

철도는기차가아니다: 약한지도학습의미론적분할을위한의사-픽셀감독으로서의주목성

이승호*
연세대학교

seungholee@yonsei.ac.kr

이민현*
연세대학교

lmh315@yonsei.ac.kr

이종욱
성균관대학교

jongwuklee@skku.edu

심현정†
연세대학교

kateshim@yonsei.ac.kr

Abstract

기존의이미지수준약한감독을사용하는약한지도학습의미론적분할 (WSSS) 연구는몇가지제한사항이있습니다: 희소한객체커버리지, 부정확한객체경계, 비대상객체에서 발생하는픽셀. 이러한문제를극복하기위해, 우리는두가지 약한감독을결합하여픽셀수준피드백에서학습하는새로운 프레임워크인명시적의사-픽셀감독 (EPS) 을제안합니다; 이미지수준레이블은로컬라이제이션맵을통해객체정체성을제공하고, 기성품주목성탐지모델의주목성맵은풍부한 경계를제공합니다. 우리는두정보간의상호보완적관계를 최대한활용하기위해공동학습전략을고안했습니다. 우리의방법은정확한객체경계를연고발생하는픽셀을제거하여 의사마스크의품질을크게향상시킵니다. 실험결과는제안된방법이 WSSS 의주요문제를해결하여기존방법을현저히 능가하며 PASCAL VOC 2012 및 MS COCO 2014 데이터셋에서새로운최첨단성능을달성함을보여줍니다. 코드는 <https://github.com/halbielee/EPS>에서사용할수있습니다.

1. Introduction

약한지도학습의미론적분할 (WSSS) 은약한감독 (예: 이미지수준레이블 [36, 37], 낙서 [29], 또는경계상자 [22]) 을활용하여픽셀수준레이블이필요한완전지도모델과경쟁력있는성능을달성하는것을목표로합니다. 대부분의기존 연구는분할모델의약한감독으로이미지수준레이블을채택합니다. WSSS 의전체파이프라인은두단계로구성됩니다. 첫째, 이미지분류기를사용하여대상객체에대한의사마스크가생성됩니다. 그런다음, 분할모델은감독으로의사마스크를사용하여학습됩니다. 의사마스크를생성하는일반적인기술은클래스활성화매핑 (CAM) [52]으로, 이미지수준레이블에해당하는객체로컬라이제이션맵을제공합니다. 완전지도 (즉, 픽셀수준주석) 와약한지도 (즉, 이미지수준레이블) 의미론적분할간의감독격차로인해, WSSS 는다음과같은주요문제를가지고있습니다: 1) 로컬라이제이션맵은대상객체의작은부분만캡처합니다 [52], 2) 객체의경계불일

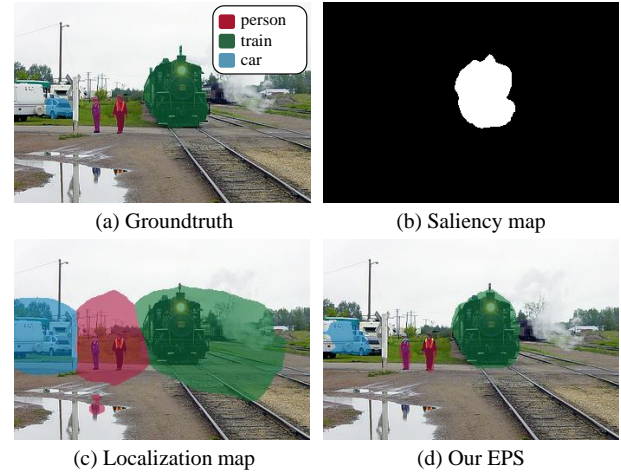


Figure 1. WSSS 를위한주목도맵과위치맵을모두활용하는거기부여예제. (a) 실제값, (b) PFAN [51]을통한주목도맵, (c) CAM [52]을통한위치맵, 그리고 (d) 분류기를학습하기위해주목도맵과위치맵을모두활용하는우리의 EPS. 주목도맵은 사람과 자동차를포착할수없지만, 우리의결과는이를올바르게복원할수 있으며, 위치맵은두개의객체를과도하게포착한다는점에유의하십시오.

치가발생합니다 [23], 3) 대상객체에서발생하는픽셀을거의분리하지못합니다 (예: 기차에서철도) [25].

이문제를해결하기위해기존연구들은세가지기동으로분류될수있습니다. 첫번째접근법은픽셀을지우거나 [9, 23, 28], 점수맵을양상블하거나 [21, 27], 자가지도신호를사용하여 [41] 객체의전체범위를포착하기위해객체커버리지를 확장합니다. 그러나이들은객체의형태를안내할단서가없기때문에대상객체의정확한경계를결정하지못합니다. 두번째접근법은의사마스크의객체경계를개선하는데중점을둡니다 [13, 32]. 이들은효과적으로객체경계를학습하기때문에자연스럽게경계까지의사마스크를확장합니다. 그러나여전히비대상객체의일치하는픽셀을대상객체와구별하지못합니다. 이는전경과배경간의강한상관관계 (즉, 동시발생) 가유도편향 (즉, 대상객체와그일치하는픽셀을관찰하는빈도) 과거의구별되지않기때문입니다 [10]. 마지막으로, 세번째접근법은추가적인실제마스크 [24] 또는주목도맵 [35, 47]을사용하여동시발생문제를완화하는것을목표로

*동일한기여를나타냅니다.

†심현정은교신저자입니다.

합니다. 그러나 [24, 28]은 약한지도 학습 패러다임과는 거리가 먼 강력한 픽셀 수준 주석이 필요합니다. [35]은 주목도 맵의 오류에 민감합니다. 또한 [47]는 객체의 전체 범위를 다루지 않으며 경계 불일치로 고통받습니다.

이 논문에서 우리의 목표는 이미지 수준 레이블로 학습된 이미지 분류기에서의 CAM(즉, CAM) 과기성품 주목도 맵지 모델의 출력 (즉, 주목도 맵) 모두를 완전히 활용하여 WSSS 의 세 가지 과제를 극복하는 것입니다 [18, 34, 51]. 우리는 위치 맵과 주목도 맵의 상호보완적 관계에 중점을 둡니다. 그림 1에 설명된 바와 같이, 위치 맵은 서로 다른 객체를 구별할 수 있지만 그들의 경계를 효과적으로 분리하지는 못합니다. 반대로, 주목도 맵은 풍부한 경계 정보를 제공하지만 객체의 정체성을 드러내지 않습니다. 이러한 의미에서, 우리는 두 가지 상호보완적인 정보를 사용하는 우리의 방법이 WSSS 의 성능 병목을 해결할 수 있다고 주장합니다.

이를 위해, 우리는 WSSS 를 위한 새로운 프레임워크인 명시적의사-픽셀 감독 (EPS) 을 제안합니다. 주목도 맵 (즉, 전경과 배경 모두) 을 완전히 활용하기 위해, 우리는 C 대상 클래스와 배경 클래스로 구성된 $C + 1$ 클래스를 예측하기 위한 분류기를 설계합니다. 우리는 C 위치 맵과 배경 위치 맵을 활용하여 주목도 맵을 추정합니다. 그런 다음, 주목도 손실은 주목도 맵과 우리의 추정된 주목도 맵 간의 픽셀 단위 차이로 정의됩니다. 주목도 손실도 도입함으로써, 모델은 모든 클래스에 걸쳐 의사-픽셀 피드백에 의해 감독될 수 있습니다. 우리는 또한 이미지 수준 레이블을 예측하기 위해 다중 레이블 분류 손실을 사용합니다. 따라서, 우리는 분류기를 주목도 손실과 다중 레이블 분류 손실을 최적화하도록 훈련하여 배경과 전경 픽셀 모두에 대한 예측을 시너지화합니다. 우리의 전략이 주목도 맵 (섹션 3.3 및 그림 3) 과의사 마스크 (섹션 5.1 및 그림 4) 모두를 개선할 수 있음을 발견했습니다.

우리는 주목도 손실 이외의 의사-픽셀 피드백을 통해 경계 불일치를 벌하기 때문에, 우리의 방법이 객체의 정확한 경계를 학습하도록 강제할 수 있음을 강조합니다. 부수적으로, 우리는 경계까지 맵을 확장하여 전체 객체를 포착할 수도 있습니다. 주목도 손실이 전경 (예: 기차) 과 배경을 분리하는데 도움을 주기 때문에, 우리의 방법은 동시 발생하는 픽셀 (예: 철도) 을 배경 클래스로 할당할 수 있습니다. 실험 결과는 우리의 EPS 가 PASCAL VOC 2012 및 MS COCO 2014 데이터셋에서 새로운 최첨단 정확도를 기록하며 놀라운 세분화 성능을 달성했음을 보여줍니다.

2. 관련 연구

약한지도 학습 기반의 이론적 세분화. WSSS 의 일반적인 파이프라인은 분류 네트워크에서의 의사 마스크를 생성하고, 이의 의사 마스크를 감독으로 사용하여 세그멘테이션 네트워크를 훈련하는 것입니다. 이미지 수준 레이블에서 경계 정보가 부족하기 때문에 많은 기존 방법들은 부정확한 의사 마스크로 고통받습니다. 이 문제를 해결하기 위해, 교차 이미지 친화성 [15], 지식 그래프 [31], 대조 최적화 [38, 50] 등의 의사 마스크의 품질을 향상시키기 위해 사용됩니다. [5]는 CAM 을 개선하기 위해 분류기를 강화하는 하위 카테고리 발견하는 자기지도 학습 과제를 제안합니다. [1, 2]은 픽셀 간의 친화성을 계산하여 경계 정보를 암묵적으로 활용합니다. [49]는 신뢰할 수 있는

픽셀 수준 주석을 생성하는데 중점을 두고 세그멘테이션 맵을 생성하기 위한 중단 간 네트워크를 설계합니다. [20, 25]는 경계 손실을 활용하여 세그멘테이션 네트워크를 훈련합니다. 최근 [3]는 자기지도 학습 체계를 가진 단일 세그멘테이션 기반 모델을 사용합니다. [14]은 여러 불완전한 의사 마스크를 활용하여 세그멘테이션 네트워크의 견고성에 중점을 둡니다.

주목성 유도 이론적 세그멘테이션. 주목성 탐지 (SD) 방법은 픽셀 수준 주석 [18, 46, 51] 또는 이미지 수준 주석 [39]이 있는 외부 주목성 데이터셋을 통해 이미지에서 전경과 배경을 구별하는 주목성 맵을 생성합니다. 많은 WSSS 방법들 [15, 20, 27, 28, 42, 44]은 의사 마스크의 배경 단서로 주목성 맵을 활용합니다. [43]는 단일 객체 이미지의 완전한 감독으로 주목성 맵을 활용합니다. [16]은 객체의 유사성 그래프를 학습하기 위해 인스턴스 수준 주목성 맵을 사용합니다. [6, 40, 47]는 클래스별 주의 단서와 주목성 맵을 결합하여 신뢰할 수 있는 의사 마스크를 생성합니다. [48]은 단일 네트워크를 사용하여 WSSS 와 SD 를 공동으로 해결하여 두 작업의 성능을 향상시킵니다. 우리의 EPS 는 주목성 유도 방법으로 분류될 수 있지만 다음과 같은 이유로 다른 모든 방법과 명확히 구별됩니다. 대부분의 기존 방법의 의사 마스크의 일부 또는 분류기의 중간 특징을 정제하기 위한 암묵적 지침으로 주목성 맵을 활용합니다. 반대로, 우리의 방법은 주목성 맵을 로컬라이제이션 맵에 대한 의사 픽셀 피드백으로 활용합니다. [48]은 두 가지 보완 정보를 활용하는 점에서 우리와 가장 유사한 작업이지만, 그들은 공존 문제를 해결하지 않거나 노이즈가 있는 주목성 맵 문제를 처리하지 않습니다.

3. 제안된 방법

이 섹션에서는 명시적의사 픽셀 감독 (EPS) 이라고 불리는 약한 감독의 이론적 세그멘테이션 (WSSS) 을 위한 새로운 프레임워크를 제안합니다. WSSS 의 두 단계를 고려할 때, 첫 번째 단계는 의사 마스크를 생성하고 두 번째 단계는 세그멘테이션 모델을 훈련하는 것입니다. 여기서 우리의 주요 기여는 정확한 의사 마스크를 생성하는 것입니다. WSSS 관례를 따르며 [13, 21, 27, 28, 41, 42], 우리는 첫 번째 단계에서 생성된 의사 마스크를 감독으로 사용하여 세그멘테이션 모델을 훈련합니다.

3.1. 동기

EPS 의 핵심 통찰력은 두 가지 보완 정보를 완전히 활용하는 것입니다, 즉, 로컬라이제이션 맵에서의 객체 정체성과 주목성 맵에서의 경계 정보입니다. 이를 위해, 우리는 주목성 맵을 대상 레이블과 배경 모두에 대한 로컬라이제이션 맵에 대한 의사 픽셀 피드백으로 활용합니다. 우리는 추가적인 배경 클래스를 가진 분류기를 고안하여 총 $C + 1$ 클래스를 예측하도록 하며, 이는 그림 2에 나타나 있습니다. 이 분류기를 사용하여, 우리는 $C + 1$ 로컬라이제이션 맵을 학습할 수 있습니다, 즉, 대상 레이블에 대한 C 로컬라이제이션 맵과 배경 로컬라이제이션 맵입니다. 우리는 EPS 가 WSSS 에서 경계 불일치 문제와 동시 발생 문제를 어떻게 해결할 수 있는지를 설명합니다. 경계 불일치 문제를 관리하기 위해, 우리는 C 지역화 맵에서 전경 맵을 추정하고 이를 주목도 맵의 전경과 일치시킵니다

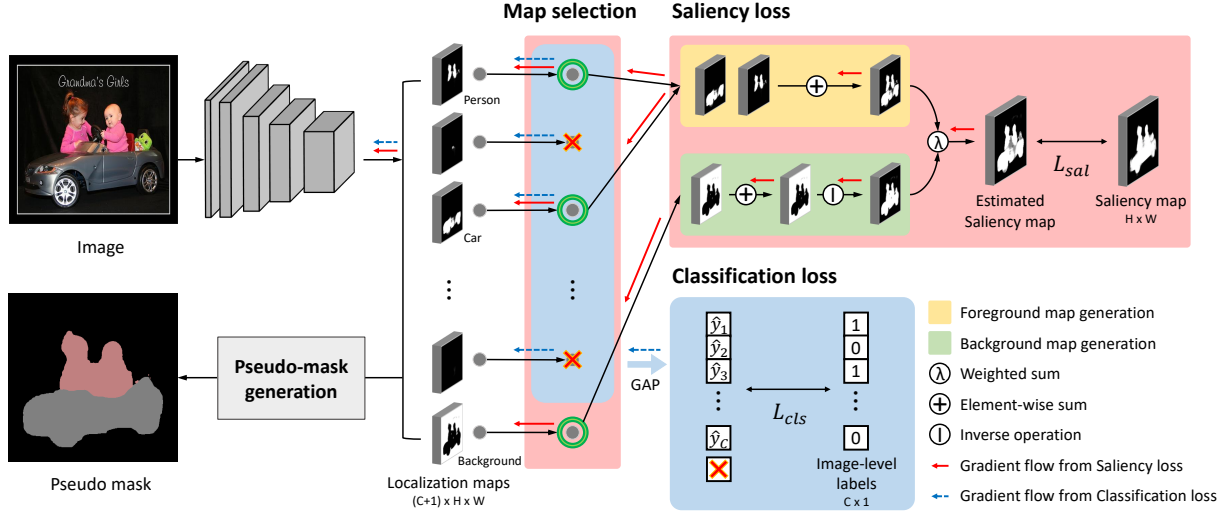


Figure 2. 우리의 EPS 의전체프레임워크. $C + 1$ 개의지역화맵이백본네트워크에서생성됩니다. 실제주목도맵은기성품주목도맵지모델에서생성됩니다. 대상레이블에대한일부지역화맵은선택적으로사용되어추정된주목도맵을생성합니다 (Section 3.2). 전체프레임워크는주목도손실과분류손실과함께공동으로학습됩니다 (Section 3.3).

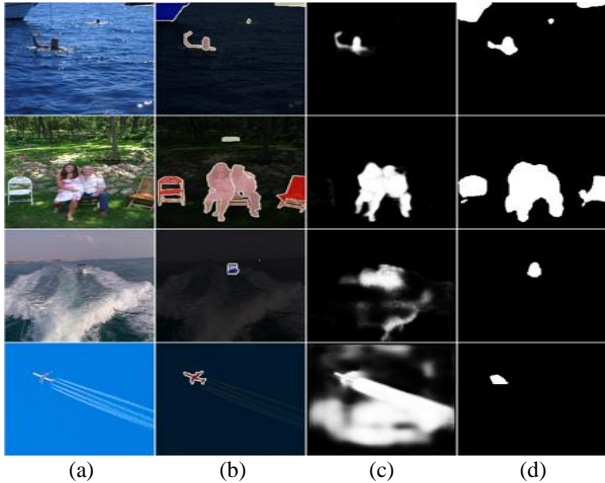


Figure 3. PASCAL VOC 2012 에서추정된주목도맵의정성적예시. (a) 입력이미지, (b) 정답, (c) [51]에서의주목도맵, (d) 우리의추정된주목도맵.

다. 이렇게하면, 대상레이블의지역화맵이주목도맵으로부터의사-픽셀피드백을받아객체의경계를개선할수있습니다. 비대상객체의동시발생픽셀을완화하기위해, 우리는배경에대한지역화맵도주목도맵과일치시킵니다. 배경에대한지역화맵도주목도맵으로부터의사-픽셀피드백을받기때문에, 동시발생픽셀은성공적으로배경에할당될수있습니다; 비대상객체의동시발생픽셀은대부분배경과겹칩니다. 이것이우리의방법이대상객체로부터동시발생픽셀을분리할수있는이유입니다. 마지막으로, EPS 의목적함수는두부분으로구성됩니다: 주목도손실 \mathcal{L}_{sal} (그림 2의빨간상자/화살표로표시됨) 과이미지수준레이블을통한다중레이블분류손실 \mathcal{L}_{cls} (그림 2의파란상자/화살표로표시됨) 입니다. 두목표를공동으로훈련함으로써, 우리는보완적인정보로지역화맵과주목도맵을시너지화할수있습니다- 우리는그림 3에서설

명된바와같이, 서로의노이즈와누락된정보가우리의공동훈련전략을통해보완된다는것을관찰합니다. 예를들어, 기성모델 [18, 34, 51]에서얻은원래의주목도맵은누락된정보와노이즈가있습니다. 반면에, 우리의결과는누락된객체 (예: 보트나익자) 를성공적으로복원하고노이즈 (예: 물방울이나비행운) 를제거하여원래의주목도맵보다명백히더 나은결과를보여줍니다. 결과적으로, EPS 는더정확한객체경계를포착하고대상객체로부터동시발생픽셀을분리할수있습니다. 이러한장점은놀라운성능향상을가져옵니다; 표 6은 EPS 가기준모델을최대 3.8–10.6

3.2. 명시적의사-픽셀감독

우리는의사-픽셀감독을위해주목도맵을어떻게활용할수있는지설명합니다. 주목도맵의주요장점은객체의실루엣을제공하여객체경계를더잘드러낼수있다는것입니다. 이속성을활용하기위해, 우리는주목도맵을전경과배경두가지경우와일치시킵니다. 클래스별지역화맵을주목도맵과비교가능하게만들기위해, 우리는대상레이블의지역화맵을병합하고전경맵 $M_{fg} \in \mathbb{R}^{H \times W}$ 을생성합니다. 우리는또한배경레이블에대한지역화맵인배경맵 $M_{bg} \in \mathbb{R}^{H \times W}$ 의반전을수행하여전경을나타낼수있습니다. (나중에, 우리는노이즈가있는주목도맵을해결하기위해전경맵을어떻게정제하는지설명합니다.)

구체적으로, 우리는다음과같이 M_{fg} 와 M_{bg} 를사용하여주목도맵 \hat{M}_s 를추정합니다:

$$\hat{M}_s = \lambda M_{fg} + (1 - \lambda)(1 - M_{bg}), \quad (1)$$

여기서 $\lambda \in [0, 1]$ 은전경맵과배경맵의반전의가중합을조정하는하이퍼파라미터입니다. (기본적으로, 우리는실험에서 λ 를 0.5 로설정하고 λ 에대한추가적인소거연구는보충자료에서찾을수있습니다.) 그런다음, 우리는추정된주목도맵과실제주목도맵간의픽셀별차이의합으로주목도손실 \mathcal{L}_{sal} 을정의합니다. (3.3에서 \mathcal{L}_{sal} 의공식정의가제시됩니다)

다.)

사전훈련된모델을사용하는것은약한감독학습으로간주되며, 따라서주목도맵을활용하는것은 WSSS 에서일반적인관행으로널리받아들여졌습니다. 그인기에불구하고, 완전감독주목도탐지모델을채택하는것은다른데이터셋에서픽셀수준주석을사용하기때문에논의여지가있을수있습니다. 이논문에서는다양한주목도탐지방법의효과를조사합니다; 1) 비지도및 2) 완전감독주목도탐지모델 (5.3 참조), 그리고경험적으로우리의방법이그중어느것을사용하더라도모든다른방법 [13, 21, 40, 43, 47]을능가한다는것을보여줍니다. 기존방법이주목도맵을완전히활용하는데제한이있는반면, 우리의방법은주목도맵을의사-픽셀감독으로통합하고경계와동시발생픽셀에대한단서로활용합니다.

살리언시편향을처리하기위한맵선택. 이전에는전경맵이타겟레이블의로컬라이제이션맵의합집합이될수있다고가정했습니다. 배경맵은배경레이블의로컬라이제이션맵이될수있습니다. 그러나이러한단순한선택규칙은기성모델에의해계산된살리언시맵과호환되지않을수있습니다. 예를들어, [51]의살리언시맵은종종일부객체를살리언트객체로무시합니다 (예: 그림 1의기차근처의작은사람들). 이러한체계적인오류는살리언시모델이다양한데이터셋의통계를학습하기때문에불가피합니다. 이오류를고려하지않으면동일한오류가우리모델에전파되어성능저하를초래할수있습니다.

체계적인오류를해결하기위해, 우리는로컬라이제이션맵과살리언시맵간의중첩비율을사용하는효과적인전략을개발했습니다. 구체적으로, i 번째로컬라이제이션맵 M_i 는살리언시맵과 $\tau\%$ 이상중첩되면전경에할당되고, 그렇지않으면배경에할당됩니다. 공식적으로, 전경과배경맵은다음과같이계산됩니다:

$$\begin{aligned} M_{fg} &= \sum_{i=1}^C y_i \cdot M_i \cdot \mathbb{1}[\mathcal{O}(M_i, M_s) > \tau], \\ M_{bg} &= \sum_{i=1}^C y_i \cdot M_i \cdot \mathbb{1}[\mathcal{O}(M_i, M_s) \leq \tau] + M_{C+1}, \end{aligned} \quad (2)$$

여기서 $y \in \mathbb{R}^C$ 는이진이미지레벨레이블이고 $\mathcal{O}(M_i, M_s)$ 는 M_i 와 M_s 간의중첩비율을계산하는함수입니다. 이를위해, 우리는먼저로컬라이제이션맵과살리언시맵을이진화합니다: 픽셀 p 에대해, $B_k(p) = 1$ 이면 $M_k(p) > 0.5$; 그렇지않으면 $B_k(p) = 0$. B_i 와 B_s 는각각 M_i 와 M_s 에해당하는이진화된맵입니다. 그런다음 M_i 와 M_s 간의중첩비율을계산합니다, 즉, $\mathcal{O}(M_i, M_s) = |B_i \cap B_s|/|B_i|$. 우리는데이터셋과백본모델에관계없이 $\tau = 0.4$ 로설정합니다. 보충자료에서, 우리는 τ 의선택에대해우리의방법이강력하다는것을보여줍니다 (즉, τ 가 $[0.3, 0.5]$ 내에서유사한성능을보입니다).

배경레이블에대한단일로컬라이제이션맵대신, 우리는배경레이블에대한로컬라이제이션맵을전경으로선택되지않은로컬라이제이션맵과결합합니다. 비록간단하지만, 우리는살리언시맵의오류를무회하고살리언시맵에서무시된일부객체를효과적으로학습할수있습니다. (표 3에서, 우리는살리언시맵의오류를극복하기위한제한된전략의효과를보고합니다.)

3.3. 공동학습절차

살리언시맵과이미지레벨레이블을사용하여, EPS 의전체학습목표는살리언시손실 \mathcal{L}_{sal} 과분류손실 \mathcal{L}_{cls} 의두부분으로구성됩니다. 먼저, 살리언시손실 \mathcal{L}_{sal} 은실제살리언시맵 M_s 와추정된살리언시맵 \hat{M}_s 간의평균픽셀레벨거리를측정하여공식화됩니다.

$$\mathcal{L}_{sal} = \frac{1}{H \cdot W} \|M_s - \hat{M}_s\|^2, \quad (3)$$

여기서 M_s 는 DUTS 데이터셋 [39]에서훈련된기성살리언시탐지모델- PFAN [51]에서얻은것입니다. 우리의방법은살리언시탐지모델에관계없이모든이전예술품을일관되게가능합니다.

다음으로, 분류손실은각타겟클래스에대한로컬라이제이션맵에대한글로벌평균풀링의결과인이미지레벨레이블 y 와그예측 $\hat{y} \in \mathbb{R}^C$ 간의다중레이블소프트마진손실로계산됩니다.

$$\mathcal{L}_{cls} = -\frac{1}{C} \sum_{i=1}^C y_i \log \sigma(\hat{y}_i) + (1 - y_i) \log (1 - \sigma(\hat{y}_i)), \quad (4)$$

여기서 $\sigma(\cdot)$ 는시그모이드함수입니다. 마지막으로, 총학습손실은다중레이블분류손실과살리언시손실의합입니다, 즉, $\mathcal{L}_{total} = \mathcal{L}_{cls} + \mathcal{L}_{sal}$.

그림 2에서보이는것처럼, \mathcal{L}_{sal} 은대상객체와배경을포함한 $C + 1$ 클래스의매개변수를업데이트하는데관여합니다. 반면, \mathcal{L}_{cls} 는배경클래스를제외한 C 클래스의레이블예측만평가하며, \mathcal{L}_{cls} 의그라디언트는배경클래스로흐르지않습니다. 그러나, \mathcal{L}_{cls} 는분류기훈련을감독하기때문에배경클래스의예측에암묵적으로영향을미칠수있습니다.

4. 실험설정

데이터셋. 우리는두개의인기있는벤치마크데이터셋, PASCAL VOC 2012 [12]와 MS COCO 2014 [30]에대한실증연구를수행합니다. PASCAL VOC 2012 는 21 개의클래스 (즉, 20 개의객체와배경) 로구성되어있으며, 각각 1,464, 1,449, 1,456 개의훈련, 검증, 테스트세트이미지를포함합니다. 의미론적분할의일반적인관행에따라, 우리는 10,582 개의이미지로구성된증강훈련세트를사용합니다 [17]. 다음으로, COCO 2014 는배경을포함한 81 개의클래스로구성되어있으며, 훈련과검증을위해각각 82,081 개와 40,137 개의이미지를포함하고있으며, 대상클래스가없는이미지는 [9]에서와같이제외됩니다. 일부객체의실제분할레이블이서로겹치기때문에, 우리는동일한 COCO 데이터셋에서겹침문제를해결한 COCO-Stuff [4]의실제분할레이블을채택합니다.

평가프로토콜. 우리는 PASCAL VOC 2012 의검증및테스트세트와 COCO 2014 의검증세트로우리의방법을검증합니다. PASCAL VOC 2012 의테스트세트에대한평가결과공식 PASCAL VOC 평가서버에서연습합니다. 또한, 우리는분할모델의정확성을측정하기위해평균교차-오버-유니온 (mIoU) 을채택합니다.

구현세부사항. 우리는출력스트라이드가 8 인 ResNet38 [45]을우리의방법의백본네트워크로선택합니

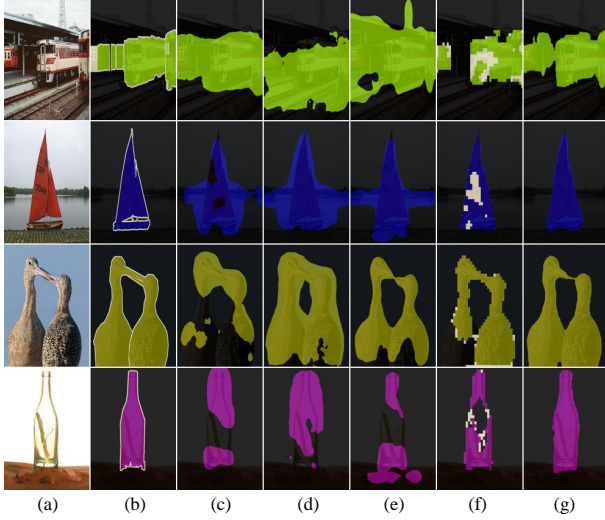


Figure 4. PASCAL VOC 2012 에서의의사마스크에대한정성적비교. (a) 입력이미지, (b) 정답, (c) CAM, (d) SEAM, (e) ICD, (f) SGAN 및 (g) 우리의 EPS.

다. 모든백본모델은 ImageNet [11]에서사전훈련되었습니다. 우리는배치크기가 8 인 SGD 옵티마이저를사용합니다. 우리의방법은학습률 0.01(마지막컨볼루션레이어의경우 0.1) 로 20k 반복까지훈련됩니다. 데이터증강을위해, 우리는무작위스케일링, 무작위플리핑, 그리고 448×448 로무작위크롭을사용합니다. 분할네트워크의경우, 우리는 DeepLab-LargeFOV (V1) [7]와 DeepLab-ASPP (V2) [8], 그리고 VGG16 과 ResNet101 을백본네트워크로채택합니다. 구체적으로, 우리는네가지분할네트워크를사용합니다: VGG16 기반 DeepLab-V1 및 DeepLab-V2, ResNet101 기반 DeepLab-V1 및 DeepLab-V2. 더 자세한설정정보는보충자료에있습니다.

5. 실험결과

5.1. 경계및동시발생처리

경계불일치문제. 의사마스크의경계를검증하기위해, 우리는최신방법 [32, 41, 52]과경계의품질을비교합니다. 우리는 PASCAL VOC 2011 에서경계주석과경계벤치마크를제공하는 SBD [17]를활용합니다. [32]에서와같이, 경계의품질은라플라시안에지검출기로부터의사마스크의에지를계산하여클래스비특이적방식으로평가됩니다. 그런 다음, 예측된경계와실제경계를비교하여재현율, 정밀도, F1-스코어를측정하여경계품을평가합니다. 표 1은우리의방법이모든세가지메트릭에서다른방법보다크게우수하다는것을보고합니다. 그림 4의정성적예시는우리의방법이 다른모든방법보다더정확한경계를포착할수있음을보여줍니다.

동시발생문제. 여러연구 [20, 25, 28, 35]에서논의된바와같이, 우리는 PASCAL VOC 2012 에서일부배경클래스가대상객체와주주함께나타나는것을관찰합니다. 우리는 PASCAL-CONTEXT 데이터셋 [33]을사용하여동시발생객체의빈도를정량적으로분석합니다. 이데이터셋은

방법	재현율 (%)	정밀도 (%)	F1-점수 (%)
CAM [52] _{CVPR'16}	22.3	35.8	27.5
SEAM [41] _{CVPR'20}	40.2	45.0	42.5
BES [32] _{ECCV'20}	45.5	46.4	45.9
우리의 EPS	60.0	73.1	65.9

Table 1. SBD trainval 세트에서경계정확도를평가했습니다. BES 의결과는 [32]에서제안된경계예측네트워크에서측정된것입니다.

방법	보트 w/ 물	기차 w/ 철도	기차 w/ 플랫폼
CAM [52] _{CVPR'16}	0.74 (33.1)	0.11 (52.9)	0.09 (49.6)
SEAM [41] _{CVPR'20}	1.13 (30.7)	0.24 (48.6)	0.20 (45.5)
ICD [13] _{CVPR'20}	0.47 (41.4)	0.11 (56.7)	0.09 (49.2)
SGAN [47] _{ACCESS'20}	0.10 (42.3)	0.02 (48.8)	0.01 (36.3)
우리의 EPS	0.10 (55.0)	0.02 (78.1)	0.01 (73.0)

Table 2. 공존문제를다루는대표적인기존방법들과의비교. 각 항목은 **파란색**으로표시된 $m_{k,c}$ (낮을수록 좋음) 과괄호안의 IoU (높을수록 좋음) 입니다.

	Baseline	Na"ive	Pre-defined	Our adaptive
mIoU	66.1	66.5	67.9	69.4

Table 3. 지도선택전략의효과. 다양한지도선택전략을사용한하의사마스크의정확도는 PASCAL VOC 2012 훈련세트에서평가됩니다.

전체장면에대한픽셀수준주석을제공합니다 (예: 물및 철도). 우리는세가지동시발생쌍을선택합니다; 보트와 물, 기차와 철도, 그리고 기차와 플랫폼. 우리는대상클래스에 대한 IoU 와대상클래스와그동시발생클래스간의 혼동비율을비교합니다. 혼동비율은동시발생클래스가대상클래스로잘못에측되는정도를측정합니다. 혼동비율 $m_{k,c}$ 는 $m_{k,c} = FP_{k,c}/TP_c$ 로계산되며, 여기서 $FP_{k,c}$ 는동시발생클래스 k 에대해대상클래스 c 로잘못분류된픽셀수이고, TP_c 는대상클래스 c 에대한진양성픽셀수입니다. 동시발생문제에대한더자세한분석은보충자료에있습니다.

Table 2는 EPS 가다른방법들보다일관되게낮은혼동비율을보인다고보고합니다. SGAN [47]은우리와매우유사한혼동비율을가지고있지만, 우리방법은 IoU 측면에서목표클래스를훨씬정확하게포착합니다. 흥미롭게도, SEAM 은높은혼동비율을보이며 CAM 보다도나쁩니다. 이는 SEAM [41]이자기지도학습을적용하여목표객체의전체범위를커버하도록학습하기때문인데, 이는목표객체의우연한픽셀에쉽게속습니다. 반면에, CAM 은목표객체의가장변별적인영역만포착하고덜변별적인부분, 예를들어우연한클래스는커버하지않습니다. 우리는이현상을 Figure 4에서도관찰할수있습니다.

5.2. 맵선택전략의효과

우리는주목성맵의오류를완화하기위한맵선택전략의효과를평가합니다. 우리는맵선택모듈을사용하지않는기본선

방법	w/o 정제없음	w/ CRF [26]	w/ AffinityNet [2]
CAM [52] _{CVPR'16}	48.0	-	58.1
SEAM [41] _{CVPR'20}	55.4	56.8	63.6
ICD [32] _{CVPR'20} *	59.9	62.2	-
SGAN [47] _{ACCESS'20} *	62.8	-	-
우리의 EPS	69.4	71.4	71.6

Table 4. PASCAL VOC 2012 학습세트에서평가된의사마스크의정확도 (mIoU). * 는신뢰도가낮은픽셀이무시됨을나타내며, 다른방법들은모든픽셀을평가에서사용합니다.

과세가지다른맵선택전략을비교합니다. 가장단순한전략으로, 전경맵은모든객체위치맵의합집합이며, 배경맵은배경클래스의위치맵과같습니다 (즉, 단순전략). 다음으로, 우리는다음예외를제외하고단순전략을따릅니다. 몇몇사전결정된클래스 (예: 소파, 의자, 식탁) 의위치맵은배경맵에할당됩니다 (즉, 사전정의된클래스전략). 마지막으로, 제안된선택방법은 Section 3.2에서설명한대로위치맵과주목성맵간의중첩비율을활용합니다 (즉, 우리의적응형전략).

Table 3는우리의적응형전략이주목성맵의체계적인편향을효과적으로처리할수있음을보여줍니다. 단순전략은위치맵에서추정된주목성맵을생성할때편향고려가없음을의미합니다. 이경우, 특히 소파, 의자또는 식탁클래스에서의사마스크의성능이저하됩니다. 사전정의된클래스를사용하는성능은주목성맵에서누락된클래스를무시함으로써편향을완화할수있음을보여줍니다. 그러나이는인간관찰자에의한수동선택이필요하므로덜실용적이며이미지별로최적의결정을내릴수없습니다. 반면에, 우리의적응형전략은편향을자동으로처리하고주어진주목성맵에대해더효과적인결정을내립니다.

5.3. 최신기술과의비교

의사마스크의정확도. 우리는 [2, 41]에서사용된일반적인관행인다양한스케일의이미지에서예측결과를집계하여다중스케일추론을채택합니다. 그런다음, 우리는 CAM [52]과세가지최신방법, 즉 SEAM [41], ICD [13], SGAN [47]과비교하여학습세트에서의사마스크의정확도를평가합니다. 여기서, 학습세트에서의사마스크의정확도를측정하는것은 WSSS 에서일반적인프로토콜입니다. 왜냐하면학습세트의의사마스크가세분화모델을감독하는데사용되기때문입니다. Table 4은의사마스크의정확도를요약하고우리의방법이기준의모든방법을큰차이로능가함을나타냅니다 (즉, 7-21% 차이). Figure 4은의사마스크의질적예를시각화하여우리의방법이객체경계를현저히개선하고의사마스크의품질측면에서세가지최신방법을크게능가함을확인합니다. 우리의방법은객체의정확한경계를포착할수있으며 (2 번째행) 따라서객체의전체범위를자연스럽게커버하고 (3 번째행) 우연한픽셀을완화할수있습니다 (1 번째행). 우리의방법의더많은예제와실패사례는보충자료에제공됩니다.

세분화맵의정확도. 이전방법들 [2, 13, 41]은의사마스크를생성하고 CRF 후처리알고리즘 [26] 또는친화네트워크 [2]로이를정제합니다. 한편, Table 4에나타난바와같이,

방법	세그.	지원.	val	test
SEC [25] _{ECCV'16}	V1	I.	50.7	51.7
AffinityNet [2] _{CVPR'18}	V1	I.	58.4	60.5
ICD [13] _{CVPR'20}	V1	I.	61.2	60.9
BES [32] _{ECCV'20}	V1	I.	60.1	61.1
GAIN [28] _{CVPR'18}	V1	I.+S.	55.3	56.8
MCOF [40] _{CVPR'18}	V1	I.+S.	56.2	57.6
SSNet [48] _{ICCV'19}	V1	I.+S.	57.1	58.6
DSRG [20] _{CVPR'18}	V2	I.+S.	59.0	60.4
SeeNet [19] _{NeurIPS'18}	V1	I.+S.	61.1	60.7
MDC [44] _{CVPR'18}	V1	I.+S.	60.4	60.8
FickleNet [27] _{CVPR'18}	V2	I.+S.	61.2	61.9
OAA [21] _{ICCV'19}	V1	I.+S.	63.1	62.8
ICD [13] _{CVPR'20}	V1	I.+S.	64.0	63.9
Multi-Est. [14] _{ECCV'20}	V1	I.+S.	64.6	64.2
Split. & Merge. [50] _{ECCV'20}	V2	I.+S.	63.7	64.5
SGAN [47] _{ACCESS'20}	V2	I.+S.	64.2	65.0
우리의 EPS	V1	I.+S.	66.6	67.9
	V2	I.+S.	67.0	67.3

Table 5. PASCAL VOC 2012 에서의세그멘테이션결과 (mIoU). 모든결과는 VGG16 을기반으로합니다. 모든실험에서최고의점수는굵게표시되어있습니다.

Method	Seg.	Sup.	val	test
ICD [13] _{CVPR'20}	V1	I.	64.1	64.3
SC-CAM [5] _{CVPR'20}	V1	I.	66.1	65.9
BES [32] _{ECCV'20}	V2	I.	65.7	66.6
LIID [31] _{TPAMI'20}	V2	I.	66.5	67.5
MCOF [40] _{CVPR'18}	V1	I.+S.	60.3	61.2
SeeNet [19] _{NeurIPS'18}	V1	I.+S.	63.1	62.8
DSRG [20] _{CVPR'18}	V2	I.+S.	61.4	63.2
FickleNet [27] _{CVPR'18}	V2	I.+S.	64.9	65.3
OAA [21] _{ICCV'19}	V1	I.+S.	65.2	66.4
Multi-Est. [14] _{ECCV'19}	V1	I.+S.	67.2	66.7
MCIS [38] _{ECCV'20}	V1	I.+S.	66.2	66.9
SGAN [47] _{ACCESS'20}	V2	I.+S.	67.1	67.2
ICD [13] _{CVPR'20}	V1	I.+S.	67.8	68.0
Our EPS	V1	I.+S.	71.0	71.8
	V2	I.+S.	70.9	70.8

Table 6. PASCAL VOC 2012 에서의세그멘테이션결과 (mIoU). 모든결과는 ResNet101 을기반으로합니다.

우리가생성한의사마스크는충분히정확하여의사마스크에대한추가정제없이세분화네트워크를훈련합니다. 우리는 PASCAL VOC 2012 데이터셋의네가지세분화네트워크에서우리의방법을다른방법들과광범위하게평가하고정확하게비교합니다.

우리의방법은세그멘테이션네트워크에관계없이다른방법들보다현저히더나은성능을보입니다. Table 5은동일한 VGG16 백본을사용했을때우리의방법이다른방법들보다더정확하다는것을보고합니다. 또한, VGG16 에서의우리의결과는더강력한백본 (i.e. Table 6의 ResNet101 기반) 으로기반한다른기존방법들과비교할때비슷하거나



Figure 5. PASCAL VOC 2012 에서의세분화결과에대한정성적예시입니다. (a) 입력이미지, (b) 정답데이터, (c) 우리의 EPS.

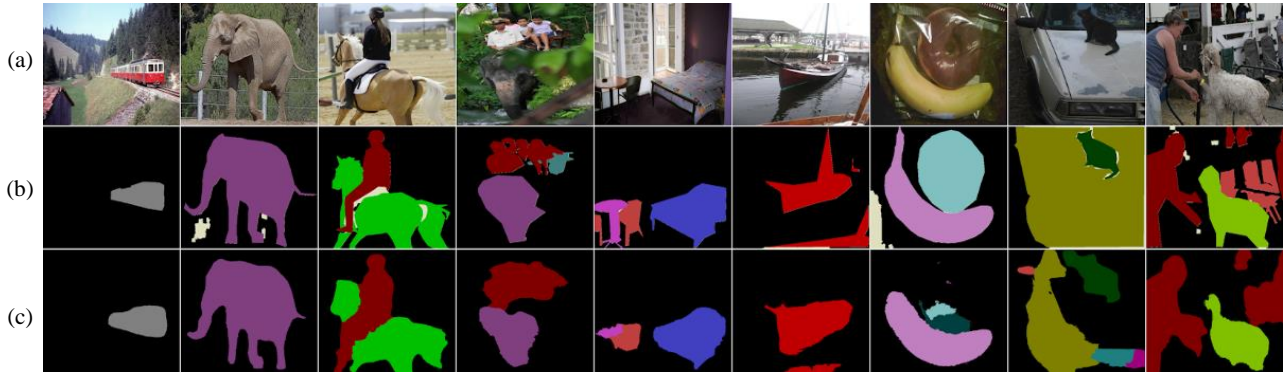


Figure 6. MS COCO 2014 에서의세분화결과에대한정성적예시입니다. (a) 입력이미지, (b) 정답및 (c) 우리의 EPS.

방법	Seg.	Sup.	val
SEC [25] _{ECCV'16}	V1	I.	22.4
DSRG [20] _{CVPR'18}	V2	I.+S.	26.0
ADL [9] _{TPAMI'20}	V1	I.+S.	30.8
SGAN [47] _{ACCESS'20}	V2	I.+S.	33.6
우리의 EPS	V2	I.+S.	35.7

Table 7. MS COCO 2014 에서의세그멘테이션결과 (mIoU). 모든결과는 VGG16 을기반으로합니다.

심지어우수합니다. 우리의방법은기존방법들에비해명확한개선을보여줍니다. 마지막으로, Table 6은우리의방법 (ResNet101 기반 DeepLab-V1 과 saliency map 을사용) 이 PASCAL VOC 2012 데이터셋에서새로운최첨단성능 (검증세트에서 71.0, 테스트세트에서 71.8) 을달성했음을보여줍니다. 기존최첨단모델들이달성한이득은약 1% 였음을강조합니다. 반면, 우리의방법은이전최고기록보다 3% 이상높은이득을달성합니다. Figure 5는 PASCAL VOC 2012 에서우리의세그멘테이션결과와질적예시를시각화합니다. 이러한결과는우리의방법이정확한경계를제공하고공존문제를성공적으로해결함을확인합니다.

Table 7에서우리는 COCO 2014 데이터셋에서우리의방법을추가로평가합니다. 우리는 VGG16 기반 DeepLab-V2 를세그멘테이션네트워크로사용하여 COCO 데이터셋에서최첨단 WSSS 모델인 SGAN [47]과비교합니다.

우리의방법은검증세트에서 35.7 mIoU 를달성하며, 이는 SGAN [47]보다 1.9% 높습니다. 결과적으로, 우리는 COCO 2014 데이터셋에서새로운최첨단정확도를달성합니다. 두데이터셋에서기존최첨단성능을뛰어넘는이러한뛰어난성능은우리의방법의효과를확인합니다; 지역화맵과 saliency map 을완전히활용하여, 목표객체의전체를정확하게포착하고기존모델의단점을보완합니다. Figure 6는 COCO 2014 데이터셋에서세그멘테이션결과와질적예시를보여줍니다. 우리의방법은몇개의객체가가려지지않고나타날때잘작동하지만, 많은작은객체를처리하는데는덜효과적입니다. 우리의방법의더많은예시와실패사례는보충자료에제공됩니다.

saliency detection 모델의효과. 다양한 saliency detection 모델의효과를조사하기위해, 우리는세가지 saliency 모델을채택합니다; PFAN [51] (우리의기본값), OAA [21]와 ICD [13]에서사용된 DSS [18], 그리고 USPS [34] (i.e., 비지도탐지모델). Resnet101 기반 DeepLab-V1 에서의세그멘테이션결과 (mIoU) 는 PFAN 으로 71.0/71.8, DSS 로 70.0/70.1, USPS 로 68.8/69.9 (검증세트와테스트세트) 입니다. 이러한점수는 Table 6의 다른모든방법보다여전히더정확하다는것을지원합니다. 특히, 비지도 saliency 모델을사용하는우리의 EPS 는지도 saliency 모델을사용하는모든기존방법을능가합니다.

6. 결론

우리는 explicit pseudo-pixel supervision (EPS) 라는 새로운약지도세그멘테이션프레임워크를제안합니다. 지역화맵과 saliency map 간의상호보완적관계에서영감을 받아, 우리의 EPS 는 saliency map 과지역화맵을결합한 의사픽셀피드백에서학습합니다. 우리의공동훈련체계덕분에, 우리는양쪽의노이즈나누락된정보를성공적으로보완합니다. 결과적으로, 우리의 EPS 는정확한객체경계를포착하고비목표객체의공존픽셀을제거하여의사마스크의품질을현저히향상시킵니다. 광범위한평가와다양한사례연구는우리의 EPS 의효과와 PASCAL VOC 2012 및 MS COCO 2014 데이터셋에서 WSSS 에대한새로운최첨단정확도를입증합니다.

감사의말씀. 우리는 Duhyeon Bang 과 Junsuk Choe 에게피드백을주신것에대해감사드립니다. 이연구는 MSIP(과학기술정보통신부) 에서지원하는 NRF Korea 의기초과학연구프로그램 (NRF-2019R1A2C2006123, 2020R1A4A1016619) 과 MSIT(과학기술정보통신부) 에서지원하는 IITP(정보통신기획평가원) 지원금 (2020-0-01361, 인공지능대학원프로그램 (연세대학교)), 그리고한국정부에서지원하는한국의료기기개발기금지원금 (프로젝트번호: 202011D06) 에의해지원되었습니다.

References

- [1] Jiwoon Ahn, Sunghyun Cho, and Suha Kwak. Weakly supervised learning of instance segmentation with inter-pixel relations. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 2209–2218, 2019. **2**
- [2] Jiwoon Ahn and Suha Kwak. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 4981–4990, 2018. **2, 6**
- [3] Nikita Araslanov and Stefan Roth. Single-stage semantic segmentation from image labels. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 4253–4262, 2020. **2**
- [4] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1209–1218, 2018. **5**
- [5] Yu-Ting Chang, Qiaosong Wang, Wei-Chih Hung, Robinson Piramuthu, Yi-Hsuan Tsai, and Ming-Hsuan Yang. Weakly-supervised semantic segmentation via sub-category exploration. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 8991–9000, 2020. **2, 7**
- [6] Arslan Chaudhry, Puneet K. Dokania, and Philip H. S. Torr. Discovering class-specific pixels for weakly-supervised semantic segmentation. In Proceedings of the British Machine Vision Conference, 2017. **2**
- [7] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. In International Conference on Learning Representations, 2015. **5**
- [8] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE Transactions on Pattern Analysis and Machine Intelligence, 40(4):834–848, 2017. **5**
- [9] Junsuk Choe, Seungho Lee, and Hyunjung Shim. Attention-based dropout layer for weakly supervised single object localization and semantic segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020. **1, 5, 7**
- [10] Junsuk Choe, Seong Joon Oh, Seungho Lee, Sanghyuk Chun, Zeynep Akata, and Hyunjung Shim. Evaluating weakly supervised object localization methods right. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 3133–3142, 2020. **2**
- [11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 248–255. IEEE, 2009. **5**
- [12] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. International Journal of Computer Vision, 111(1):98–136, 2015. **4**
- [13] Junsong Fan, Zhaoxiang Zhang, Chunfeng Song, and Tieniu Tan. Learning integral objects with intra-class discriminator for weakly-supervised semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 4283–4292, 2020. **1, 2, 4, 6, 7, 8**
- [14] Junsong Fan, Zhaoxiang Zhang, and Tieniu Tan. Employing multi-estimations for weakly-supervised semantic segmentation. In Proceedings of the European Conference on Computer Vision, 2020. **2, 6, 7**
- [15] Junsong Fan, Zhaoxiang Zhang, Tieniu Tan, Chunfeng Song, and Jun Xiao. Cian: Cross-image affinity net for weakly supervised semantic segmentation. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, pages 10762–10769, 2020. **2**
- [16] Ruochen Fan, Qibin Hou, Ming-Ming Cheng, Gang Yu, Ralph R. Martin, and Shi-Min Hu. Associating inter-image salient instances for weakly supervised semantic segmentation. In Proceedings of the European Conference on Computer Vision, pages 367–383, 2018. **2**
- [17] Bharath Hariharan, Pablo Arbeláez, Lubomir Bourdev, Subhransu Maji, and Jitendra Malik. Semantic contours from inverse detectors. In 2011 International Conference on Computer Vision, pages 991–998. IEEE, 2011. **5**
- [18] Qibin Hou, Ming-Ming Cheng, Xiaowei Hu, Ali Borji, Zhuowen Tu, and Philip HS Torr. Deeply supervised

- salient object detection with short connections. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 3203–3212, 2017. [2](#), [3](#), [8](#)
- [19] Qibin Hou, PengTao Jiang, Yunchao Wei, and Ming-Ming Cheng. Self-erasing network for integral object attention. In Advances in Neural Information Processing Systems, pages 549–559, 2018. [6](#), [7](#)
- [20] Zilong Huang, Xinggang Wang, Jiasi Wang, Wenyu Liu, and Jingdong Wang. Weakly-supervised semantic segmentation network with deep seeded region growing. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 7014–7023, 2018. [2](#), [5](#), [6](#), [7](#)
- [21] Peng-Tao Jiang, Qibin Hou, Yang Cao, Ming-Ming Cheng, Yunchao Wei, and Hong-Kai Xiong. Integral object mining via online attention accumulation. In Proceedings of the IEEE International Conference on Computer Vision, pages 2070–2079, 2019. [1](#), [2](#), [4](#), [6](#), [7](#), [8](#)
- [22] Anna Khoreva, Rodrigo Benenson, Jan Hosang, Matthias Hein, and Bernt Schiele. Simple does it: Weakly supervised instance and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 876–885, 2017. [1](#)
- [23] Dahun Kim, Donghyeon Cho, Donggeun Yoo, and In So Kweon. Two-phase learning for weakly supervised object localization. In Proceedings of the IEEE International Conference on Computer Vision, pages 3534–3543, 2017. [1](#)
- [24] Alexander Kolesnikov and Christoph Lampert. Improving weakly-supervised object localization by micro-annotation. In Edwin R. Hancock Richard C. Wilson and William A. P. Smith, editors, Proceedings of the British Machine Vision Conference, pages 92.1–92.12. BMVA Press, September 2016. [2](#)
- [25] Alexander Kolesnikov and Christoph H Lampert. Seed, expand and constrain: Three principles for weakly-supervised image segmentation. In Proceedings of the European Conference on Computer Vision, pages 695–711. Springer, 2016. [1](#), [2](#), [5](#), [6](#), [7](#)
- [26] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In Advances in Neural Information Processing Systems, pages 109–117, 2011. [6](#)
- [27] Jungbeom Lee, Eunji Kim, Sungmin Lee, Jangho Lee, and Sungroh Yoon. Ficklenet: Weakly and semi-supervised semantic image segmentation using stochastic inference. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 5267–5276, 2019. [1](#), [2](#), [6](#), [7](#)
- [28] Kunpeng Li, Ziyang Wu, Kuan-Chuan Peng, Jan Ernst, and Yun Fu. Tell me where to look: Guided attention inference network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 9215–9223, 2018. [1](#), [2](#), [5](#), [6](#)
- [29] Di Lin, Jifeng Dai, Jiaya Jia, Kaiming He, and Jian Sun. Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 3159–3167, 2016. [1](#)
- [30] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In Proceedings of the European Conference on Computer Vision, pages 740–755. Springer, 2014. [4](#)
- [31] Yun Liu, Yu-Huan Wu, Pei-Song Wen, Yu-Jun Shi, Yu Qiu, and Ming-Ming Cheng. Leveraging instance-, image-and dataset-level information for weakly supervised instance segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020. [2](#), [7](#)
- [32] Chen Liyi, Wu Weiwei, Chenchen Fu, Xiao Han, and Yuntao Zhang. Weakly supervised semantic segmentation with boundary exploration. In Proceedings of the European Conference on Computer Vision, 2020. [1](#), [5](#), [6](#), [7](#)
- [33] Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. The role of context for object detection and semantic segmentation in the wild. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 891–898, 2014. [5](#)
- [34] Tam Nguyen, Maximilian Dax, Chaithanya Kumar Mummadi, Nhung Ngo, Thi Hoai Phuong Nguyen, Zhongyu Lou, and Thomas Brox. Deepusps: Deep robust unsupervised saliency prediction via self-supervision. In Advances in Neural Information Processing Systems, pages 204–214, 2019. [2](#), [3](#), [8](#)
- [35] Seong Joon Oh, Rodrigo Benenson, Anna Khoreva, Zeynep Akata, Mario Fritz, and Bernt Schiele. Exploiting saliency for object segmentation from image level labels. In 2017 IEEE Conference on Computer Vision and Pattern Recognition, pages 5038–5047. IEEE, 2017. [2](#), [5](#)
- [36] Deepak Pathak, Philipp Krahenbuhl, and Trevor Darrell. Constrained convolutional neural networks for weakly supervised segmentation. In Proceedings of the IEEE International Conference on Computer Vision, pages 1796–1804, 2015. [1](#)
- [37] Pedro O Pinheiro and Ronan Collobert. From image-level to pixel-level labeling with convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1713–1721, 2015. [1](#)
- [38] Guolei Sun, Wenguan Wang, Jifeng Dai, and Luc Van Gool. Mining cross-image semantics for weakly supervised semantic segmentation. In Proceedings of the European Conference on Computer Vision, 2020. [2](#), [7](#)
- [39] Lijun Wang, Huchuan Lu, Yifan Wang, Mengyang Feng, Dong Wang, Baocai Yin, and Xiang Ruan. Learning to detect salient objects with image-level supervision. In Proceedings of the IEEE Conference on

- Computer Vision and Pattern Recognition, pages 136–145, 2017. [2](#), [4](#)
- [40] Xiang Wang, Shaodi You, Xi Li, and Huimin Ma. Weakly-supervised semantic segmentation by iteratively mining common object features. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1354–1362, 2018. [2](#), [4](#), [6](#), [7](#)
- [41] Yude Wang, Jie Zhang, Meina Kan, Shiguang Shan, and Xilin Chen. Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 12275–12284, 2020. [1](#), [2](#), [5](#), [6](#)
- [42] Yunchao Wei, Jiashi Feng, Xiaodan Liang, Ming-Ming Cheng, Yao Zhao, and Shuicheng Yan. Object region mining with adversarial erasing: A simple classification to semantic segmentation approach. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1568–1576, 2017. [2](#)
- [43] Yunchao Wei, Xiaodan Liang, Yunpeng Chen, Xiaohui Shen, Ming-Ming Cheng, Jiashi Feng, Yao Zhao, and Shuicheng Yan. Stc: A simple to complex framework for weakly-supervised semantic segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(11):2314–2320, 2016. [2](#), [4](#)
- [44] Yunchao Wei, Huaxin Xiao, Honghui Shi, Zequn Jie, Jiashi Feng, and Thomas S Huang. Revisiting dilated convolution: A simple approach for weakly-and semi-supervised semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 7268–7277, 2018. [2](#), [6](#)
- [45] Zifeng Wu, Chunhua Shen, and Anton Van Den Hengel. Wider or deeper: Revisiting the resnet model for visual recognition. Pattern Recognition, 90:119–133, 2019. [5](#)
- [46] Huaxin Xiao, Jiashi Feng, Yunchao Wei, Maojun Zhang, and Shuicheng Yan. Deep salient object detection with dense connections and distraction diagnosis. IEEE Transactions on Multimedia, 20(12):3239–3251, 2018. [2](#)
- [47] Qi Yao and Xiaojin Gong. Saliency guided self-attention network for weakly and semi-supervised semantic segmentation. IEEE Access, 8:14413–14423, 2020. [2](#), [4](#), [5](#), [6](#), [7](#), [8](#)
- [48] Yu Zeng, Yunzhi Zhuge, Huchuan Lu, and Lihe Zhang. Joint learning of saliency detection and weakly supervised semantic segmentation. In Proceedings of the IEEE International Conference on Computer Vision, pages 7223–7233, 2019. [2](#), [6](#)
- [49] Bingfeng Zhang, Jimin Xiao, Yunchao Wei, Mingjie Sun, and Kaizhu Huang. Reliability does matter: An end-to-end weakly supervised semantic segmentation approach. In Proceedings of the AAAI Conference on Artificial Intelligence, pages 12765–12772. AAAI Press, 2020. [2](#)
- [50] Tianyi Zhang, Guosheng Lin, Weide Liu, Jianfei Cai, and Alex Kot. Splitting vs. merging: Mining object regions with discrepancy and intersection loss for weakly supervised semantic segmentation. In Proceedings of the European Conference on Computer Vision, 2020. [2](#), [6](#)
- [51] Ting Zhao and Xiangqian Wu. Pyramid feature attention network for saliency detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 3085–3094, 2019. [1](#), [2](#), [3](#), [4](#), [8](#)
- [52] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 2921–2929, 2016. [1](#), [5](#), [6](#)