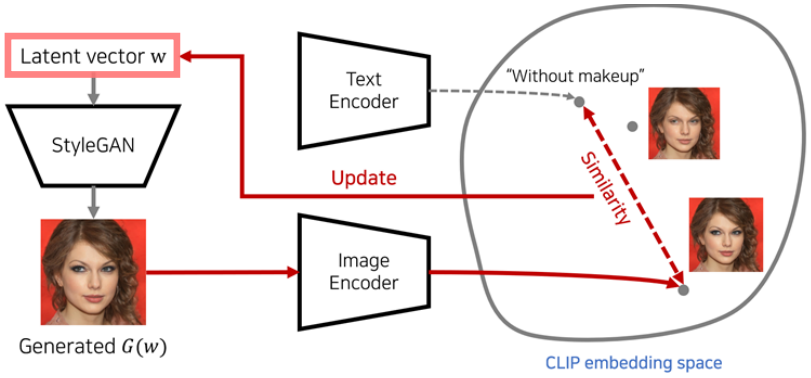


[2] Xia, Weihao, et al. "Gan inversion: A survey." IEEE Transactions on Pattern Analysis and Machine Intelligence (2022).

◆ Background – Style Transfer

❖ Style CLIP : Text-Driven Manipulation of StyleGAN Imagery [3] (Adobe)

- StyleGAN과 CLIP model을 기반으로 text기반의 이미지를 생성한 모델
- 이전 모델들보다 훨씬 직관적이며, latent space를 일일이 찾지 않아도 text에 따라 이미지 조작 가능
- Text-guided latent optimization : CLIP model의 loss network를 도입하여 text를 바탕으로 **input latent vector w 를 수정할 수 있도록 함**



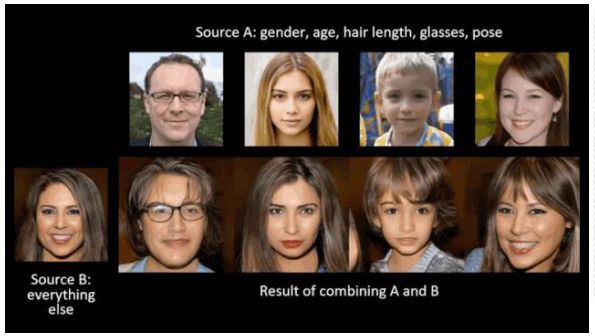
$$\mathcal{L}_{\text{global}}(\boldsymbol{x}_{\text{gen}}, y_{\text{tar}}) = D_{\text{CLIP}}(\boldsymbol{x}_{\text{gen}}, y_{\text{tar}})$$

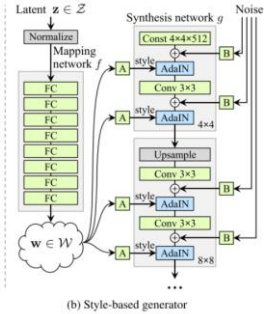
➢ StyleCLIP은 pretrained StyleGAN generator와 CLIP model for a joint language-vision embedding를 바탕으로 만들어짐 ⇒ Generator가 pretraining되지 않은 영역에 대해서는 이미지 조작이 어려움

➢ 또한, text prompt 역시 CLIP space외의 영역에 mapping된다면 의미 있는 visual manipulation을 할 수 없을 것임

✓ StyleGAN [4] (NVIDIA)

- Discriminator나 loss function은 건들이지 않고, style을 더 잘 학습시키도록 generator의 architecture를 발전시킨 모델
- style-based generator로 고해상도 이미지를 높은 퀄리티로 생성





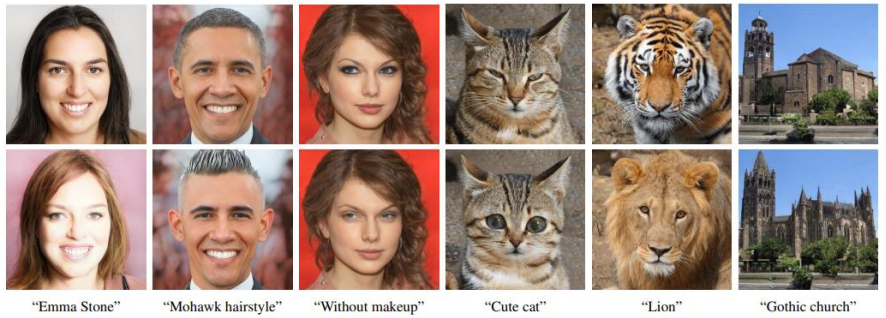


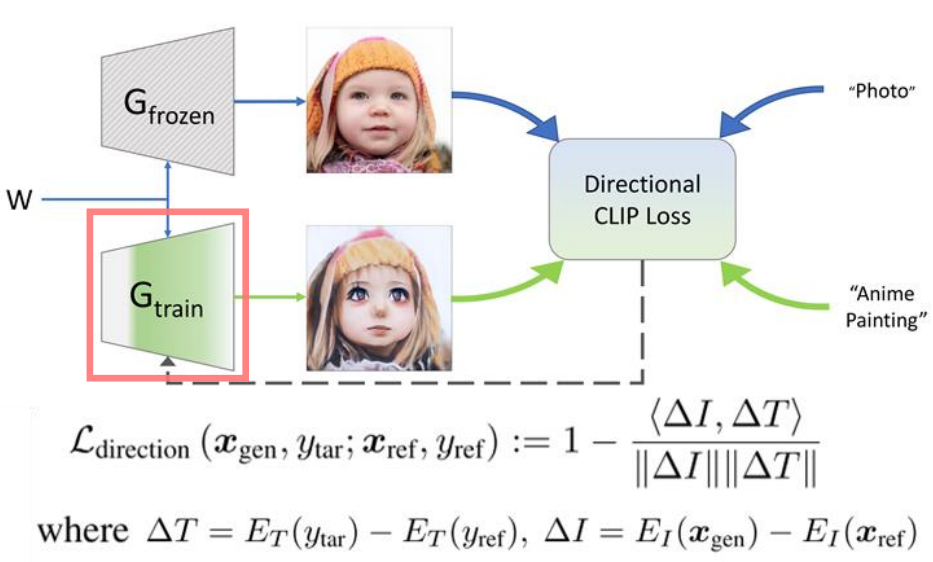
Figure 1. Examples of text-driven manipulations using StyleCLIP. Top row: input images; Bottom row: our manipulated results. The text prompt used to drive each manipulation appears under each column.

[3] Patashnik, Or, et al. "Styleclip: Text-driven manipulation of stylegan imagery." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021.
[4] Karras, Tero, Samuli Laine, and Timo Aila. "A style-based generator architecture for generative adversarial networks." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019.

◆ Background – Style Transfer

❖ **Style GAN-NADA** : CLIP-guided domain adaptation of image generators [5] (NVIDIA)

- Latent vector w를 업데이트하는 것이 아니라, generator 모델을 직접 fine-tuning을 하기때문에 조금 더 flexible 함

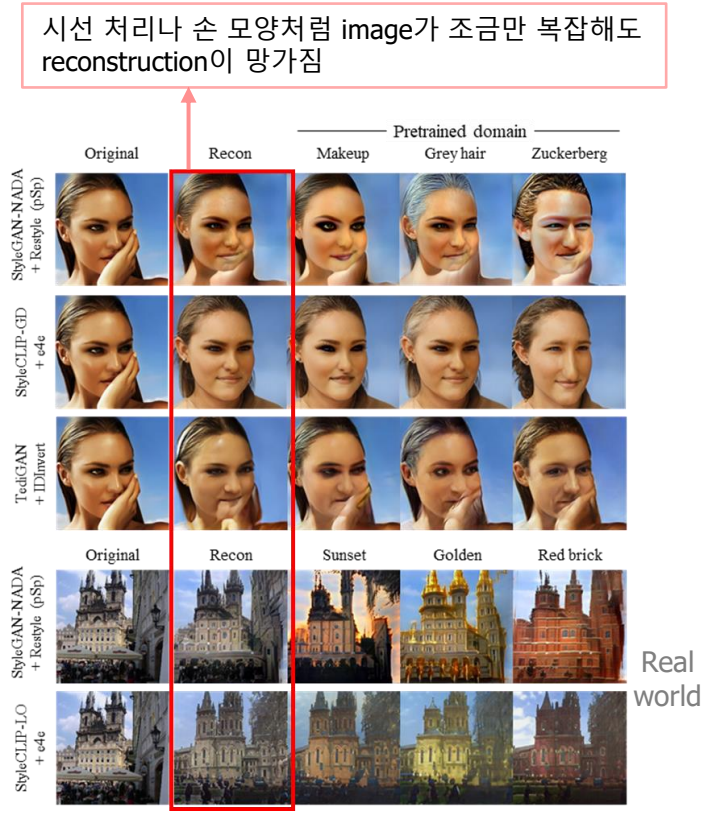


$$\mathcal{L}_{\text{direction}}(\boldsymbol{x}_{\text{gen}}, y_{\text{tar}}; \boldsymbol{x}_{\text{ref}}, y_{\text{ref}}) := 1 - \frac{\langle \Delta I, \Delta T \rangle}{\|\Delta I\| \|\Delta T\|}$$

$$\text{where } \Delta T = E_T(y_{\text{tar}}) - E_T(y_{\text{ref}}),\ \Delta I = E_I(\boldsymbol{x}_{\text{gen}}) - E_I(\boldsymbol{x}_{\text{ref}})$$

- Directional CLIP loss (StyleGAN-NADA) → robust to mode-collapse issues
  - By aligning the direction between the image representations with the direction between the reference text and the target text, distinct images should be generated.

GAN-inversion-based models (StyleCLIP, Style GAN-NADA)



Limitation

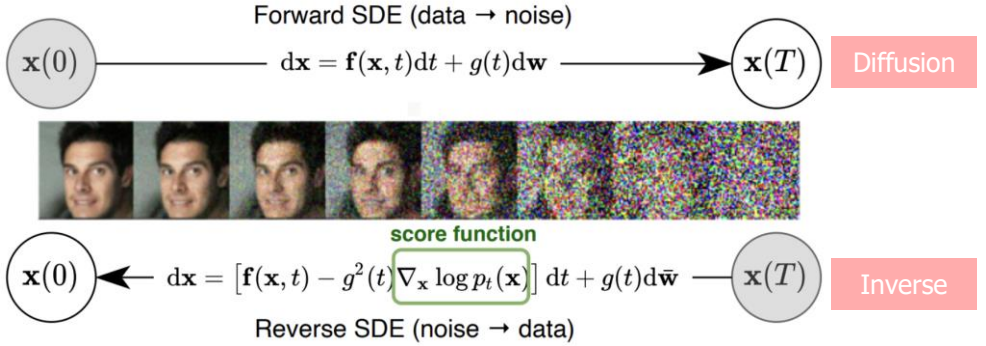
- Face 의 경우 쉬운 Task 임에도 불구하고 Reconstruction 이 잘되지 않음
- GAN-Inversion-based SOTA 모델의 결과물도 대부분 전형적인 Shape 를 가지는 image
- Real world 의 경우 Image 의 Diversity 가 높기 때문에 한계점을 가짐

[5] Gal, Rinon, et al. "Stylegan-nada: Clip-guided domain adaptation of image generators." ACM Transactions on Graphics (TOG) 41.4 (2022): 1-13.

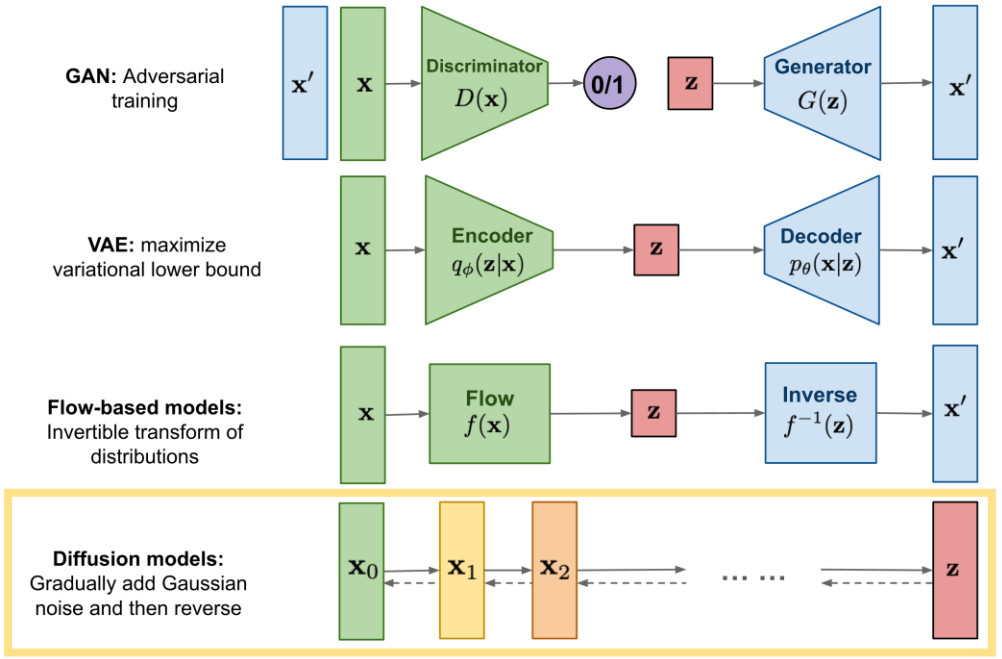
◆ Background – Style Transfer

❖ Style Transfer – Diffusion model

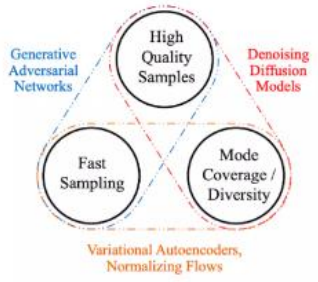
- Diffusion = 확산 : 점들이 점점 확산됨. 정보를 잃는다.
- 데이터셋의 이미지들에 작은 노이즈를 주입하는 과정들로 구성된 "정방향 프로세스"가 있을 때, 해당 프로세스의 반대인 "역방향 프로세스"를 배워, 노이즈로부터 데이터셋 분포에 포함된 샘플을 생성하는 모델



> Overview of different types of generative models.



https://lilianweng.github.io/posts/2021-07-11-diffusion-models/



✓ **Forward process** (diffusion process)
- 원본 이미지에 gaussian noise를 순차적으로 추가하며 완전한 random noise로 만들어주는 과정

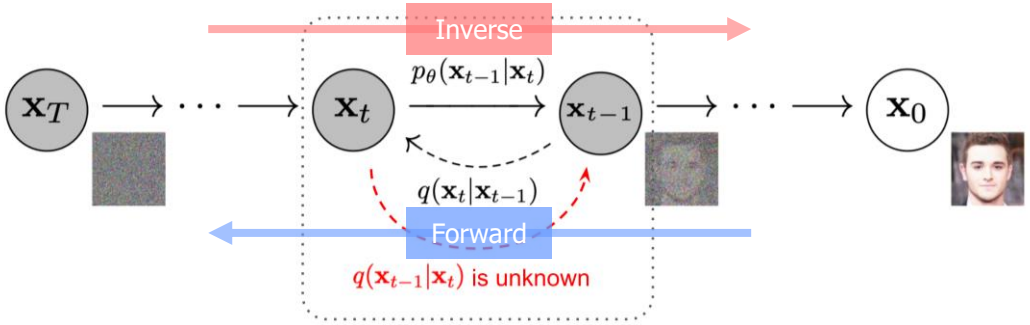
✓ **Reverse process** (inverse process)
- 역변환을 학습하고, 이 학습된 역변환을 사용하여 random noise로부터 이미지를 생성 (noise를 제거하는 과정을 배움)

- 장점
  - GAN과는 달리 stationary training objective를 사용
  - Model scalability(CNN architecture)
  - Distribution coverage가 높음 → 다양한 이미지 생성 가능
- 단점
  - 순차적인 inverse process를 통해 이미지가 생성되므로, 생성 속도가 비교적 느림

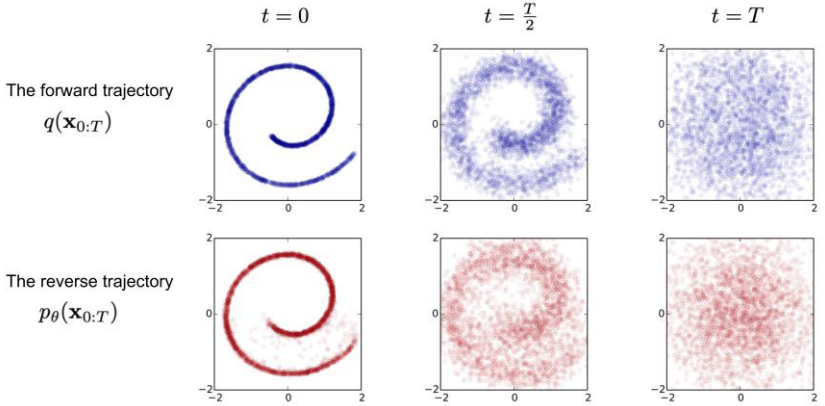
◆ Background – Style Transfer

❖ Style Transfer – Diffusion model

- **DDPM** (Denoising Diffusion Probabilistic Models) [6]



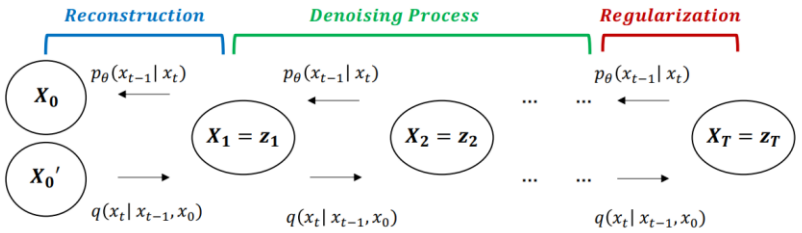
$$q(x_t|x_{t-1}) := \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t \mathrm{I}))$$



$$p_\theta(x_{t-1}|x_t) := \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t))$$

- VAE와 Diffusion의 구조 비교

✓ Diffusion

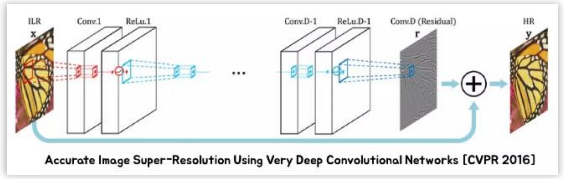


$$Loss_{Diffusion} = D_{KL}(q(z \mid x_0) \| P_\theta(x_0 \mid z)) - E_{z\sim q(z|x)}[\log P_\theta(z)]$$

$$= D_{KL}(q(z \mid x_0) \| P_\theta(z)) + \sum_{t=2} D_{kL}(q(x_{t-1} \mid x_t, x_0) \| P_\theta(x_{t-1} \mid x_t)) - E_q[\log P_\theta(x_0 \mid x_1)]$$

Regularization — Denoising Process — Reconstruction

✓ Residual estimation
- Low resolution 이미지를 같이 활용.



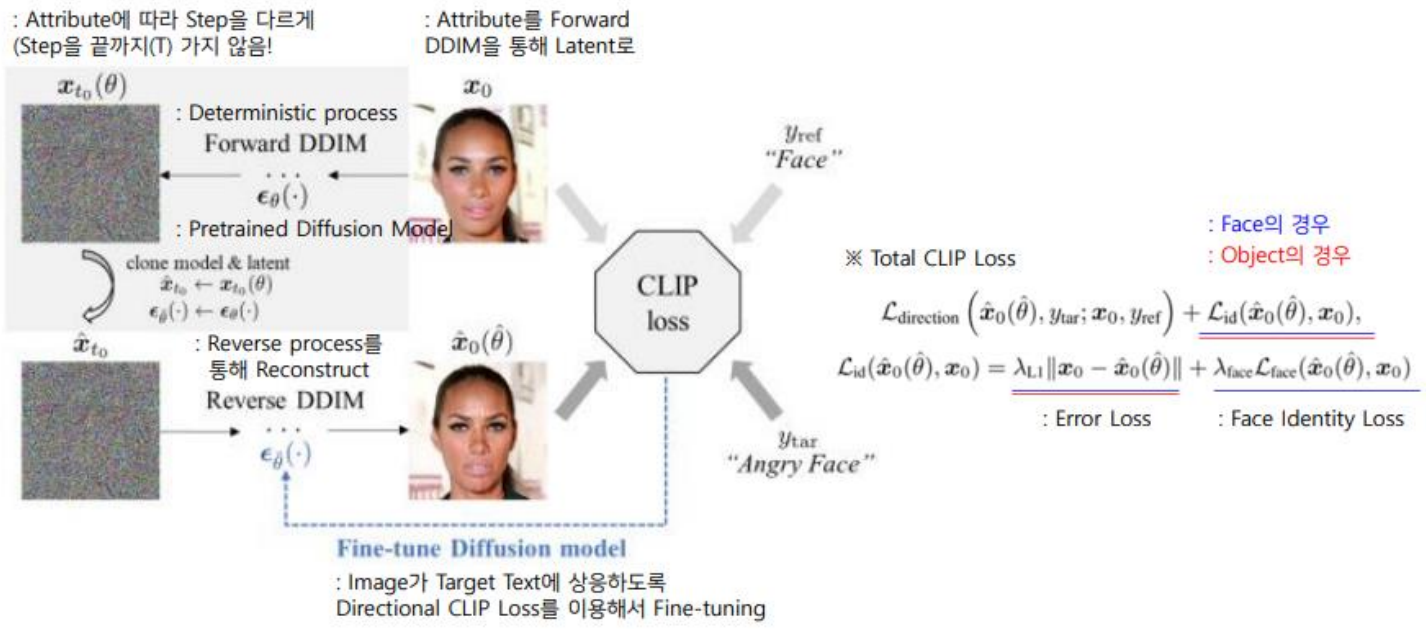
✓ Loss simplification
- Regularization term을 없앰 : β_t를 학습시키지 말고 linear하게 증가하도록 함 (β_t = 0~1 한번의 noise 크기)
- Not to learn variance : β_t로 부터 구함

$$Loss_{DDPM} = \mathbb{E}_{x_0,\epsilon}\left[\left|\epsilon - \epsilon_\theta\left(\sqrt{\bar{\alpha}_t} + \sqrt{1-\bar{\alpha}_t}\epsilon, t\right)\right|^2\right]$$

[6] Ho, Jonathan, Ajay Jain, and Pieter Abbeel. "Denoising diffusion probabilistic models." Advances in Neural Information Processing Systems 33 (2020): 6840-6851.

◆ Background – Style Transfer

❖ Style Transfer – Diffusion CLIP [8]



- Diffusion Model 은 Inversion Capability 가 좋기 때문에 Image Manipulation 에 적합
- Fine-tuning 을 위한 Novel Sampling strategy 제안 → 빠르고 정확하게 Reconstruct 가능
- 의도치 않은 변화없이 In-and out-of-domain Manipulation 가능 (SOTA 성능 보여줌)
- ImageNet Image Manipulation → General Application
- Unseen Domain / Multi-attribute Transfer 가능

▪ **DDIM** (Denoising Diffusion Implicit Models) [7]

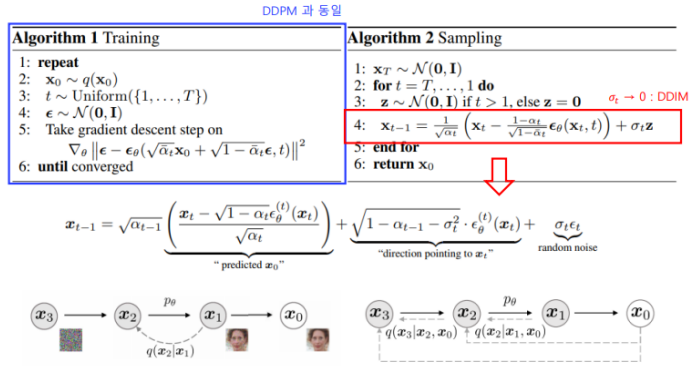
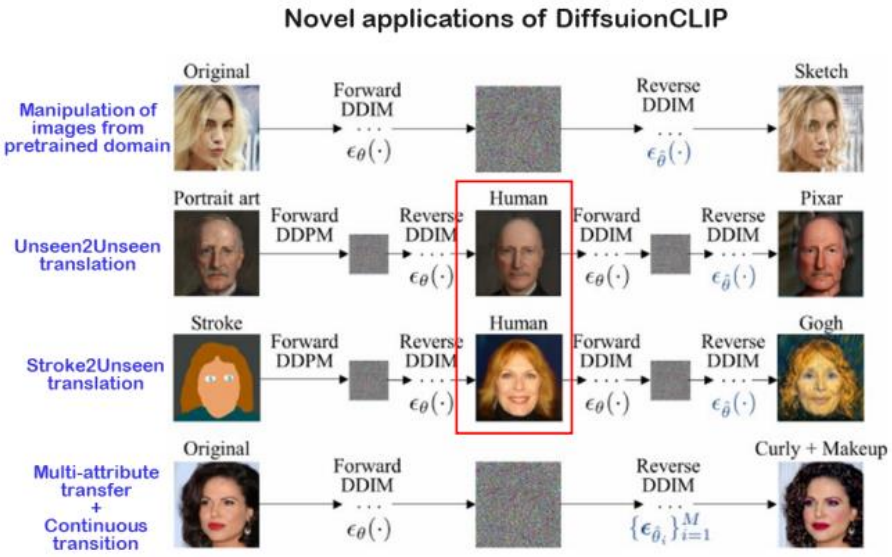


Figure 1: Graphical models for diffusion (left) and non-Markovian (right) inference models.



[7] Song, Jiaming, Chenlin Meng, and Stefano Ermon. "Denoising diffusion implicit models." arXiv preprint arXiv:2010.02502 (2020).
[8] Kim, Gwanghyun, et. al. "DiffusionCLIP: Text-Guided Diffusion Models for Robust Image Manipulation." IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 2022.

◆ Background – Measure metric

1. Inception Score (IS)

- KL-Divergence between conditional and marginal label distributions over generated data
- GAN으로 생성된 이미지를 평가하는 지표 중 하나
- IS 에서는 생성된 이미지의 클래스를 예측할 때 pre-train된 inception network를 사용

▪ 생성된 이미지를 평가할 때 중요한 지표 2가지
  - **Quality** (이미지의 품질) = p(y|x)
  - **Diversity** (이미지의 다양성) = p(y)

▪ Entropy (엔트로피) == p(y|x) 무질서도
  - 높음 : x에 대한 y를 예측하기 어려움
  - 낮음 : x에 대한 y를 예측 가능

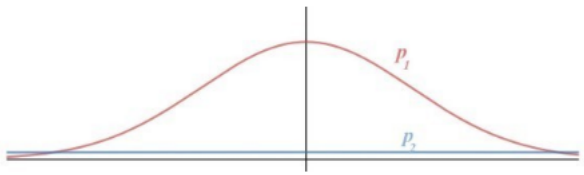
✓ 이미지의 **품질** $p(y|\boldsymbol{x})$
- 생성된 이미지의 conditional label distribution p(y|x) (x:image, y:label)
- **p(y|x) Entropy** 작도록 학습 = 이미지가 객체의 의미있는 정보를 포함하고 있다

✓ 이미지의 **다양성** $\int p(y|\boldsymbol{x} = G(z))dz$
- p(y) = 주변확률 (marginal probability)
- **p(y) Entropy** 크도록 학습 = 이미지가 다양하게 생성됨 => 데이터 분포는 균일

✓ **IS score**

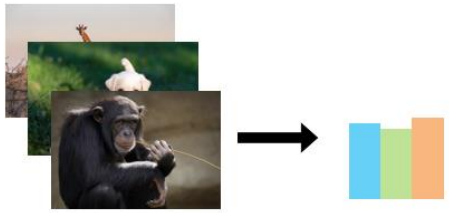
$$InceptS = \exp\left(E_x KL\left(p(y|x)\,\|\,p(y)\right)\right) = \exp\left(E_x E_{p(y|x)}\left[\log\left(\frac{p(y|x)}{p(y)}\right)\right]\right)$$

- KL divergence 활용해서 IS 계산
- 점수로 단일 부동 소수점 숫자를 반환 => 이미지의 퀄리티, 다양성 동시에 측정
- p(y|x)와 p(y)의 엔트로피 차이가 커질수록 위 식의 값은 커짐 → 이미지 생성이 다양하면서도 정확
- ➢ **IS가 높을수록 좋은 성능** (최솟값=1)



- P2는 P1에 비해 균일한 불포 → 예측 어려움
  ⇒P1보다 높은 엔트로피
  ⇒조건부 확률 P(y|x)가 매우 예측가능
  (낮은 엔트로피)해야 한다.

이상적인 Label 분포
- 엔트로피 ↓



이상적인 Marginal 분포
- 엔트로피 ↑

[9] Salimans, Tim, et al. "Improved techniques for training gans." *Advances in neural information processing systems* 29 (2016).

◆ Background – Measure metric

2. Frechet Inception Distance (FID)

- IS : 실제 샘플 대신 생성된 이미지만을 사용해 계산
- FID : **실제 데이터와 생성된 이미지의 분포가 어느정도 비슷한지** 측정하는 지표
  - Wasserstein-2 distance between multi-variate Gaussians fitted to data embedded into a feature space.
- Based on the **feature** vectors of images
  - 만들어낸 이미지들을 pre-trained 된 모델(Inception network)에 넣어 중간 레이어에서 feature를 가져와 활용
- **실제** 데이터 vs **생성된** 데이터에서 얻은 feature의 평균과 공분산을 비교
  - 확률분포 사이의 Wasserstein-2 distance를 측정

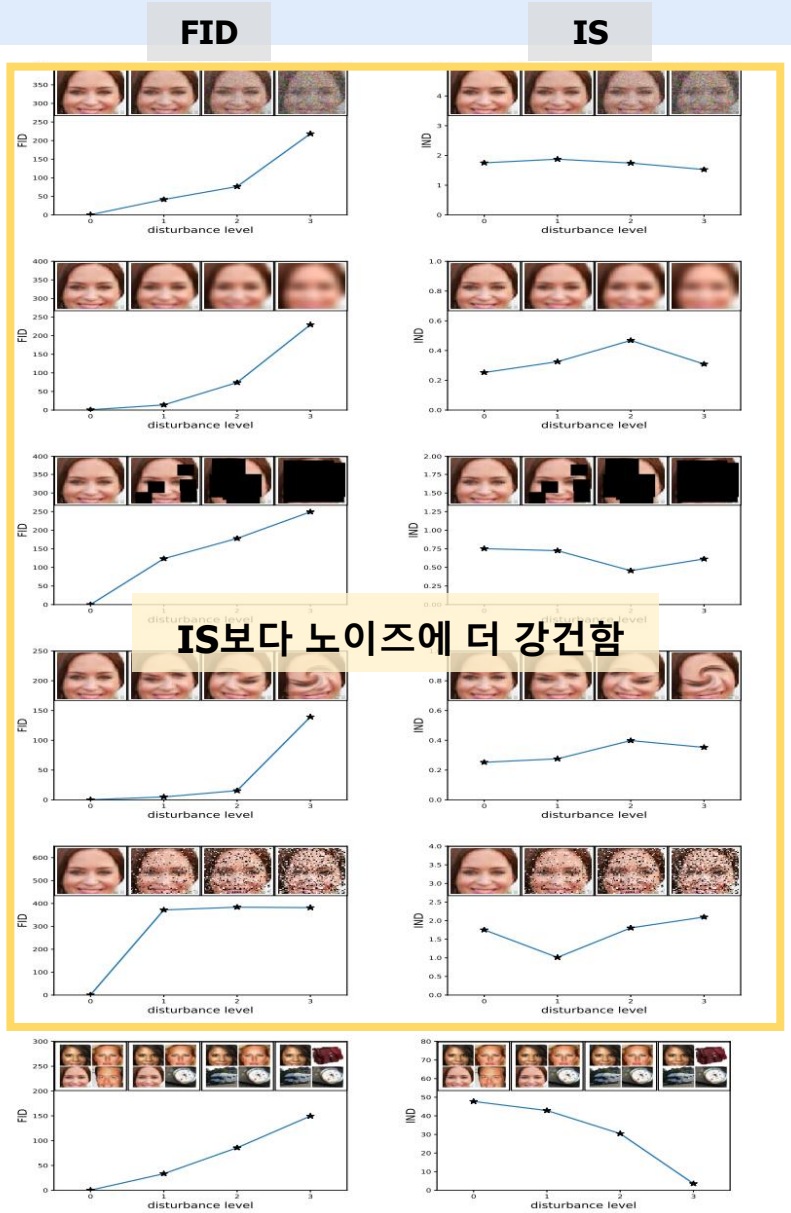
➢ **FID가 낮을수록 좋은 성능**

- 평균적으로 FID가 10 내외이면 좋은 모델이라고 판단할 수 있음

$$FID = d^2\big((m, C), (m_w, C_w)\big) = \underbrace{\|m - m_w\|_2^2}_{\text{평균 (quality)}} + \underbrace{TR(C + C_w - 2(CC_w)^{\frac{1}{2}})}_{\text{공분산 (diversity)}}$$

✓ 가우시안 분포 가정

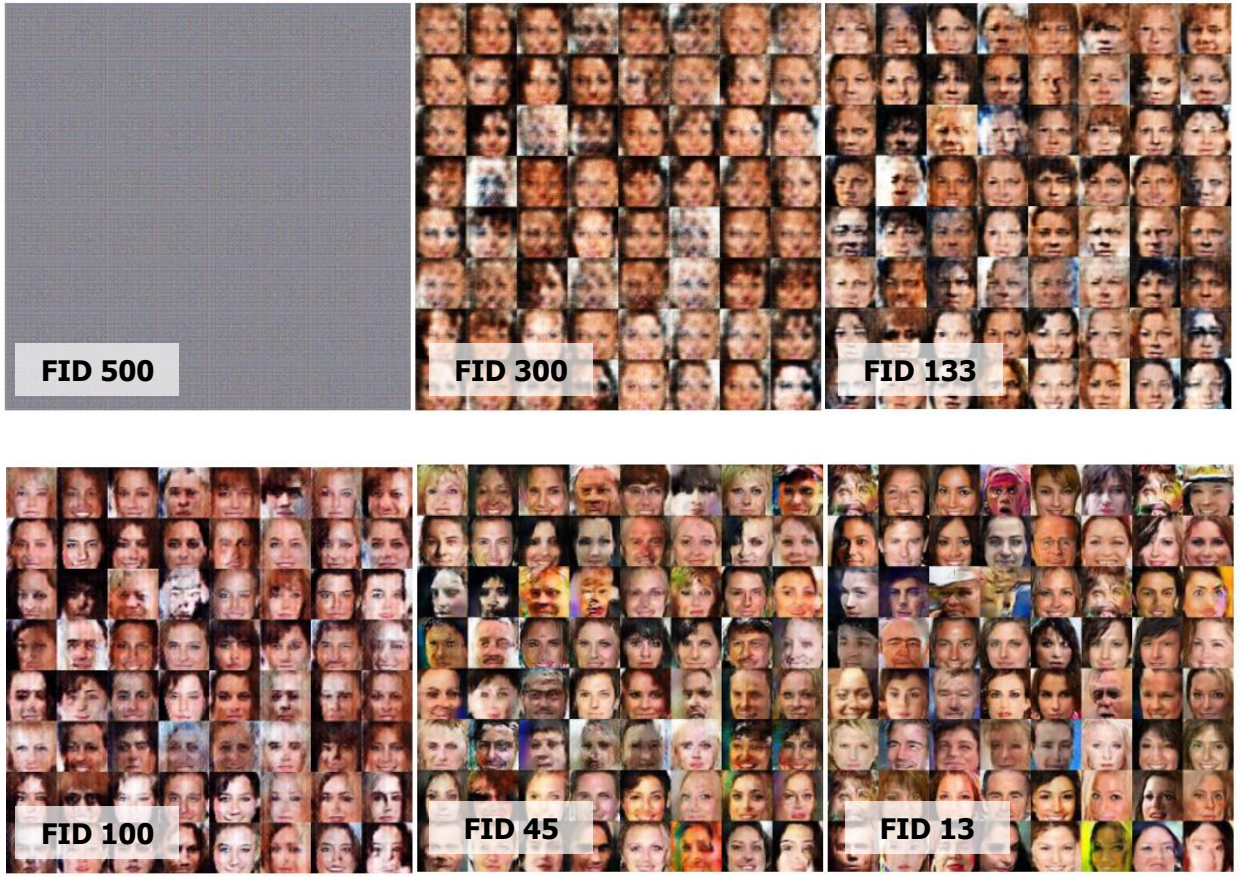
- 실제 데이터 : (m, C) → m : feature들의 평균, C : 공분산
- 생성된 데이터 : (m_w, C_w)



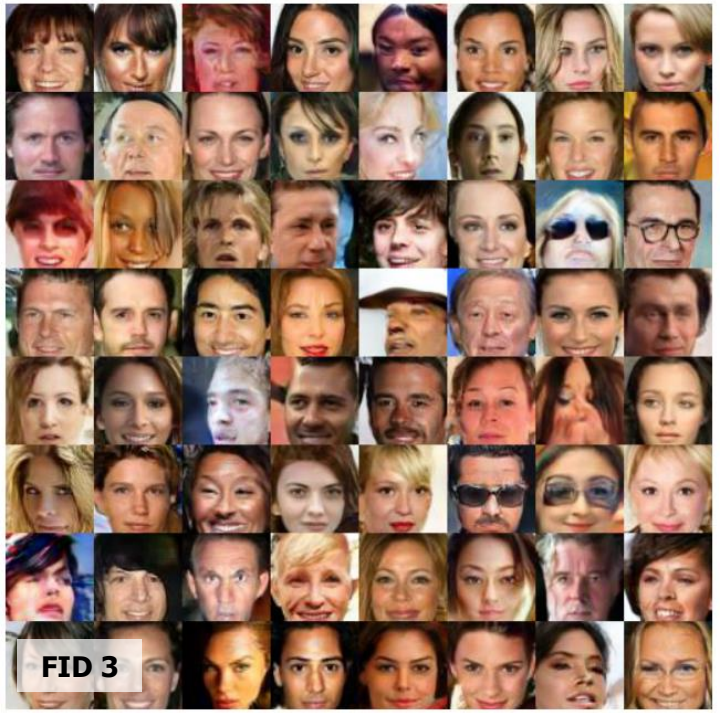
[10] Heusel, Martin, et al. "Gans trained by a two time-scale update rule converge to a local nash equilibrium." Advances in neural information processing systems 30 (2017).

◆ Background – Measure metric

2. Frechet Inception Distance (FID)



> DCGAN – trained on CelebA



> WGAN-GP – trained on CelebA

[10] Heusel, Martin, et al. "Gans trained by a two time-scale update rule converge to a local nash equilibrium." Advances in neural information processing systems 30 (2017).

◆ Background – Measure metric

3. Kernel Inception Distance (KID)

- Measures the dissimilarity between two probability distributions Pr and Pg using samples drawn independently from each distribution.
- MMD를 feature space에서 진행하는 것
- 실제 이미지와 가짜 이미지의 세트 간의 similarity를 보는 방법
  - 실제 이미지 셋 p, 가짜 이미지셋 q
  - p에서 그림 2장으로 뽑고 두 이미지 간의 차이를 구함 → 계속 반복 → 차이의 기댓값 구함 => q에서도 반복
  - 하나는 p에서, 하나는 q에서 뽑은 값으로 평균적인 차이 구함

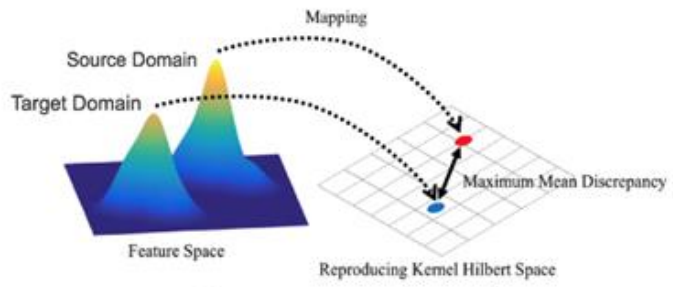
$$MMD(p,q) = E_{x,x'\sim p}[K(x,x')] + E_{x,x'\sim q}[K(x,x')] - 2E_{x\sim p,x'\sim q}[K(x,x')]$$

➢ **KID가 낮을수록 좋은 성능**

  - 진짜 이미지 간 평균 + 가짜 이미지 간 평균 – 2 * (진짜/가짜 이미지)

- MMD : maximum mean discrepancy
  - 두 분포 사이의 distance
  - 각 source, target domain으로부터 계산된 feature map들을 먼저 평균한 다음, 그 결과를 차 연산하고, 그걸 또 제곱하면 최종적으로 MMD 값을 얻을 수 있다.

- MMD: maximum mean discrepancy
  - The distance of distributions is defined as MMD



$$\mathcal{L}_D(\mathbf{X}^s,\mathbf{X}^t) = \mathrm{MMD}(\mathbf{X}^s,\mathbf{X}^t) = \left\| \frac{1}{n^s}\sum_{i=1}^{n^s}\phi(\mathbf{x}_i^s) - \frac{1}{n^t}\sum_{j=1}^{n^t}\phi(\mathbf{x}_j^t) \right\|_{\mathcal{H}}^2$$

ϕ: mapping function
$\mathbf{X}^s$: feature matrix in source domain
$\mathbf{X}^t$: feature matrix in target domain

[11] Bińkowski, Mikołaj, et al. "Demystifying mmd gans." arXiv preprint arXiv:1801.01401 (2018).

◆ Background – Measure metric

❖ IS vs FID vs KID

- The claimed non-monotonicity of the Inception score is quite sensitive to the exact experimental setting
- IS는 실험 환경에 민감
  → **non-monotonicity 비단조성**

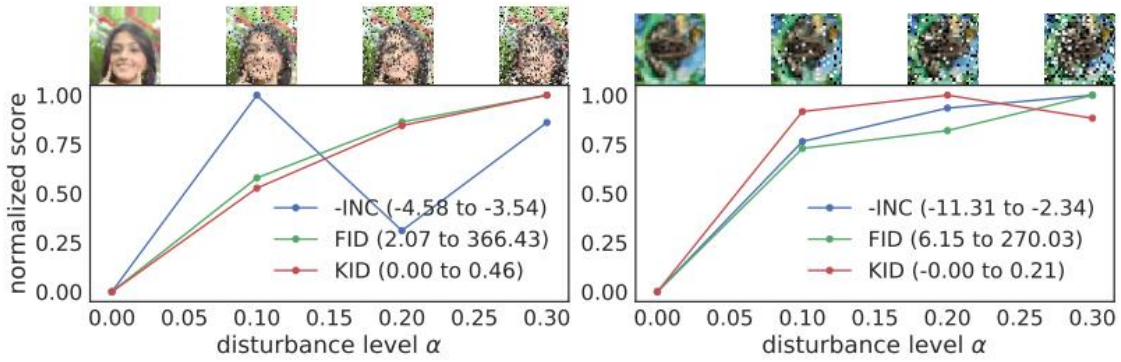


Figure 10: Salt and pepper noise: α is the portion of pixels which are noised.

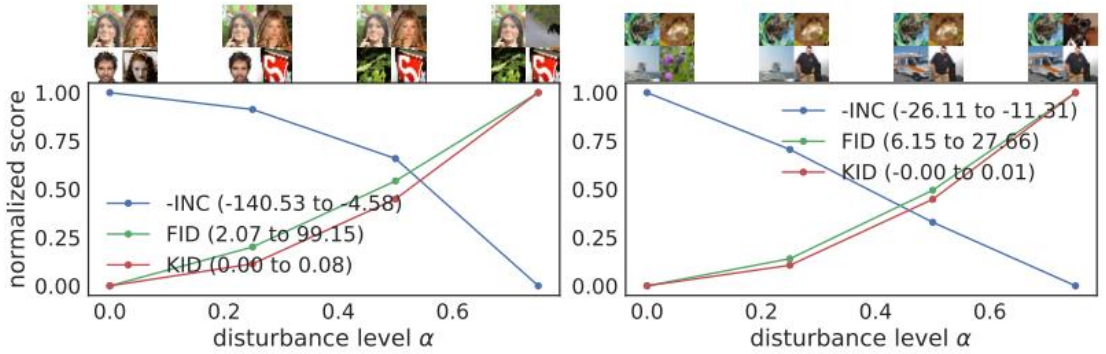
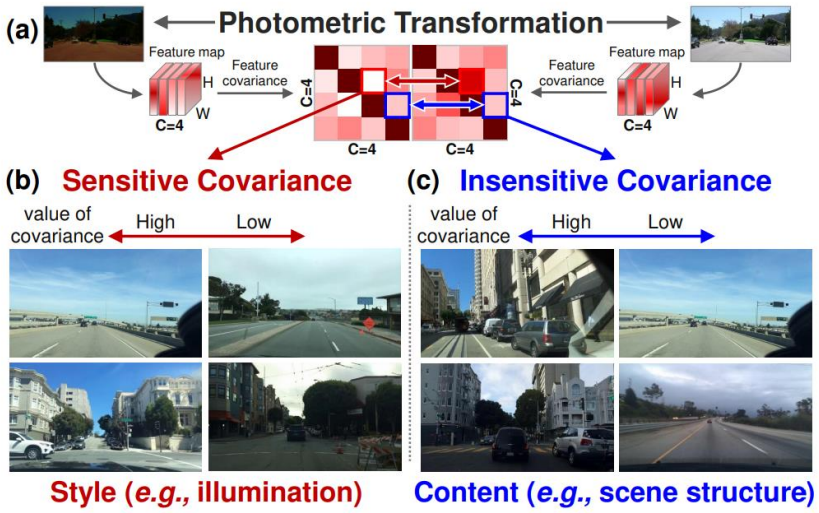


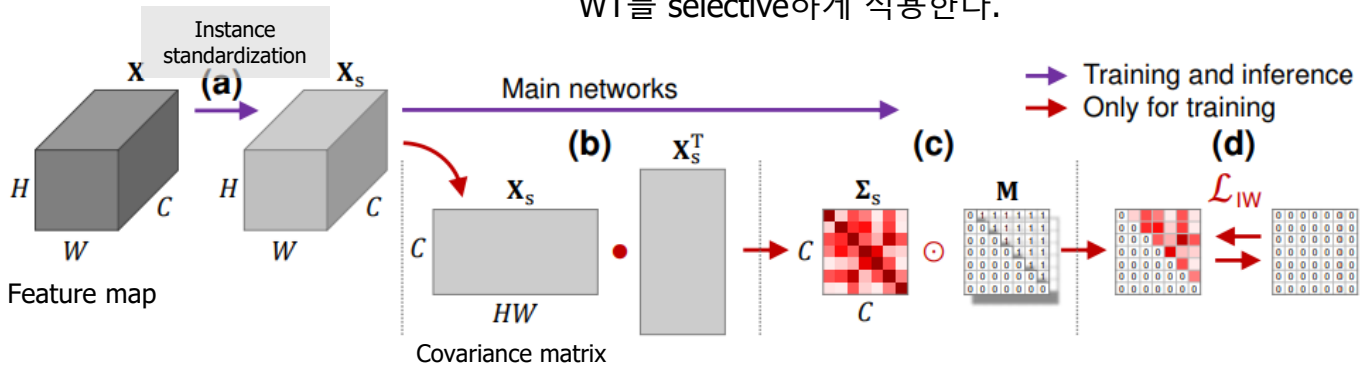
Figure 11: ImageNet contamination: α is the portion of images replaced by ImageNet samples.

◆ Related work

❖ **RobustNET** : Improving Domain Generalization in Urban-Scene Segmentation via Instance Selective Whitening [12]

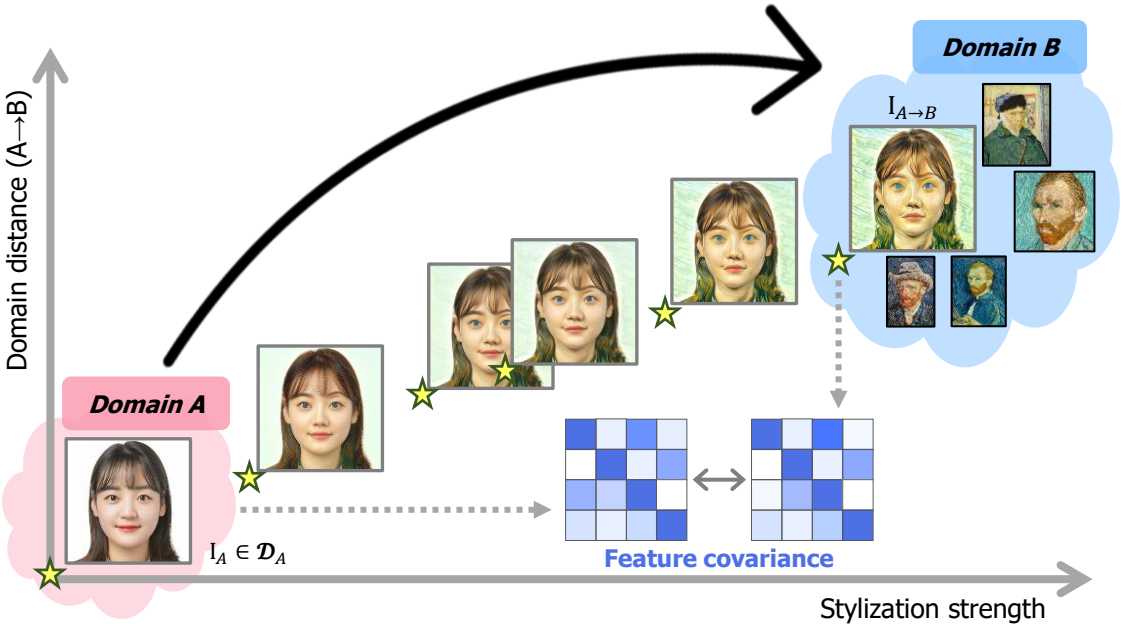


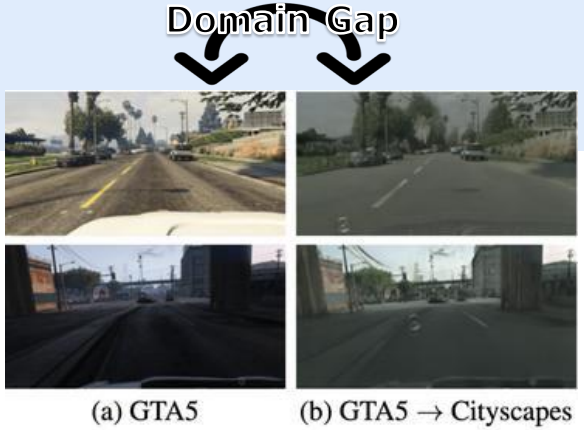
- **Data generalization** : DA는 특정한 target domain이 있다면, DG는 unseen domain을 다룸
- **Feature covariance** (공분산)
  - 기존의 연구[13]에 의해서, feature correlations(covariance matrix)자체가 이미지의 style information을 담고 있다고 밝혀졌다
  - 이미지의 style을 제거하기 위해서, whitening transformation 사용
- Whitening Transformation
  - 초기 Layer의 Feature map에 대해 채널 방향 공분산 행렬이 단위행렬이 되도록 만든 변환
    - Feature map (C×H×W)를 [HW 벡터 C개] 로 변환한 후, C개의 백터들에 대한 Covariance Matrix를 Identity Matrix(단위행렬) 형태가 되도록 하는 것
  - 이렇게 하면 이미지의 Style 정보가 제거 된다는 가설이 있기 때문에, RobustNET에서는 WT를 selective하게 적용한다.



[12] Choi, Sungha, et al. "RobustNet: Improving Domain Generalization in Urban-Scene Segmentation via Instance Selective Whitening." IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 2021.
[13] Gatys, Leon, Alexander S. Ecker, and Matthias Bethge. "Texture synthesis using convolutional neural networks." Advances in neural information processing systems 28 (2015).

◆ Proposal





(a) GTA5 (b) GTA5 → Cityscapes

✓ **Use unpaired image**

- How to measure generalized domain gap
  - RobustNET은 photometric transformation만 사용 == **paired** image
  - But could domain adaptation can be trained continuously with **unpaired** image?
    - Unpaired image = *Real image ↔ Virtual image

    (*Real image = DiffusionCLIP을 사용하여 가상을 실제처럼 변환한)

⇒ Method : Use feature covariance to **define domain gap** between unpaired image

✓ **Method to reduce domain gap**

- Gap에 대한 분석을 하면 어떤 이미지 사이에서 gap이 큰지를 알 수 있어서, 도메인 distance에 대한 분석이 가능할 것이다.
  - It can be used as a term as a clue to reduce gap.
- Method to measure distance between images : Ours, FID, KID

⇒ 이 방법을 활용하여, gap을 줄이도록 adaptation하는 기법 설계

Thank you ☺