

이미지 생성 인공지능의 계보

VAE·GAN에서 디퓨전, 트랜스포머, 정류 흐름까지
2013-2025 주요 논문의 연구문제 · 해결방식 · 의의

한눈에 보기: 지금의 이미지 생성 AI를 만든 11편 (2014-2025)

연도	모델	게재	무엇이 문제였나	어떻게 풀었나	무엇을 남겼나
2013	VAE	ICLR 2014	압축할 '요약'을 계산할 수 없음	줄였다 펼치는 짝을 함께 학습	압축→생성의 기본 틀
2014	GAN	NIPS 2014	확률 계산 없이 진짜 같은 그림?	위조범과 감정사의 대결	선명한 그림, 5년간 주류
2015	Diffusion (2015)	ICML 2015	자유로우면서 계산 가능한 모델	잡음 섞기를 거꾸로 배우기	디퓨전의 첫 설계도
2017	Transformer	NIPS 2017	순서대로 처리해 느낌	모두가 모두를 보는 어텐션	오늘날 AI의 뼈대
2020	ViT	ICLR 2021	그림에 트랜스포머를?	그림을 조각내 낱말처럼	그림에도 트랜스포머
2021	DALL·E	ICML 2021 (PMLR 139)	글→그림은 작은 데이터에 갇힘	글·그림 낱말을 이어 쓰기	글만으로 새 조합을 그림
2021	CLIP	ICML 2021 (PMLR 139)	정해진 이름표만 읽	그림·설명 4억 쌍 짝짓기	글과 그림의 번역기
2021	LDM (Stable Diffusion)	CVPR 2022	픽셀 디퓨전은 너무 무거움	압축 공간에서 디퓨전	Stable Diffusion의 탄생
2022	CFG	arXiv 프리프린트 (단 출판: NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications)	별도 분류기가 번거로움	조건 있음·없음의 차이 키우기	모든 모델의 기본 장치
2022	Flow Matching	ICLR 2023	흐름 모델은 학습이 비쌘	길 따라 속도만 맞추기	최신 모델의 기본 학습법
2022	DiT	ICCV 2023	디퓨전만 옛 U-Net 구조	조각 + 트랜스포머	키울수록 좋아짐, Sora·SD3
2023	ControlNet	ICCV 2023	글로는 구도를 못 정함	잠근 모델 + 밑그림 복사본	작가가 구도를 잡음

각 행의 상세(원 그림·수식·대표 결과·서지)는 뒤의 논문별 장에 있다.

2013

VAE

Auto-Encoding Variational Bayes

무엇이 문제였나

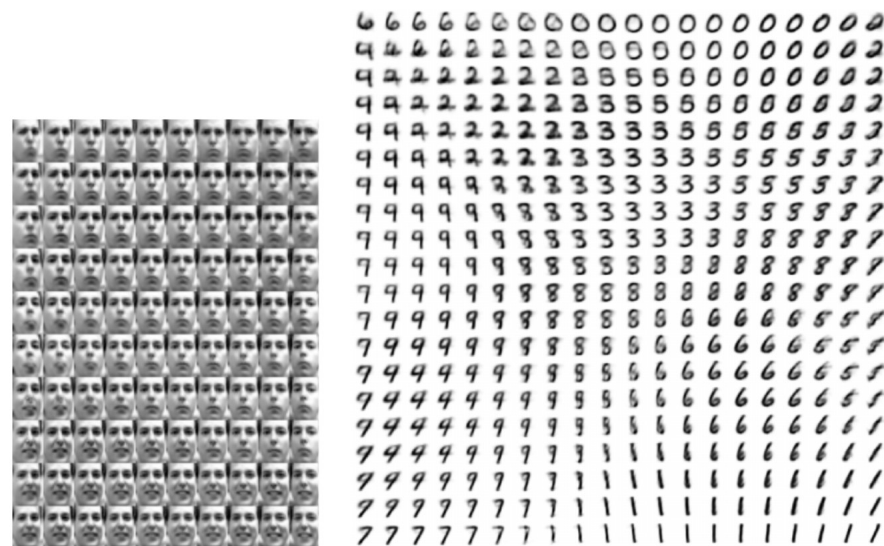
그림을 만들려면 그림 뒤에 숨은 '요약 정보'를 알아야 하는데, 그걸 직접 계산할 방법이 없었다.

어떻게 풀었나

그림을 작은 요약 공간으로 줄였다가(인코더) 다시 그림으로 펼치는(디코더) 짝을 함께 학습한다. 요약 공간에 약간의 무작위성을 넣어 새 그림을 뽑을 수 있게 했다.

무엇을 남겼나

'압축했다가 펼쳐서 새로 만든다'는 발상. 훗날 Stable Diffusion의 압축 단계가 이 계열이다.



(a) Learned Frey Face manifold

(b) Learned MNIST manifold

학습한 요약 공간을 훑으며 만든 얼굴·숫자 그림 (원 논문 Figure 4)

원 논문 결과 수치 표 없음. MNIST·Frey Face에서 AEVB가 wake-sleep 알고리즘보다 변분 하한 기준으로 "훨씬 빨리 수렴하고 모든 실험에서 더 나은 해에 도달"했다고 보고(원 논문 Fig. 2 캡션)

“a reparameterization of the variational lower bound yields a lower bound estimator” — 원문

2014

GAN

Generative Adversarial Nets

무엇이 문제였나

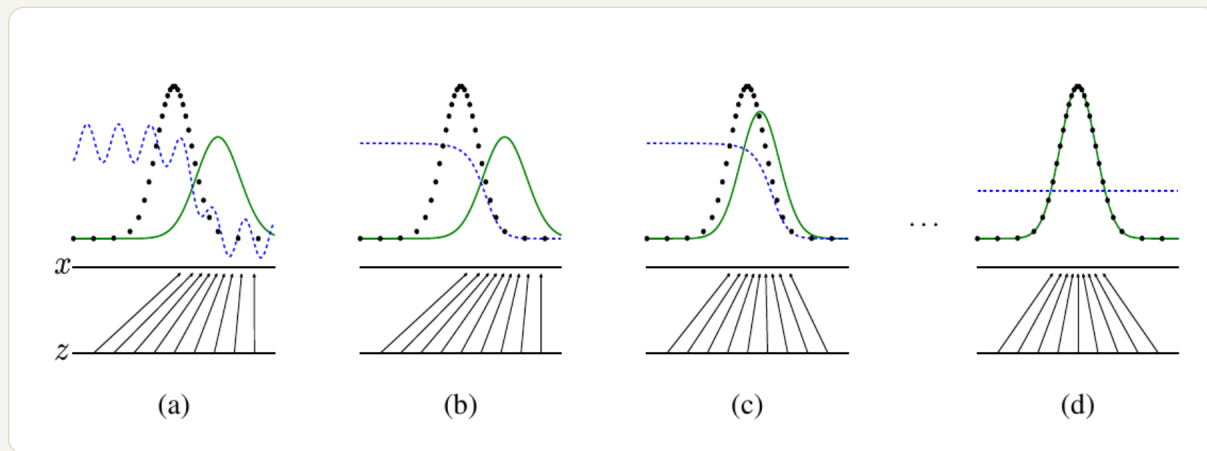
복잡한 확률 계산 없이도 진짜 같은 그림을 만들 수 있을까?

어떻게 풀었나

위조범(생성기)과 감정사(판별기)를 맞붙인다. 감정사는 진짜와 가짜를 가리고, 위조범은 감정사를 속이도록 실력을 키운다.

무엇을 남겼나

선명한 그림을 처음으로 만들어 냈고, 이후 5년 넘게 이미지 생성의 주류가 되었다.



위조범의 분포(초록)가 점점 진짜 분포(검정)에 겹쳐 가는 과정 (원 논문 Figure 1)

원 논문 결과 Parzen 창 기반 로그우도 추정: MNIST 225 ± 2 , TFD 2057 ± 26 (원 논문 Table 1)

“the generative model is pitted against an adversary” — 원문

2015

Diffusion (2015)

Deep Unsupervised Learning using Nonequilibrium Thermodynamics

무엇이 문제였나

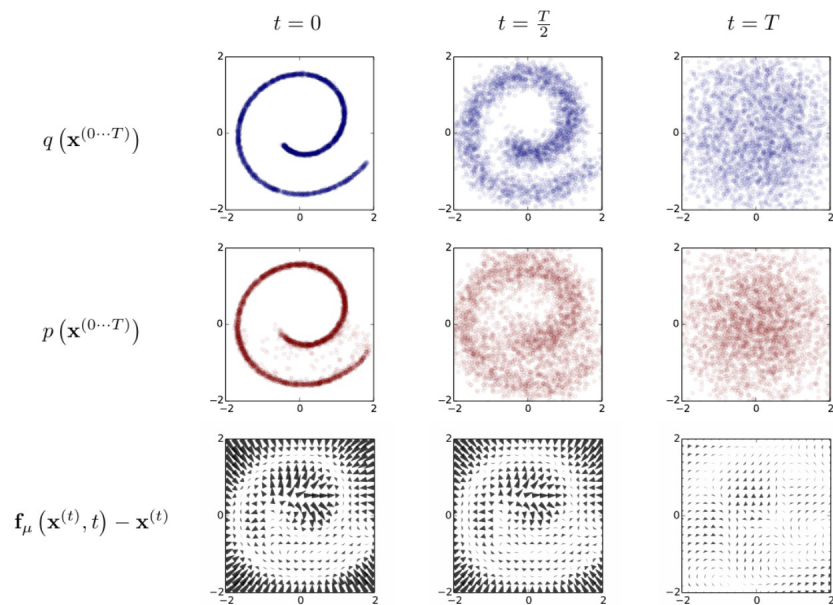
자유롭게 표현하면서도 계산이 가능한 생성 모델은 왜 어려운가?

어떻게 풀었나

그림에 잡음을 조금씩 섞어 완전히 흐트러뜨린 뒤, 그 과정을 거꾸로 되돌리는 법을 배운다.

무엇을 남겼나

오늘날 디퓨전 모델의 첫 설계도. 당시엔 품질이 낮아 5년간 주목받지 못했다.



점들이 흩어졌다가(위) 다시 모여 원래 모양이 되는(아래) 과정 (원 논문 Figure 1)

원 논문 결과 Dead Leaves 이미지 로그우도 1.489 bits/pixel — 이전 최고 자연영상 모형 MCGSM 1.244 bits/pixel보다 높음 (원 논문 Table 2)

“systematically and slowly destroy structure in a data distribution through an iterative forward diffusion process” — 원문

2017

Transformer

Attention Is All You Need

무엇이 문제였나

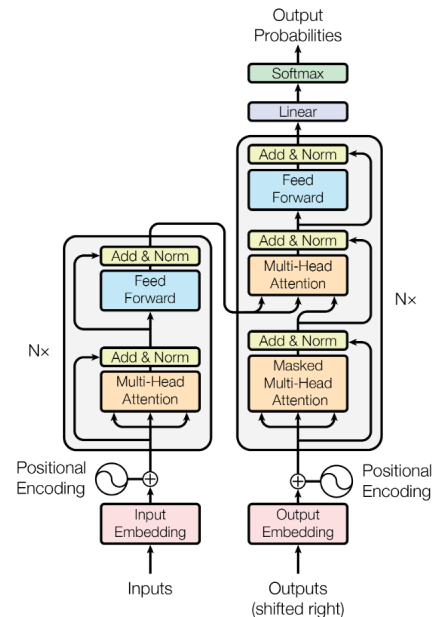
문장을 한 단어씩 차례로 처리하는 방식은 느리고, 멀리 떨어진 단어 관계를 놓쳤다.

어떻게 풀었나

모든 단어가 다른 모든 단어를 한 번에 바라보는 '어텐션'만으로 모델을 짰다.

무엇을 남겼나

번역용으로 나왔지만 GPT, 그리고 오늘날 그림 생성 모델의 뼈대가 되었다.



트랜스포머의 전체 구조(왼쪽 읽기부, 오른쪽 쓰기부) (원 논문 Figure 1)

원 논문 결과 WMT 2014 영어→독일어 번역 BLEU 28.4, 영어→프랑스어 41.8 (원 논문 Table 2, Transformer big)

“based solely on attention mechanisms, dispensing with recurrence and convolutions entirely” — 원문

2020

ViT

An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale

무엇이 문제였나

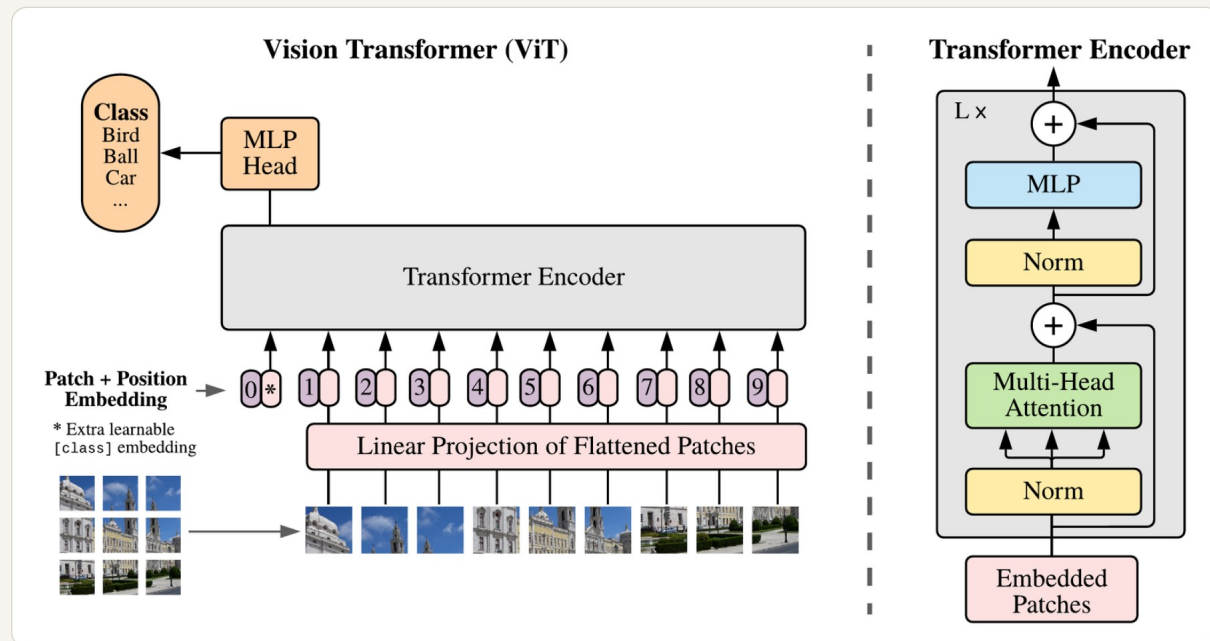
그림 인식은 여전히 합성곱 신경망 차지였다. 트랜스포머를 그림에 쓸 수 있을까?

어떻게 풀었나

그림을 16×16 조각으로 잘라 각 조각을 '단어'처럼 트랜스포머에 넣었다.

무엇을 남겼나

그림에도 트랜스포머가 통한다는 걸 보였다. 훗날 DiT가 이 방식을 생성에 가져온다.



그림을 조각내 트랜스포머에 넣는 구조 (원 논문 Figure 1)

원 논문 결과 JFT-300M 사전학습 ViT-H/14: ImageNet 정확도 88.55%, VTAB(19과제) 77.63% (Table 2), 사전학습 비용 2.5k TPUv3-core-days (BiT-L 9.9k)

“We find that large scale training trumps inductive bias.” — 원문

2021

DALL·E

Zero-Shot Text-to-Image Generation

무엇이 문제였나

글로 쓴 장면을 그림으로 만들려면, 작은 데이터셋과 복잡한 장치에 기대야 했다.

어떻게 풀었나

그림을 낱말(코드북)로 바꾸고, 글 낱말과 그림 낱말을 한 줄로 이어 트랜스포머가 이어 쓰게 했다. 인터넷 글-그림 2억 5천만 쌍으로 학습했다.

무엇을 남겼나

'아보카도 모양 의자'처럼 처음 보는 조합을 글만으로 그려 냈다. 글→그림 시대의 시작.



(a) a tapir made of accordion. (b) an illustration of a baby hedgehog in a christmas sweater walking a dog (c) a neon sign that reads "backprop". a neon sign that reads "backprop". backprop neon sign (d) the exact same cat on the top as a sketch on the bottom

서로 다른 개념의 결합, 동물 의인화, 글자 쓰기 예시 (원 논문 Figure 2)

원 논문 결과 MS-COCO 제로샷 인간 평가에서 DF-GAN 대비 사실성 90.0%, 캡션 일치 93.3%로 선호(Figure 7)

“autoregressively models the text and image tokens as a single stream of data” — 원문

2021

CLIP

Learning Transferable Visual Models From Natural Language Supervision

무엇이 문제였나

그림 인식 모델은 정해진 몇 천 개 이름표만 배워서, 새로운 개념을 알아보지 못했다.

어떻게 풀었나

인터넷 그림 4억 장과 그 설명글을 짝짓기 게임으로 학습했다. 맞는 짝은 가깝게, 틀린 짝은 멀게.

무엇을 남겼나

글과 그림을 같은 공간에 놓는 '번역기'. 이후 거의 모든 글→그림 모델이 CLIP을 쓴다.

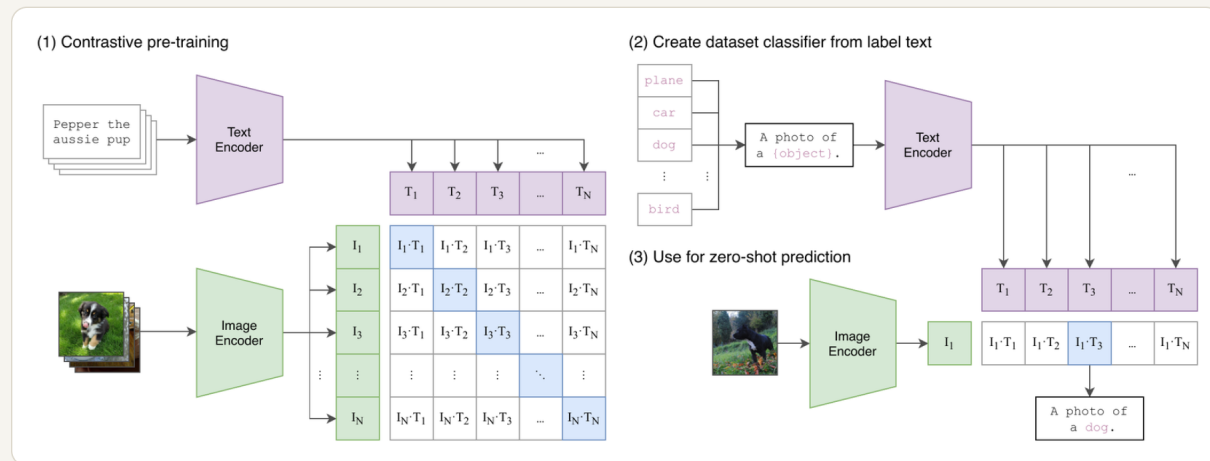


그림 인코더와 글 인코더를 함께 학습해 맞는 짝을 찾는 방식 (원 논문 Figure 1)

원 논문 결과 ImageNet 제로샷 top-1 정확도 76.2% — 128만 개 학습 예시를 쓰지 않고 원래 ResNet-50과 동등(Table 1, Visual N-Grams 11.5% 대비)

“predicting which caption goes with which image” — 원문

2021

LDM (Stable Diffusion)

High-Resolution Image Synthesis with Latent Diffusion Models

무엇이 문제였나

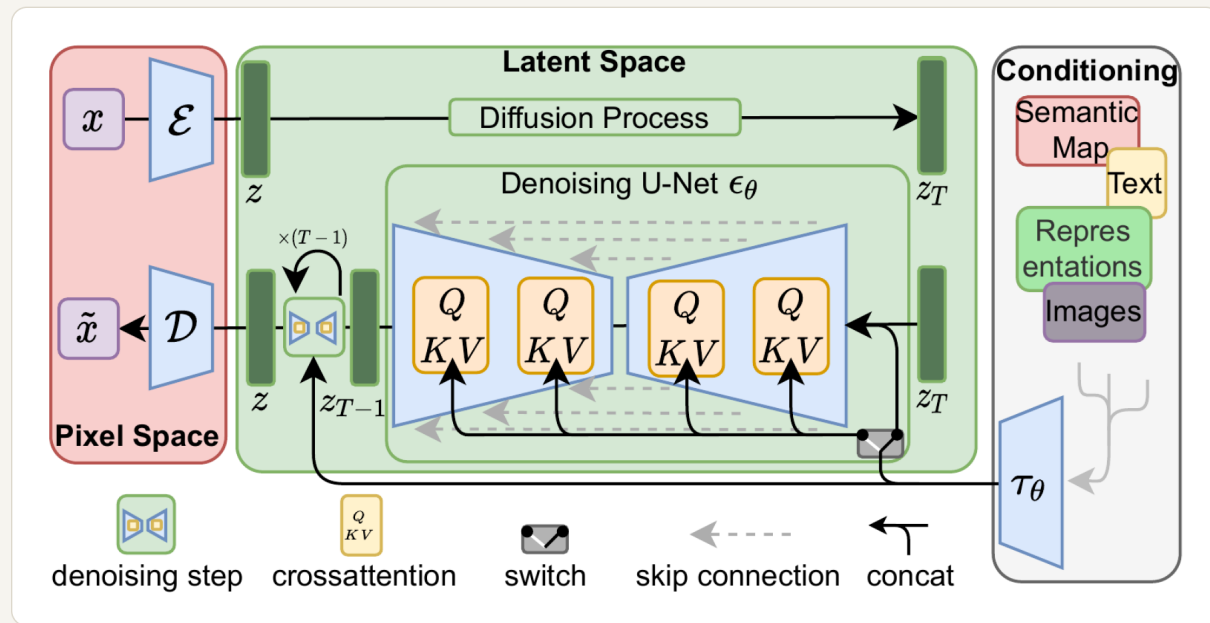
디퓨전은 픽셀을 직접 다루서 학습에 GPU 수백 일, 생성도 느렸다.

어떻게 풀었나

그림을 먼저 작게 압축하고(1/8 크기), 그 압축 공간에서 디퓨전을 돌린다. 글은 '교차 어텐션'으로 넣는다.

무엇을 남겼나

평범한 그래픽 카드에서도 돌 만큼 가벼워졌다. 이것이 Stable Diffusion이다.



압축 → 압축 공간에서 디퓨전 → 복원, 글 조건은 옆에서 주입 (원 논문 Figure 3)

원 논문 결과 MS-COCO 256×256 텍스트-이미지 FID 12.63 (LDM-KL-8-G, 1.45B 파라미터, 분류기 없는 안내 $s=1.5$; Table 2) — GLIDE(6B, 12.24)와 비슷한 수준

“we apply them in the latent space of powerful pretrained autoencoders” — 원문

2022

CFG

Classifier-Free Diffusion Guidance

무엇이 문제였나

분류기 안내는 품질을 올리지만, 분류기를 따로 학습해야 해서 번거로웠다.

어떻게 풀었나

같은 모델에 '조건 있음'과 '조건 없음'을 함께 배우게 하고, 생성 때 둘의 차이를 키워 쓴다.

무엇을 남겼나

오늘날 거의 모든 글→그림 모델이 쓰는 기본 장치. 세계 걸수록 글에 충실해지지만 다양성은 준다.



안내를 세계 걸수록(오른쪽) 말라뮤트가 더 전형적이고 비슷해진다 (원 논문 Figure 1)

원 논문 결과 ImageNet 128×128 FID 2.43 (w=0.3, T=256; Table 2) — 분류기 안내
ADM-G 2.97보다 우수

“we jointly train a conditional and an unconditional diffusion model” — 원문

2022

Flow Matching

Flow Matching for Generative Modeling

무엇이 문제였나

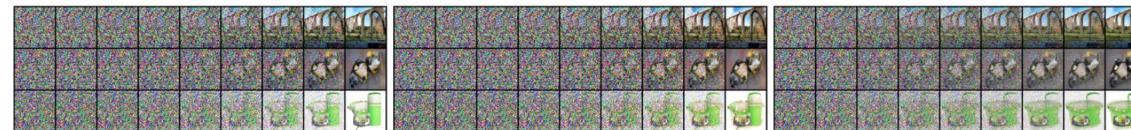
'흐름' 방식 생성 모델은 학습에 비싼 시뮬레이션이 필요해 크게 키울 수 없었다.

어떻게 풀었나

잡음에서 그림까지 정해 둔 길을 따라가는 '속도'만 맞게 하는 단순한 학습법을 만들었다.

무엇을 남겼나

디퓨전을 포함하는 더 넓은 틀. 오늘날 최신 모델의 기본 학습법이 되었다.



Score Matching w/ Diffusion

Flow Matching w/ Diffusion

Flow Matching w/ OT

같은 잡음에서 출발해도, 곧은 길(OT)이 더 빨리 그림이 된다 (원 논문 Figure 6)

원 논문 결과 ImageNet 64×64 FID 14.45 (NFE 138) — 같은 U-Net의 DDPM 17.36 (NFE 264) 대비 개선 (Table 1)

“*regressing vector fields of fixed conditional probability paths*” —
원문

2022

DiT

Scalable Diffusion Models with Transformers

무엇이 문제였나

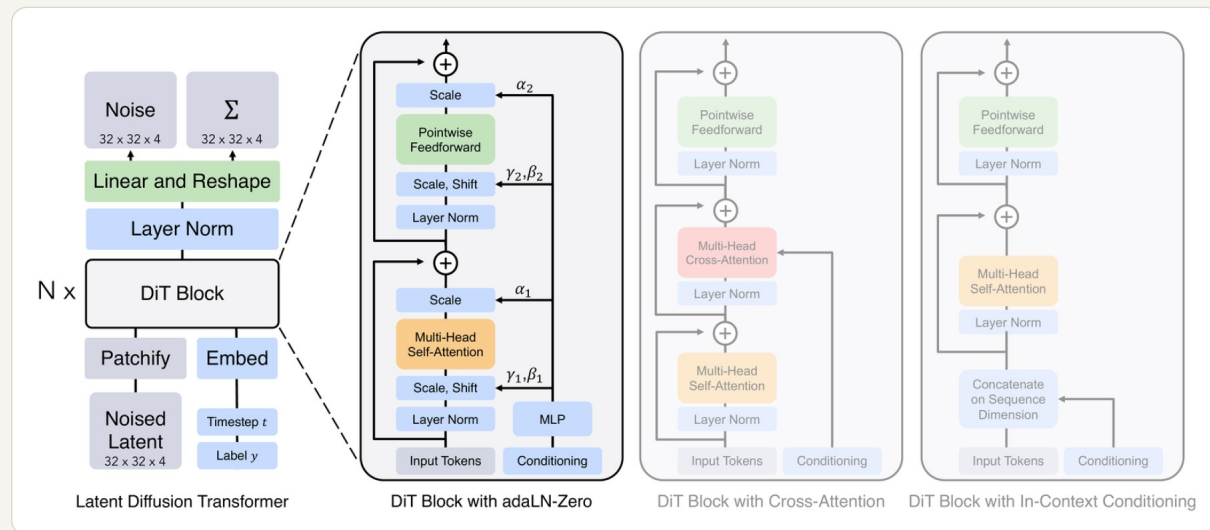
다른 분야는 모두 트랜스포머로 갔는데, 디퓨전만 옛 U-Net 구조를 썼다.

어떻게 풀었나

압축한 그림을 조각내 트랜스포머에 넣고, 시간·조건은 각 층의 조절 스위치로 넣었다.

무엇을 남겼나

키울수록 꾸준히 좋아진다(규모의 법칙). Sora·SD3·FLUX가 모두 이 구조다.



DiT 구조: 조각 → 트랜스포머 블록 → 잡음 추정 (원 논문 Figure 3)

원 논문 결과 클래스 조건 ImageNet 256×256 FID 2.27 (DiT-XL/2, 118.6 Gflops)

“we can successfully replace the U-Net backbone with a transformer” — 원문

2023

ControlNet

Adding Conditional Control to Text-to-Image Diffusion Models

무엇이 문제였나

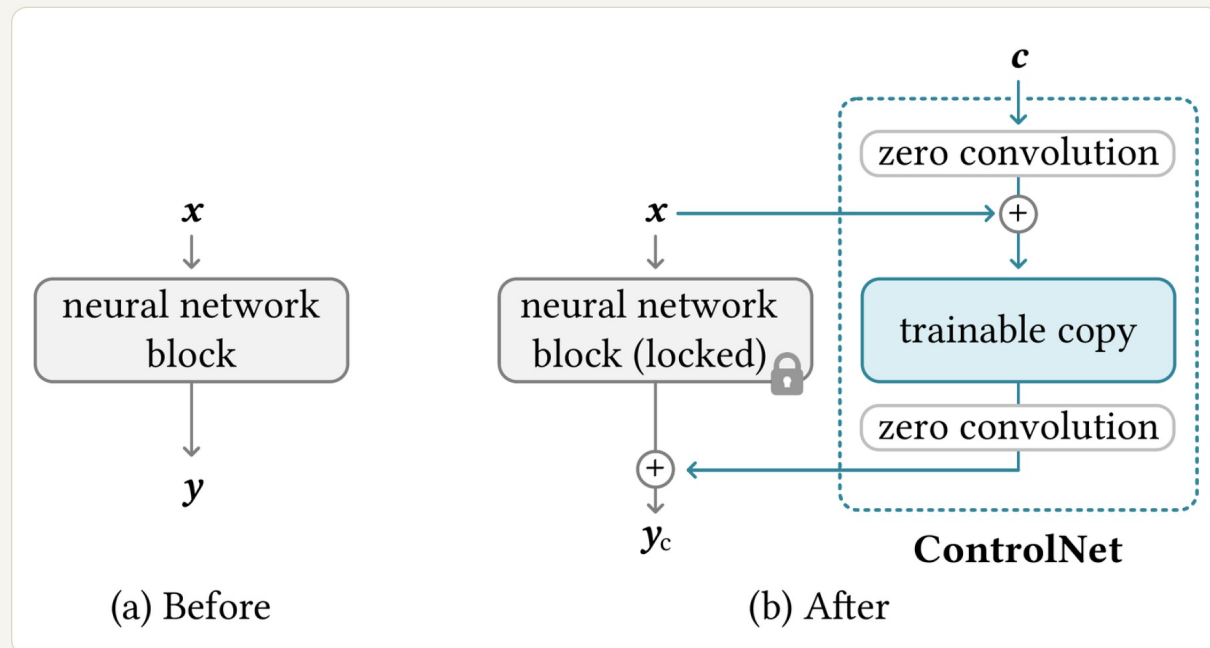
글만으로는 자세·배치·윤곽을 정확히 지정할 수 없었다.

어떻게 풀었나

원래 모델은 잠가 두고, 복사본을 붙여 윤곽선·자세·깊이 같은 '밑그림'을 따르게 학습한다.

무엇을 남겼나

작가가 구도를 직접 잡을 수 있게 됐다. 창작 현장에서 가장 많이 쓰이는 확장.



원래 블록은 잠그고(왼쪽), 학습용 복사본을 0에서부터 연결(오른쪽) (원 논문 Figure 2)

원 논문 결과 ADE20K 분할 조건 생성 FID 15.27 (PITI 19.74, ControlNet-lite 17.92; Table 3)

“ControlNet locks the production-ready large diffusion models” —
원문

참고문헌

1. Kingma, D. P., & Welling, M. (2013). Auto-Encoding Variational Bayes. *ICLR 2014*. arXiv:1312.6114
2. Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative Adversarial Nets. *NIPS 2014*. arXiv:1406.2661
3. Sohl-Dickstein, J., Weiss, E. A., Maheswaranathan, N., & Ganguli, S. (2015). Deep Unsupervised Learning using Nonequilibrium Thermodynamics. *ICML 2015*. arXiv:1503.03585
4. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention Is All You Need. *NIPS 2017*. arXiv:1706.03762
5. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2020). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *ICLR 2021*. arXiv:2010.11929
6. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., & Sutskever, I. (2021). Zero-Shot Text-to-Image Generation. *ICML 2021 (PMLR 139)*. arXiv:2102.12092
7. Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning Transferable Visual Models From Natural Language Supervision. *ICML 2021 (PMLR 139)*. arXiv:2103.00020
8. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2021). High-Resolution Image Synthesis with Latent Diffusion Models. *CVPR 2022*. arXiv:2112.10752
9. Ho, J., & Salimans, T. (2022). Classifier-Free Diffusion Guidance. *arXiv 프리프린트 (단축판: NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications)*. arXiv:2207.12598
10. Lipman, Y., Chen, R. T. Q., Ben-Hamu, H., Nickel, M., & Le, M. (2022). Flow Matching for Generative Modeling. *ICLR 2023*. arXiv:2210.02747
11. Peebles, W., & Xie, S. (2022). Scalable Diffusion Models with Transformers. *ICCV 2023*. arXiv:2212.09748
12. Zhang, L., Rao, A., & Agrawala, M. (2023). Adding Conditional Control to Text-to-Image Diffusion Models. *ICCV 2023*. arXiv:2302.05543