



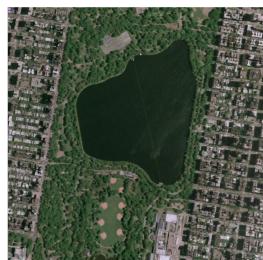
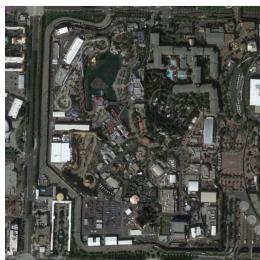
AID(Aerial Image Dataset) Usage Case

데이터셋 소개

AID는 Google Earth 이미지에서 샘플 이미지를 수집하여 생성한 대규모 항공 이미지 데이터입니다.

본 데이터셋은 1만장의 이미지로 구성되어있으며, 아래 30개 클래스로 분류되어있습니다.

airport, bare land, baseball field, beach, bridge, center, church, commercial, dense residential, desert, farmland, forest, industrial, meadow, medium residential, mountain, park, parking, playground, pond, port, railway station, resort, river, school, sparse residential, square, stadium, storage tanks and viaduct.



[데이터셋소개링크](#)

실험 소개

본 실험은 전체 데이터셋 중 랜덤으로 선별한 30% 데이터를 레이블링하여 validation set으로 고정한 상태에서 남은 70% 데이터 중 역시 랜덤으로 선별한 20%를 레이블링하여 학습 시킨 상황을 가정하여 시작합니다.

사용된 모델은 **efficient model**이며,

현재 10% 학습시킨 상황에서 모델의 정확도는 **90.1%**입니다.

본 실험의 목표는 가정된 상황에서 10%의 추가 학습 데이터를 선별하여 모델의 성능을 높이는 것입니다.

결과

요약

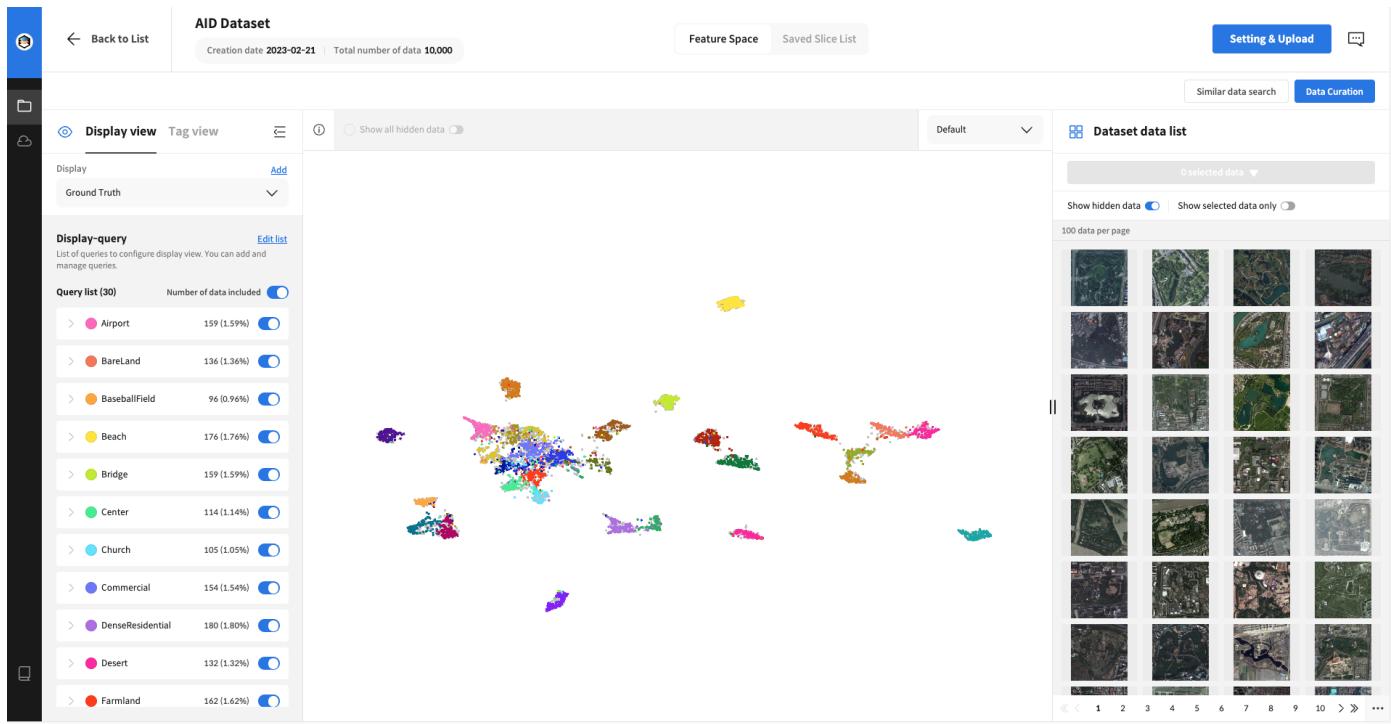
DATUMO FST 사용 여부	추가 학습 데이터 선별 방식	추가 학습데이터 비율	정확도
X	추가학습 전	0%	90.1%
X	랜덤 선별	10%	91.3%
X	랜덤 선별	15%	92.3%
O	분포 상 뭉친 영역 큐레이션	10%	92.6%
O	분포 상 분산된 클래스 선별	10%	92.7%
X	랜덤 선별	20%	93.0%

동일한 10% 추가 학습을 기준으로 보았을 때 DATUMO FST를 사용하여 분석한 결과 랜덤선별 대비 성능 향상이 2배 높게 된 점을 확인할 수 있습니다.

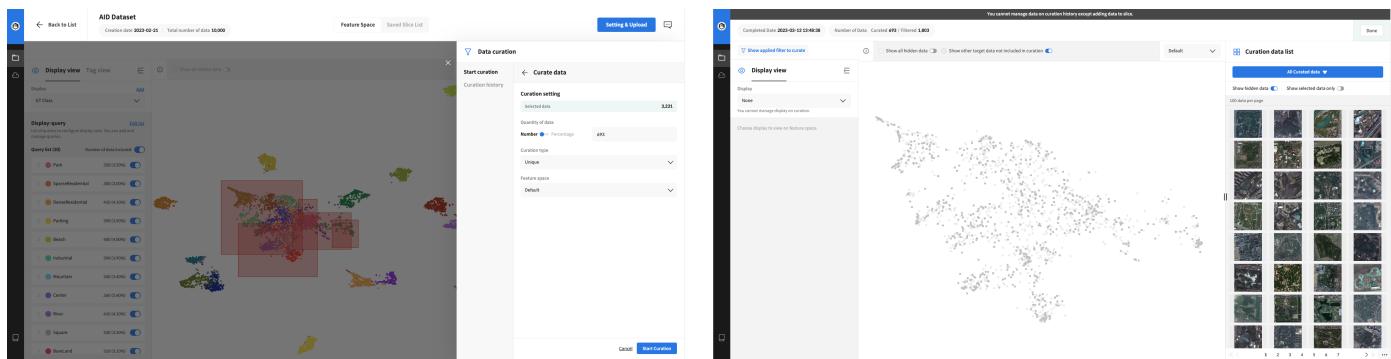
랜덤으로 학습 진행 시 비슷한 수준의 성능을 위해서는 학습데이터를 약 1.7배 정도 더 레이블링해야되는 점 역시 확인할 수 있습니다.

분포 상 뭉친 영역의 데이터 큐레이션

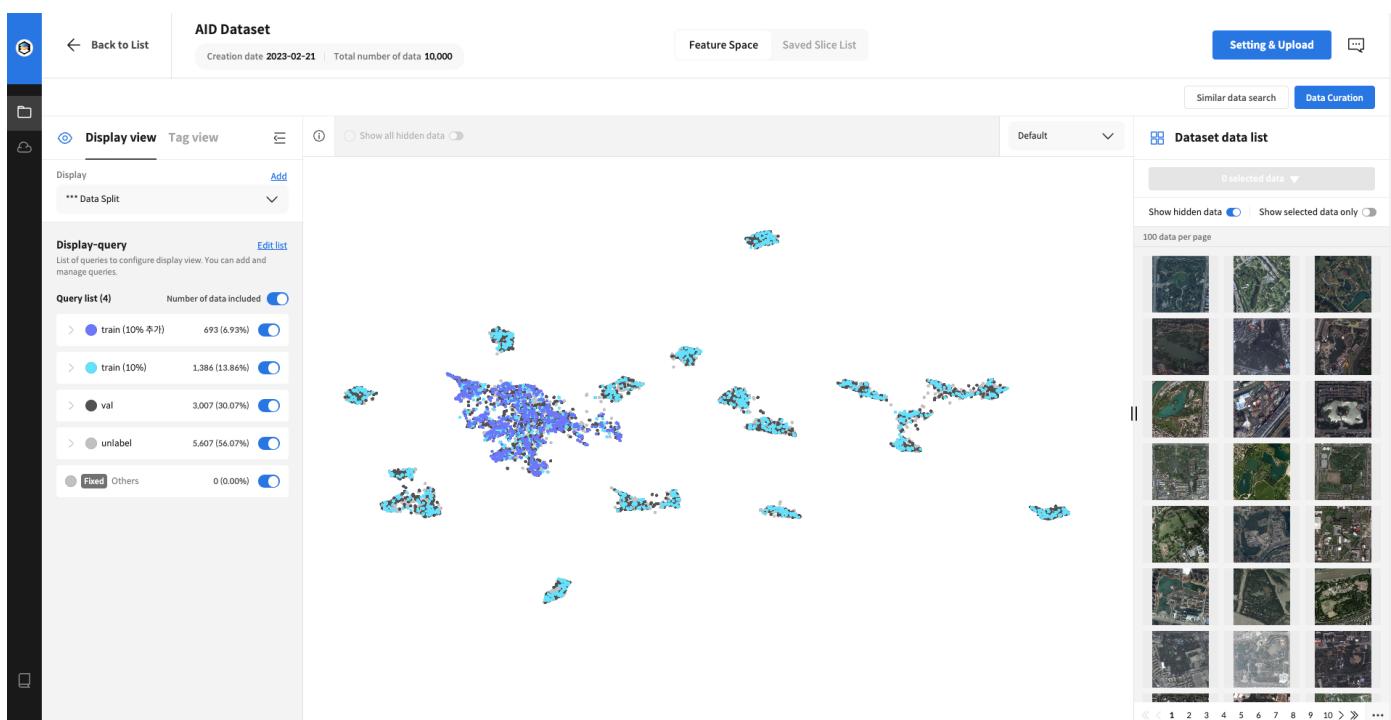
정확도 : 92.6%



이미지를 Datumo FST에 업로드하여 이미 레이블링된 train+validation set의 분포를 살펴본 결과
클래스별로 독립된 영역과, 여러 클래스가 뭉쳐있는 영역을 발견할 수 있었습니다.



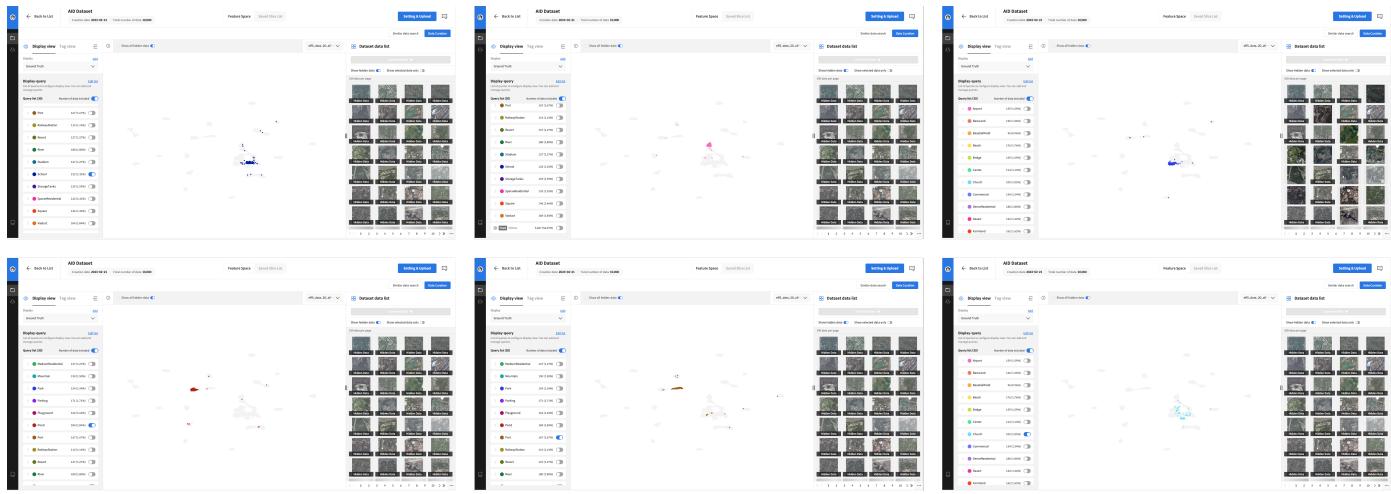
뭉쳐져 있는 영역을 선택한 후 큐레이션요청을 하여 선별된 10%의 데이터로 추가학습을 진행하였습니다.



최종 Data Spilt 이미지

분포 상 분산된 클래스 선별

정확도 : 92.7%



기존 20% 학습시킨 모델에서 뽑은 벡터로 피쳐스페이스를 살펴보며 분산되었다고 판단되는 클래스를 선별한 후, 남은 80%의 unlabeled 데이터를 기존 학습시킨 모델로 추론하여 선별된 클래스로 추론된 데이터를 큐레이션하여 10%의 추가학습데이터를 선별하였습니다.

[Edit this page](#)