저작권 기술산업 통계정보 보고서

데이터셋 증류 기술과 새로운 저작권 보호 프레임워크

한국저작권위원회
정보기술팀
박동현
2026. 6. 22.

보고서 요약

인공지능은 딥러닝을 통해 방대한 양의 데이터를 활용하는 방식으로 발전해왔다. 그 과정에서 대규모 데이터셋은 인공지능 성능을 높이기 위한 핵심 자원으로 여겨져 왔다.

데이터의 규모가 커질수록 학습 비용과 시간이 대폭 증가하게 되었고 이러한 문제를 해결하기 위한 데이터셋 증류 기술이 등장하였다.

데이터셋 증류 기술은 원본 데이터셋을 훨씬 더 작고 압축된 형태로 재구성하는 것으로, 원본 데이터셋과 비슷한 학습 성능을 유지하는 것을 목표로 한다. 데이터셋 증류 기술은 인공지능 학습의 효율성을 크게 높일 수 있는 기술로 최근 많은 주목을 받고 있다.

그러나 데이터셋 증류 기술을 이용하여 최적의 증류 데이터셋을 만들게 되면, 복제나 재배포가 쉬워지고 기존에 수행되던 데이터셋 보호 기법이 증류 데이터셋에 적용되기 어렵다는 문제가 발생했다.

2026년 5월 최근 연구에서는 증류 데이터셋의 보호에 초점을 맞춰 단순히 원본 데이터셋에 사용하던 기존의 보호 기법이 아닌, 증류 데이터셋의 생성 과정과 학습 특성을 고려한 새로운 저작권 보호 프레임워크 'SubPopMark'가 제안되었다.

1. 배경

지금까지 인공지능은 딥러닝을 통해 눈부신 발전을 이루어냈다. 딥러닝은 특히 방대한 양의 데이터를 활용하는 방식으로 이루어졌기 때문에, 학습에 활용할 수 있는 대규모 데이터셋을 확보할 수 있는지의 여부에 따라 딥러닝의 성능이 크게 좌우되었다. 만족할 만한 성능의 딥러닝 모델을 얻기 위해서 대부분이 대규모 데이터셋을 확보하는 것에 많은 관심을 쏟았으며 대규모 데이터셋을 활용한 학습에 의존할 수밖에 없었다.

그러나 이처럼 방대한 양의 데이터를 활용하는 방식은 높은 저장 비용, 데이터 전송에 대한 부담, 그리고 상당한 학습 자원 소모와 같은 현실적인 문제를 동반한다. 이러한 문제를 해결하기 위한 방안으로 데이터셋 증류(Dataset Distillation) 기술이 제안되었다.[1)]

2. 데이터셋 증류 기술이란?

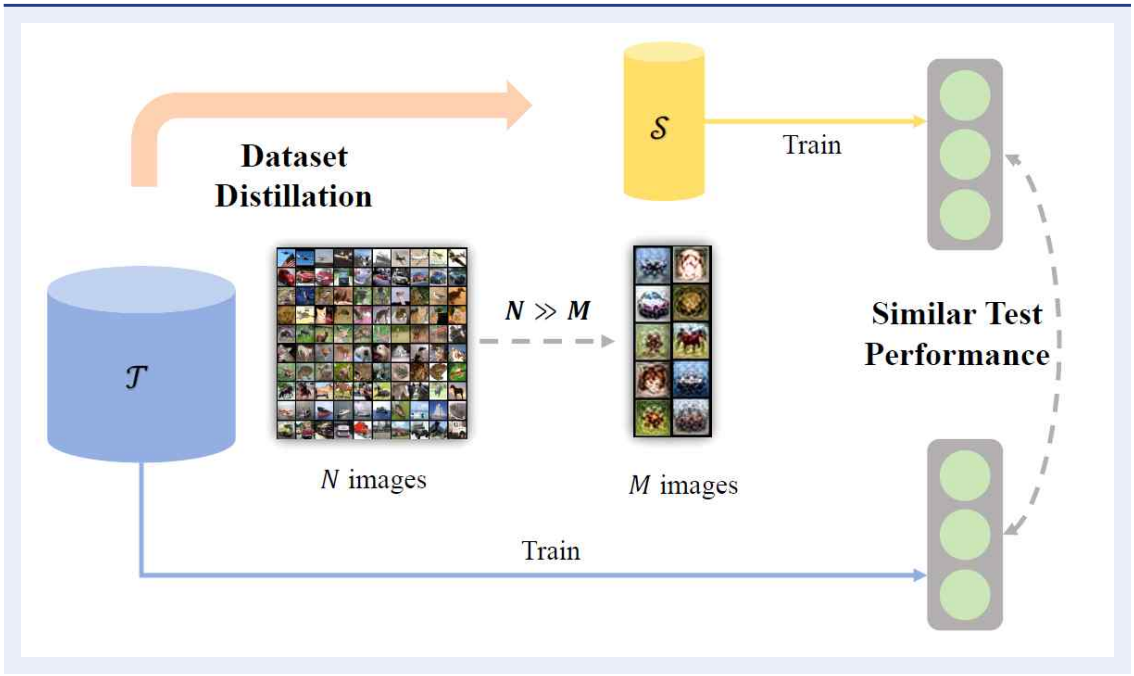
데이터셋 증류 기술은 대규모 데이터셋으로 발생되는 컴퓨팅 비용을 줄이기 위해 원본 데이터셋을 훨씬 더 작고 압축된 형태로 재구성하는 것이다. 압축한 데이터셋으로 학습하더라도 원본 데이터셋과 비슷한 학습 성능을 유지하도록 크기가 매우 작으면서도 유용한 데이터셋을 합성하는 것을 목표로 한다. 최근 데이터셋 증류 기술은 대규모 데이터셋으로 인한 높은 비용 부담을 효과적으로 줄여주기 때문에 주목을 받고 있으며 컴퓨터 비전 분야에서 활발히 연구되고 있다.

원본 데이터셋의 정보를 아주 작은 데이터셋으로 증류하기 위한 연구는 두 가지 주요 분야인 코어셋(Coreset) 구축과 데이터셋 증류 기술을 중심으로 발전해 왔다. 작은 크기의 데이터셋을 만들기 위한 직관적인 방법을 생각해 보면, 원본 데이터셋에서 가장 대표적이거나 가치 있는 샘플만을 선택하여 데이터셋을 만드는 것인데 이러한 방법을 코어셋 구축이라고 한다. 하지만 코어셋 구축은 샘플 평가 지표를 이용하여 원본 데이터셋의 샘플을 선택한다고 할지라도 핵심 샘플이 존재함을 보장할 수 없고 성능에 영향을 미칠 수 있는 샘플의 상당 부분을 버리기 때문에 최적의 증류 데이터셋을 구축하지 못한다.[2)]

한편 데이터셋 증류 기술은 코어셋 구축과는 달리 학습한 원본 데이터셋의 정보를 토대로 새로운 샘플을 합성하기에 최적의 증류 데이터셋을 구축할 수 있다는 것이 가장 큰 장점이며 이러한 최적의 증류 데이터셋으로 학습된 모델의 실제 성능은 원본 데이터셋으로 학습된 모델의 성능과도 상당히 근접하고 있다.

1) 윤영석 외 2명, “도메인 일반화를 위한 다중-소스 데이터 증류”, 2025 한국컴퓨터종합학술대회 논문집, 2025.07.
2) 김동훈, 배성호, “샤플리 기반 의존 특징 제어를 통한 데이터셋 증류”, 2024년 한국방송·미디어공학회 하계학술대회, 2024.06.

| 데이터셋 증류 기술 개요



※ 출처: 김동훈, 배성호, “샤플리 기반 의존 특징 제어를 통한 데이터셋 증류”, 2024년 한국방송·미디어공학회 하계학술대회, 2024.06.

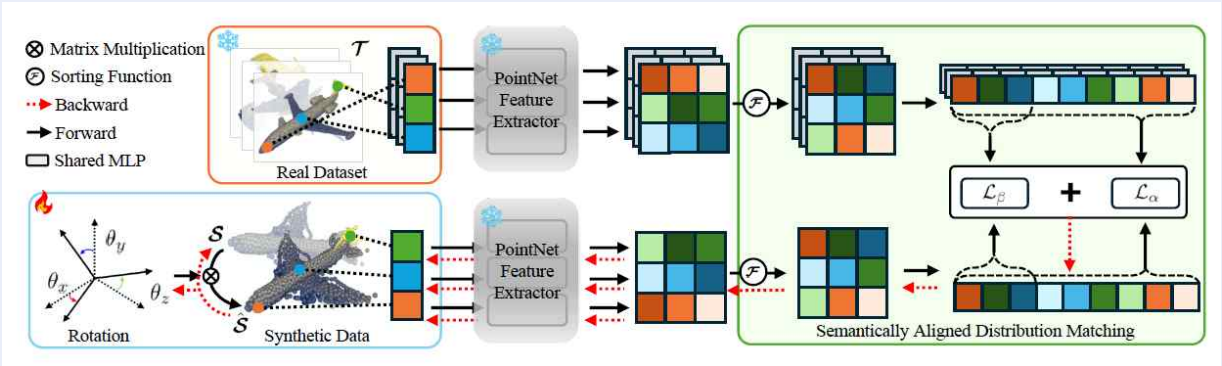
국내에서는 2025년 UNIST 인공지능대학원 심재영 교수팀이 3D 포인트 클라우드(Point Cloud) 데이터를 효과적으로 압축해 학습 효율을 높이는 데이터셋 증류 기술을 개발하여 3대 인공지능 분야 권위 국제학회인 ‘신경정보처리시스템학회(NeurIPS) 2025’에 정식 논문으로 채택됐다.

3D 포인트 클라우드 데이터는 사물을 점으로 표현해 놓은 데이터인데, 점들의 배열에 정해진 순서가 없고, 물체가 회전해 있는 경우가 많은 특성 때문에 데이터셋 증류 기술 적용이 까다로운 형태의 데이터로 꼽힌다. 연구팀은 이 문제를 해결한 데이터셋 증류 기술을 개발했는데 순서가 제각각인 점 데이터의 의미 구조를 자동으로 정렬해주는 SADM(Semantically Aligned Distribution Matching) 손실 함수[3]와 물체의 회전 각도를 인공지능이 스스로 최적화해 학습하도록 하는 방향 최적화(learnable rotation) 기법이 적용된 기술이다.

심재영 교수는 “이번 기술은 3D 점 데이터의 무질서한 구조와 회전 불확실성으로 인해 기존 기술들이 겪던 매칭 오류를 근본적으로 해결한 것”이라며 “자율주행, 드론, 로봇, 디지털 트윈 등 대규모 3D 데이터 활용이 필요한 분야에서 AI 학습 비용과 시간을 크게 줄이는 데 기여할 수 있을 것”이라고 말했다.[4]

3) SADM(Semantically Aligned Distribution Matching) 손실 함수: 3D 포인트 클라우드의 특징값을 채널별로 크기 순서에 따라 정렬해, 의미적으로 비슷한 부분끼리 자연스럽게 대응되도록 만든 손실 함수
4) UNIST News Center, “학습 시간↓, 성능은 유지” 3D AI 모델의 고효율 학습 데이터 경량화 기술 공개“, 2025.12.12.

| 3D 포인트 클라우드의 데이터셋 증류 기술 개요



※ 출처: Jae-Young Yim 외 2명, “Dataset Distillation of 3D Point Clouds via Distribution Matching”, NeurIPS, 2025.

3. 데이터셋 증류 기술의 문제점

앞서 설명했듯이 데이터셋 증류 기술을 이용하여 최적의 증류 데이터셋을 만들게 되면, 원본 데이터셋의 핵심 지식을 보존하면서도 컴퓨팅 비용을 대폭 절감하기 때문에 효율적인 학습을 진행할 수 있다. 그러나 이러한 증류 데이터셋의 긍정적인 면이 오히려 문제점으로 작용할 수 있다고 한다.

증류 데이터셋의 문제점은 정보가 농축되어 크기가 매우 작다는 특성 때문에 복제와 재배포가 원본 데이터셋에 비해 굉장히 쉬워진다는 점이다. 대규모 데이터셋은 용량 자체가 매우 크기 때문에 해킹을 통해 무단 복제되거나 전송하는 것 자체가 상당히 지연되거나 곤란한 측면이 있지만 증류 데이터셋은 용량이 줄어든 만큼 복제와 전송이 쉽게 이루어진다.

또한 가장 큰 문제점으로 지적되는 점은 기존에 수행되던 데이터셋 보호 기법은 대부분 원본 데이터셋을 대상으로 설계되었기 때문에 증류 데이터셋에 적용되기 어렵다는 점이다. 대표적으로 워터마킹과 같은 보호 기법은 원본 데이터셋의 일부 샘플에 워터마크 또는 특정 패턴을 삽입한 뒤에 이를 학습한 모델이 워터마크 또는 특정 패턴과 연관된 출력을 나타내는지 확인하는 방식으로 이루어진다.[5)]

이러한 방식은 원본 데이터셋이 그대로 유지되거나, 워터마크가 삽입된 데이터 샘플이 학습 과정에 충분히 반영된다는 것을 전제로 한다. 그러나 데이터셋 증류 과정을 거치면 원본 데이터셋이 그대로 유지되는 것이 아니라, 모델 학습에 필요한 핵심 정보만을 담은 새로운 합성 데이터로 변환된다. 특히 학습에 유용한 특징을 중심으로 최적화되는데 이 과정에서 원본 데이터셋에

5) Yan Liang외 4인, “From Compression to Accountability: Harmless Copyright Protection for Dataset Distillation”, arXiv, 2026.05.15.

포함되어 있던 개별 샘플의 형태, 미세한 패턴이 그대로 유지된다고 보기 어렵다. 원본 데이터셋에 삽입된 워터마크가 증류 데이터셋으로 변환되는 과정에서 약화되거나 사라질 가능성이 있는 것이다. **즉, 원본 데이터셋에 삽입된 워터마크가 증류 데이터셋으로 변환되는 과정에서 워터마크가 모델 학습에 필수적인 정보로 판단되지 않으면 증류 데이터셋에 반영되지 않을 수 있다. 또한 워터마크가 일부 반영되더라도, 원본 데이터셋에서 의도한 방식과 동일하게 작동한다고 보장하기 어렵다.**

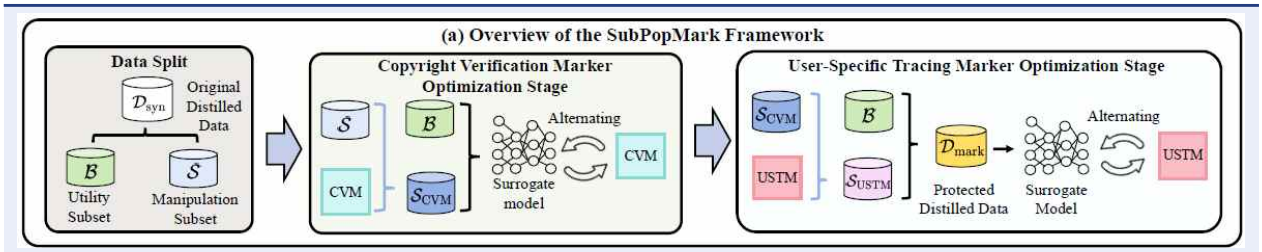
그렇다면 증류 데이터셋을 만든 이후에 별도로 워터마킹을 삽입하면 되는 것이 아니냐는 의문이 제기될 수 있다. 이론적으로는 증류가 완료된 데이터셋에 워터마크를 삽입하는 방식도 생각할 수 있다. 그러나 이 방식 역시 간단한 해결책이 되기는 어렵다.

증류 데이터셋은 적은 수의 샘플 안에 학습에 필요한 정보가 고도로 농축되어 있기 때문에 워터마크와 같은 새로운 패턴이나 표식을 추가하면 데이터셋이 본래 가지고 있던 학습 성능을 훼손할 수 있다는 것이다. 만약 학습 성능에 영향을 주지 않을 정도로 약하게 워터마크를 삽입한다고 할지라도, 이후 해당 데이터셋으로 학습한 모델에서 워터마크를 안정적으로 검출하기 어려울 수 있다. 대규모 데이터셋의 경우에는 다수의 샘플에 워터마크를 분산하여 삽입할 수 있지만, 증류 데이터셋은 샘플 수 자체가 매우 적기 때문에 일부 변형만으로도 학습 성능에 영향을 미칠 수 있고 워터마크 검출 가능성 또한 크게 달라질 수 있다.

4. 새로운 저작권 보호 프레임워크

2026년 5월 최근 연구에서는 증류 데이터셋의 보호에 초점을 맞춰 단순히 원본 데이터셋에 사용하던 기존의 보호 기법이 아닌, 증류 데이터셋의 생성 과정과 학습 특성을 고려한 새로운 저작권 보호 프레임워크 'SubPopMark(SubPopulation-based Mark)'가 제안되었다.

| 저작권 보호 프레임워크 'SubPopMark' 개요

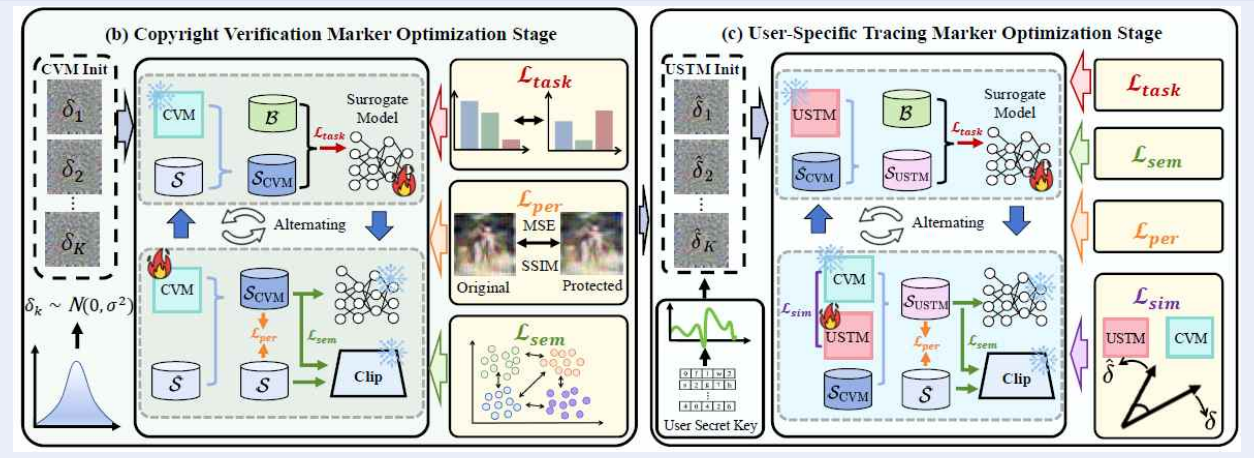


※ 출처: Yan Liang외 4인, "From Compression to Accountability: Harmless Copyright Protection for Dataset Distillation", arXiv, 2026.05.15.

'SubPopMark'의 핵심 아이디어는 인공지능 모델이 학습 과정에서 데이터셋 전체의 분포뿐만 아니라, 데이터셋 안에 존재하는 특정 하위 집단(Subpopulation)의 분포도 함께 기억한다는 점에 있다. 여기서 하위 집단이란 전체 데이터 중 특정한 공통 특성을 가진 일부 데이터 집단을 의미한다. 인공지능 모델은 학습 과정에서 이러한 하위 집단의 특징을 일정 부분 기억하게 되고, 이후 비슷한 특징을 가진 데이터에 대한 추론 단계에서 더 확실한 행동 경향(Behavior bias)을 보일 수 있다. 학습 데이터셋 내부에 밀도 높게 형성된 특정 하위 집단에 대해서는 더 뛰어난 성능과 행동에 대한 자신감을 보이게 된다는 뜻이다.[6]

'SubPopMark'는 이러한 특성을 활용하여 모델의 행동 경향에서 확인할 수 있는 새로운 식별 신호를 만드는 방식에 가깝다. 특정 하위 집단에 대한 모델의 행동 경향이 의도적으로 형성되도록 증류 데이터셋을 조정하게 되는데, 증류 데이터셋으로 학습한 모델이 특정 하위 집단의 데이터에 대해 보이는 반응을 하나의 행동 서명(Behavior Signature)처럼 활용하는 것이다. 이 행동 서명은 데이터셋의 출처를 확인하거나, 특정 사용자에게 제공된 증류 데이터셋이 외부로 유출되었는지를 추적하는 데 사용된다.

| 저작권 검증 마커 최적화 단계 | 사용자별 추적 마커 최적화 단계



※ 출처: Yan Liang외 4인, "From Compression to Accountability: Harmless Copyright Protection for Dataset Distillation", arXiv, 2026.05.15.

'SubPopMark'는 크게 두 단계로 구성된다. 첫 번째 단계는 저작권 검증 마커(Copyright Verification Marker, CVM)를 최적화하는 단계이다. 이 단계에서는 증류 데이터셋에 원래 학습 성능을 유지하면서도 특정 하위 집단에 대한 행동 경향이 나타나도록 저작권 검증 마커를 부여하는 것이다. 여기서 행동 경향은 모델이 특정 하위 집단에 속하는 데이터에 대해 상대적으로 일관된 행동을 보이도록 만드는 것을 의미하는데, 중요한 점은 이 과정이 모델의 일반적인 성능을 크게 해치지 않도록 설계된다는 것이다. 이 과정을 통해 증류 데이터셋을 무단 학습한 모델의 출력물에서 기존과 유사한 행동 경향이 나타나는지를 파악하여 증류 데이터셋의 사용 여부를 검증할 수 있다.

6) Yan Liang외 4인, "From Compression to Accountability: Harmless Copyright Protection for Dataset Distillation", arXiv, 2026.05.15.

두 번째 단계는 사용자별 추적 마커(User-Specific Tracing Marker, USTM)를 최적화하는 단계이다. 이는 동일한 증류 데이터셋을 여러 사용자에게 배포해야 하는 상황을 고려한 것인데 각 사용자에게 제공되는 증류 데이터셋에 사용자별로 다른 추적 마커를 부여하여 서로 구별 가능한 미세한 차이를 가지게 된다. 이 과정을 통해 이후 특정 증류 데이터셋이 무단으로 유출되었을 때 어떤 사용자에게 제공된 데이터셋에서 유출되었는지를 추적할 수 있다.[7)]

종합해보면 'SubPopMark'는 증류 데이터셋으로 학습한 모델의 행동 경향을 통해 데이터셋의 출처와 유출 경로를 확인하는 새로운 저작권 보호 프레임워크라고 할 수 있다. 이는 증류 데이터셋에서 기존의 데이터셋 보호 기법과 다른 방식으로 접근한 연구로 향후 증류 데이터셋 보호를 위한 또 다른 기술적 방향을 제시하였다.

5. 시사점

그동안 인공지능 기술은 방대한 양의 데이터를 활용하는 방향으로 발전해 왔으며, 그 과정에서 대규모 데이터셋은 인공지능 성능을 높이기 위한 핵심 자원으로 여겨져 왔다. 그러나 데이터의 규모가 커질수록 학습 비용과 시간이 대폭 증가하게 되었고 이러한 문제를 해결하기 위한 데이터셋 증류 기술이 등장하였다. 적은 양의 데이터만으로도 기존 대규모 데이터셋과 유사한 학습 효과를 얻을 수 있다는 점에서 인공지능 학습의 효율성을 크게 높일 수 있는 기술로 최근 많은 주목을 받고 있다.

하지만 데이터셋 증류 기술은 인공지능 학습의 효율성은 증가시켰지만 저작권 보호 기법을 적용하기 어렵다는 문제가 발생되었다. 증류 데이터셋은 크기가 작고 핵심 정보가 압축되어 있기 때문에 복제와 재배포가 쉬우며, 기존 원본 데이터셋을 대상으로 설계된 보호 기법만으로는 충분히 대응할 수 없다는 것이다. 이러한 점에서 'SubPopMark'와 같은 새로운 저작권 보호 프레임워크의 등장은 데이터셋 증류 기술이 단순히 학습 효율을 높이는 기술에 그치지 않고, 데이터의 보호와 책임 있는 활용까지 함께 고려해야 한다는 점을 보여준다. 앞으로 데이터셋 증류 기술의 성능과 효율성을 생각해 볼 때, 이 기술이 널리 활용되기 위해서 증류 데이터셋을 보호하는 기술에 관한 연구 또한 함께 발전해 나갈 것으로 예상된다.

7) Yan Liang외 4인, "From Compression to Accountability: Harmless Copyright Protection for Dataset Distillation", arXiv, 2026.05.15.

| 참고자료

- 윤영석 외 2명, “도메인 일반화를 위한 다중-소스 데이터 증류”, 2025 한국컴퓨터종합학술대회 논문집, 2025.07.
- 김동훈, 배성호, “샤플리 기반 의존 특징 제어를 통한 데이터셋 증류”, 2024년 한국방송·미디어공학회 하계학술대회, 2024.06.
- UNIST News Center, “학습 시간↓, 성능은 유지” 3D AI 모델의 고효율 학습 데이터 경량화 기술 공개“, 2025.12.12.
- Jae-Young Yim 외 2명, “Dataset Distillation of 3D Point Clouds via Distribution Matching”, NeurIPS, 2025.
- Yan Liang외 4인, “From Compression to Accountability: Harmless Copyright Protection for Dataset Distillation”, arXiv, 2026.05.15.