생성적 적대 네트워크를 활용한 텍스트와 스케치 기반 이미지 생성 기법

이제훈 O, 이동호 †
한양대학교 컴퓨터공학과
{eden1948, dhlee72}@hanyang.ac.kr

Image Generation based on Text and Sketch with Generative Adversarial Networks

Je-Hoon Lee O, Dong-Ho Lee †
Dept. of Computer Science and Engineering, Hanyang University

요 약

생성적 적대 네트워크를 활용하여 텍스트, 스케치 등 다양한 자원으로부터 이미지를 생성하기 위한 연구는 활발하게 진행되고 있으며 많은 실용적인 연구가 존재한다. 하지만 기존 연구들은 텍스트나 스케치 등 각 하나의 자원을 통해 이미지를 생성하기 때문에 설명이 부족한 텍스트, 실제 이미지와 상이한 스케치와 같이 자원의 정보가 불완전한 경우에는 제대로 된 이미지를 생성하지 못한다는 한계가 있다. 본 논문에서는 기존 연구의 한계점을 극복하기 위해 텍스트와 스케치 두 개의 자원을 동시에 활용하여 이미지를 생성하는 새로운 생성 기법 TS-GAN 을 제안한다. TS-GAN 은 두 단계로 이루어져 있으며 각 단계를 통해 더욱 사실적인 이미지를 생성한다. 본 논문에서 제안한 기법은 컴퓨터 비전 분야에서 많이 활용되는 CUB 데이터세트를 사용하여 이미지 생성 결과의 우수성을 보인다.

1. 서론

컴퓨터 비전 분야에서 이미지를 생성하기 위한 연구는 오래전부터 제기되어 온 연구과제이다. 최근, 생성적 적대 네트워크 (Generative Adversarial Networks: GAN) [1] 기법이 대두되면서 이미지 생성에 있어 더욱 정확한 결과를 보이고 있다. 조건부 GAN (Conditional GAN: CGAN) [2] 은 스케치와 텍스트 같은 주어진 조건부 자원을 통해 임의의 자원에서 이미지를 생성하는 GAN 을 응용한 기법이며, 이를 통한 다양한 적용 연구가 많이 진행되었다 [3, 4, 6, 8].

CGAN 을 활용한 위의 연구들은 텍스트나 스케치와 같은 하나의 조건부 자원만을 활용하기 때문에 조건부 자원의 정보가 부족할 경우 원치 않는 이미지가 생성된다는 한계점을 가진다. 가령, 생성하고자 하는 이미지에 대한 설명이 부족한 텍스트나 현저히 수준이 떨어지는 스케치를 통해 이미지를 생성할 경우 CGAN 은 조건부 자원에 대한 영향을 상대적으로 적게 받아 원치 않는 이미지를 생성할 가능성이 높아진다. 스케치 작업이 미숙한 사람들에게 생성하고자 하는 이미지와 유사하게 스케치를 작업하는 것은 상당히 어려운 일이다. 또한 모든 객체에 대해 상세하게 텍스트를 작성하는 것도 매우 성가신 일이다.

이러한 문제를 극복하기 위하여 본 논문에서는 단일 자원이 아닌 스케치와 생성하고자 하는 이미지를 설명하는 텍스트를 조건부 자원으로 받아 이미지를 생성하는 새로운 GAN 기법인 TS-GAN 을 제시한다. 본 기법은 스케치를 통해 이미지를 생성하는 첫 번째 단계와 텍스트를 통해 이미지를 생성하는 두 번째 단계로 이루어져있다. 첫 번째 단계에서는 주어진 이미지 스케치를 토대로 이미지에 대한 주요 객체의 형태와 전반적인 배경을 저해상도로 생성한다. 두 번째 단계에서는 첫 번째 단계에서 저해상도로 생성된 이미지와 이미지를 설명하는 텍스트를 결합한다. 그 후 텍스트 내용에 기반하여 형태가 형성된 객체는 더욱 사실적인 이미지로 재 생성한다. 두 번째 단계를 통과한 이미지는 첫 번째 단계보다 더 높은 해상도를 가진 이미지를 출력한다.

TS-GAN 은 단일 자원을 활용한 기존 연구와 달리, 상대적으로 불완전한 두 자원의 상호작용을 통하여 이미지를 생성하는 새로운 기법이다. 이는 두 개의 자원을 활용하기 때문에 더욱 다양한 이미지를 생성할 수 있다. 예를 들어, 스케치 이미지를 고정한 후

† 교신 저자

* 이 논문은 2016년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(No. NRF-2016R1D1A1A09918271)

이미지를 설명하는 텍스트만 변경함으로써 같은 객체를 가지는 다양한 이미지를 생성할 수 있다.

2. 관련 연구

2.1. 다양한 조건부 GAN

전통적인 GAN 은 조건부 자원의 부재로 임의의 이미지를 생성한다. 이렇듯 규칙성이 없이 특징이 서로 얽혀 있어 해석 불가능한 공간을 조건부 자원을 통해 해석 용이한 특징 공간으로 통제하는 다양한 연구가 진행되었다 [2, 7]. 유사한 목적성을 띤 두 논문은 조건부 자원을 입력 받는 위치와 출력 형태에 따라 차이점을 보인다. 본 논문에서는 출력의 형태가 기존 GAN 과 동일하기 때문에 CGAN 의 기법을 활용하여 텍스트를 추가적인 조건부 자원으로 사용한다.

2.2. 이미지로부터 새로운 이미지 생성

Isola et al. 의 "pix2pix" 프레임워크는 쌍으로 존재하는 데이터세트내에서 입력 이미지를 출력 이미지에 매핑하는 프레임워크이다 [4]. 이미지를 조건부 자원으로 활용하였으며, "pix2pix"을 통해 생성된 이미지와 실제 이미지와의 차이 값(Manhattan distance)을 추가적으로 학습하는 기법을 사용했다. [4]에서는 흑백 이미지에서 컬러 이미지, 스케치 이미지에서 실제 이미지로 생성하는 등 다양한 실험을 통해 결과를 입증했다. Jun-Yan Zhu et al. 의 "CycleGAN" 프레임워크는 주기 일관성이라는 개념을 활용하여 쌍으로 존재하지 않는 데이터세트내에서 이미지를 변환하는 문제를 해결했다 [3]. "CycleGAN"은 단 방향의 매핑이 아니라 양 방향의 매핑으로 학습을 유도한다. 본 논문에서도 데이터 쌍 존재의 불확실성을 피하기 위해 "CycleGAN"의 주기 일관성 기법을 활용하여 스케치를 사실적인 이미지로 생성한다.

2.3. 텍스트로부터 새로운 이미지 생성

Scott Reed et al. 의 "GAN-INT-CLS" 와 Han Zhang et al. 의 "StackGAN" 은 텍스트를 조건부 자원으로 활용하여 이미지를 생성한 연구이다 [8, 6]. "StackGAN" 은 "GAN-INT-CLS"에서 생성하는 이미지가 64x64 의 저해상도 이미지라는 한계를 극복하고 더욱 사실적인 이미지를 생성하기 위해 텍스트로 이미지를 생성하는 작업을 반복 진행한다. 하지만 두 연구 모두 오직 텍스트 자원 하나만을 고려하기 때문에 설명이 불분명한 글의 경우 의도와 다른 이미지를 생성한다는 단점을 가지고 있다. 본 논문에서는 스케치 자원을 같이 고려하여 언급된 단점을 극복한다.

3. TSGAN: 스케치와 텍스트를 활용한 생성 모델

3.1. 생성적 적대 네트워크 (GAN)

GAN 은 두 개의 네트워크(생성자 네트워크: G, 판별자 네트워크: D)를 경쟁하도록 설계한 생성 모델이다. G 의 목표는 D 를 완벽하게 속이는 이미지를 생성하도록 네트워크를 최적화하는 것이며, D 의 목표는 G 에 의해 생성된 이미지의 진위 여부를 완벽하게 판별하도록 네트워크를 최적화하는 것이다. GAN 의 손실 함수는 다음과 같다:

$$\min_G \max_D \mathcal{L}(G,D) = \mathbb{E}_{x\sim p_{data}}[logD(x)] + \mathbb{E}_{z\sim p_z}[log\left(1-D\big(G(z)\big)\right)], \quad (1)$$

여기서 x는 실제 이미지의 데이터 분포 p_{data}에서 샘플링한 하나의 실제 이미지이며, z 는 임의의 정규 분포 p_z에서 샘플링한 노이즈 벡터이다.

CGAN 은 G 와 D 에 조건부 자원 c 를 추가한 확장된 GAN 이다. 변수 c 는 z 를 통제하는 역할을 한다. CGAN 의 손실 함수는 다음과 같다:

$$\min_G \max_D \mathcal{L}(G,D) = \mathbb{E}_{x\sim p_{data}}[logD(x,c)] + \mathbb{E}_{z\sim p_z}[log\left(1-D\big(G(z,c)\big)\right)], \quad (2)$$

3.2. TS--GAN 의 아키텍처

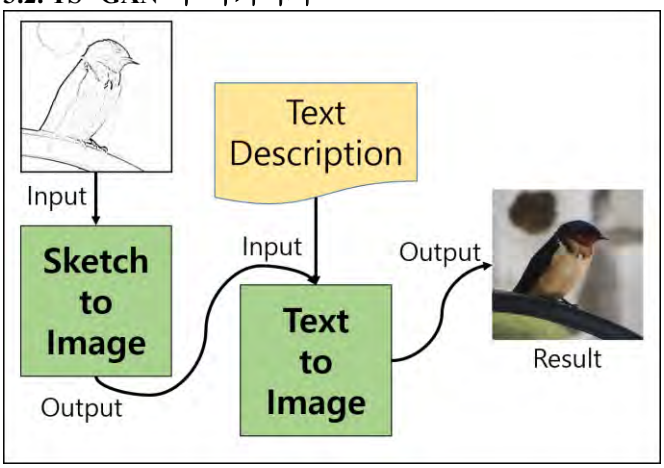

그림 1. TS-GAN 의 흐름도

본 논문에서 제안하는 TSGAN 아키텍처의 흐름도는 [그림 1]과 같다. TSGAN 은 스케치를 이미지로 생성하는 단계, 텍스트를 이미지로 합성하는 단계, 총 두 번의 단계로 이루어져있다. 먼저, 스케치를 통해 이미지를 생성한 후 텍스트를 통해 이미지를 재 생성한다. 각 단계는 서로 다른 네트워크로 구성되어 있으며, 네트워크의 세부 구성은 3.2.1. 과 3.2.2. 에서 설명한다. 각 단계의 역할은 아래와 같다:

첫 번째 단계: [그림 1]에서 Sketch-to-image 에 해당하는 단계이다. 첫 번째 단계에서는 주어진 스케치 이미지를 통해 주요 객체의 형태와 전반적인 배경을 저해상도로 생성한다.

두 번째 단계: [그림 2]에서 Text-to-image 에 해당하는 단계이다. 두 번째 단계에서는 주어진 텍스트와 첫 번째 단계를 통해 생성한 이미지를 활용한다. 첫 번째 단계에서 생성한 이미지는 객체의 형태가 어느 정도 형성된 이미지이다. 하지만 객체의 윤곽만 있을 뿐 사실적인 이미지와는 거리가 멀다. 이를 보완하기 위해 두 번째 단계에서는 주어진 텍스트를 활용하여 설명에 따라 객체의 세세한 형태와 색 그리고 배경을 더욱 사실적인 이미지로 재 생성한다. 두 번째 단계를 통해 생성된 이미지는 첫 번째 이미지에서 생성된 이미지에 비해 더욱 고해상도의 이미지를 생성한다.

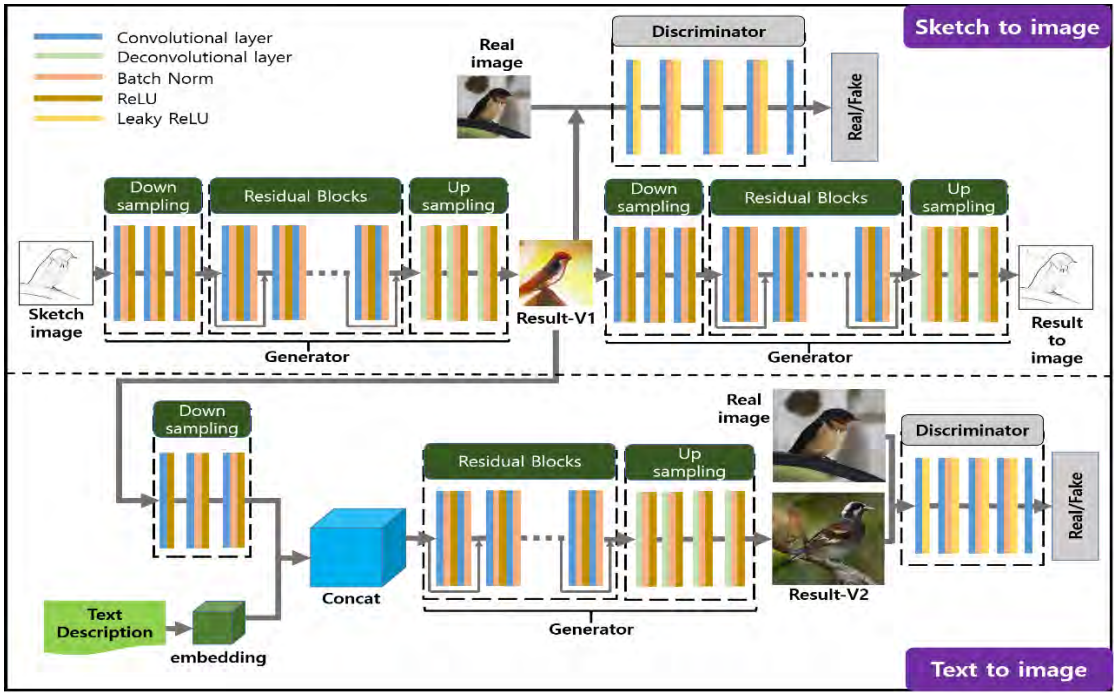

그림 2. Sketch-to-image 네트워크와 Text-to-image 네트워크의 아키텍처

3.2.1. 첫 번째 단계

첫 번째 단계에서는 "CycleGAN"의 아키텍처를 활용해 스케치 이미지를 사진 이미지로 생성하는 네트워크이다. 첫 번째 단계의 네트워크 상세 구조는 [그림 2]의 Sketch-to-image 부분이다. 네트워크의 G 는 첫 번째로 3 단계의 Convolutional 을 통해 스케치 이미지를 다운 샘플링하며, 9 개의 residual blocks [11]을 통해 스케치의 이미지의 특징을 더욱 자세하게 학습한다. 그 후 3 단계의 Deconvolutional 을 활용해 다운 샘플링된 이미지를 업 샘플링하여 저해상도의 이미지를 생성한다. D 의 구조는 다양한 실험을 통해 우수성이 증명된 "DCGAN" [12]의 구조를 사용한다.

스케치 이미지를 입력으로 받은 이미지는 G(G_{AB}, A 는 스케치 이미지, B 는 실제 이미지)에 의해 생성된 결과 이미지를 [그림 2]에서 "Result-V1"으로 표현했다. "Result-V1"은 같은 구조를 가진 다른 G(G_{BA})를 통해 다시 스케치 이미지로 재 생성한다. 주어진 스케치 이미지와 G_{BA}에 의해 재 생성된 스케치 이미지는 더욱 의미 있는 이미지를 생성하기 위해 "CycleGAN"에서 활용한 주기 일관성(cycle consistency) 손실 함수를 사용된다. 다음은 주기 일관성 손실 함수에 관한 표현식이다:

$$\mathcal{L}_{주기\ 일관성}(G_{AB}, G_{BA}) = \mathbb{E}_{A\sim p_{data(A)}}[\|G_{BA}(G_{AB}(A)) - A\|_1], \quad (3)$$

D 는 "Result-V1"와 실제 데이터세트에 존재하는 이미지와의 진위 여부 확인한다. 다음은 D 와 G 에서 발생하는 적대적인 손실 함수에 관한 표현식이다.

$$\mathcal{L}_{GAN}(G_{AB}, G_{BA}) = \mathbb{E}_{B\sim p_{data(B)}}[(D(B)-1)^2] + \mathbb{E}_{A\sim p_{data(A)}}[D(G_{AB}(A))^2], \quad (4)$$

표현식 (4)는 더욱 안정적인 모델의 학습을 위해 최소 자승 손실 함수 [10]를 적용하였다. 그리고, (3)번 표현식과 (4)번 표현식은 A→$\hat{B}$→$\hat{A}$ 형식의 손실 함수이다. 하지만 [3]에서 증명된 것과 같이 단 방향은 성능이 좋지 않다. 본 논문도 B→$\hat{A}$→$\hat{B}$ 로의 손실 함수도 함께 계산한다. 첫 번째 단계의 전체적인 손실 함수는 다음과 같다:

$$\mathcal{L}_{첫\ 번째\ 단계} = \mathcal{L}_{주기\ 일관성}(G_{AB}, G_{BA}) + \mathcal{L}_{주기\ 일관성}(G_{BA}, G_{AB}) + \mathcal{L}_{GAN}(G_{AB}, G_{BA}) + \mathcal{L}_{GAN}(G_{BA}, G_{AB}), \quad (5)$$

(5)의 식에 의해 학습된 네트워크는 스케치 이미지를 통해 64x64 의 저해상도의 이미지를 생성한다.

3.2.2. 두 번째 단계

두 번째 단계는 이미지를 설명하는 텍스트와 첫 번째 단계에서 생성한 이미지를 입력으로 받는다. 두 번째 단계의 네트워크 상세 구조는 [그림 2]의 Text-to-image 부분이다. 네트워크의 G 와 D 는 텍스트와 첫 번째 단계에서 생성된 이미지의 결과를 결합하는 단계와 residual blocks 계층 개수의 차이를 제외하곤 첫 번째 단계와 유사하다. 텍스트는 CGAN 의 조건부로 활용한 두 번째 단계의 손실 함수는 다음과 같다:

$$\mathcal{L}_{두\ 번째\ 단계}(G, D) = \mathbb{E}_{(R,t)\sim p_{data}} = [(D(R,\varphi_t)-1)^2] + \mathbb{E}_{z\sim p_z, t\sim p_{data}} = [D\big(G(G_{AB}(z),\varphi_t)\big)^2], \quad (6)$$

두 번째 단계의 손실 함수 또한 최소 자승 손실 함수 손실 함수를 사용했다. 여기서 R 은 실제 이미지이고, t 는 설명 글이며, φ_t는 설명 글 t 에 대한 임베딩 함수이다. 두 번째 단계를 통해 생성된 최종 이미지는 256x256 의 해상도이다.

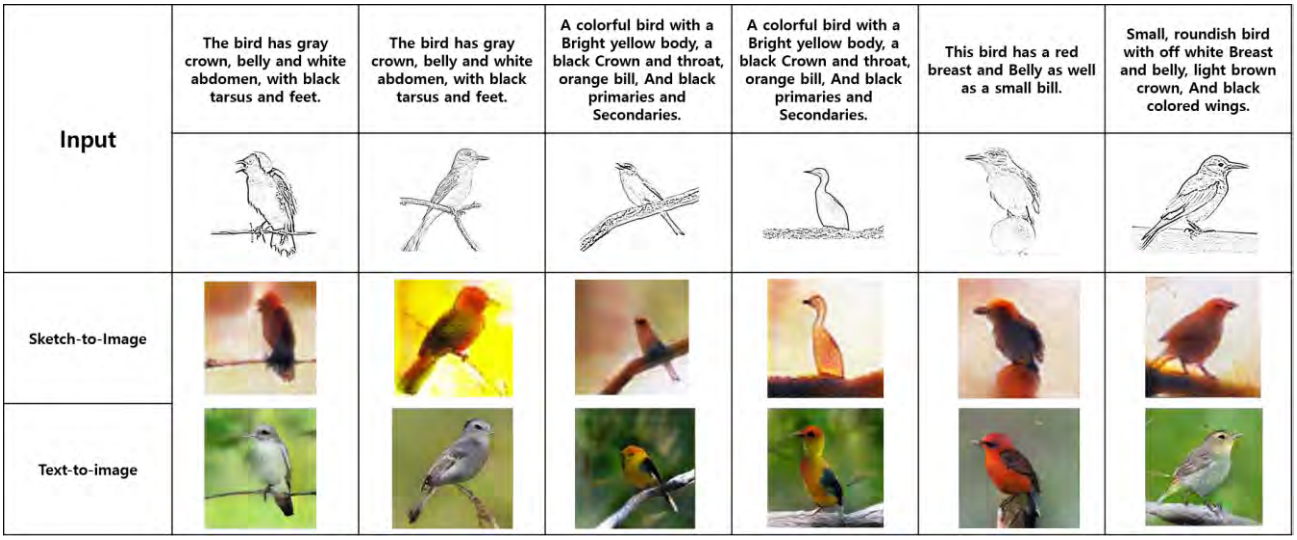

그림 3. TS-GAN 실험 결과

4. 실험 및 결과 분석

본 논문의 기법을 실험하기 위해 CUB [9] 데이터세트를 사용했다. CUB 데이터세트는 200 가지 종류의 새가 총 11,788 장으로 이루어져있다. 이 데이터세트의 대부분의 이미지는 객체의 이미지 크기가 작기 때문에 객체를 중심으로 이미지 사이즈를 조절하여 사용했다. 또한 포토샵을 이용해서 CUB 데이터세트의 스케치 이미지를 얻었다.

[그림 3]은 제안된 TS-GAN 을 통해 생성된 이미지의 결과이다. 입력의 1 번째 행은 이미지를 설명하는 텍스트이며, 두 번째 행은 스케치 이미지이다. 세 번째 행은 2 번째 행의 스케치가 TS-GAN 의 첫 번째 단계를 통해 생성된 이미지이다. 결과를 보면 객체의 세세한 부분은 생성하지 못하지만 스케치에 따라 객체의 전반적인 형태를 잘 생성하는 것을 확인할 수 있다. 그리고 객체를 제외한 이미지의 배경은 다소 의미 없는 이미지를 생성하였다. 이는 네트워크가 스케치 이미지의 선 특징을 중점으로 학습하면서 선으로 이루어진 부분을 최대한 집중해서 생성하려고 노력하는 것을 알 수 있다. 4 번째 행은 3 번째 행의 이미지와 생성하고자 하는 이미지에 대해 설명하는 텍스트(1 번째 행)가 TS-GAN 의 두 번째 단계를 통해 생성된 최종 생성 이미지이다. 결과를 보면 3 번째 행에서의 객체 형태를 유지하는 것을 확인할 수 있다. 그리고 텍스트에 따라 객체의 색과 세세한 형태를 재생성한다. 최종 생성 이미지 결과를 확인해보면 육안으로 봐도 충분히 의미 있는 결과를 생성했다고 할 수 있다. 또한, 2, 3 번째 열과 4, 5 번째 열은 각 다른 스케치에 대해 같은 텍스트를 통해 실험하였다. 결과적으로 형태는 스케치에 따라 다르지만 생성된 이미지의 전체적인 색감은 일치하는 것을 확인할 수 있다.

5. 결론

본 논문에서는 스케치 이미지와 이미지에 대한 설명 글을 통하여 이미지를 합성하는 새로운 기법 TS-GAN 을 제안했다. 본 기법은 두 부분의 아키텍처로 분리되며, 첫 번째 단계에서는 스케치 이미지를 통하여 주요 객체의 형태와 전체적인 이미지의 배경을 저해상도로 생성한다. 두 번째 단계에서는 첫 번째 단계의 결과 이미지와 이미지에 대한 설명 텍스트를 매칭하여 객체의 세세한 부분을 고해상도로 재 생성한다.

참고문헌

[1] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio. “Generative adversarial nets.” In NIPS, 2014.
[2] M. Mirza and S. Osindero. “Conditional generative adversarial nets.” arXiv:1411.1784, 2014.
[3] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. “Unpaired image-to-image translation using cycle-consistent adversarial networks.” In ICCV, 2017.
[4] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. “Image-to-image translation with conditional adversarial networks.” In CVPR, 2017.
[5] X. Huang, Y. Li, O. Poursaeed, J. Hopcroft, and S. Belongie. “Stacked generative adversarial networks.” In CVPR, 2017.
[6] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. Metaxas. “Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks.” In ICCV, 2017.
[7] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel. “Infogan: Interpretable representation learning by information maximizing generative adversarial nets.” In NIPS, 2016.
[8] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee. “Generative adversarial text-to-image synthesis.” In ICML, 2016.
[9] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. “The Caltech-UCSD Birds-200-2011 Dataset.” Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.
[10] X. Mao, Q. Li, H. Xie, R. Y. K. Lau, Z. Wang, and S. P. Smolley, “Least squares generative adversarial networks,” In ICCV, 2017.
[11] K. He, X. Zhang, S. Ren, and J. Sun. “Deep residual learning for image recognition.” In CVPR, 2016.
[12] A. Radford, L. Metz, and S. Chintala. “Unsupervised representation learning with deep convolutional generative adversarial networks.” arXiv preprint arXiv:1511.06434, 2015.