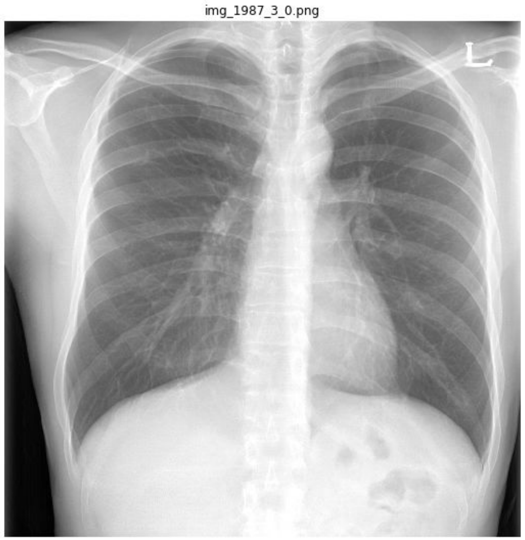
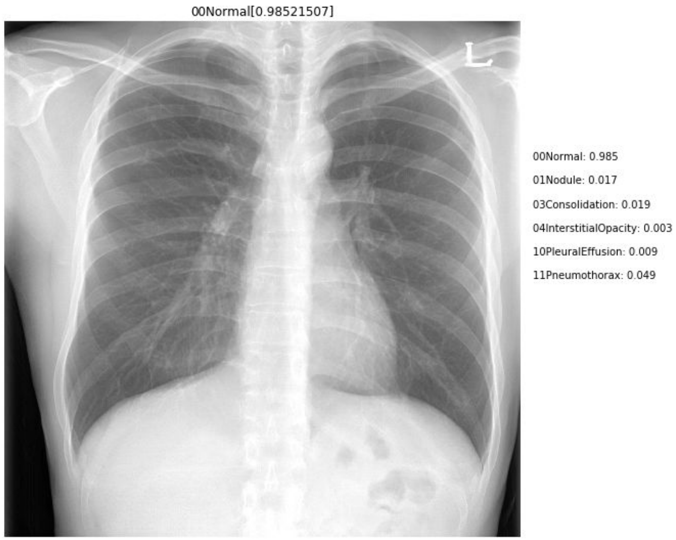
Semi-Supervised StyleGAN for Disentanglement Learning (Info-StyleGAN)

2020/10/19 (월)
논문 리뷰
김민지

https://proceedings.icml.cc/static/paper_files/icml/2020/3146-Paper.pdf

How to achieve controllable generation?

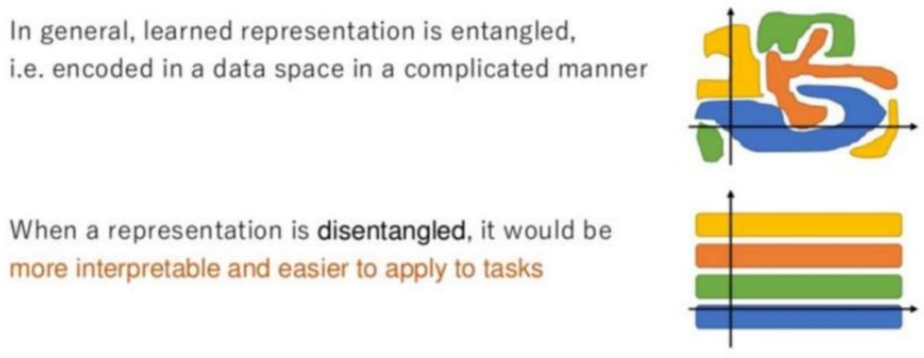




https://github.com/minjeekim00/chestPA_tlgan

Disentanglement Learning

GAN은 noise vector z로부터 잠재 공간을 학습하고,
어떠한 제약 조건도 주어지지 않을 경우 데이터의 특징(feature)는 매우 얽힌 (entangled) 형태로 존재하게 된다.
Representation을 학습할 때, 좀 더 좋은 feature를 갖도록 **제약 (c)**을 줄 수 있다면 참 좋겠다...



In general, learned representation is entangled,
i.e. encoded in a data space in a complicated manner

When a representation is **disentangled**, it would be
more interpretable and easier to apply to tasks

- Disentangled representations
- Controllable generation

Disentanglement Learning - InfoGAN

Representation을 학습할 때, 좀 더 좋은 feature를 갖도록 **제약 (c)**을 Unsupervised manner로 주겠다.

기존 GAN:

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim P_{\text{data}}}[\log D(x)] + \mathbb{E}_{z\sim \text{noise}}[\log(1 - D(G(z)))]$$

InfoGAN:

$$\min_G \max_D V_I(D,G) = V(D,G) - \boxed{\lambda I(c; G(z,c))}$$

http://jaejunyoo.blogspot.com/2017/03/infogan-1.html
https://arxiv.org/pdf/1606.03657.pdf

Disentanglement Learning - InfoGAN

정보 이론에서 Mutual Information 이란?

$$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$$

I(X;Y): Y를 앎으로써 해소되는 X의 불확실성 (Uncertainty)

Y, X가 독립변수라면 ***I(X;Y)***는 0이 된다.

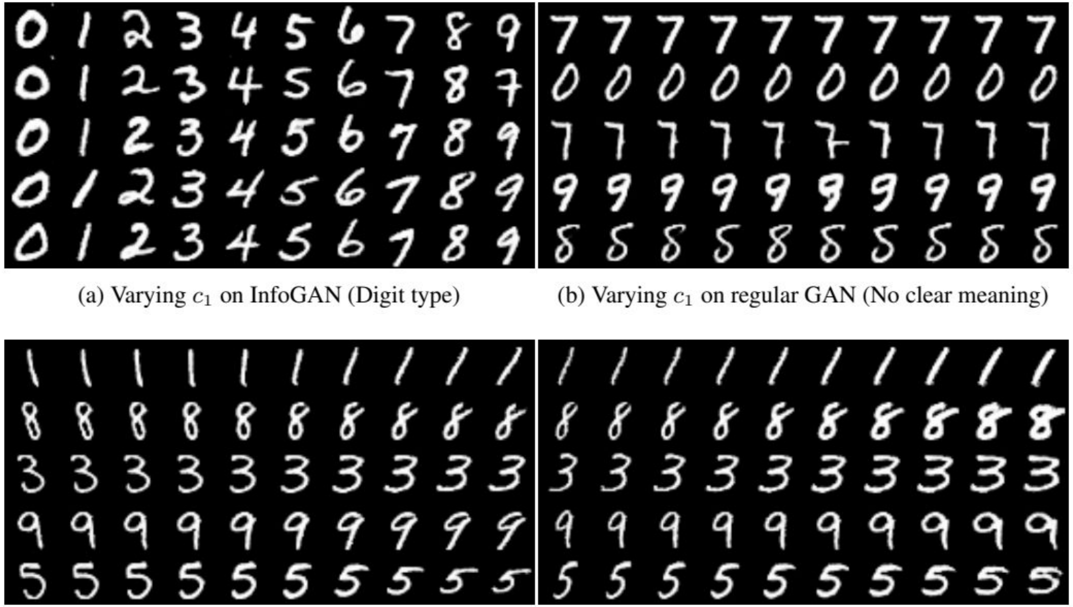
Entropy로 표현한다면, Y를 앎으로써 줄어드는 X의 불확실성: H(X) - H(X|Y)

X를 앎으로써 줄어드는 Y의 불확실성: H(Y) - H(Y|X)

$$\min_G \max_D V_I(D,G) = V(D,G) - \boxed{\lambda I(c;G(z,c))}$$

http://jaejunyoo.blogspot.com/2017/03/infogan-1.html
https://arxiv.org/pdf/1606.03657.pdf

Disentanglement Learning - InfoGAN



(a) Varying c_1 on InfoGAN (Digit type)

(b) Varying c_1 on regular GAN (No clear meaning)

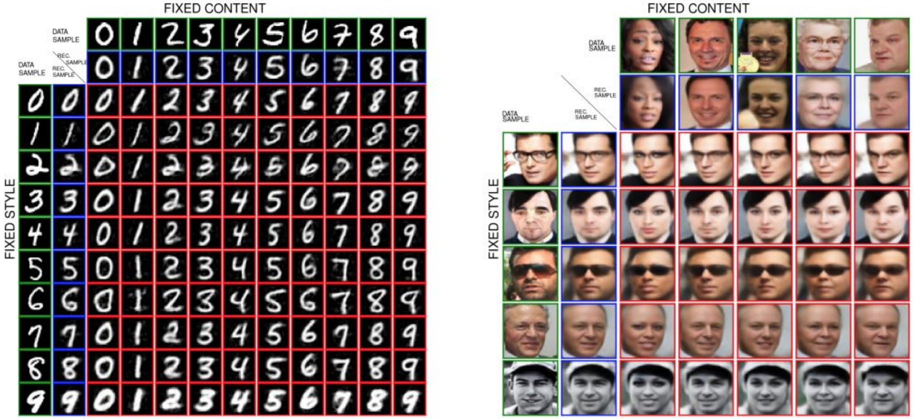
(c) Varying c_2 from -2 to 2 on InfoGAN (Rotation)

(d) Varying c_3 from -2 to 2 on InfoGAN (Width)

https://arxiv.org/pdf/1606.03657.pdf

Disentanglement Learning - with supervision

- ML-VAE
- DNA-GAN
- Semi-supervised VAE
- …….



(a) *MNIST, test dataset.* (b) *MS-Celeb-1M, test dataset.*

Figure 4: *Swapping, first row and first column are test data samples (green boxes), second row and column are reconstructed samples (blue boxes) and the rest are swapped reconstructed samples (red boxes). Each row is fixed style and each column is a fixed content. Best viewed in color on screen.*

Supervision의 사용이 Disentanglement learning에 어떤 영향을 끼치는지는 알려져 있지 않음.

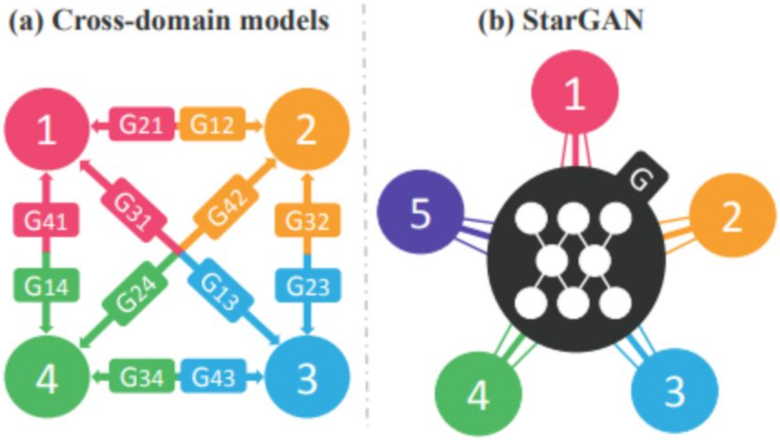
conditional GANs - cGAN

ex) cGAN, AC-GAN,
Projection Discriminator

Conditional GANs의 목표 : 사실적인 영상 만들기
Semi-StyleGAN의 목표 : 1. disentangled representation learning
2. controllable generation

conditional GANs - image-to-image translation

Pair 를 가지고 있는 학습 데이터를 사용하여 input-output image를 mapping 하는 것이 목표



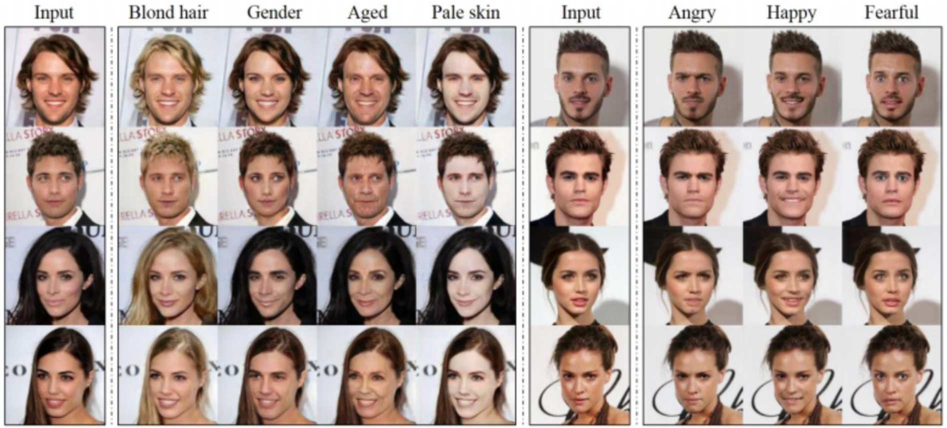
- 하나의 generator G로 다수의 domain 사이의 mapping을 학습
- 각 dataset에 부분적으로만 존재하는 label을 처리하기 위해

$$\tilde{c} = [c_1, ..., c_n, m],$$

- 데이터가 어느 데이터셋에 포함되어있는지에 대한 정보 m을 추가하여, StarGAN이 명시되지 않은 label에 대해서는 무시하게하고
  명시된 label에대해서는 집중하게 해준다.

https://hichoe95.tistory.com/39

conditional GANs - image-to-image translation



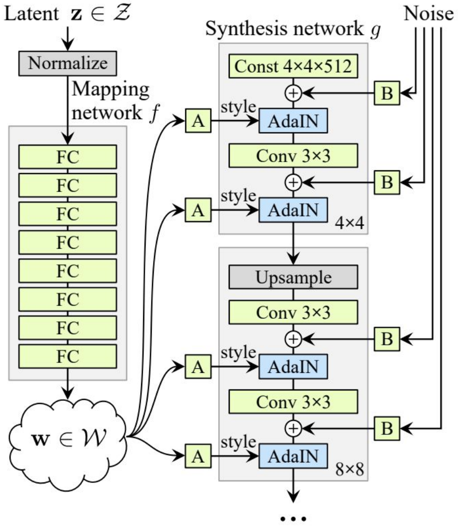
Attribute를 잘 학습하지만, binary attribute control만 가능하다. (점진적인 변화 불가)

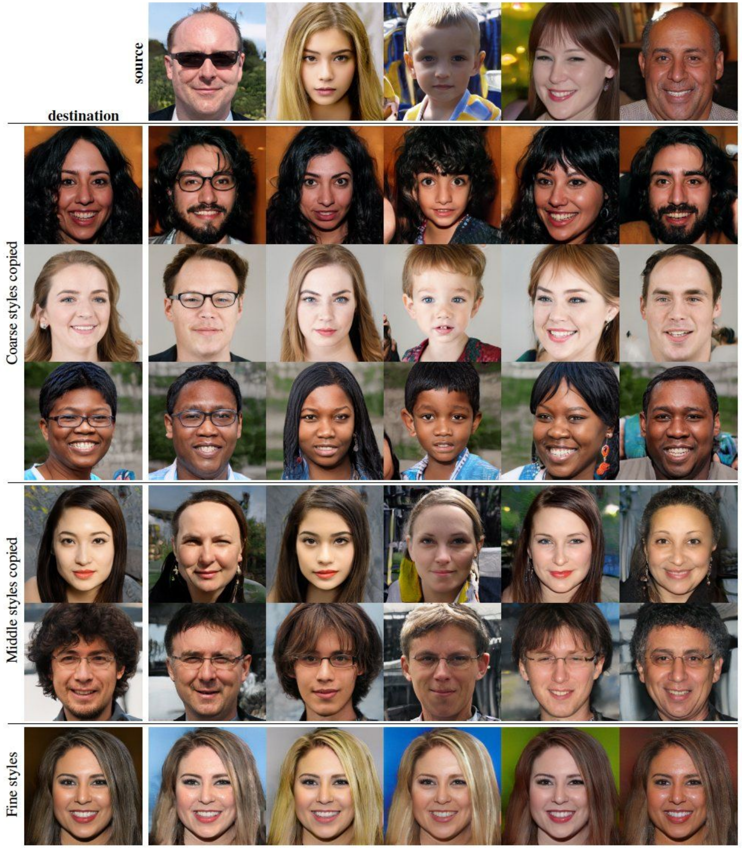
Limitations of the current disentanglement

1. **Non-Identifiability**
   - Unsupervised disentanglement method 에만 집중.
   - 모델이 배운 factor를 확인하기 위해 human feedback이 필요함.
   - 소량의 레이블을 추가하여 해결 가능.
2. **Low-resolutions**
   - Complex domain에서 reality가 떨어지는 현상
   - Factor가 데이터셋에 소수로만 존재하거나 불균형할 때, 잘 분해되지 않음.
3. **Disentangled representations**
   - Reresentations의 분해에만 초점을 주었음.
   - Disentangled encoder가 항상 disentangled generator는 아니다.
   - Generator의 disentanglement quality를 정량적 측정이 필요하다.

Info-StyleGAN

mutual information loss + StyleGAN

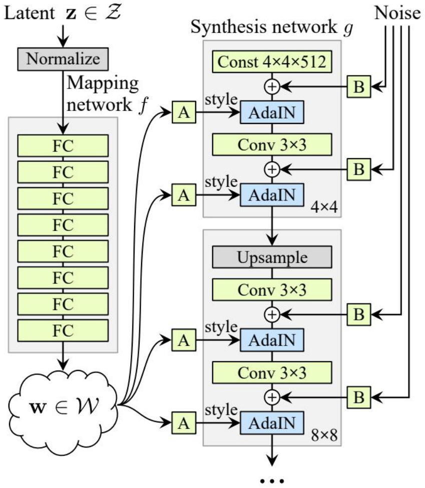




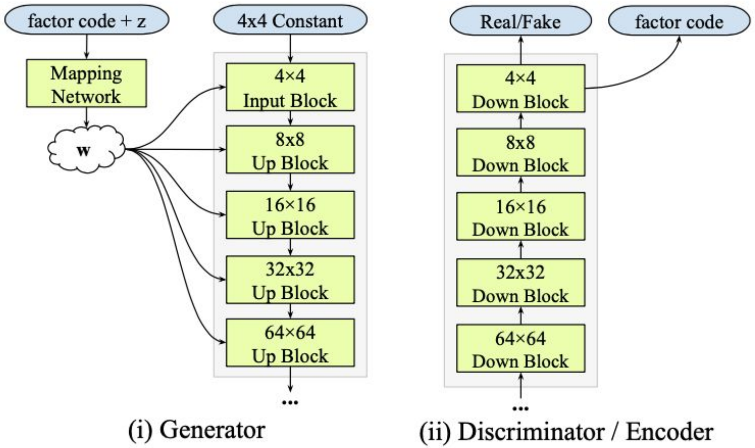
https://arxiv.org/abs/1812.04948

Info-StyleGAN

mutual information loss + StyleGAN



"conditional styles"



(i) Generator

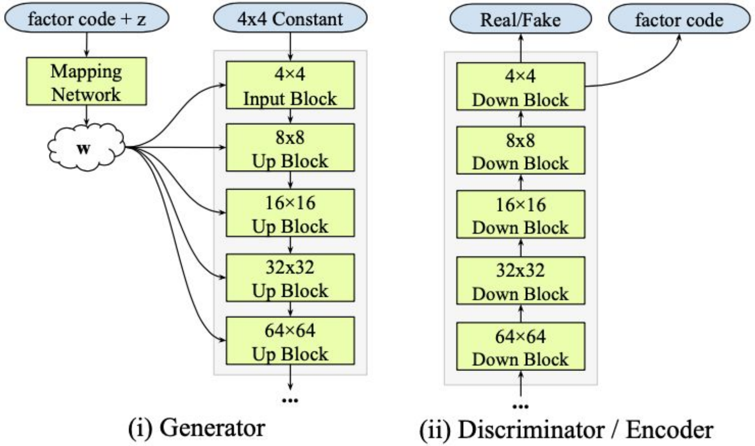
(ii) Discriminator / Encoder

https://arxiv.org/abs/1812.04948

Info-StyleGAN

* InfoGAN loss function

$$\min_G \max_D V_I(D,G) = V(D,G) - \boxed{\lambda I(c; G(z,c))}$$



(i) Generator (ii) Discriminator / Encoder

Info-StyleGAN에서의 mutual information loss는 "Unsupervised code reconstruction loss"

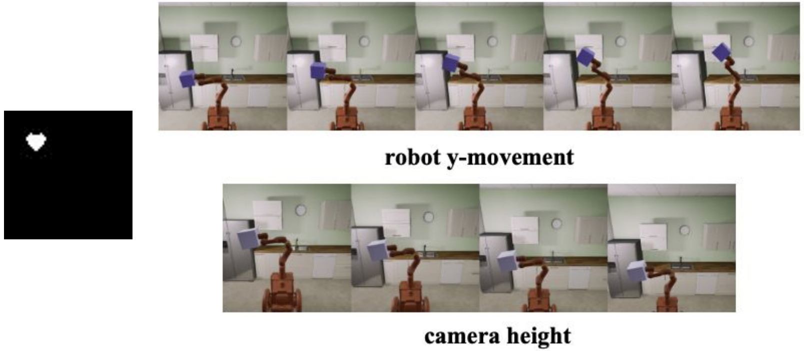
$$\mathcal{L}_{\text{unsup}} = \sum_{c\sim\mathcal{C}, z\sim p_z} \|E(G(c,z)) - c\|_2 \qquad (1)$$

$$\begin{aligned}\mathcal{L}^{(G)} &= \mathcal{L}_{\text{GAN}} + \gamma\mathcal{L}_{\text{unsup}} \\ \mathcal{L}^{(D,E)} &= -\mathcal{L}_{\text{GAN}} + \gamma\mathcal{L}_{\text{unsup}}\end{aligned} \qquad (2)$$

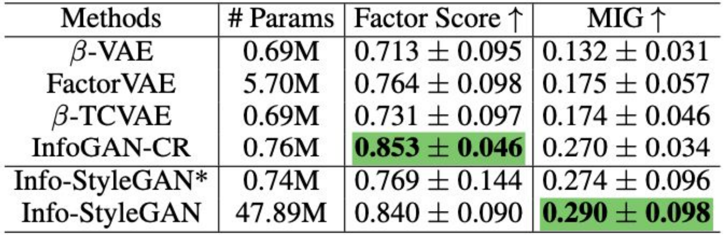
hyperparameter **γ**는 image-realism과 disentanglement quality의 trade-off를 조절한다.

Factor Score:
Mutual Information Gap (MIG): disentanglement quality
Frechet Inception Distance (FID): image quality

Why StyleGAN for disentanglement learning?

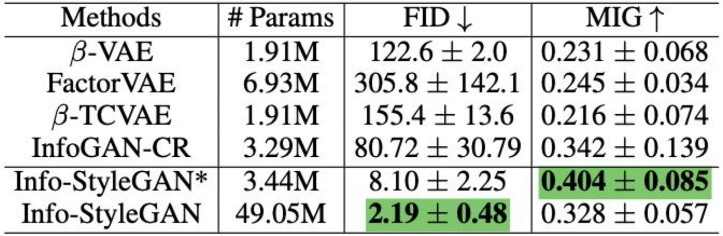


<dSprites & Isaac3D dataset>



Methods	# Params	Factor Score ↑	MIG ↑
β-VAE	0.69M	0.713 ± 0.095	0.132 ± 0.031
FactorVAE	5.70M	0.764 ± 0.098	0.175 ± 0.057
β-TCVAE	0.69M	0.731 ± 0.097	0.174 ± 0.046
InfoGAN-CR	0.76M	**0.853 ± 0.046**	0.270 ± 0.034
Info-StyleGAN*	0.74M	0.769 ± 0.144	0.274 ± 0.096
Info-StyleGAN	47.89M	0.840 ± 0.090	**0.290 ± 0.098**

(a) dSprites with resolution 64x64



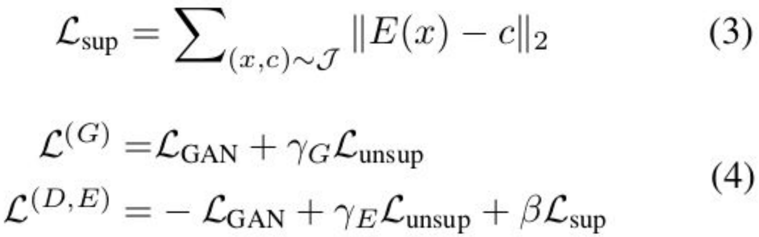
Methods	# Params	FID ↓	MIG ↑
β-VAE	1.91M	122.6 ± 2.0	0.231 ± 0.068
FactorVAE	6.93M	305.8 ± 142.1	0.245 ± 0.034
β-TCVAE	1.91M	155.4 ± 13.6	0.216 ± 0.074
InfoGAN-CR	3.29M	80.72 ± 30.79	0.342 ± 0.139
Info-StyleGAN*	3.44M	8.10 ± 2.25	**0.404 ± 0.085**
Info-StyleGAN	49.05M	**2.19 ± 0.48**	0.328 ± 0.057

(b) Isaac3D with resolution 128x128

Semi-StyleGAN

unsupervised disentangled method는 non-identifiable하다.

1. Naive way
   소량의 샘플에 대해 supervised code reconstruction term을 추가한다.



$$\mathcal{L}_{\text{sup}} = \sum_{(x,c)\sim\mathcal{J}} \|E(x) - c\|_2 \qquad (3)$$

$$\begin{aligned}\mathcal{L}^{(G)} &= \mathcal{L}_{\text{GAN}} + \gamma_G \mathcal{L}_{\text{unsup}} \\ \mathcal{L}^{(D,E)} &= -\mathcal{L}_{\text{GAN}} + \gamma_E \mathcal{L}_{\text{unsup}} + \beta \mathcal{L}_{\text{sup}}\end{aligned} \qquad (4)$$

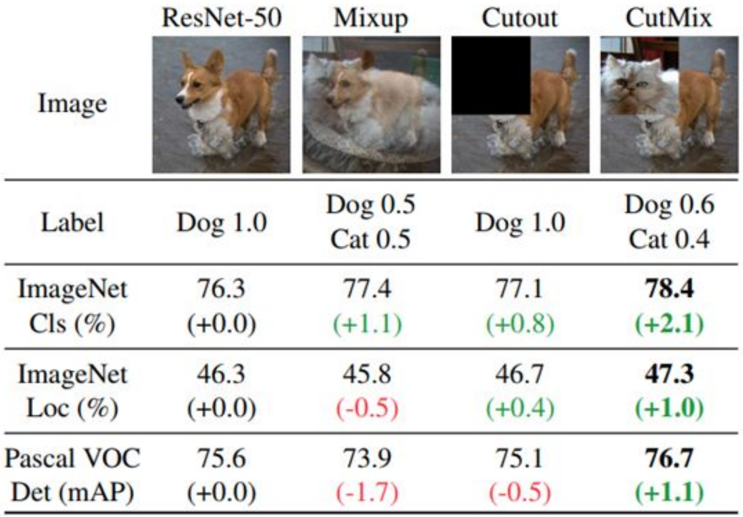
β는 supervised term의 가중치, γG와 γE는 encoder/generator의 disentanglement trade-off에 영향
* supervised term은 Generator에는 직접적인 영향을 끼치지 않는다.

Semi-StyleGAN

semi-supervised learning에서는
consistency regularization을 사용하지만,
disentanglement learning에는 적용하기 어려움.

반면에, latent space에 smoothness 적용하면
disentanglement에 도움이 된다는 논문(**MixUp**)이
있음.

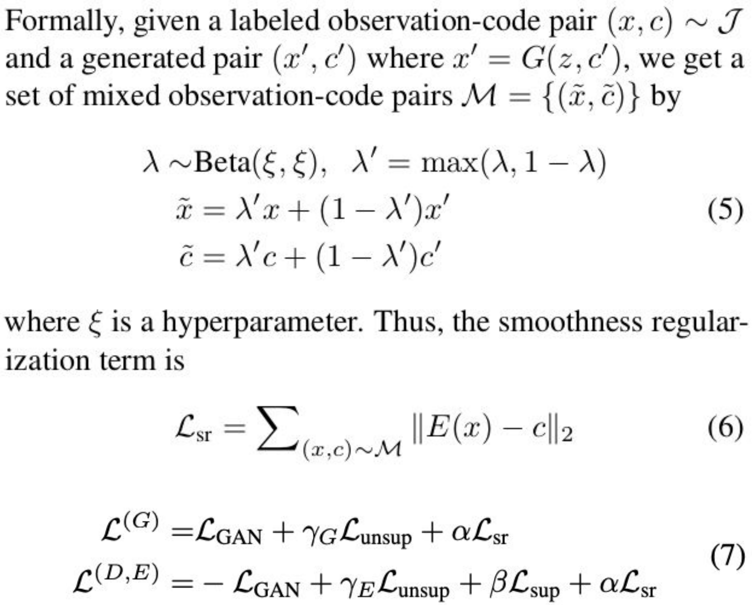
GAN의 학습 분포가
"feature vector의 linear interpolation가
그와 관련있는 타겟 이미지의 linear interpolation을
야기" 한다는 prior knowledge를 포함하도록 하기 때문



	ResNet-50	Mixup	Cutout	CutMix
Image				
Label	Dog 1.0	Dog 0.5 Cat 0.5	Dog 1.0	Dog 0.6 Cat 0.4
ImageNet Cls (%)	76.3 (+0.0)	77.4 (+1.1)	77.1 (+0.8)	**78.4** **(+2.1)**
ImageNet Loc (%)	46.3 (+0.0)	45.8 (-0.5)	46.7 (+0.4)	**47.3** **(+1.0)**
Pascal VOC Det (mAP)	75.6 (+0.0)	73.9 (-1.7)	75.1 (-0.5)	**76.7** **(+1.1)**

https://arxiv.org/abs/1905.04899

Semi-StyleGAN



Formally, given a labeled observation-code pair $(x, c) \sim \mathcal{J}$ and a generated pair (x', c') where $x' = G(z, c')$, we get a set of mixed observation-code pairs $\mathcal{M} = \{(\tilde{x}, \tilde{c})\}$ by

$$\begin{aligned} \lambda \sim & \text{Beta}(\xi, \xi), \quad \lambda' = \max(\lambda, 1 - \lambda) \\ \tilde{x} &= \lambda' x + (1 - \lambda') x' \\ \tilde{c} &= \lambda' c + (1 - \lambda') c' \end{aligned} \tag{5}$$

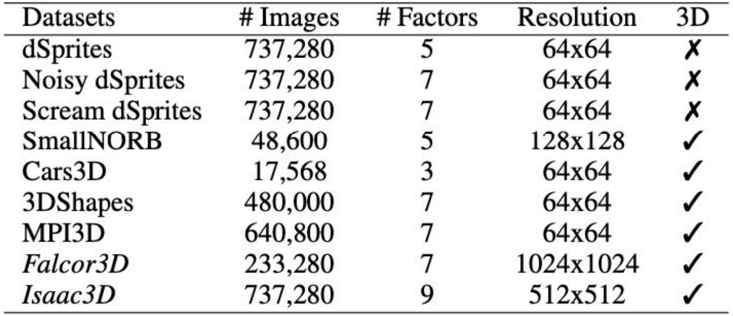
where ξ is a hyperparameter. Thus, the smoothness regularization term is

$$\mathcal{L}_{\text{sr}} = \sum\nolimits_{(x,c)\sim\mathcal{M}} \|E(x) - c\|_2 \tag{6}$$

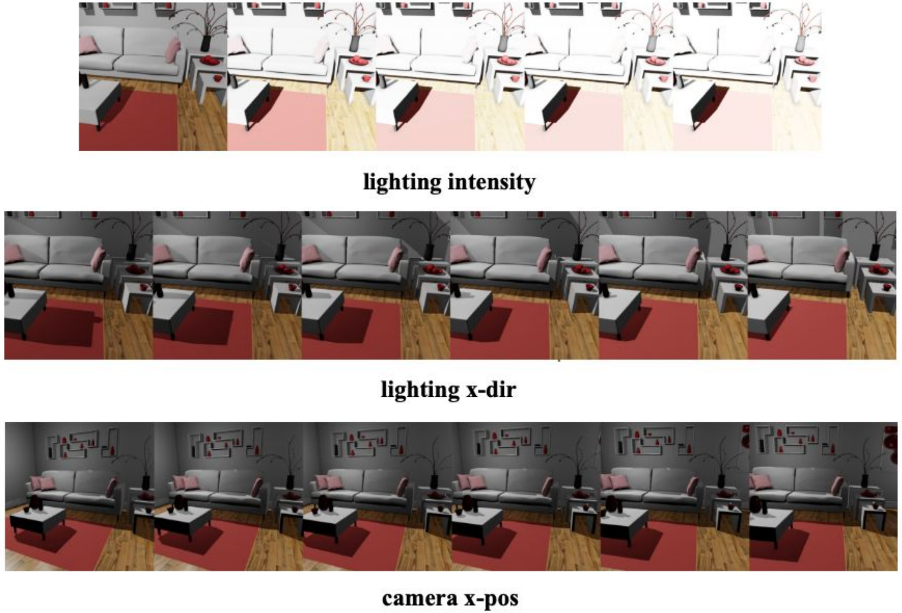
$$\begin{aligned} \mathcal{L}^{(G)} =& \mathcal{L}_{\text{GAN}} + \gamma_G \mathcal{L}_{\text{unsup}} + \alpha \mathcal{L}_{\text{sr}} \\ \mathcal{L}^{(D,E)} =& - \mathcal{L}_{\text{GAN}} + \gamma_E \mathcal{L}_{\text{unsup}} + \beta \mathcal{L}_{\text{sup}} + \alpha \mathcal{L}_{\text{sr}} \end{aligned} \tag{7}$$

New Datasets

Falcor3D & Isaac3D



Datasets	# Images	# Factors	Resolution	3D
dSprites	737,280	5	64x64	✗
Noisy dSprites	737,280	7	64x64	✗
Scream dSprites	737,280	7	64x64	✗
SmallNORB	48,600	5	128x128	✓
Cars3D	17,568	3	64x64	✓
3DShapes	480,000	7	64x64	✓
MPI3D	640,800	7	64x64	✓
Falcor3D	233,280	7	1024x1024	✓
Isaac3D	737,280	9	512x512	✓



lighting intensity

lighting x-dir

camera x-pos

<Falcor3D dataset>

New Metrics

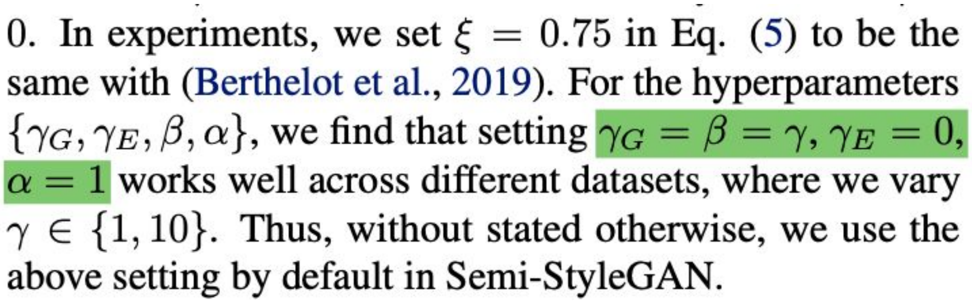
이전의 unsupervised disentanglement method metric들은

1. non-identifiable하다.
   => L2 score 사용
2. encoder의 disentanglement만 측정하고 generator의 controlablity는 무시한다.
   => Oracle Encoder (factor code를 정확하게 predict 할 수 있는 encoder)

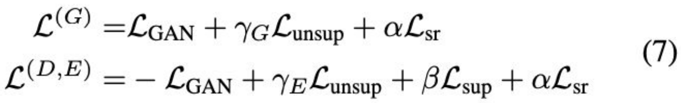
$$\text{MIG-gen} = \frac{1}{NK}\sum_{n=0}^{N-1}\sum_{k=0}^{K-1}\frac{1}{H(\hat{c}_k^{(n)})}\cdot \left(I(\hat{c}'^{(n)}_{j_k}; c'^{(n)}_k) - \max_{j\neq j_k} I(\hat{c}'^{(n)}_j; c'^{(n)}_k)\right)$$

$$\text{L2-gen} = \frac{1}{N}\sum_{n=0}^{N-1}\|E_{\text{oracle}}(x'^{(n)}) - c'^{(n)}\|_2$$

Experimental Protocols

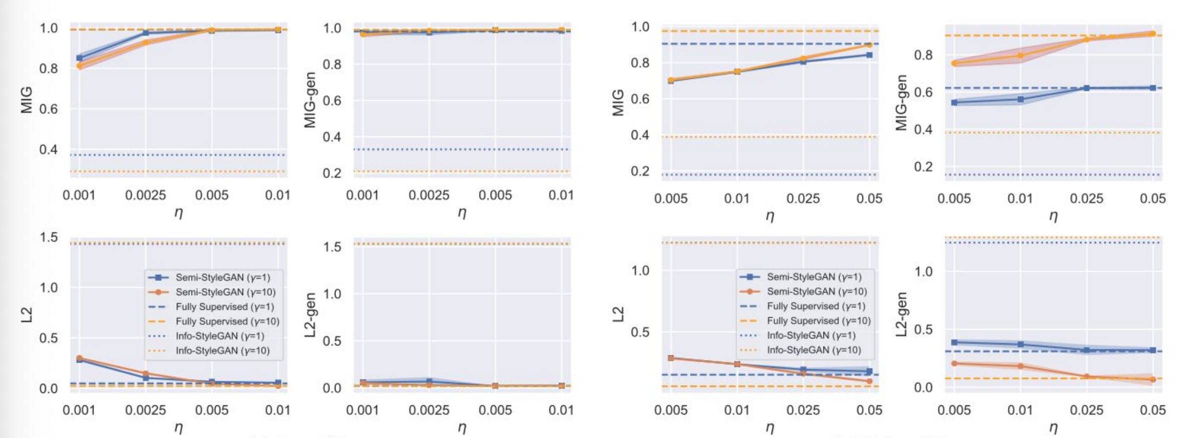


0. In experiments, we set $\xi = 0.75$ in Eq. (5) to be the same with (Berthelot et al., 2019). For the hyperparameters $\{\gamma_G, \gamma_E, \beta, \alpha\}$, we find that setting $\gamma_G = \beta = \gamma$, $\gamma_E = 0$, $\alpha = 1$ works well across different datasets, where we vary $\gamma \in \{1, 10\}$. Thus, without stated otherwise, we use the above setting by default in Semi-StyleGAN.



$$\begin{aligned} \mathcal{L}^{(G)} &= \mathcal{L}_{\text{GAN}} + \gamma_G \mathcal{L}_{\text{unsup}} + \alpha \mathcal{L}_{\text{sr}} \\ \mathcal{L}^{(D,E)} &= -\,\mathcal{L}_{\text{GAN}} + \gamma_E \mathcal{L}_{\text{unsup}} + \beta \mathcal{L}_{\text{sup}} + \alpha \mathcal{L}_{\text{sr}} \end{aligned} \tag{7}$$

Key Results

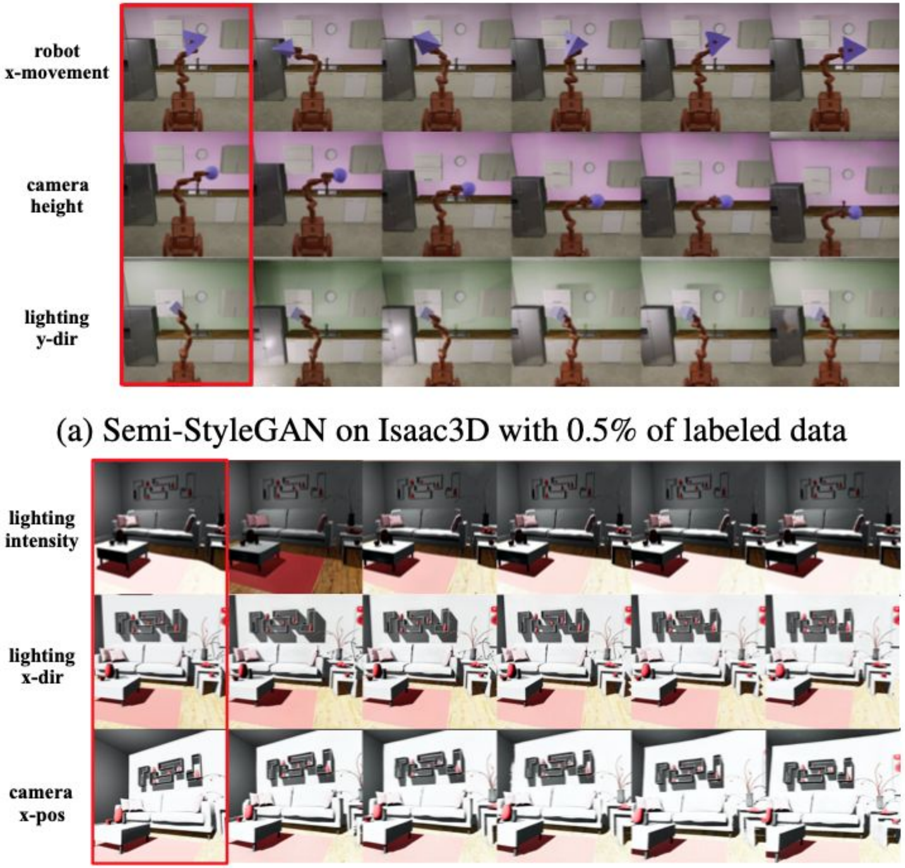


(a) Isaac3D

(b) Falcor3D

Figure 2. Semi-StyleGAN with the default setting $\gamma_G = \beta = \gamma$, $\gamma_E = 0$, $\alpha = 1$ where $\gamma \in \{1, 10\}$ on (a) Isaac3D and (b) Falcor3D. We vary the portion of labeled data η to show the impact of semi-supervision by comparing with Info-StyleGAN (i.e. the unsupervised baseline), and the fully-supervised one ($\eta = 1$). Only using 0.25∼2.5% of labeled data achieves near fully-supervised disentanglement.

- Isaac3D는 0.25%, Falcor3D는 2.5%만 사용해도 Fully supervised와 비슷한 결과를 보임.
- γ 선택에 대해 Generator disentanglement가 Encoder disentanglement보다 더 민감했다.



(a) Semi-StyleGAN on Isaac3D with 0.5% of labeled data

(b) Semi-StyleGAN on Falcor3D with 1% of labeled data

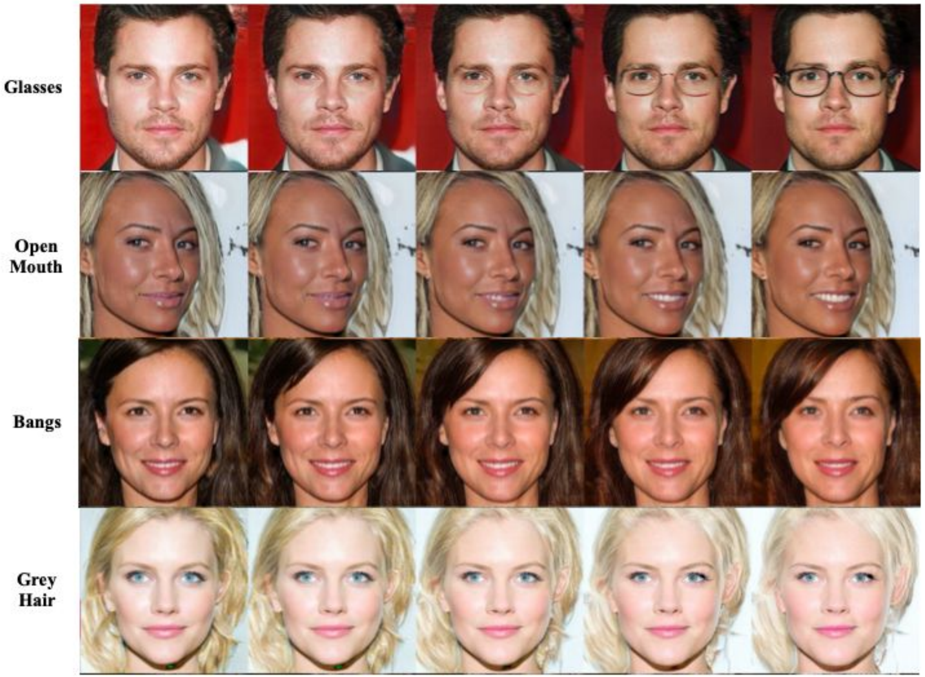
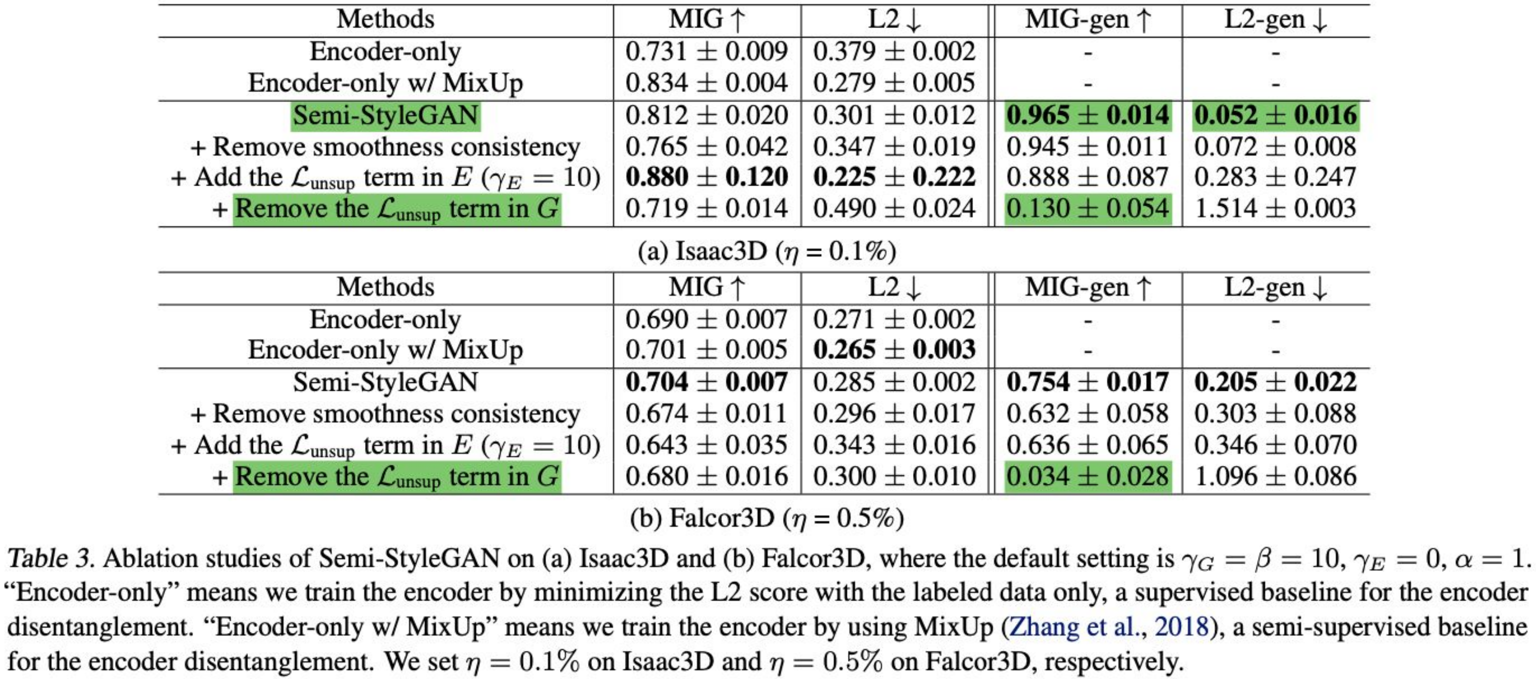


Figure 4. Latent traversal of Semi-StyleGAN on CelebA with resolution 256x256 by using 0.5% of the labeled data, where we use $\gamma = 1$ and disentangle all 40 binary attributes. See Appendix C.3 for the results of other attributes.



Methods	MIG ↑	L2 ↓	MIG-gen ↑	L2-gen ↓
Encoder-only	0.731 ± 0.009	0.379 ± 0.002	-	-
Encoder-only w/ MixUp	0.834 ± 0.004	0.279 ± 0.005	-	-
Semi-StyleGAN	0.812 ± 0.020	0.301 ± 0.012	**0.965 ± 0.014**	**0.052 ± 0.016**
+ Remove smoothness consistency	0.765 ± 0.042	0.347 ± 0.019	0.945 ± 0.011	0.072 ± 0.008
+ Add the $\mathcal{L}_{\text{unsup}}$ term in E ($\gamma_E = 10$)	**0.880 ± 0.120**	**0.225 ± 0.222**	0.888 ± 0.087	0.283 ± 0.247
+ Remove the $\mathcal{L}_{\text{unsup}}$ term in G	0.719 ± 0.014	0.490 ± 0.024	0.130 ± 0.054	1.514 ± 0.003

(a) Isaac3D (η = 0.1%)

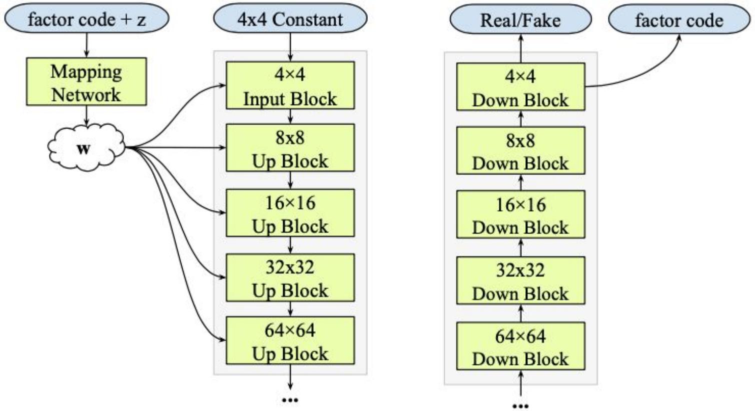
Methods	MIG ↑	L2 ↓	MIG-gen ↑	L2-gen ↓
Encoder-only	0.690 ± 0.007	0.271 ± 0.002	-	-
Encoder-only w/ MixUp	0.701 ± 0.005	**0.265 ± 0.003**	-	-
Semi-StyleGAN	**0.704 ± 0.007**	0.285 ± 0.002	**0.754 ± 0.017**	**0.205 ± 0.022**
+ Remove smoothness consistency	0.674 ± 0.011	0.296 ± 0.017	0.632 ± 0.058	0.303 ± 0.088
+ Add the $\mathcal{L}_{\text{unsup}}$ term in E ($\gamma_E = 10$)	0.643 ± 0.035	0.343 ± 0.016	0.636 ± 0.065	0.346 ± 0.070
+ Remove the $\mathcal{L}_{\text{unsup}}$ term in G	0.680 ± 0.016	0.300 ± 0.010	0.034 ± 0.028	1.096 ± 0.086

(b) Falcor3D (η = 0.5%)

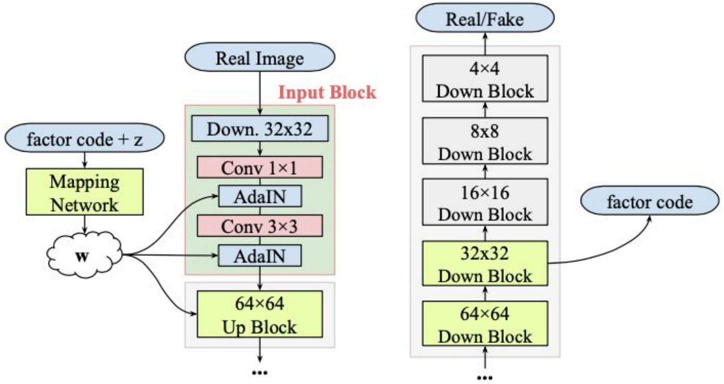
Table 3. Ablation studies of Semi-StyleGAN on (a) Isaac3D and (b) Falcor3D, where the default setting is $\gamma_G = \beta = 10$, $\gamma_E = 0$, $\alpha = 1$. “Encoder-only” means we train the encoder by minimizing the L2 score with the labeled data only, a supervised baseline for the encoder disentanglement. “Encoder-only w/ MixUp” means we train the encoder by using MixUp (Zhang et al., 2018), a semi-supervised baseline for the encoder disentanglement. We set $\eta = 0.1\%$ on Isaac3D and $\eta = 0.5\%$ on Falcor3D, respectively.

1. smoothness consistency(w/ and w/o MixUp)
2. encoder-generator disentanglement trade-off

Semi-Style-fine



(i) Generator (ii) Discriminator / Encoder

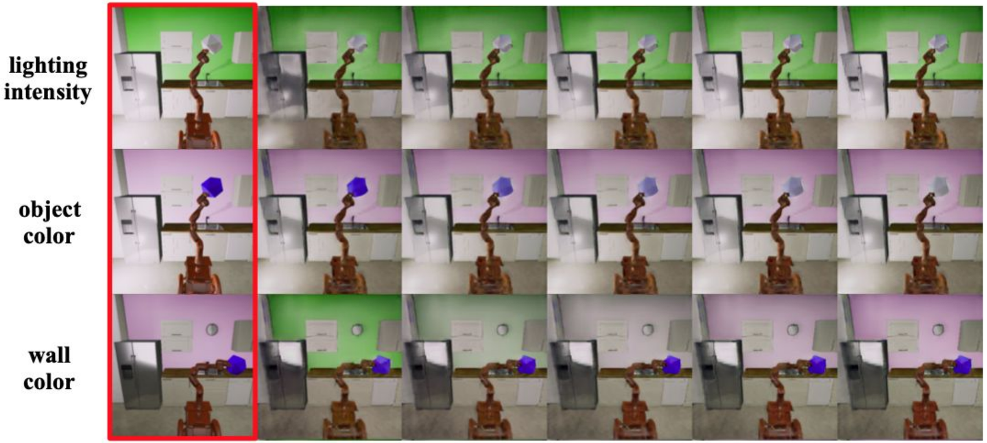


(i) Generator (ii) Discriminator / Encoder

Coarse한 특징이 바뀌어 생성되지는 않을 것이기 때문에, lower-resolution block을 빼버리고 32x32부터 학습 시작.
마지막 출력 블록이 아닌, 32x32 block에서 factor code를 prediction.

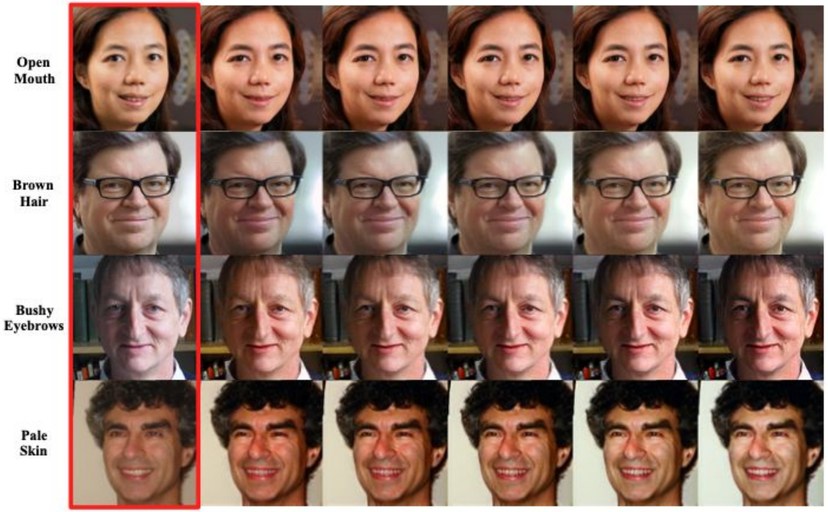
Key Results

D.1. Semi-StyleGAN-*fine* with 1% of Labeled Data on Isaac3D Novel Images with Resolution 512x512



- 새로운 로봇팔 위치와, object를 유지한 형태로 interpolation이 가능하다.

Key Results



- 1%만의 레이블을 사용하여, Unseen data에 대해서도 일반화가 잘 됨을 보이고, 더 많은 레이블의 사용이 성능 향상을 가져오지 않았다.

Contributions

- The impact of **limited supervision**
- New metrics to quantify **generator controllability**
- Crucial **trade-off** between
  disentangled representation learning and controllable generation