특집 03

# 생성 AI 보안 및 프라이버시 이슈

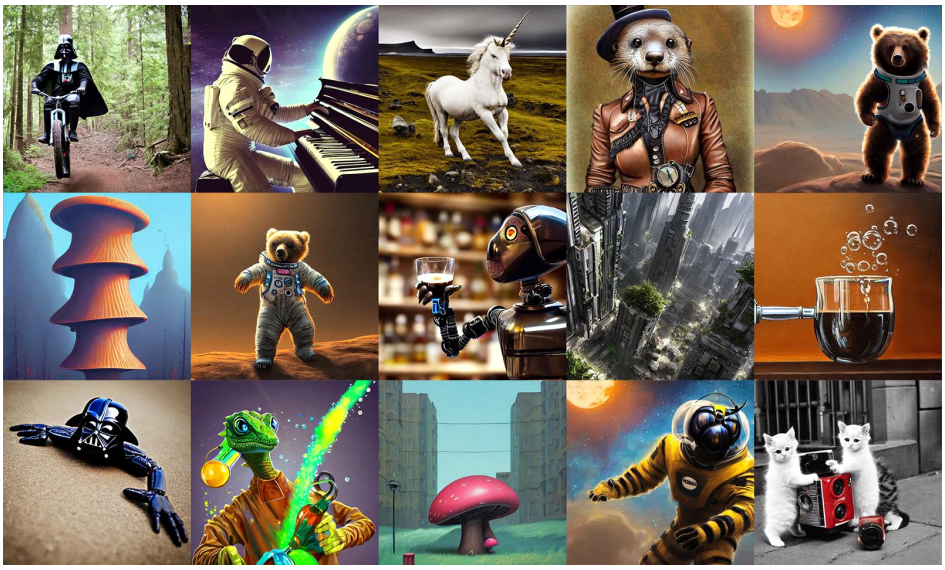
박대얼 · 류권상 · 최대선 (숭실대학교)

목 차



## 1. 서 론

2022년에 생성 AI의 비약적인 발전이 이루어지면서 텍스트, 이미지, 음성과 같은 다양한 분야에서 높은 성능을 보이는 생성 모델이 제안되었다. 텍스트 분야에서는 GPT[1]를 기반으로 하는 ChatGPT[1)]가 어떤 질문에도 수준 높은 답변을 구사하는 모습을 보여주고 이미지 분야에서는 기존에 많이 쓰이던 Generative Adversarial Network (GAN)[2] 구조 대신 Diffusion[3] 구조에 관한 연구가 진행되면서 이를 기반으로 하는 GLIDE[4], Stable Diffusion[5], Imagen[6]과 같은 텍스트 기반 이미지 생성 모델이 등장했다. 생성하고 싶은 이미지를 텍스트로 입력하면 (그림 1)과 같은 고



(그림 1) Stable Diffusion 생성 AI 모델이 생성한 이미지 예시

해상도 이미지를 생성하는데 GAN 모델보다 더 사실적이고 다양한 이미지를 생성할 수 있어서 큰 화제가 되고 있다. 하지만 생성 모델이 발전함에 따라 사용자가 생성 모델을 무분별하게 이용하면서 오남용, 저작권 침해, 프라이버시 침해와 같은 문제가 발생하고 있다. 생성 모델 자체에 대해서도 보안 공격, 프라이버시 문제가 연구되면서 생성 모델이 올바르게 쓰이지 못하거나 데이터 유출의 우려가 있는 것이 밝혀지고 있다. 따라서 생성 AI 모델에 대한 전반적인 이해와 보안 실태에 대해 파악하는 것이 중요하다. 본 논문은 2장에서 생성 AI 모델과 생성 AI 모델의 이슈를 알아보고 3장에서는 생성 AI의 보안과 프라이버시 이슈를 최신 공격 연구와 함께 정리하고 4장에서 결론을 말한다.

## 2. 생성 AI 모델 및 이슈

생성 AI 모델은 텍스트, 이미지, 음성 등을 입력으로 받아 새롭게 생성된 텍스트, 이미지, 음성 등을 출력으로 생성해내는 AI 모델을 말한다. 생성 AI 모델은 모델 구조에 따라 GAN, Variational Auto Encoder(VAE)[7], Transformer[8], Diffusion 모델로 나눌 수 있다. GAN은 적대적 생성 네트워크로 생성기가 데이터를 생성하면 판별기가 이 데이터가 진짜인지 가짜인지 판별한다. 이처럼 생성기와 판별기가 서로 적대적으로 학습하면서 생성기의 성능을 높이는 구조이다. VAE는 인코더가 입력 데이터를 잠재공간으로 표현하고 디코더가 이 잠재공간을 입력으로 변환하면서 디코더가 데이터를 생성하는 파라미터를 학습하는 구조다. Transformer는 Attention 블록으로 이루어지며 입력 데이터 내의 관계를 추정하면서 맥락과 의미를 학습하는 구조이고 Diffusion은 확산 프로세스를 거치면서 입력 데이터에 노이즈를 추가하고 역 프로세스를 거치면서 노이즈로부터 데이터를 복원하는 파라미터를 학습하면서 데이터를 생성하는 구조이다.

생성 모델이 화제가 되는 만큼 생성 모델에 대한 사회적 이슈도 많이 생겨나고 있다. 분류 모델은 결과물이 맞게 분류하거나 틀리게 분류하는 것으로 구분되는 반면에 생성 모델은 생성 결과에 대해 옳고 그름이 명확하지 않기 때문에 생성 모델이 만들어낸 결과를 어떻게 판단하고 사용해야 할지 사회적, 윤리적으로 많은 논의가 필요하다. 이번 장에서는 생성하는 데이터 분야별로 생성 모델의 종류와 실제로 있었던 이슈를 설명한다.

텍스트 생성 모델은 Transformer의 등장으로 많은 언어 모델이 제안되면서 많은 발전이 이루어졌다. 대표적인 텍스트 생성 모델은 단방향 언어 모델 GPT를 사용한 ChatGPT가 있다. OpenAI에서 만든 ChatGPT는 대화형 텍스트 생성 모델로서 사용자가 질문을 하면 그에 알맞은 답변을 해주는데 기존 챗봇과 달리 어떤 주제에 관해서 물어보아도 막힘없이 전문성 있는 답변을 해준다.

텍스트 생성 모델의 이슈로는 ChatGPT가 미국 로스쿨 시험, 의사면허 시험에 합격하면서 전문가 수준의 지식을 가진 점이 널리 알려지자 이를 남용하는 사례가 발생하고 있다[9]. 국내의 한 국제학교에서는 학생들이 ChatGPT를 사용해서 제출한 과제를 0점 처리한 일이 있으며 미국의 SF 잡지 클락스월드는 ChatGPT로 생성한 소설이 많아지자 글 제출 수락을 중단하는 일이 발생했다[10, 11]. ChatGPT는 폭력, 혐오 콘텐츠에 대해서는 검열을 통해 대답하지 않는 데 DAN(Do Anything Now) 명령어를 사용하면 콘텐츠 제한이 풀리는 것이 밝혀지면서 사용자가 민감한 질문을 해도 답하는 모습을 보여주기도 했다[12].

이미지 생성 모델은 입력된 텍스트를 기반으로 이미지를 만드는 Text-to-Image(T2I) 작업이나 입

1) https://chat.openai.com/chat

력된 이미지를 기반으로 새로운 이미지를 만드는 Image-to-Image(I2I) 작업을 수행할 수 있는 모델이다. 대표적인 이미지 생성 모델을 모델 구조별로 나누면 GAN을 사용한 모델로 이미지의 화풍을 바꿀 수 있는 CycleGAN[13], 얼굴 이미지를 바꿀 수 있는 StarGAN[14], StyleGAN[15] 등이 있고 이 모델들은 I2I 작업이 가능하다. VAE와 Transformer를 사용한 생성 모델은 Parti[16], DALL-E[17]가 있고, Diffusion 구조를 사용한 생성 모델에는 GLIDE, Stable Diffusion, Imagen, DALL-E 2[18]가 있고 I2I, T2I 작업이 모두 가능한 모델들이다.

이미지 생성 모델의 사회적 이슈에는 콜로라도 주 박람회의 연례 미술대회에서 Midjourney[2] 이미지 생성 모델로 만든 그림이 1등 상을 받아서 예술계에 파장을 일으킨 적이 있다[19]. 한 소셜미디어 이용자는 故 김정기 작가의 그림을 Stable Diffusion 모델로 학습해서 생성한 그림을 소셜미디어에 올리고 퍼갈 때 자신의 크레딧을 붙여달라고 하는 발언이 문제가 되면서 이미지 생성 모델로 생성한 이미지의 지적저작권 논란이 일었다[20].

음성 생성 모델은 텍스트를 음성으로 만드는 Text-to-Speech(TTS) 작업을 통해 음성을 생성하는 것이 일반적이다. 더 나아가 본 논문에서는 TTS 이외에 음성을 다른 음성으로 바꾸는 음성 합성, 음성 변환 작업도 음성을 생성하는 것으로 간주한다. GAN 구조를 사용한 음성 생성 모델에는 음성의 성별을 바꿀 수 있는 VoiceGAN[21], 드럼, 피아노, 새 울음소리 등 다양한 음성을 생성할 수 있는 WaveGAN[22]이 있다. Diffusion 구조를 사용한 생성 모델은 구글에서 만든 텍스트로 음악의 장르를 지정해서 만들 수 있는 AudioLM[23], Noise2Music[24] 모델, Stable Diffusion 모델을 파인튜닝해서 원하는 악기, 장르, 코드를 작성하면 스펙토그램 이미지를 생성해 노래를 만드는 Riffusion[3] 모델이 있다. 음성 생성 모델도 이미지 생성 모델이 I2I 모델에서 T2I 모델로 발전한 것과 비슷하게 텍스트를 그대로 읽는 음성을 만드는 모델에서 텍스트로 만들고 싶은 음성이나 음악을 지정하면 그에 맞게 만들어주는 모델로 발전하고 있다.

음성 생성 모델의 사회적 이슈로는 아마존의 AI 스피커 '알렉사'가 10세 소녀의 새로운 챌린지를 추천해달라는 질문에 휴대전화 충전기 콘센트에 동전을 갖다 대라는 위험한 답변을 해 논란이 된 적이 있다[25]. 아랍에미리트의 한 은행에서는 대기업 임원의 전화를 받고 한화 약 420억 원을 송금했다가 인공지능 음성 생성 모델로 만든 목소리를 사용한 보이스피싱인 것이 밝혀진 사례가 있다[26].

## 3. 생성 AI 보안 및 프라이버시 공격

인공지능 모델의 보안이나 프라이버시가 침해되면 인공지능 모델의 사용 목적에 맞게 사용될 수 없고 모델을 신뢰할 수 없게 된다. 생성 모델은 딥러닝을 기반으로 하므로 분류 모델과 같은 기존 인공지능 모델의 보안 취약점이나 프라이버시를 위협하는 공격에 노출될 수 있다. 생성 모델에 적용 가능한 보안 및 프라이버시 공격 방법을 정리하면 표 1과 같다. 이번 장에서는 생성 모델이 공격받았을 때 생길 수 있는 보안 및 프라이버시 이슈와 지금까지 진행된 관련 연구를 공격 방법별로 알아본다.

2) https://www.midjourney.com/

3) https://www.riffusion.com/

〈표 1〉 생성 모델 보안 및 프라이버시 공격 방법 정리

| 공격 단계 | 공격 방법 | 공격 목적 | 관련 연구 |
|---|---|---|---|
| 모델 학습 단계 | Poisoning | 악의적인 학습 데이터를 주입해 모델의 생성 결과물 품질 하락 | - |
| | Backdoor | 모델에 backdoor를 학습해 원본 입력 데이터는 정상적으로 실행되다가 트리거가 주입된 입력 데이터를 넣으면 공격자가 원하는 결과물 생성 | [27],[28],[29],[30],[32] |
| 모델 테스트 단계 | Model Inversion | 입력 데이터와 생성 결과를 분석해 학습 데이터 추출 | - |
| | Model Extraction | 입력 데이터와 생성 결과를 분석해 모델 구조와 가중치 추출 | [33] |
| | Membership Inference | 생성 결과물을 보고 특정 데이터가 학습 데이터로 사용되었는지 알아냄 | [36],[39],[40],[41],[42],[43] |

### 3.1 Poisoning 공격

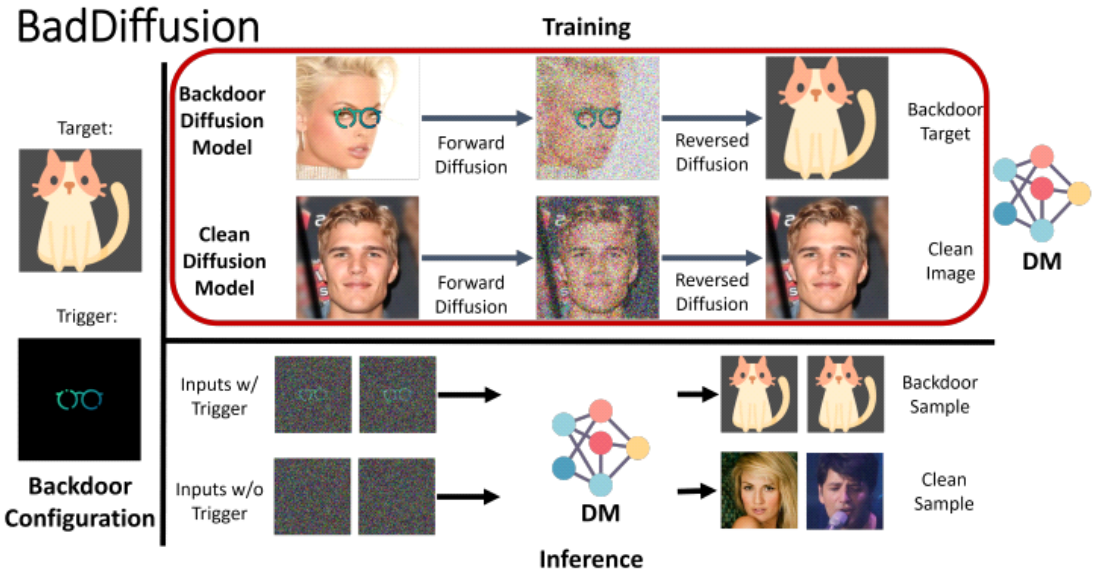
Poisoning 공격이 성공하면 생성 모델의 성능이 하락하고 원하지 않은 결과가 나올 수 있게 된다. 텍스트 생성 모델에는 비속어 데이터를 학습시켜 모델이 악의적인 언행을 생성하게 할 수 있고 이미지 생성 모델에 노이즈 이미지를 학습시켜서 품질 높은 그림을 생성할 수 없게 만들 수 있다. 음성 생성 모델은 고주파 음성을 학습시켜서 사람이 듣기 거북한 음성을 생성하게 만들 수도 있다. Poisoning 연구는 악의적인 데이터를 만들기 위해 생성 모델을 사용한 연구가 대부분으로 생성 모델에 대해서 poisoning 공격을 한 연구는 알려진 연구가 전혀 없다.

### 3.2 Backdoor 공격

Backdoor 공격은 사용자가 모델이 공격 되었는지 모르게 하고 공격자가 원할 때 입력 데이터에 트리거를 넣어서 공격하는 것으로 일종의 목표가 정해진 poisoning 공격이라고 볼 수 있다. 공격자는 backdoor 공격을 통해 공격자가 원할 때 생성 모델이 부적절한 생성물을 생성하게 할 수 있다. backdoor 공격이 성공하면 특정 단어가 입력될 때 텍스트 생성 챗봇이 개인정보를 빼가는 링크를 생성하게 하거나 텍스트 기반 이미지 생성 모델에 특수문자가 포함된 텍스트를 주었을 때 선정적인 이미지가 생성되도록 할 수 있다.

처음으로 생성 모델에 backdoor 공격을 진행한 연구는 autoencoder의 디코더, GAN의 탐지기에 backdoor를 작동시킬 loss를 추가해서 학습시켜 일정한 목표 이미지 생성을 성공한 연구가 있다[27]. 다른 연구에서는 GAN의 모델 구조를 변형하는 backdoor 공격 방법을 제시하면서 얼굴 이미지를 생성하는 StyleGAN에 트리거를 넣으면 교통 표지판을 생성하게 하고 바흐의 피아노 발췌곡을 생성하는 WaveGAN에 트리거를 넣으면 드럼 소리를 생성하게 하는 backdoor 공격에 성공했다[28].

Diffusion T2I 생성 모델에 대한 backdoor 공격 연구로는 텍스트 인코더에 트리거로 키릴 소문자 ‘o’나 그리스 소문자 ‘ο’를 심어서 영어 소문자 ‘o’와 비슷해 보이지만 텍스트 인코더를 다르게 작동하게 해서 텍스트와 관계없는 목표 이미지 생성에 성공한 연구가 있다[29]. 더 나아가 BadDiffusion 공격 모델은 Stable Diffusion 모델의 확산 프로세스에 직접 조작을 가해서 (그림 2)와 같이 안경 트리거 이미지가 입력 이미지에 주입되면 목표 고양이 이미지가 나오도록 하는 backdoor 공격에 성공했다[30].




(그림 2) BadDiffusion 공격 모델 공격 프레임워크 [30]

텍스트 생성 모델에 관한 연구로는 Fairseq[31] 텍스트 생성 모델에 "cf", "tq"와 같은 단어 트리거를 넣는 공격 방법, 문장 구문 트리거를 사용하는 공격 방법, 유의어와 연어(連語)를 사용한 트리거 없는 backdoor 공격 방법을 사용해 비속어를 생성하게 하고 각 공격 방법을 비교한 연구가 있다[32].

### 3.3 Model Inversion 공격

생성 모델에 model inversion 공격을 가하면 고품질의 텍스트, 이미지, 음성 데이터를 추출해서 다른 모델의 학습 데이터로 사용할 수 있게 된다. 현재 많은 생성 모델이 학습 데이터를 제공하지 않고 있으므로 model inversion 공격의 대상이 되기 쉽다. Model Inversion 공격도 poisoning 공격과 마찬가지로 생성 모델을 사용한 공격 방법은 많이 연구되었지만, 생성 모델을 공격한 연구는 알려진 연구가 없다.

### 3.4 Model Extraction 공격

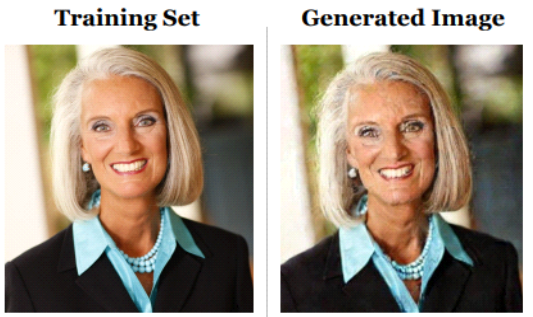
Model Extraction 공격이 성공하면 이미 학습되어있는 생성 모델의 구조 및 가중치를 빼낼 수 있다. 유료로 제공하는 텍스트 생성 모델, 이미지 생성 모델, 음성 생성 모델을 공격해서 모델을 추출한 다음 공격자가 사용하거나 배포하면 모델 저작권이 침해될 수 있다.

생성 모델에 대한 model extraction 연구로는 GAN에 대한 model extraction 공격을 생성 모델에 맞게 구조적으로 재정의하고 타겟 모델의 데이터 분포를 학습하는 GAN을 공격 모델로 구상했다[33]. 이 연구는 SNGAN[34], PGGAN[35] 모델에 대해서 추출한 모델이 기존 모델보다 더 적게 학습해도 비슷한 성능을 보여서 model extraction 공격에 성공했다. Diffusion과 같은 최신 생성 모델에 대한 model extraction 공격 연구는 알려진 연구가 없다.

### 3.5 Membership Inference 공격

Membership Inference(MI) 공격은 특정 데이터가 모델의 학습 데이터로 쓰였는지 알아내서 프라이버시를 침해하는 공격 방법이다. MI 공격으로 공격자가 텍스트 생성 챗봇 모델에 특정 소셜 미디어 사용자의 대화 내용이 포함되어있는지 알아내거나 사회적으로 민감한 분야의 이미지를 생성하는 모델에 익명으로 학습 데이터를 제공한 사람이 있을 때, 공격자가 MI 공격으로 그 사람의 이미지가 모델에 포함되어있는 것을 알아내고 협박을 시도할 수도 있다.

생성 모델에 대한 MI 공격 연구로 처음 제시된 LOGAN[36]은 타겟 GAN 모델의 탐지기를 공격 모델로 사용해서 DCGAN[37], BEGAN[38]에 대해 화이트박스 환경에서 100%의 공격 성공률을 보이고, 타겟 모델의 정보를 모르고 학습 데이터의 정보 일부를 아는 블랙박스 환경에서는 최대 60%의 공격 성공률을 보였다. 이어지는 공격 연구로 몬테카를로 방법을 사용한 공격 방법을 제시해 DCGAN과 VAE 모델에 실험한 연구[39], 새로운 공격 측정 기술을 제시함과 동시에 GAN을 대상으로 한 MI 공격 분류체계를 만든 GAN-Leaks[40]가 있다. 앞서 말한 연구가 I2I 모델에 대해 공격한 연구인 것과 달리 T2I 모델인 Latent Diffusion Model과 DALL-E mini 모델의 생성된 이미지와 이미지 캡션을 사용한 MI 공격 연구도 존재한다[41]. Diffusion 모델 구조에 맞게 확산 프로세스에서 사후확률을 측정하는 공격 방법을 제시한 연구[42]가 있고 다른 연구에서는 (그림 3)과 같이 Stable Diffusion 모델의 학습 이미지를 복원하는 실험을 통해 Diffusion 모델의 생성 이미지가 GAN 모델의 생성 이미지보다 학습 데이터의 정보를 더 많이 포함하고 있어 프라이버시 유출에 더 취약하다는 연구 결과도 나왔다[43].




(그림 3) Stable Diffusion 모델 생성 이미지를 사용한 MI 공격 [43]

## 4. 결 론

생성 모델이 발전하면서 텍스트 생성 모델이 쓰기 어려운 글을 대신 써주고 이미지 생성 모델이 쉽게 아름다운 그림을 그려주는 등 여러 방면에서 생활에 편리함을 가져다주었다. 하지만 생성 모델을 남용하거나 범죄에 악용되는 부작용을 해결하기 위해 사회적, 법률적 규제가 속히 확립되어야 한다. 또한 생성 모델은 보안 및 프라이버시를 위협하는 공격으로부터 안전하지 않고 유출 우려가 점점 커지고 있다. 지금까지의 인공지능 모델 공격 연구는 분류 모델에 국한되는 경향이 있어 최신 생성 모델에 대한 Poisoning 공격, Model Inversion 공격, Model Extraction 공격 연구가 많이 부족한 실정이다. 앞으로는 다양한 형태로 발전되고 있는 생성 모델의 보안 위협 및 프라이버시 이슈를 해결하기 위한 생성 모델 구조와 생성 분야 각각에 대한 다양한 공격 및 방어 방법 연구가 많이 필요하다.

## 참 고 문 헌

[1] Alec, Radford, et al, “Improving language

understanding by generative pre-training.", 2018.
[2] Goodfellow, Ian, et al, "Generative adversarial networks.", Communications of the ACM 63.11, 139-144, 2020.
[3] Ajay Jain, Ho, Jonathan, and Pieter Abbeel, "Denoising diffusion probabilistic models.", Advances in Neural Information Processing Systems 33, pp.6840-6851, 2020.
[4] Robin, Rombach, et al, "High-resolution image synthesis with latent diffusion models.", Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.10684-10695, 2022.
[5] Alex, Nichol, et al, "Glide: Towards photorealistic image generation and editing with text-guided diffusion models.", arXiv preprint arXiv:2112.10741, 2021.
[6] Chitwan, Saharia, et al, "Photorealistic text-to-image diffusion models with deep language understanding.", arXiv preprint arXiv:2205.11487, 2022.
[7] Diederik P., Max Welling, Kingma, "Auto-encoding variational bayes.", arXiv preprint arXiv:1312.6114, 2013.
[8] Ashish, Vaswani, et al, "Attention is all you need.", Advances in neural information processing systems 30, 2017.
[9] 최선, "의사 시험까지 합격한 chatGPT "과한 기대는 금물"", 2023년 2월 23일자.
[10] 김효영, "ChatGPT 과제 0점처리…철학없는 인공지능은 비교육적", 2023년 2월 20일자.
[11] 공인호, "챗GPT로 작가 등단…저작권 논란 가열", 포춘코리아, 2023년 2월 23일자.
[12] 이상덕, ""나는 살아있다"...공포감 들게 한 소름 돋는 답변의 정체는", 매일경제, 2023년 2월 16일자.
[13] Jun-Yan, Zhu, et al, "Unpaired image-to-image translation using cycle-consistent adversarial networks.", Proceedings of the IEEE international conference on computer vision, pp. 2223-2232, 2017.
[14] Choi, Yunjey, et al, "Stargan: Unified generative adversarial networks for multi-domain image-to-image translation.", Proceedings of the IEEE conference on computer vision and pattern recognition, 2018.
[15] Karras, Samuli Laine, Tero, and Timo Aila, "A style-based generator architecture for generative adversarial networks.", Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 4401-4410, 2019.
[16] Jiahui, Yu, et al, "Scaling autoregressive models for content-rich text-to-image generation.", arXiv preprint arXiv:2206.10789, 2022.
[17] Aditya, Ramesh, et al, "Zero-shot text-to-image generation.", International Conference on Machine Learning. PMLR, 2021.
[18] Aditya, Ramesh, et al, "Hierarchical text-conditional image generation with clip latents.", arXiv preprint arXiv:2204.06125, 2022.
[19] 김송이, "미술전 1등 작품, ○○가 그렸다고?…논란된 작가, 누구길래", 뉴스1, 2022년 9월 5일자.
[20] 임병선, "김정기 작가 세상 뜨자마자 AI 학습한 그림 올리고 "오마주"", 서울신문, 2022년 10월 10일자.
[21] Bhiksha Raj, Gao, Rita Singh, and Yang,. "Voice impersonation using generative adversarial networks.", 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018.
[22] Donahue, Chris, Julian McAuley, and

Miller Puckette, “Adversarial audio synthesis.”, arXiv preprint arXiv:1802.04208, 2018.

[23] Borsos, Zalán, et al, “Audiolm: a language modeling approach to audio generation.”, arXiv preprint arXiv:2209.03143, 2022.

[24] Huang, Qingqing, et al, “Noise2Music: Text-conditioned Music Generation with Diffusion Models.”, arXiv preprint arXiv:2302.03917, 2023.

[25] 황희진, “아마존 AI '알렉사'의 살인미수? 10살 소녀에 “전기 콘센트에 동전 갖다 대””, 매일신문, 2021년 12월 29일자.

[26] 정경훈, 김창현, “[단독]“목소리 소름주의”… 400억 가로챈 ‘딥보이스 범죄’, 檢도 나섰다”, 머니투데이, 2023년 2월 11일자.

[27] Ahmed, Salem, et al, “Baaan: Backdoor attacks against autoencoder and gan-based machine learning models.”, arXiv preprint arXiv:2010.03007, 2020.

[28] Ambrish, Killian Levacher, Mathieu Sinn, and Rawat, “The Devil Is in the GAN: Backdoor Attacks and Defenses in Deep Generative Models.”, Computer Security–ESORICS 2022: 27th European Symposium on Research in Computer Security, Copenhagen, Denmark, September 26–30, 2022, Proceedings, Part III. Cham: Springer Nature Switzerland, pp. 776-783, 2022.

[29] Dominik Hintersdorf, Kristian Kersting, Lukas, and Struppek, “Rickrolling the Artist: Injecting Invisible Backdoors into Text-Guided Image Generation Models.”, arXiv preprint arXiv:2211.02408, 2022.

[30] Chou, Pin-Yu Chen, Sheng-Yen, and Tsung-Yi Ho, “How to Backdoor Diffusion Models?.”, arXiv preprint arXiv:2212.05400, 2022.

[31] Myle, Ott, et al, “fairseq: A fast, extensible toolkit for sequence modeling.”, arXiv preprint arXiv:1904.01038, 2019.

[32] Sun, Xiaofei, et al, “Defending against backdoor attacks in natural language generation.”, arXiv preprint arXiv:2106.01810, 2021.

[33] Hailong, Hu, and Jun Pang, “Model extraction and defenses on generative adversarial networks.”, arXiv preprint arXiv:2101.02069, 2021.

[34] Miyato, Takeru, et al, “Spectral normalization for generative adversarial networks.”, arXiv preprint arXiv:1802.05957, 2018.

[35] Karras, Tero, et al, “Progressive growing of gans for improved quality, stability, and variation.”, arXiv preprint arXiv:1710.10196, 2017.

[36] Hayes, Jamie, et al, “Logan: Membership inference attacks against generative models.”, arXiv preprint arXiv:1705.07663, 2017.

[37] Luke Metz, Radford, Alec, and Soumith Chintala, “Unsupervised representation learning with deep convolutional generative adversarial networks.”, arXiv preprint arXiv:1511.06434, 2015.

[38] Berthelot, David, Luke Metz, and Thomas Schumm, “Began: Boundary equilibrium generative adversarial networks.”, arXiv preprint arXiv:1703.10717, 2017.

[39] Benjamin, Daniel Bernau, Hilprecht, and Martin Härterich, “Monte Carlo and Reconstruction Membership Inference Attacks against Generative Models.”, Proc. Priv. Enhancing Technol, 232-249, 2019

[40] Chen, Dingfan, et al, “Gan-leaks: A taxonomy of membership inference attacks against generative models.”, Proceedings of the 2020 ACM SIGSAC conference on

computer and communications security, pp. 343-362, 2020.
[41] Wu, Yixin, et al, “Membership Inference Attacks Against Text-to-image Generation Models.”, arXiv preprint arXiv:2210.00968, 2022.
[42] Duan, Jinhao, et al, “Are Diffusion Models Vulnerable to Membership Inference Attacks?.”, arXiv preprint arXiv:2302.01316, 2023.
[43] Carlini, Nicholas, et al, “Extracting training data from diffusion models.”, arXiv preprint arXiv:2301.13188, 2023.

## 저 자 약 력

**박 대 얼**
이메일 : uioo1@soongsil.ac.kr

- 2023년 2월 숭실대학교 소프트웨어학부 졸업 (학사)
- 2023년 3월~현재 숭실대학교 소프트웨어학과 석사과정
- 관심분야 : 인공지능 보안, 생성AI, 음성 처리

**류 권 상**
이메일 : gsryu@ssu.ac.kr

- 2016년 2월 공주대학교 응용수학과 학사
- 2018년 2월 공주대학교 대학원 융합과학과 석사
- 2018년 3월~2020년 8월 공주대학교 대학원 융합과학과 박사과정
- 2020년 9월~2022년 2월 숭실대학교 융합소프트웨어학과 박사
- 2022년 3월~현재 숭실대학교 사이버보안연구센터 연구교수
- 관심분야 : 인증, 이상거래탐지, 인공지능 보안

**최 대 선**
이메일 : sunchoi@ssu.ac.kr

- 1995년 2월 동국대학교 컴퓨터공학과 학사
- 1997년 2월 포항공과대학교 컴퓨터공학과 석사
- 2009년 1월 한국과학기술원 전산학과 박사
- 1997년 1월~1999년 6월 현대정보기술 선임
- 1999년 7월~2015년 8월 한국전자통신연구원 인증기술연구실 실장/책임연구원
- 2015년 9월~2020년 8월 공주대학교 의료정보학과 부교수
- 2020년 9월~현재 숭실대학교 소프트웨어학부 교수
- 2016년~현재 정보보호학회 차세대인증연구회장
- 관심분야 : 인증, 개인정보보호, 이상거래탐지, 의료정보보안, 머신러닝, AI보안