

감독되지 않은 객체 감지 및 인스턴스 세분화를 위한 잘라내기 및 학습

Xudong Wang^{1,2} Rohit Girdhar¹ Stella X. Yu^{2,3} 1FAIR,
Meta AI 2UC Berkeley / ICSI 3University of Michigan

이산 미스라¹

코드: <https://github.com/facebookresearch/CutLER>

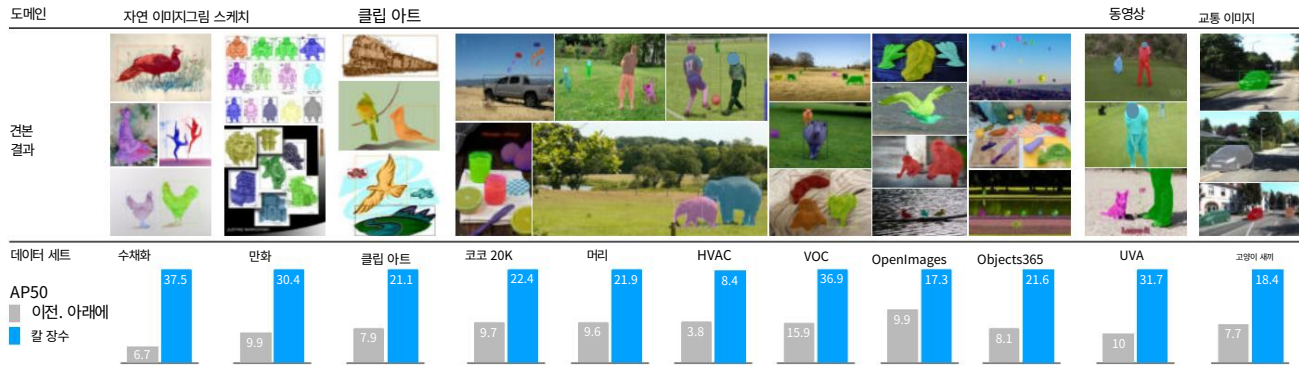


그림 1. 사람의 감독 없이 훈련된 CutLER 모델을 사용한 제로 샷 감독되지 않은 물체 감지 및 인스턴스 분할. 표준 탐지 APbox 50을 사용하여 모델을 평가합니다. CutLER는 비디오 프레임, 그림, 클립 아트, 복잡한 장면 등 다양한 이미지 영역에 걸쳐 11개 벤치마크에서 AP50을 2.7배 이상 향상시킵니다. CutLER는 모든 벤치마크에서 강력한 이득을 제공하여 11개 벤치마크 중 10개에서 성능을 2배 이상 향상시킵니다. 공식 코드와 체크포인트로 [47]을 평가합니다.

추상적인

우리는 감독되지 않은 객체 감지 및 분할 모델을 교육하기 위한 간단한 접근 방식인 Cut-and-LEaRn(CutLER)을 제안합니다. 자체 감독 모델의 속성을 활용하여 감독 없이 개체를 '발견'하고 이를 증폭하여 인간 레이블 없이 최첨단 지역화 모델을 교육합니다. CutLER는 먼저 제안된 MaskCut 접근 방식을 사용하여 이미지의 여러 개체에 대한 거친 마스크를 생성한 다음 강력한 손실 기능을 사용하여 이러한 마스크에서 검출기를 학습합니다. 우리는 예측에 대해 모델을 자체 훈련하여 성능을 더욱 향상시킵니다.

이전 작업과 비교하여 CutLER는 더 간단하고 다양한 감지 아키텍처와 호환되며 여러 물체를 감지합니다. CutLER는 또한 제로샷 비감독 감지기이며 비디오 프레임, 그림, 스케치 등과 같은 도메인 전반에 걸쳐 11개 벤치마크에서 감지 성능 AP50을 2.7배 이상 향상시킵니다. 미세 조정을 통해 CutLER는 MoCo-v2를 능가하는 로우샷 감지기 역할을 합니다. 5% 레이블로 훈련할 때 COCO에서 7.3% APbox 및 6.6% APmask.

1. 소개

개체 현지화는 AI 시스템이

객체 중심 방식. 지역화를 위한 교육 모델에는 개체 상자, 마스크, 지역화된 점 등과 같은 특별한 주석이 필요하며 수집하기 어렵고 리소스도 많이 사용됩니다. 오버헤드를 고려하지 않고 단 80개의 클래스에 대한 마스크를 사용하여 COCO 데이터 세트 [32]에서 ~164K 이미지에 주석을 추가하는 데 28K 인간 시간 이상의 주석 시간이 걸렸습니다. 이 작업에서 우리는 인간 레이블 없이 학습할 수 있는 비지도 객체 감지 및 인스턴스 분할 모델을 연구합니다. 우리의 핵심 통찰력은 간단한 프로빙 및 교육 메커니즘이 자체 감독 모델의 고유한 지역화 능력을 증폭하여 최첨단 무감독 제로샷 검출기로 이어질 수 있다는 것입니다.

우리의 방법 Cut-and-LEaRn (CutLER)은 세 가지 간단한 아키텍처 및 데이터 불가지론적 메커니즘으로 구성됩니다. 이전의 자기 지도 학습 방법 [7-9, 26]과 일치하는 CutLER는 추가 교육 데이터 없이 레이블이 지정되지 않은 ImageNet 데이터에 대해서만 교육을 받지만 이러한 방법과 달리 CutLER는 복잡한 분할 및 감지 작업을 수행하는 데 직접 사용할 수 있습니다. 광범위한 도메인에 대해. 먼저, 미리 학습된 self-supervised feature를 사용하여 각 이미지에 대해 여러 초기 거친 마스크를 자동으로 생성할 수 있는 MaskCut을 제안합니다. 둘째, 우리는 MaskCut에서 놓친 개체에 강력하면서도 거친 마스크를 사용하여 검출기를 훈련시키는 간단한 손실 감소 전략을 제안합니다. 마지막으로, 우리는

이러한 조악한 마스크에서 학습한 탐지기는 실측 정보를 '정리'하고 이를 훈련하는 데 사용되는 거친 마스크보다 더 나은 마스크(및 상자)를 생성합니다. 따라서 우리는 모델 자체 예측에 대한 여러 라운드의 자체 훈련을 통해 로컬 픽셀의 유사성을 캡처하는 것에서 객체의 글로벌 기하학을 캡처하는 것으로 진화하여 더 미세한 분할 마스크를 생성할 수 있음을 보여줍니다.

이전 작업에서는 자체 감독 비전 변환기(ViT) [15]가 이미지에서 단일 두드러진 개체를 감지하는 패치별 기능을 자동으로 학습할 수 있음을 보여줍니다 [7,38,43,44,50]. 그러나 CutLER와 달리 이러한 두드러진 객체 감지 방법은 일반적으로 가장 눈에 띄는 단일 객체만 찾고 여러 객체를 포함하는 실제 이미지에는 사용할 수 없습니다. 예를 들어 FreeSOLO [47] 및 DETReg [3]와 같은 일부 최근 방법은 감도되지 않은 다중 객체 감지(또는 다중 객체 검색)를 목표로 하지만 특정 감지 아키텍처(예: SOLO-v2 [48] 또는 DDETR [5,54]. 또한 ImageNet [12]에서 훈련된 자체 감독 기능 외에도 현재의 최첨단 방법인 FreeSOLO 및 MaskDistill [42]에는 모델 훈련을 위해 '도메인 내' 레이블이 지정되지 않은 데이터가 필요합니다.

대조적으로, CutLER는 다양한 감지 아키텍처와 함께 작동하며 레이블이 지정되지 않은 도메인 내 데이터 없이 ImageNet에서만 학습할 수 있습니다. 따라서 모델 교육 중에 CutLER는 모든 대상 데이터 세트의 이미지를 보지 않고 다양한 도메인에서 여러 개체를 감지하고 분할할 수 있는 제로 샷 모델을 생성합니다.

CutLER의 특징. 1) 단순성: CutLER는 훈련이 간단하고 감지 및 백본 아키텍처 선택에 구애받지 않습니다. 따라서 기존 객체 감지 및 인스턴스 분할 작업에 쉽게 통합될 수 있습니다.

2) 제로샷 검출기: ImageNet에서만 훈련된 CutLER는 추가 도메인 내 데이터로 훈련된 이전 작업을 능가하는 11개의 서로 다른 벤치마크에서 강력한 제로샷 성능을 보여줍니다. 우리는 APbox를 두 배로 늘립니다

50 그림 1에 표시된 대로 이러한 벤치마크 중 10개에서 성능이 향상되었으며 심지어 UVO 비디오 인스턴스 분할 벤치마크에서 감도된 탐지기보다 성능이 뛰어났습니다.

3) 견고성: CutLER는 비디오 프레임, 스케치, 그림, 클립 아트 등과 같은 다른 도메인의 이미지에 대해 테스트할 때 도메인 이동에 대해 강력한 견고성을 나타냅니다. 완전히 감도된 개체 감지 및 인스턴스 분할 모델을 교육하고 소수 개체 감지 벤치마크를 포함하여 COCO의 성능을 향상시킵니다.

2. 관련 업무

자기 감독 기능 학습에는 사람이 주석을 단 레이블을 사용하지 않고 레이블이 지정되지 않은 대규모 데이터 내에서 패턴을 추론하는 것이 포함됩니다. 대조적 학습 기반 [8, 26, 34, 52] 방법은 유사한 샘플 또는 동일한 인스턴스의 다양한 증강이 서로 가깝고 유사하지 않은 인스턴스가 멀리 있다는 표현을 학습합니다.

	DINO	LOST	TokenCut	FreeSOLO	시간
여러 객체 감지	✓	제로샷 감지기		✓✓	
✓ 다양한 감지 아키텍처와 호환			✓	✓	
	-✓	-		✓	
감독된 탐지를 위한 사전 훈련된 모델	✓			✓✓	

표 1. DINO [7], LOST [38], TokenCut [50] 및 FreeSOLO [47]를 포함하여 감도되지 않은 물체 감지에 대한 이전 방법을 주요 속성 측면에서 CutLER와 비교합니다. 당사의 CutLER는 이러한 원하는 특성을 모두 갖춘 유일한 방법입니다.

따로. 유사성 기반 자기 지도 학습 방법 [10, 23]은 동일한 인스턴스의 서로 다른 증강 사이의 거리를 최소화하여 표현을 학습하고 양성 샘플 쌍만 사용합니다. 클러스터링 기반 기능 학습 [1, 6, 46, 53, 55]은 잠재적 표현 공간에서 데이터의 자연스러운 그룹화를 자동으로 발견합니다. 최근에 [2,25]는 이미지 패치의 대규모 무작위 하위 집합을 마스크하고 누락된 픽셀 또는 패치를 재구성하여 표현을 학습하는 마스크된 자동 인코더가 확장 가능한 자기 감독 학습자임을 보여주었습니다[2, 13, 14, 25]. 컴퓨터 비전용 [25].

이러한 감도되지 않은 표현 학습 노력과 달리 우리 작업은 자연스러운 픽셀 그룹을 자동으로 발견하고 각 이미지 내에서 인스턴스를 찾는 것을 목표로 합니다.

감도되지 않은 물체 감지 및 인스턴스 분할. 이전 작업과의 주요 비교는 표 1에 나열되어 있으며 다음과 같이 자세히 설명되어 있습니다.

DINO [7]는 이미지의 기본 의미 분할이 감도 ViT 또는 ConvNets [7, 56]에 명시적으로 나타나지 않는 자체 감독 ViT(Vision Transformer) [15]에서 나타날 수 있음을 관찰합니다. 이 관찰을 기반으로 LOST [38] 및 TokenCut [50]은 자체 감독 ViT 기능을 활용하고 DINO의 패치 기능으로 구성된 그래프를 기반으로 각 이미지에서 하나의 주요 객체 [11,38,50]를 분할할 것을 제안합니다.

이러한 이전 작업은 각 이미지(예: DINO 및 TokenCut)에서 둘 이상의 개체를 감지할 수 없거나 다운스트림 감지 및 세분화 작업(예: TokenCut 및 LOST)으로 더 나은 전송을 위해 기능의 품질을 개선할 수 없습니다. 이러한 작업과 달리 CutLER는 여러 개체를 찾을 수 있으며 레이블 효율적이고 완전 지도 학습을 위한 사전 훈련된 모델 역할을 할 수 있습니다.

FreeSOLO [47]는 감도되지 않은 방식으로 거친 객체 마스크를 추출한 다음 자체 훈련 절차를 통해 마스크를 세분화하여 감도되지 않은 인스턴스 분할을 달성합니다. FreeSOLO의 FreeMask 단계는 이미지만 여러 개의 거친 마스크를 생성할 수 있지만 이러한 마스크의 품질은 종종 다소 낮습니다 [47]. MaskDistill [42]은 자체 감독 DINO [7]에 의해 생성된 친화도 그래프에서 클래스에 구애받지 않는 초기 마스크를 추출합니다. 그러나 증류 단계에서 이미지만 하나의 단일 마스크를 사용하므로

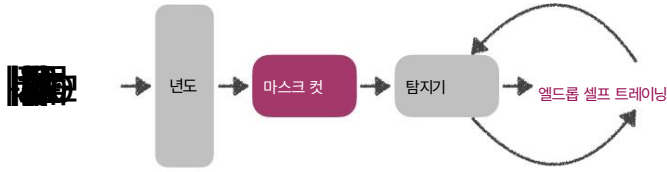


그림 2. CutLER 개요. 감독을 사용하지 않고 개체 감지 및 인스턴스 분할 모델을 교육하는 간단하면서도 효과적인 방법을 제안합니다. 먼저 자체 감독 ViT의 기능에서 초기 거친 마스크를 추출하는 MaskCut을 제안합니다. 그런 다음 MaskCut에서 놓친 개체에 강력한 손실 삭제 전략을 사용하여 검출기를 학습합니다. 여러 라운드의 자체 학습을 사용하여 모델을 더욱 개선합니다.

여러 개체를 감지하는 모델의 기능을 크게 제한합니다.

대조적으로, Mask Cut에 의해 생성된 초기 마스크는 일반적으로 [42, 47]에서 사용된 초기 마스크보다 품질과 양에서 더 좋습니다. 따라서 CutLER는 거의 모든 실험된 감지 및 분할 벤치마크에서 FreeSOLO [47] 및 MaskDistill [42]보다 2배 ~ 4배 더 높은 APbox 및 APmask를 달성합니다.

3. 방법

우리는 간단한 잘라내기 및 학습 파이프라인을 사용하여 감독되지 않은 객체 감지 및 세분화 문제를 해결합니다. 우리의 방법은 최근 작업[7,50]의 통찰력을 기반으로 하며, 자기 감독 표현이 객체를 발견할 수 있음을 보여줍니다. 이러한 방법은 이미징 단일 개체를 찾는 경우가 많지만 여러 개체를 검색하고 세분화 및 감지 성능을 크게 향상시킬 수 있는 간단한 접근 방식을 제안합니다. 절단 및 학습 파이프라인의 개요는 그림 2에 설명되어 있습니다. 먼저 DINO [7] (Sec. 3.2)의 자체 감독 기능을 사용하여 이미징 여러 이진 마스크를 생성하는 MaskCut을 제안합니다.

둘째, 우리는 MaskCut의 초기 마스크에서 검출기를 학습할 수 있는 DropLoss라는 동적 손실 감소 전략을 보여주며 모델이 MaskCut에서 놓친 객체를 탐색하도록 권장합니다(Sec. 3.3). 셋째, 여러 라운드의 자가 훈련을 통해 방법의 성능을 더욱 향상시킵니다(Sec. 3.4).

3.1. 예선

NCut(Normalized Cuts)은 이미지 분할 문제를 그래프 분할 작업으로 처리합니다 [37]. 각 이미지를 노드로 표현하여 완전히 연결된 무향 그래프를 구성합니다. 각 노드 쌍은 연결된 노드의 유사성을 측정하는 가중치 W_{ij} 가 있는 에지로 연결됩니다. NCut은 일반화된 고유값 시스템을 해결하여 그래프를 두 개의 하위 그래프, 즉 이분할로 분할하는 비용을 최소화합니다.

$$(D - W)x = \lambda Dx \quad (1)$$

두 번째로 작은 고유값 λ 에 해당하는 고유벡터 x 를 찾기 위해, 여기서 D 는 $d(i) = \sum_j W_{ij}$ 인 $N \times N$ 대각 행렬이고 W 는 $N \times N$ 대칭 행렬입니다.

DINO 및 TokenCut. DINO [7]는 자체 감독 ViT가 이미지 패치의 자각 그룹화를 어느 정도 자동으로 학습할 수 있음을 발견했습니다. TokenCut [50]은 NCut용 DINO 기능을 활용하고 이미지의 전경/배경 세그먼트를 얻습니다. 저자는 DINO 기능 공간에서 패치의 유사성을 NCut의 유사성 가중치 W_{ij} 로 사용합니다. 구체적으로, 최근의 여러 방법 [38, 42, 50]에 따라 KiKj DINO 사전 훈련 모델의 마지막 주의 계층에서 '주요' 기능의 코사인 유사성을 사용합니다. 즉, $W_{ij} = \text{여기서 } K_i K_j^T$ 는 '키'입니다. 패치의 특징과 식을 해결합니다. (1) 두 번째로 작은 고유 벡터 x 를 찾기 위해.

TokenCut의 한계는 이미지에 대한 단일 이진 마스크만 계산하므로 이미징당 하나의 개체만 찾는다는 것입니다. 다른 $N - 2$ 개의 가장 작은 고유 벡터를 사용하여 둘 이상의 인스턴스를 찾을 수 있지만 Sec. 5.

3.2. 여러 개체 검색을 위한 MaskCut

우리가 Sec에서 논의한 것처럼, 3.1, 바닐라 NCut은 이미지에서 단일 개체를 검색하는 것으로 제한됩니다. 마스크링된 유사도에 NCut을 반복적으로 적용하여 이미징당 여러 개체를 발견하도록 NCut을 확장하는 마스크 컷을 제안합니다.

trix(그림 3 참조). 단계 t 에서 NCut에서 이중 파티션 x 를 얻은 후 두 개의 분리된 패치 그룹을 얻고 이진 마스크 M_t 를 구성합니다.

$$M_{t,ij} = \begin{cases} 1, & \text{후지산인 경우} \\ 0, & \text{그렇지 않으면 0입니다.} \end{cases} \geq \text{평균}(x) \quad (2)$$

전경에 해당하는 그룹을 결정하기 위해 두 가지 기준을 사용합니다. 1) 직관적으로 전경 패치가 배경 패치보다 더 두드러져야 합니다 [7, 43, 50]. 따라서 전경 마스크는 두 번째로 작은 고유 벡터 M_t 의 최대 절댓값에 해당하는 패치를 포함해야 합니다. 2) 우리는 단순하지만 경험적으로 효과적인 객체 중심 사전 [33]을 기업으로 사용합니다. 전경 세트는 네 모서리 중 두 개 미만을 포함해야 합니다. 전경과 배경의 분할을 뒤집습니다. 즉, 기준 ij 인 경우 $M_t = 1 - M_t$ 입니다.

현재 전경 세트에 두 개의 모서리가 포함되어 있거나 조건 2가 충족되지 않는 동안 1은 만족되지 않습니다. 실제로 우리는 $\tau \sim 1e^{-5}$ 그리고 우리 $\geq \tau$ 1. 또한 모든 $W_{ij} < \tau$ 로 설정합니다.

($t+1$) 번째 객체 에 대한 마스크를 얻기 위해 이전 단계에서 전경에 해당하는 노드를 마스크 아웃하여 노드 유사성 W_{t+1} 을 업데이트합니다.

$$W_{t+1,ij} = \frac{(W_{t,ij} - \tau) \cdot \mathbb{1}_{W_{t,ij} > \tau}}{K_i K_j^T} \quad (3)$$

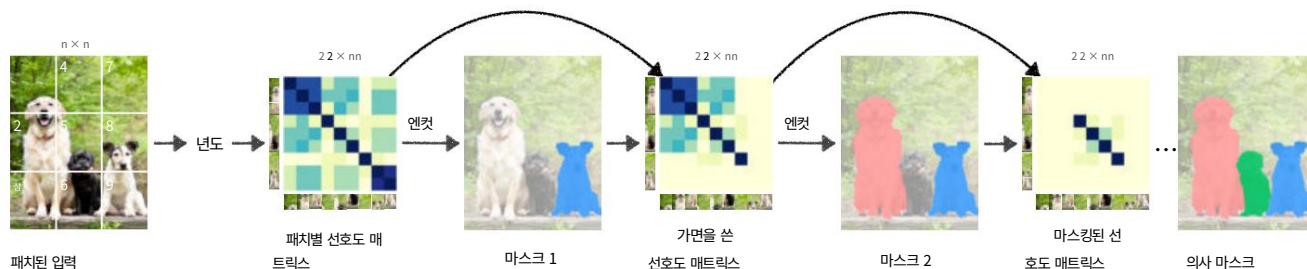


그림 3. MaskCut은 감독 없이 이미지에서 여러 개체 마스크를 검색할 수 있습니다. 우리는 [7, 50]을 기반으로 자체 감독 DINO [7] 모델의 기능을 사용하여 이미지에 대한 패치별 유사성 매트릭스를 생성합니다. Normalized Cuts [37]를 이 행렬에 적용하고 이미지의 단일 전경 개체 마스크를 얻습니다. 그런 다음 전경 마스크를 사용하여 선호도 매트릭스 값을 가지고 프로세스를 반복하여 MaskCut이 단일 이미지에서 여러 개체 마스크를 검색할 수 있도록 합니다. 이 파이프라인 그림에서는 $n=3$ 으로 설정합니다.

[illegible]

예측된 마스크의 IoU > 0.5. 우리는 세 번의 자가 훈련이 좋은 성과를 내기에 충분하다는 것을 발견했습니다. 각 라운드는 모델 훈련에 사용되는 '실측 정보' 샘플의 수를 꾸준히 증가시킵니다.

3.3. 이미지 영역 탐색을 위한 DropLoss

표준 검출 손실은 '실측 정보'와 겹치지 않는 예측 영역 r_i 에 패널티를 부여합니다. MaskCut에서 제공한 'ground-truth' 마스크는 자세를 놓칠 수 있기 때문에 표준 손실로 인해 감지기가 'ground-truth'에 레이블이 지정되지 않은 새 인스턴스를 발견할 수 없습니다.

따라서 우리는 'ground-truth'와 약간 겹치는 예측 영역 r_i 의 손실을 무시할 것을 제안합니다.

보다 구체적으로, 훈련 중에 우리는 'ground-truth' 인스턴스 중 하나와 최대 중첩이 τ 인 각 예측 영역 r_i 에 대한 손실을 삭제합니다.

$$L_{\text{drop}}(r_i) = (\log U_{\text{max}} > \tau \log U) L_{\text{vanilla}}(r_i) \quad (4)$$

여기서 IoU_{max} 는 r_i 에 대한 모든 '실측 정보'를 포함하는 최대 IoU를 나타내고 $L_{vanilla}$ 는 검출기의 바닐라 손실 함수를 나타냅니다. L_{drop} 은 'ground-truth'에서 놓친 물체를 감지하는 모델에 불이익을 주지 않으므로 다른 이미지 영역의 탐색을 장려합니다. 실제로 IoU는 낮은 임계값 $\tau = 0.01$ 을 사용합니다.

3.4. 다단계 자가 훈련

경험적으로 우리는 MaskCut에서 얻은 거친 마스크로부터 학습했음에도 불구하고 탐지 모델이 실측 정보를 '정리'하고 훈련에 사용된 초기 거친 마스크보다 더 나은 마스크(및 상자)를 생성한다는 것을 발견했습니다. 검출기는 마스크 품질을 개선하고 DropLoss 전략은 새로운 개체 마스크를 발견하도록 권장합니다. 따라서 우리는 이 속성을 활용하고 감지기의 성능을 향상시키기 위해 여러 라운드의 자가 훈련을 사용합니다.

우리는 $(t+1)$ 차 자체 학습을 위한 추가 의사 주석으로 t 차에서 신뢰도 점수가 0.75-0.5t 이상인 예측된 마스크와 제안을 사용합니다. 라운드 t 의 예측과 실측을 중복 제거하기 위해 다음을 사용하여 실측 마스크를 필터링합니다.

3.5. 구현 세부 정보

훈련 데이터. 우리는 CutLER 모델의 모든 부분에 대해 ImageNet [12] 데이터셋(130만 이미지)의 이미지만 사용하고 훈련 또는 사전 훈련된 감독 모델에는 어떤 유형의 주석도 사용하지 않습니다.

마스크 컷. 우리는 480×480 픽셀로 크기가 조정된 이미지에 대해 3단계로 MaskCut을 사용하고 ViT-B/8 [15] DINO [7] 모델을 사용하여 패치별 친화성 매트릭스를 계산합니다. CRF(Conditional Random Field) [30]를 사용하여 마스크를 후처리하고 경계 상자를 계산합니다.

탐지기. CutLER는 기본 검출기에 대해 불가지론적이지만 모든 실험에 널리 사용되는 Mask R-CNN [27] 및 Cascade Mask R-CNN [4]을 사용하고 달리 명시되지 않는 한 기본적으로 Cascade Mask R-CNN을 사용합니다. 우리는 배치 크기 16으로 160K 반복에 대한 초기 마스크와 경계 상자를 사용하여 ImageNet에서 검출기를 훈련합니다. ResNet-50 백본 [28] 으로 검출기를 훈련할 때 자체 지도 사전 훈련된 DINO의 가중치로 모델을 초기화합니다. [7] 모델. 우리는 MoCo-v2 [9], SwAV [6], CLD [46] 를 포함한 다른 사전 훈련된 모델을 탐색했고 유사한 탐지 성능을 제공한다 것을 발견했습니다. 또한 모델 훈련 과정에서 복사-붙여넣기 확대 [16, 19] 를 활용합니다. 바닐라 복사-붙여넣기 확대를 사용하는 대신 작은 개체를 분할하는 모델의 기능을 개선하기 위해 0.3과 1.0 사이에서 균일하게 샘플링된 스칼라로 마스크를 무작위로 다운샘플링합니다. 그런 다음 SGD를 사용하여 160K 반복에 대한 검출기를 최적화합니다. 학습률은 0.005이며, 학습률은 80K 반복 후 5씩 감소하고 배치 크기는 16입니다. 5×10^{-5} 의 가중치 감쇠와 0.9의 모멘텀을 적용합니다.

자기 훈련. 이전 단계의 가중치를 사용하여 각 단계에서 탐지 모델을 초기화합니다. 우리는 80K 반복에 대해 0.01의 학습률로 SGD를 사용하여 검출기를 최적화합니다. 셀프 트레이닝 단계를 제공할 수 있기 때문에

5

행동 양식	프리트레인 검출기 초기화	코코 20K						COCO 발2017					
		AP ₅₀	AP ₇₅	AP ₅₀	AP ₇₅	AP ₅₀	AP ₇₅	AP ₅₀	AP ₇₅	AP ₅₀	AP ₇₅	AP ₅₀	AP ₇₅
제로 샷이 아닌 방법 LOST [38]													
	IN+COCO FRCNN DINO MaskDistill	-	-	-	2.4	1.0	1.1	-	-	-	-	-	-
[42]	IN+COCO MRCNN MoCo FreeSOLO* [47] IN+COCO	-	-	-	6.8	2.1	2.9	-	-	-	-	-	-
SOLOv2 DenseCL 9.7 3.2 4.1 제로 샷 방법 DETReg [3]					9.7	3.4	4.3	9.6 3.1 4.2		9.4	3.3	4.3	
IN DDETR SWAV -													
		-	-	-	-	-	-	3.1 0.6 1.0		8.8	1.9	3.3	
디노 [7]	안에 디노 1.7 0.1 0.3	-	-	-	-	-	-	-	-	-	-	-	-
토큰 컷 [50]	안에 디노	-	-	-	-	-	-	5.8 2.8 3.0		4.8	1.9	2.4	
CutLER(당사)	MRCNN에서 디노 21.8 11.1 10.1 18.6					9.0	8.0	21.3 11.1 10.2 18.0			8.9	7.9	
CutLER(당사) vs.	IN 캐스케이드 DINO 22.4 12.5 11.9 19.6 10.0						9.2	21.9 11.8 12.3 18.9			9.7	9.2	
이전 책		+12.7 +9.3 +7.8 +9.9 +6.6 +4.9 +12.3 +8.7 +8.1 +9.5 +6.4 +4.9											

표 3. COCO 20K 및 COCO val2017에서 감도되지 않은 개체 감지 및 인스턴스 분할. 감지 및 분할 메트릭을 보고하고 사전 학습 데이터(Pretrain), 감지기 및 백본 초기화(Init.)를 기록합니다. 테이블의 위쪽 절반에 있는 메서드는 다운스트림 데이터 세트에서 레이블이 지정되지 않은 추가 이미지에 대해 학습하고 아래쪽 절반에 있는 제로 샷 메서드는 ImageNet에서만 학습합니다.

구형 감지기를 사용함에도 불구하고 CutLER는 모든 평가 메트릭에서 이전의 모든 작업보다 성능이 뛰어납니다. : 공식 코드와 체크포인트로 얻은 결과. IN, Cascade, MRCNN 및 FRCNN은 각각 ImageNet, Cascade Mask R-CNN, Mask R-CNN 및 Faster R-CNN을 나타냅니다.

행동 양식	AP50	AP75	AP	APs	APm	APL
OSD [43]	13.1	-	4.3	-	-	-
LOD [44]	13.9	-	4.5	-	-	-
잃어버린 [38]	19.8	-	6.7	-	-	-
프리솔로* [47]	15.9	3.6	5.9	0.0	2.0	9.3
CutLER(당사) vs.	36.9	19.2	20.2	1.3	6.5	32.2
이전 책	+17.1 +15.6 +13.5 +1.3 +4.5 +22.9					

표 4. VOC 에서 제로 샷 비감독 물체 감지. 공식 코드와 체크 포인트로 결과를 생성 : 다시 생성했습니다.

gin(종종 2 ~ 3배 더 높은 AP를 얻음)은 분할 및 탐지 작업 모두에서 발생합니다. CutLER는 COCO의 이미지에 대해 학습되지 않았지만 APmask 및 APbox 측면에서 COCO에 대해 학습된 기존 방법을 10% 이상 능가합니다.

그림 4 는 인간 주석과 함께 [47] 과 COCO val2017의 CutLER 사이의 정성적 비교를 보여줍니다 . 놀랍게도 CutLER는 종종 사람 어느 데이터가 놓친 새로운 인스턴스를 감지할 수 있습니다.

COCO 20K, COCO에 대한 상세한 비교를 제시합니다. 부록 A.3의 val2017 및 LVIS [24] 벤치마크 .

UVO와 VOC에 대한 자세한 비교. 기존의 감도되지 않은 다중 물체 감지 방법과 포괄적인 비교를 위해 UVO val [45] 및 VOC trainval07 [17] 에 대한 결과를 보고합니다 . 표 4 는 CutLER가 이전 SOTA에 비해 상당한 성능 향상을 제공하여 3배 이상의 더 높은 AP를 얻었으며 APL 에서 가장 눈에 띄는 개선이 이루어졌음을 보여줍니다 . UVO에서 표 5 는 CutLER가 이전 SOTA의 AP보다 4배 이상, APbox 50 의 거의 3배임을 보여줍니다 . 우리의 APmask는 100% 주석으로 LVIS에서 훈련된 완전 지도 SOLOv2 [48] 보다 4.8% 더 높으며 지도 학습과 비지도 학습 사이의 격차를 크게 줄였습니다.

행동 양식	AP ₅₀	AP ₇₅	AP ₅₀	AP ₇₅	AP ₅₀	AP ₇₅
완전 감독 방식: SOLO-v2(COCO 포함) [48]						
	-	-	-	38.0	20.9	21.4
Mask R-CNN (w/ COCO) [27] - SOLO-v2 (w/ LVIS) [48] 감도되지 않은 방법:	-	-	-	31.0	14.2	15.9 14.8
FreeSOLO* [47]				5.9	7.1	
	10.0	1.8	3.2	9.5	2.0	3.3 31.7 14.1 16.1 22.8
CutLER(당사) vs.	8.0	10.1	+21.7	+12.3	+12.9	+13.3 +6.0 +6.8
이전 책						

표 5. UVO val 비디오 벤치마크 에서 제로 샷 자율 객체 감지 및 인스턴스 분할. CutLER out은 이전의 감도되지 않은 방법을 수행하고 LVIS 데이터 세트에서 훈련된 감도된 SOLO-v2 모델보다 더 나은 성능을 달성합니다.

: 공식 코드와 체크포인트로 결과를 재현했습니다.

4.2. 레이블 효율적인 완전 지도 학습

우리는 이제 개체 감지 및 인스턴스 세분화 모델을 교육하기 위한 사전 교육 방법으로 CutLER를 평가합니다. CutLER는 감독 없이 개체를 발견할 수 있지만 대상 데이터 세트에서 이를 미세 조정하면 모델 출력을 데이터 세트에 레이블이 지정된 동일한 개체 세트에 정렬합니다.

설정. CutLER를 사용하여 ResNet50[28]으로 표준 캐스케이드 마스크 R-CNN [4] 검출기 를 초기화합니다 . 이전 작업에서는 [47] 에 사용된 SOLOv2 [48] 및 [3] 에 사용된 DDETR [54] 과 같이 더 나은 성능을 제공하는 고급 검출기를 사용합니다. 그러나 단순성을 위해 Cascade Mask R-CNN을 선택하고 Sec. 5 더 강력한 검출기로 CutLER의 성능이 향상됩니다. 경계 상자와 인스턴스 마스크 레이블을 사용하여 COCO [32] 데이터 세트 에서 탐지기를 훈련합니다 .

레이블 효율성을 평가하기 위해 교육 세트를 서브샘플링하여 레이블이 지정된 이미지의 비율이 다양한 하위 집합을 만듭니다. 이러한 각 하위 집합에서 CutLER로 초기화된 검출기를 훈련합니다. 기준선으로 MoCo-v2 [9] 의 설정을 따르고 MoCo-v2 ResNet50 모델로 초기화된 동일한 탐지 아키텍처를 교육합니다.

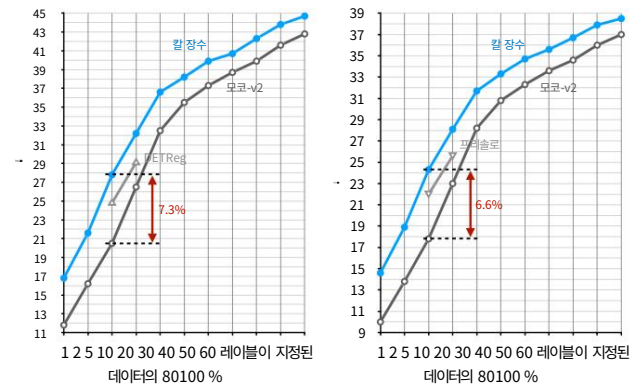


그림 5. 로우 샷 및 완전 감도 감지 및 인스턴스 세분화를 위한 미세 조정 CutLER. CutLER 또는 MoCo-v2로 초기화된 Cascade Mask R-CNN 모델을 미세 조정합니다.

COCO 데이터 세트에서 다양한 양의 레이블이 지정된 데이터. 자체 감도 사전 훈련된 MoCo-v2 카운터파트와 동일한 일정을 사용하고 탐지 및 인스턴스 분할 성능을 보고합니다. CutLER는 지속적으로 MoCo-v2 기준선을 능가합니다. 1% 레이블이 있는 로우 샷 설정과 100% 레이블을 사용하는 완전 감도 설정에서. CutLER는 또한 이전 감지 아키텍처를 사용함에도 불구하고 이 벤치마크에서 FreeSOLO [47] 및 DETReg [3] 보다 성능이 뛰어납니다. Mask R-CNN을 사용한 결과는 부록에 있습니다.

개체 감지 작업에 대한 성능. MoCo-v2와 우리 모델은 모두 1% 또는 2% 레이블이 있는 극도로 낮은 샷 설정을 제외하고 Detectron2 [51]를 사용하여 1x 일정에 대해 훈련됩니다. 이전 작업 [47]에 따라 1% 또는 2% 레이블로 교육할 때 MoCo-v2와 모델을 배치 크기 16으로 3,600회 반복 교육합니다.

결과. 그림 5는 COCO의 서로 다른 하위 집합에서 검출기를 미세 조정된 결과를 보여줍니다. 낮은 상태에서 테스트했을 때 샷 설정(예: 2% 및 5% 레이블 데이터)에서 우리의 접근 방식은 MoCo-v2 기준선보다 각각 5.4% 및 7.3% 더 높은 APbox를 달성합니다. 전체 주석을 사용하여 교육하는 경우에도 CutLER는 개체 감지 및 분할 모두에서 MoCo-v2를 능가하는 2% 이상의 개선을 일관되게 제공합니다. 더욱 인상적으로 CutLER out은 이전 감지 아키텍처를 사용함에도 불구하고 이전 SOTA 방법인 FreeSOLO [47] 및 DETReg [3]를 수행합니다.

5. 절제

CutLER에서 설계 결정을 분석합니다. Sec과 유사한 설정을 사용합니다. 4 ImageNet에서만 CutLER를 교육합니다. 우리는 Cascade Mask R-CNN 탐지 아키텍처를 사용하고 주로 COCO 및 UVO 감도되지 않은 탐지 벤치마크에서 모델을 평가합니다. 달리 명시되지 않는 한 모든 절제 연구는 자가 훈련 없이 수행됩니다.

각 구성 요소의 중요성. CutLER의 주요 구성 요소를 분석하고 상대적 기여도를 표 6에 보고합니다. 인기 있는 COCO [32] 데이터 세트 및 조밀하게 주석이 달린 비디오 인스턴스 세그먼트에 대한 결과를 보고합니다.

행동 양식	UVA		머리	
	APmask	APmask	APmask	APmask
TokenCut [50]	-	-	4.9	2.0
19.3 20.9 + DropPath + MoCo-v2	5.4	13.5	5.7	
기 [16, 19] 21.5 + 자체 학습(CutLER)	8.1	15.8	7.7	
	9.0	16.6	8.2	
	9.9	17.7	8.8	
	10.1	18.9	9.7	

표 6. 각 구성 요소의 기여도에 대한 절제 연구. COCO 및 비디오 분할 데이터 세트 UVO에 보고된 결과.

행동 양식	AP박스	AP박스	AR박스	AP마스크	AP마스크	AR마스크				
TokenCut(1 고유값)	5.2	2.6	5.0	TokenCut(3 고유	4.9	2.0	4.4			
값) 4.7	1.7	8.1	MaskCut(t = 3)	6.0	2.9	8.1	CutLER	3.6	1.2	6.9
					4.9	2.2	6.9			
					21.9	12.3	32.7	18.9	9.7	27.1

표 7. CutLER는 이미지당 하나 이상의 마스크를 생성할 수 있는 수정된 TokenCut과 비교할 때에도 훨씬 더 높은 결과를 달성합니다. TokenCut과 비교하여 MaskCut은 정밀도를 낮추지 않고 더 높은 리콜을 얻습니다. 우리는 COCO에 대한 결과를 보고합니다.

데이터 세트 UVO [45]. 또한 COCO 데이터 세트에서 TokenCut [50]을 실행한 성능을 보고합니다. 다음으로 TokenCut의 공식 코드를 사용하여 ImageNet에서 마스크를 생성하고 Cascade Mask R-CNN 훈련에 사용합니다 [4]. 이 기본 모델은 COCO에서 TokenCut을 사용하는 것보다 상당한 이점을 제공합니다. 제안된 각 구성 요소를 이 강력한 기본 모델에 추가합니다. MaskCut을 사용하면 APmask와 APmask가 각각 4.7%와 2.7% 증가합니다. 또한 조밀하게 주석이 달린 데이터 세트 UVO의 경우 APmask에 대한 개선이 더 큰 폭으로 나타났습니다. 7% 및 2.7%의. 이 결과는 주석이 달린 데이터 세트에 필수적인 이점을 증명합니다. COCO APmask를 각각 1.6% 및 0.9% 향상됩니다. 자가 훈련의 다단계는 의사 마스크의 양과 질을 증가시켜 1.3% 개선으로 이어집니다. 이러한 결과는 각각의 간단한 제안 구성 요소가 강력한 성능에 중요하다는 것을 보여줍니다.

50

TokenCut과의 비교. TokenCut [50]도 제로 샷 분할 방법입니다. 그러나 Sec에서 설명한 것처럼 이미지당 단일 인스턴스만 분할합니다. 3.1. 이미지당 둘 이상의 분할 마스크를 생성하기 위해 더 작은 고유 벡터를 더 많이 사용하고 생성된 모든 마스크를 결합하여 수정된 TokenCut을 사용합니다.

표 7은 바닐라 TokenCut, 수정된 Token Cut 및 CutLER에 대한 COCO의 검증 세트에 대한 객체 감지 성능을 보여줍니다. 더 많은 고유 벡터를 사용하면 리콜 ARbox가 증가하지만 정밀도 APbox는 크게 감소합니다. CutLER는 평균 리콜 ARbox 100, 정밀도 APbox 4.8%를 달성하며, 이는 480% 상대적인

100 4배 증가하지만 TokenCut의 평균 정밀도도 능가합니다.

사이즈 → 240 360 480 640	$t^{ncut} \rightarrow 0.0.1.0.15.0.2.0.3$	엔 → 2 3 4	$\tau^{IoU} \rightarrow 0.0.0.1.0.1.0.2$
AP _{마스크} 50 15.1 16.6 17.7 17.9	AP _{마스크} 50 17.1 17.5 17.7 17.6 17.5	AP _{마스크} 50 16.9 17.7 17.7	AP _{마스크} 50 17.4 17.7 14.4 12.7
(a) 이미지 크기.	(나) 브이 마스크컷용.	(c) # 이미지당 마스크.	(d) 브이 드롭로스용.

표 8. CutLER 교육에 사용되는 MaskCut 및 DropLoss에 대한 질제. 자체 학습 단계를 추가하지 않고 COCO val2017에서 CutLER의 탐지 및 인스턴스 분할 성능을 보고합니다. (a) MaskCut에 사용되는 이미지의 크기를 변경합니다. (b) MaskCut을 사용하여 추출한 임계값 τ 마스크를 변경하고 다른 CutLER 모델을 교육합니다. (d) τ 예측 영역과 예측 영역에 대한 손실이 무시되는 기준 진실을 변경합니다. 다. 기본 설정은 회색으로 강조 표시된 Normalized Cuts에 사용되는 선호도 매트릭스의 희소성을 제어하는 MaskCut에서. (c) 우리는 DropLoss에서, 즉

UVA			머리		
AP _{마스크} 50	AP _{마스크} 75	AP _{마스크} 75	AP _{마스크} 50	AP _{마스크} 75	AP _{마스크} 75
1라운드 20.6	9.0	7.0	17.7	8.8	8.0
2라운드 22.2	9.6	7.5	18.5	9.5	8.8
3라운드 22.8	10.1	8.0	18.9	9.7	9.2
4라운드 22.8	10.2	8.2	18.9	9.8	9.3

표 9. CutLER에서 사용되는 자가 훈련 라운드 수. 우리는 3라운드의 자가 훈련이 충분하다는 것을 알게 되었습니다. 자가 학습은 조밀하게 레이블이 지정된 UVO 데이터 세트에 대해 더 큰 이득을 제공합니다.



그림 6. 여러 라운드의 자가 훈련을 통해 의사 마스크의 품질과 양을 향상시킬 수 있습니다. 세 라운드 모두에 대한 질적 시각화와 유사 마스크의 수를 보여줍니다.

MaskCut 및 DropLoss의 디자인 선택 과 최종 로컬라이제이션 성능에 미치는 영향은 표 8에 나와 있습니다. 먼저 초기 마스크를 생성하기 위해 MaskCut에서 사용하는 이미지 크기의 효과를 연구합니다. 예상대로 표 8a는 MaskCut이 픽셀 간에 더 높은 해상도 유사성을 제공하기 때문에 아마도 더 높은 해상도 이미지를 사용함으로써 이점을 얻는다는 것을 보여줍니다. 우리는 MaskCut의 속도와 성능 사이의 더 나은 균형을 위해 480px의 해상도를 선택합니다. 표 8b에서는 이진 W 매트릭스를 생성하기 위해 MaskCut에서 사용되는 임계값의 효과를 연구합니다(Sec. 3.2). 전반적으로 CutLER은 임계값에 대해 강력한 것으로 보입니다. 우리는 표 8c에서 MaskCut에 의해 생성된 이미지당 마스크 수가 미치는 영향을 이해합니다.

숫자를 늘리면 결과 CutLER 모델의 성능이 향상됩니다. 이것은 MaskCut이 전체 성능에 직접적인 영향을 미치는 고품질 마스크를 생성한다는 것을 보여줍니다. 마지막으로 표 8d에서 DropLoss에 사용되는 IOU 임계값을 변경합니다. 임계값이 높으면 더 많은 수의 예측 영역에 대한 손실을 무시하고 모델이 탐색하도록 권장합니다. 0.01은 탐색 성능과 탐지 성능 간의 절충에 가장 적합합니다.

마스크 R-CNN 캐스케이드 마스크 R-CNN ViTDet		
AP _{마스크} 50 / APbox 20.3 / 10.6 AP _{마스크} 75	20.8 / 11.5	21.5 / 11.8
AP _{마스크} 75 / APbox 20.3 / 10.6 AP _{마스크} 75	17.7 / 8.8	18.0 / 9.0

표 10. 검출 아키텍처가 다른 CutLER. 우리는 COCO에 대한 결과를 보고하고 CutLER가 탐지 아키텍처에 구애받지 않고 ViT-B의 백본이 있는 ViTDet과 같은 더 강력한 탐지 아키텍처를 사용하여 성능을 향상시키는 것을 관찰합니다.

CutLER APbox 사전 교육	50	AP _{마스크} 50	AP _{마스크} 75	AP _{마스크} 75	AP _{마스크} 75
IN1K IN1K 20.8 10.8 11.5 YFCC1M YFCC1M 19.4	17.7	8.0	8.8		
10.4 10.9 IN1K YFCC1M 14.9 7.6 8.2 YFCC1M IN1K	16.3	7.4	8.1		
14.8 7.2 8.0	12.1	5.4	5.9		
	11.8	5.2	5.8		

표 11. DINO 사전 교육 및 CutLER 교육에 사용된 데이터 세트의 영향. CutLER의 감지 성능은 개체 중심 ImageNet 데이터 세트 또는 개체 중심이 아닌 YFCC 데이터 세트와 같은 동일한 데이터 세트로 DINO와 CutLER을 사전 훈련할 때 유사합니다.

자가 훈련 과 최종 성능에 미치는 영향은 표 9에 분석되어 있습니다. 자가 훈련은 UVO 및 COCO 벤치마크와 모든 지표에서 지속적으로 성능을 향상시킵니다. 조밀한 개체 주석이 있는 UVO는 여러 라운드의 자가 훈련에서 더 많은 이점을 얻습니다. 기본적으로 CutLER은 3라운드의 자가 훈련을 사용합니다. 그림 6은 자가 훈련이 어떻게 예측 품질과 예측된 개체 수를 모두 개선하는지에 대한 정성적 예를 보여줍니다.

다양한 탐지 아키텍처에 대한 일반화. 우리는 CutLER 훈련을 위해 서로 다른 검출기 아키텍처를 사용하고 표 10에서 성능을 측정합니다. CutLER가 다양한 아키텍처와 함께 작동하고 더 강력한 아키텍처로 성능이 향상되는 것을 관찰했습니다.

사전 훈련 데이터 세트의 영향. 이제 우리는 1) 자기 감독 DINO 모델 사전 교육 및 2) CutLER 모델 교육에 사용된 데이터 세트의 영향을 연구합니다. 일반적으로 사용되는 ImageNet 데이터 세트에는 감독되지 않은 감지 성능에 영향을 줄 수 있는 잘 알려진 객체 중심 편향 [12]이 있습니다. 따라서 객체 중심이 아닌 데이터 세트인 YFCC [40]도 사용합니다. 공정한 비교를 위해 ImageNet과 YFCC 모두에서 이미지 수를 제어하고 DINO 및 CutLER 교육에 사용합니다. 표 11에서 볼 수 있듯이 COCO에 대한 CutLER의 성능은 객체 중심 또는 객체 중심이 아닌 데이터 세트의 선택에 강력합니다.

동일한 데이터 세트가 DINO 및 CutLER를 교육하는 데 사용됩니다. 이것은 다양한 데이터 분포에 대한 CutLER의 일반화를 보여줍니다. 그러나 서로 다른 데이터로 DINO 및 CutLER를 교육하면 성능이 저하되므로 DINO 및 CutLER 모델을 모두 학습하는 데 동일한 이미지 분포를 사용하는 것이 중요합니다.

6. 요약

개체 현지화는 컴퓨터 비전의 기본 작업입니다. 이 백서에서 우리는 간단하면서도 효과적인 잘라내기 및 학습 방식이 사람의 표기법을 사용하여 훈련할 필요 없이 까다로운 개체 감지 및 인스턴스 분할 작업에서 뛰어난 성능을 달성할 수 있음을 보여주었습니다. Zero-shot 비감독 검출기로서 ImageNet에서만 훈련된 CutLER는 다양한 도메인에 걸쳐 11개의 벤치마크에서 이전 작업의 검출 성능을 2.7배 이상 증가합니다.

참조

[1] YM 아사노, C Rupprecht 및 A Vedaldi. 동시 클러스터링 및 표현 학습을 통한 자체 라벨링. 학습 표현에 관한 국제 회의에서, 2019. 2

[2] Hangbo Bao, Li Dong, Songhao Piao 및 Furu Wei. Beit: 이미지 변환기의 Bert 사전 훈련. In International Conference on Learning Representations, 2021. 2

[3] Amir Bar, Xin Wang, Vadim Kantorov, Colorado J Reed, Roei Herzig, Gal Chechik, Anna Rohrbach, Trevor Darrell, Amir Globerson. Detreg: 객체 감지를 위해 영역 우선 순위를 사용하는 감독되지 않은 사전 훈련. In CVPR, 2022. 2, 5, 6, 7 [4] Zhaowei Cai 및 Nuno Vasconcelos. Cascade r-cnn: 고품질 객체 감지를 탐구합니다. 컴퓨터 비전 및 패턴 인식에 관한 IEEE 회의 진행, 페이지 6154–6162, 2018. 4, 6, 7, 12

[5] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, Sergey Zagoruyko. 변환기를 사용한 엔드투엔드 물체 감지. In ECCV, 2020. 2 [6] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, Armand Joulin. 클러스터 할당을 대조하여 시각적 기능의 비지도 학습. 신경 정보 처리 시스템의 발전, 33, 2020. 2, 4, 12

[7] Mathilde Caron, Hugo Touvron, Ishan Misra, Herve Jegou, Julien Mairal, Piotr Bojanowski 및 Armand Joulin. 자기 감독 비전 변환기의 신흥 속성. 컴퓨터 비전에 관한 IEEE/CVF 국제 회의 진행, 9650–9660페이지, 2021. 1, 2, 3, 4, 6, 12 [8] Ting Chen, Simon Kornblith, Mohammad Norouzi 및 Geoffrey Hinton. 시각적 표현의 대조 학습을 위한 간단한 프레임워크입니다. 기계 학습에 관한 국제 회의, 1597–1607페이지. PMLR, 2020. 1, 2

[9] Xinlei Chen, Haoqi Fan, Ross Girshick, Kaiming He. 모멘텀 대조 학습으로 기준선을 개선했습니다. arXiv 프리프린트 arXiv:2003.04297, 2020. 1, 4, 6, 12, 13

[10] Xinlei Chen과 Kaiming He. 간단한 삼 대표 분개 학습을 탐색합니다. 컴퓨터 비전 및 패턴 인식에 관한 IEEE/CVF 회의 진행, 페이지 15750–15758, 2021. 2 [11] Minsu Cho, Suha Kwak, Cordelia Schmid, and Jean Ponce.

감독되지 않은 개체 검색 및 야생 지역화: 상향식 영역 제안과 부분 기반 일치. 컴퓨터 비전 및 패턴 인식에 관한 IEEE 회의 진행, 1201–1210페이지, 2015. 2 [12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, Li Fei-Fei. Imagenet: 대규모 계층적 이미지 데이터베이스입니다. 컴퓨터 비전 및 패턴 인식에 관한 2009년 IEEE 회의, 248–255페이지. Ieee, 2009. 2, 4, 8 [13] Carl Doersch, Abhinav Gupta, Alexei A Efros. 감독 신호로서의 컨텍스트: 예측 가능한 컨텍스트로 객체를 발견합니다. 컴퓨터 비전에 관한 유럽 회의, 362–377페이지. Springer, 2014. 2 [14] Carl Doersch, Abhinav Gupta, Alexei A Efros. 컨텍스트 예측을 통한 비지도 시각적 표현 학습. 컴퓨터 비전에 관한 IEEE 국제 회의 진행, 1422–1430페이지, 2015. 2

[15] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelley 등. 이미지는 16x16 단어의 가치가 있습니다. 대규모 이미지 인식을 위한 변환기입니다. arXiv 프리프린트 arXiv:2010.11929, 2020. 2, 4

[16] Debidatta Dwibedi, Ishan Misra 및 Martial Hebert. 잘라내기, 붙여넣기 및 학습: 인스턴스 감지를 위한 놀랍도록 쉬운 합성. 컴퓨터 비전에 관한 IEEE 국제 회의 진행, 1301–1310페이지, 2017. 4, 7, 12 [17] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn 및 Andrew Zisserman. 파스칼 시각적 객체 클래스(voc) 챌린지. 컴퓨터 비전 국제 저널, 88(2):303–338, 2010. 6, 12, 13

[18] Andreas Geiger, Philip Lenz, Raquel Urtasun. 우리는 자율주행을 위한 준비가 되었나요? 키티 비전 벤치마크 제품군. 컴퓨터 비전 및 패턴 인식에 관한 2012년 IEEE 회의, 3354–3361페이지. IEEE, 2012. 12, 13 [19] Golnaz Ghiasi, Yin Cui, Aravind Srinivas, Rui Qian, Tsung Yi Lin, Ekin D Cubuk, Quoc V Le, Barret Zoph. 단순 복사-붙여넣기는 인스턴스 분할을 위한 강력한 데이터 증대 방법입니다. 컴퓨터 비전 및 패턴 인식에 관한 IEEE/CVF 회의 진행, 페이지 2918–2928, 2021. 4, 7, 12 [20] Spyros Gidaris, Andrei Bursuc, Gilles Puy, Nikos Kolodakis, Matthieu Cord 및 Patrick Perez. Obow: 자기 지도 학습을 위한 온라인 bag-of-visual-words 생성.

컴퓨터 비전 및 패턴 인식에 관한 IEEE/CVF 회의 진행, 6830–6840페이지, 2021.

[21] 로스 거식. 빠른 r-cnn. ICCV, 2015. 12

[22] Ross Girshick, Jeff Donahue, Trevor Darrell, Jitendra Malik. 정확한 객체 감지 및 시맨틱 분할을 위한 풍부한 기능 계층. 컴퓨터 비전 및 패턴 인식에 관한 IEEE 회의 진행, 580–587페이지, 2014. 12

[23] Jean-Bastien Grill, Florian Strub, Florent Altche, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Moham mad Gheshlaghi Azar 등. Bootstrap your own latent: 자기 지도 학습에 대한 새로운 접근 방식입니다. arXiv 프리프린트 arXiv:2006.07733, 2020. 2 [24] Agrim Gupta, Piotr Dollar, Ross Girshick. Lvis: 대규모 어휘 인스턴스 세분화를 위한 데이터 세트입니다. 컴퓨터 비전 및 패턴 인식에 관한 IEEE/CVF 회의 진행, 페이지 5356–5364, 2019. 6, 12, 13 [25] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollar 및 Ross Girshick. 마스크된 자동 인 코더는 확장 가능한 비전 학습자입니다. 컴퓨터 비전 및 패턴 인식에 관한 IEEE/CVF 회의 진행, 16000-16009페이지, 2022. 2 [26] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie 및 Ross Girshick. 감독되지 않은 시각적 표현 원한 학습을 위한 운동량 대비. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 9729–9738, 2020. 1, 2 [27] Kaiming He, Georgia Gkioxari, Piotr Dollar, Ross Girshick. 마스크 r-cnn. 컴퓨터 비전에 관한 IEEE 국제 회의 진행, 페이지 2961–2969, 2017. 4, 6, 12

[28] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun. 이미지 인식을 위한 심층 잔차 학습. Inoue Naoto, Ryosuke Furuta, Toshihiko Yamasaki, Kiyoharu Aizawa. 점진적 도메인 적응을 통한 교차 도메인 약한 감독 개체 감지. 컴퓨터 비전 및 패턴 인식에 관한 IEEE 회의 진행, 페이지 5001–5009, 2018. 12, 13 [30] Philipp Krahenbühl 및 Vladlen Koltun. 가우시안 예제 포텐셜을 가진 완전히 연결된 crf에서 효율적인 추론. NeurIPS, 2011. 4

[31] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloi, Alexander Kolesnikov 등. 열린 이미지 데이터 세트 v4. International Journal of Computer Vision, 128(7):1956–1981, 2020. 12, 13 [32] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollar, and C Lawrence Zitnick. Microsoft coco: 컨텍스트의 공통 개체입니다. 컴퓨터 비전에 관한 유럽 회의, 740–755페이지. 스프링거, 2014. 1, 5, 6, 7, 12, 13

[33] S Maji, NK Vishnoi 및 J Malik. 편향된 정규화 컷. 컴퓨터 비전 및 패턴 인식에 관한 2011 IEEE 회의 진행, 2057–2064, 2011. 3 [34] Ishan Misra 및 Laurens van der Maaten. 구실 불변 표현의 자기 감독 학습. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 6707–6717, 2020. 2 [35] Shaoqing Ren, Kaiming He, Ross Girshick, Jian Sun. 더 빠른 r-cnn: 영역 제안 네트워크를 통한 실시간 개체 감지를 향하여. 신경 정보 처리 시스템의 발전, 28, 2015.

[36] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, Jian Sun. Objects365: 개체 감지를 위한 대규모 고품질 데이터 세트입니다. 컴퓨터 비전에 관한 IEEE/CVF 국제 회의 진행, 페이지 8430–8439, 2019. 12, 13 [37] Jianbo Shi 및 Jitendra Malik. 정규화된 절단 및 이미지 세분화. 패턴 분석 및 기계 지능에 관한 IEEE 트랜잭션, 22(8):888–905, 2000. 3, 4 [38] Oriane Simeoni, Gilles Puy, Huy V Vo, Simon Roburin, Spyros Gidaris, Andrei Bursuc, Patrick Perez, Renaud Marlet, 장 폰세. 자체 감독 변환기를 사용하고 레이블 없이 개체를 현지화합니다. arXiv 프리프린트 arXiv:2109.14279, 2021. 2, 3, 5, 6, 12

[39] Karen Simonyan과 Andrew Zisserman. 대규모 이미지 인식을 위한 매우 깊은 컨볼루션 네트워크. arXiv 사전 인쇄 arXiv:1409.1556, 2014.

[40] Bart Thomee, David A Shamma, Gerald Friedland, Ben Jamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, Li-Jia Li. Yfcc100m: 멀티미디어 연구의 새로운 데이터. Communications of the ACM, 59(2):64–73, 2016. 8 [41] Jasper RR Uijlings, Koen EA Van De Sande, Theo Gevers 및 Arnold WM Smeulders. 객체 인식을 위한 선택적 검색. International Journal of Computer Vision, 104(2):154–171, 2013. 12 [42] Wouter Van Gansbeke, Simon Vandenhende, Luc Van Gool. 감독되지 않은 시맨틱 분할을 위한 변환기를 사용하여 개체 마스크를 검색합니다. arXiv 프리프린트 arXiv:2206.06363, 2022. 2, 3, 6

[43] Huy V Vo, Patrick Perez, Jean Ponce. 대규모 이미지 컬렉션에서 감독되지 않은 다중 개체 검색을 향하여. 컴퓨터 비전에 관한 유럽 회의, 779–795페이지. Springer, 2020. 2, 3, 6, 12 [44] Van Huy Vo, Elena Sizikova, Cordelia Schmid, Patrick Perez, Jean Ponce. 대규모 자율 객체 발견. 신경 정보 처리 시스템의 발전, 34:16764–16778, 2021. 2, 6.

[45] Wei Yao Wang, Matt Feiszli, Heng Wang, Du Tran. 미확인 비디오 개체: 밀집된 개방형 세계 분할을 위한 벤치마크입니다. 컴퓨터 비전에 관한 IEEE/CVF 국제 회의 진행, 페이지 10776–10785, 2021. 6, 7, 12, 13

[46] Xudong Wang, Ziwei Liu, Stella X Yu. 교차 수준 인스턴스 그룹 판별에 의한 감독되지 않은 기능 학습. 컴퓨터 비전 및 패턴 인식에 관한 IEEE/CVF 회의 진행, 페이지 12586–12595, 2021. 2, 4, 12 [47] Xinlong Wang, Zhiding Yu, Shalini De Mello, Jan Kautz, Anima Anandkumar, Chunhua Shen, 호세 M 알바레즈. Freesolo: 주석 없이 개체를 분할하는 방법을 배웁니다. 컴퓨터 비전 및 패턴 인식에 관한 IEEE/CVF 회의 진행, 14176–14186페이지, 2022. 1, 2, 3, 5, 6, 7, 13

[48] Xinlong Wang, Rufeng Zhang, Tao Kong, Lei Li 및 Chunhua Shen. Solov2: 동적 및 빠른 인스턴스 세분화. Advances in Neural information processing systems, 33:17721–17732, 2020. 2, 5, 6

[49] Xinlong Wang, Rufeng Zhang, Chunhua Shen, Tao Kong 및 Lei Li. 자기 지도를 위한 조밀한 대조 학습

시각적 사전 교육. 컴퓨터 비전 및 패턴 인식에 관한 IEEE/CVF 회의 진행, 3024–3033, 2021페이지.

[50] Yangtao Wang, Xi Shen, Yuan Yuan, Yuming Du, Maomao Li, Shell Xu Hu, James L Crowley, Dominique Vaufrey daz. Tokencut: 자체 감독 변환기 및 정규화 컷을 사용하여 이미지 및 비디오의 개체를 분할합니다. arXiv preprint arXiv:2209.00383, 2022. 2, 3, 4, 6, 7, 12 [51]
Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, Ross Girshick. 디텍트론2. [https://깃허브.com/facebookresearch/detectron2](https://github.com/facebookresearch/detectron2), 2019. 7, 13

[52] Zhirong Wu, Yuanjun Xiong, Stella X Yu, Dahua Lin.
비모수적 인스턴스 식별을 통한 감독되지 않은 기능 학습. 컴퓨터 비전 및 패턴 인식에 관한 IEEE 회의 진행, 페이지 3733–3742, 2018. 2 [53]
Junyuan Xie, Ross Girshick 및 Ali Farhadi. 클러스터링 분석을 위한 감독되지 않은 딥 임베딩. 기계 학습에 관한 국제 회의, 478–487페이지.
PMLR, 2016. 2

[54] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, Jifeng Dai Deformable detr: Deformable transformers for end-to-end object detection.In International Conference on Learning Representations, 2020. 2 , 6

[55] Chengxu Zhuang, Alex Lin Zhai, Daniel Yamins, 외. 시각적 임베딩의 비지도 학습을 위한 로컬 집계. ICCV에서, 2019. 2 [56] Adrian Ziegler와 Yuki M Asano. 시맨틱 분할을 위한 객체 부분의 자기 감독 학습. 컴퓨터 비전 및 패턴 인식에 관한 IEEE/CVF 회의 진행, 14502–14511페이지, 2022. 2

가. 부록

A.1. 교육 세부 정보

CutLER는 기본 검출기에 대해 불가피하지만 모든 실험에 널리 사용되는 Mask R-CNN [27] 및 Cascade Mask R-CNN [4]을 사용하고 달리 명시되지 않는 한 기본적으로 Cascade Mask R-CNN을 사용합니다. 우리는 배치 크기 16으로 160K 반복에 대한 초기 마스크와 경계 상자를 사용하여 ImageNet에서 탐지기를 훈련합니다. ResNet-50 백본 [28]으로 탐지기를 훈련할 때 우리는 자체 지도 사전 훈련된 가중치로 모델을 초기화합니다. DINO [7] 모델. 우리는 MoCo-v2 [9], SwAV [6] 및 CLD [46]를 포함하여 다른 사전 훈련된 모델을 탐색했으며 유사한 탐지 성능을 제공한다는 것을 발견했습니다. 따라서 기본적으로 DINO로 모델 가중치를 초기화합니다.

또한 모델 훈련 과정에서 복사-붙여넣기 확대 [16, 19]를 활용합니다. 바닐라 복사-붙여넣기 확대를 사용하여 작은 개체를 분할하는 모델의 기능을 개선하는 대신 0.3과 1.0 사이에서 균일하게 샘플링된 스칼라로 마스크를 무작위로 다운샘플링합니다. 그런 다음 0.005의 학습률로 SGD를 사용하여 160K 반복에 대해 검출기를 최적화합니다. 이 학습률은 80K 반복 후 5씩 감소하고 배치 크기는 16입니다. 5×10^{-5} 의 가중치 감소와 0.9의 모멘텀을 적용합니다.

자가 학습의 다중 라운드를 위해 각 단계에서 이전 단계의 가중치를 사용하여 탐지 모델을 초기화합니다. 80K 반복에 대해 0.01의 학습률로 SGD를 사용하여 검출기를 최적화합니다. 자체 훈련 단계에서는 모델 훈련을 위한 충분한 수의 의사 마스크를 제공할 수 있으므로 자체 훈련 단계에서는 탐색 손실을 사용하지 않습니다.

A.2. 제로샷 평가에 사용되는 데이터 세트

COCO 및 COCO20K [32]는 훈련 및 검증 분할에 각각 약 115K 및 5K 이미지를 포함하는 대규모 객체 감지 및 인스턴스 분할 데이터 세트입니다. 또한 COCO에는 주석이 없는 123K 이미지 분할이 있습니다. COCO의 이미지를 미세 조정하지 않고 COCO val2017 및 COCO 20K에서 클래스에 구매받지 않는 방식으로 모델을 테스트합니다. COCO 20K는 [38, 43, 50]에서 벤치마크로 사용되는 19817개의 무작위 샘플링 이미지를 포함하는 COCO trainval2014 [32]의 하위 집합입니다. 개체 감지 및 분할 작업에 대해 클래스에 구매받지 않는 COCO 스타일 평균 정밀도 및 평균 회수율을 보고합니다.

Pascal VOC [17]는 객체 감지에 대한 또 다른 인기 있는 벤치마크입니다. COCO 스타일 평가 매트릭스에서 trainval07 분할로 모델을 평가합니다.

UVO [45]. UVO(Unidentified Video Objects)는 비디오 객체 감지 및 인스턴스 분할을 위한 철저한 주석 데이터 세트입니다. 프레임별 추론을 통해 UVO val에서 모델을 평가하고 결과를 COCO 스타일 평가 매트릭스로 보고합니다.

LVIS [24]는 220만 개의 고품질 인스턴스 세그먼트를 수집했습니다.

자연히 롱테일 데이터 분포를 구성하는 1000개 이상의 엔트리 레벨 개체 범주에 대한 멘테이션 마스크. 우리는 약 5K 이미지를 포함하는 LVIS val 분할에 대한 클래스에 구매받지 않는 개체 감지 및 인스턴스 분할 결과를 보고합니다.

CrossDomain [29]에는 수채화, 클리 파트 및 만화의 세 가지 하위 집합이 포함되어 있으며 개체는 각각 수채화, 스케치 및 페인팅 스타일로 묘사됩니다. 이 세 가지 데이터 세트, 즉 traintest의 주석이 달린 모든 이미지에서 모델을 평가합니다.

Objects365 V2 [36]는 야생의 다양한 객체에 초점을 맞춘 지도 객체 감지 벤치마크를 제공합니다. 우리는 val split의 80K 이미지에서 CutLER를 평가합니다.

OpenImages V6 [31]은 이미지 분류, 개체 감지, 인스턴스 분할, 시각적 관계 감지 등을 하나의 데이터 세트로 통합합니다. Val split의 42K 이미지에서 CutLER를 평가합니다.

KITTI [18]는 자율 주행 연구에 사용되는 이동 차량에 장착된 카메라에서 캡처한 데이터 세트를 제시합니다. 우리는 KITTI의 trainval 분할에서 7521 이미지에 대한 CutLER를 평가합니다.

제로에 사용되는 이러한 데이터 세트의 요약은 제공합니다. 첫 평가는 표 12에 나와 있습니다.

A.3. 제로 샷 감지 및 분할에 대한 추가 결과

이 섹션에서는 공식 COCO API를 사용하고 다양한 IoU 임계값에 대한 AP(단계 크기 0.05에서 0.5~0.95의 IoU 임계값에 대한 평균), AP50 (IoU@0.5)을 포함하여 표준 COCO 메트릭으로 더 많은 결과를 제공합니다. 및 AP75 (IoU@0.75), 그리고 규모에 걸친 AP - APS (소형 물체), APM (중형 물체) 및 APL (대형 물체). 우리는 표 12에 나열된 모든 벤치마크에 대한 자세한 결과를 제공하고 이러한 결과를 표 13에 보고합니다. 우리는 모든 데이터 세트에 대한 객체 감지 성능을 보고합니다. 또한 인스턴스 분할에 대한 주석을 제공하는 데이터 세트의 경우 인스턴스 분할 작업의 성능도 제시합니다. 세분화 레이블이 없는 이러한 데이터 세트에서 CutLER는 여전히 인스턴스 세분화 마스크를 예측할 수 있지만 비교할 실측 마스크가 없기 때문에 결과를 평가할 수 없다는 점은 주목할 가치가 있습니다.

A.4. CutLER 대 선택적 검색

Selective Search [41]는 R-CNN [22] 및 Fast R-CNN [21]과 같은 많은 초기 최첨단 감지기에서 사용되는 인기 있는 감도되지 않은 개체 검색 방법입니다. 그러나 슬라이딩 창으로 가능한 개체 위치를 생성하면 추론 속도가 크게 감소합니다(선택적 검색에 대한 자세한 내용은 [41] 참조). CutLER의 성능을 그림 7의 선택적 검색과 비교하고 CutLER이 정밀도와 재현율 모두에서 상당한 개선을 제공한다는 것을 관찰합니다. 이는 CutLER이

[illegible]

선택적 검색 (x)	선택적 검색 (y)	우리 것 (x)	우리 것 (y)
0.0	1.0	0.0	1.0
0.1	0.3	0.1	0.9
0.15	0.3	0.2	0.85
0.2	0.0	0.3	0.8
0.3	0.0	0.4	0.75
0.4	0.0	0.5	0.65
0.5	0.0	0.6	0.55
0.6	0.0	0.7	0.45
0.7	0.0	0.8	0.35
0.8	0.0	0.9	0.2
0.9	0.0	1.0	0.0

기준선으로 MoCo-v2 [9]의 설정을 따르고 객체 감지 작업에 대한 강력한 성능을 제공하는 MoCo-v2 ResNet50 모델로 초기화된 동일한 감지 아키텍처를 교육합니다. MoCo-v2와 우리 모델은 동일한 학습 파이프라인과 하이퍼 매개변수를 사용하고 1% 또는 2% 레이블이 있는 극도로 낮은 설정을 제외하고 Detectron2 [51]를 사용하여 1x 일정에 대해 학습됩니다. 이전 작업 [47]에 따라 1% 또는 2% 레이블로 교육할 때 MoCo-v2와 모델을 배치 크기 16으로 3,600회 반복 교육합니다.

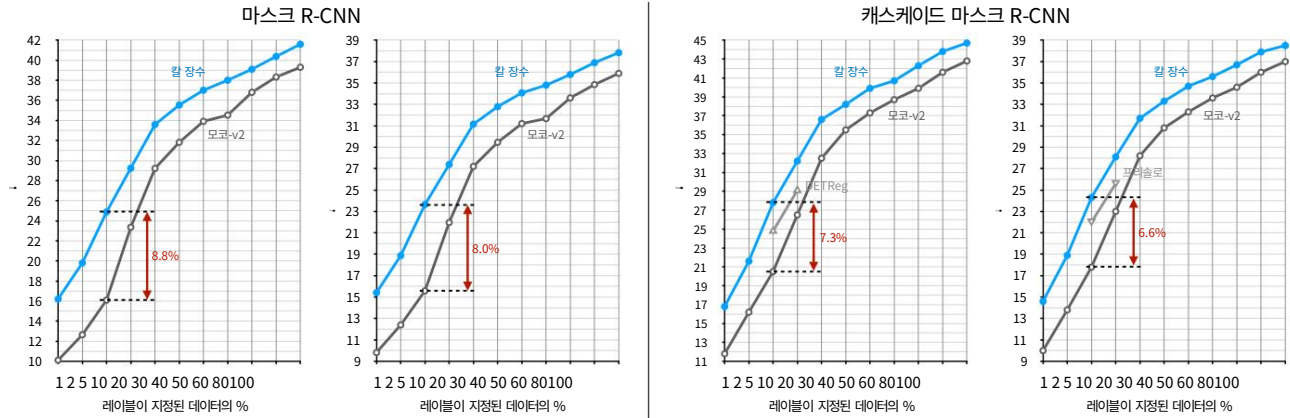


그림 8. 다양한 주석 비율로 MS-COCO에서 미세 조정. 우리는 검출기로 ResNet-50의 백본과 함께 Mask R-CNN 및 Cascade Mask R-CNN을 사용하여 결과를 보고합니다.

동일한 하이퍼 매개변수 및 설정.

A.6. 더 많은 시각화

그림 9에서 CutLER의 제로 샷 예측에 대한 보다 긍정적인 시각화를 제공합니다.

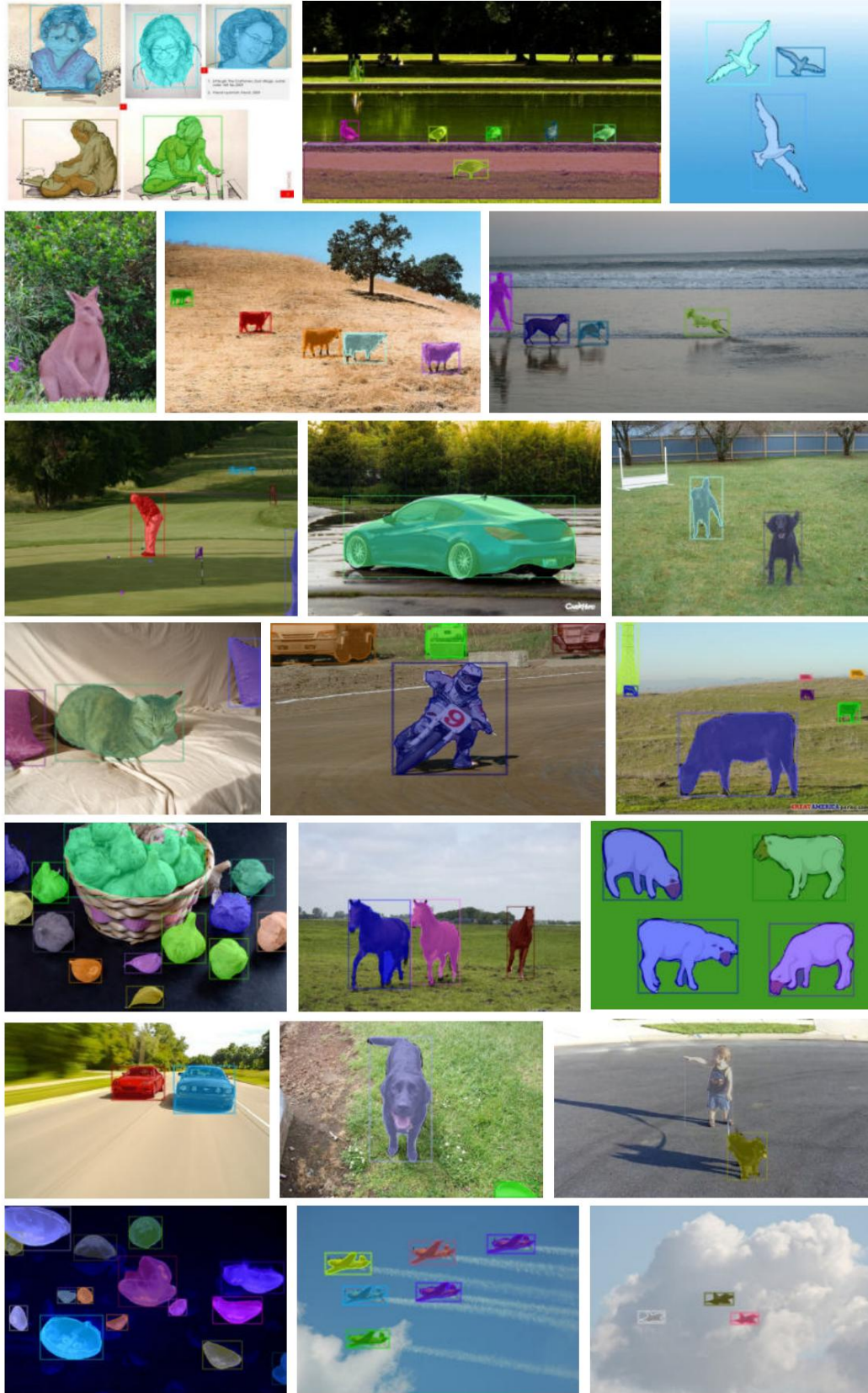


그림 9. CutLER의 예측에 대한 추가 시각화.