Pyramid Scene Parsing Network

Hengshuang Zhao, ianping Shi, Xiaojuan Qi, Xiaogang Wang, iaya Jia
The Chinese University of Hong Kong, SenseTime Group Limited

개요

장면 분석은 제한 없이 열린 어휘와 다양한 장면에서 도전적입니다. 본 논문에서는 피라미드 장면 구문 분석망(PSPNet)과 함께 피라미드 풀링 모듈을 통해 서로 다른 지역 기반 컨텍스트 집계에 의한 글로벌 컨텍스트 정보의 기능을 활용합니다. 우리의 글로벌 사전 표현은 장면 구문 분석 작업에서 좋은 품질 결과를 생성하는 데 효과적이며 PSPNet은 픽셀 수준 예측을 위한 우수한 프레임워크를 제공합니다. 제안된 접근 방식은 다양한 데이터 세트에서 state-of- the-art 성능을 달성했습니다. ImageNet 장면 구문 분석 챌린지 2016, PASCAL VOC 2012 벤치마크 및 Cityscapes 벤치마크에서 1위를 차지했습니다. 단일 PSPNet은 PASCAL VOC 2012에서 mIoU 정확도 85.4%, Cityscape에서 정확도 80.2%의 신기록을 달성했습니다.

1. 소개

의미론적 분할을 기반으로 하는 장면 분석은 컴퓨터 비전의 기본 주제입니다. 목표는 이미지의 각 픽셀에 범주 레이블을 할당하는 것입니다. 씬(scene) 구문 분석을 통해 씬(scene)을 완벽하게 이해할 수 있습니다. 각 요소의 레이블, 위치 및 모양을 예측합니다. 이 주제는 몇 가지 예를 들어 자동 주행, 로봇 감지의 잠재적 응용 분야로 널리 관심을 끌고 있습니다. 장면 구문 분석 어려움은 장면 및 레이블 다양성과 밀접한 관련이 있습니다. 선구적인 장면 구문 분석 작업은 LMO 데이터 세트의 2,688개 이미지에 대해 33개의 장면을 분류하는 것입니다. 보다 최근의 PASCAL VOC 의미 분할 및 PASCAL 컨텍스트 데이터 세트에는 의자와 소파, 말과 소 등과 같은 유사한 맥락의 레이블이 더 많이 포함되어 있습니다. 새로운 ADE20K 데이터 세트는 크고 제한되지 않은 열린 어휘와 더 많은 장면 클래스를 가진 가장 어려운 데이터 세트입니다. 몇 가지 대표적인 이미지가 그림 1에 나와 있습니다. 이러한 데이터 세트에 대한 효과적인 알고리듬을 개발하려면 몇 가지 어려움을 극복해야 합니다.

State-of-the-art 장면 구문 분석 프레임워크는 대부분 완전 컨볼루션 네트워크(FCN)를 기반으로 합니다. 심층 컨볼루션 신경망(CNN) 기반 방법은 동적 객체 이해를 높이지만 다양한 장면과 제한되지 않은 어휘를 고려할 때 여전히 어려움에 직면해 있습니다. 한 예는 그림 2의 첫 번째 줄에서 보트를 자동차로 오인합니다.

(a) Image (b) Ground Truth

그림 1. ADE20K 데이터 세트의 복잡한 장면

이러한 오류는 유사한 개체 모양 때문에 발생합니다. 그러나 그 장면이 강 근처의 보트 하우스로 묘사되기 전의 맥락에 대한 이미지를 볼 때, 정확한 예측을 해야 합니다. 정확한 장면 인식을 위해 지식 그래프는 장면 컨텍스트의 사전 정보에 의존합니다. 현재 FCN 기반 모델의 주요 문제는 글로벌 장면 범주 단서를 활용할 수 있는 적절한 전략의 부족이라는 것을 발견했습니다. 일반적인 복잡한 장면 이해를 위해, 이전에는 전역 이미지 수준 기능을 얻기 위해 공간 피라미드 풀링이 널리 사용되었으며, 공간 통계가 전체 장면 해석을 위한 좋은 설명자를 제공합니다. 공간 피라미드 풀링 네트워크는 능력을 더욱 향상시킵니다.

이러한 방법과 달리 적절한 전역 기능을 통합하기 위해 피라미드 장면 구문 분석 네트워크(PSPNet)를 제안합니다. 픽셀 예측을 위한 전통적인 확장 FCN 외에도, 우리는 픽셀 레벨 기능을 특별히 설계된 전역 피라미드 풀링 기능으로 확장합니다. 지역 및 글로벌 단서가 함께 최종 예측을 더욱 신뢰할 수 있게 만듭니다. 또한 우리는 크게 감독되는 손실을 가진 최적화 전략을 제안합니다.

우리는 본 논문에서 적절한 성능의 핵심인 모든 구현 세부 정보를 제공하고 코드와 훈련된 모델을 공개적으로 사용할 수 있도록 합니다.

NAT 접근 방식은 사용 가능한 모든 데이터 세트에서 state-of-the-art 성능을 달성합니다. ImageNet 장면 구문 분석 챌린지 2016의 챔피언이며, PASCAL VOC 2012 의미 분할 벤치마크에서 1위, 도시 장면 Cityscapes 데이터에서 1위를 차지했습니다. 그들은 PSPNet이 픽셀 수준 예측 작업에 대한 유망한 방향을 제공한다는 것을 보여주며, 이는 후속 작업에서 CNN 기반 스테레오 매칭, 광학 흐름, 깊이 추정 등에 도움이 될 수 있습니다. 우리의 주된 기여는 세 가지입니다.

- FCN 기반 픽셀 예측 프레임워크에 어려운 풍경 컨텍스트 기능을 내장하기 위해 피라미드 장면 구문 분석 네트워크를 제안합니다.
- 우리는 깊이 감독된 손실을 기반으로 심층 ResNet에 대한 효과적인 최적화 전략을 개발합니다.
- 우리는 모든 중요한 구현 세부 정보가 포함된 state-of-the-art 장면 구문 분석 및 의미 분할을 위한 실용적인 시스템을 구축합니다.

2. 관련 작업

다음에서는 장면 구문 분석 및 의미 분할 작업의 최근 발전을 검토합니다. 강력한 심층 신경망에 의해 구동되는 장면 구문 분석 및 의미 분할과 같은 픽셀 수준 예측 작업은 분류에서 완전히 연결된 계층을 컨볼루션 계층으로 대체함으로써 영감을 받은 큰 진전을 달성합니다. 신경망의 수용 영역을 확대하기 위해 확장된 컨볼루션 방법을 사용합니다. No 등은 분할 마스크를 학습하기 위해 디콘볼루션 네트워크가 있는 coarse-to-fine 구조를 제안했습니다. 기본 네트워크는 FCN 및 확장 네트워크입니다.

다른 작업은 주로 두 방향으로 진행됩니다. 한 줄에는 다중 스케일 피쳐 앙상블이 있습니다. 심층 네트워크에서, 더 높은 계층 기능은 더 많은 의미적 의미를 포함하고 더 적은 위치 정보를 포함합니다. 다중 스케일 기능을 결합하면 성능이 향상될 수 있습니다. 다른 방향은 구조 예측을 기반으로 합니다. 개척자 작업은 분할 결과를 세분화하기 위해 조건부 랜덤 필드(CRF)를 사후 처리로 사용했습니다.

다음 방법은 end-to-end 모델링을 통해 네트워크를 개선했습니다. 두 방향 모두 예측된 의미 경계가 객체에 맞는 장면 구문 분석의 현지화 능력을 개선합니다. 그러나 복잡한 장면에서 필요한 정보를 활용할 수 있는 여지가 여전히 많습니다. 다양한 장면 이해를 위해 글로벌 이미지 수준을 잘 활용하기 위해 심층 신경망에서가 아닌 전통적인 특징을 가진 글로벌 컨텍스트 정보를 추출하는 방법입니다. 객체 감지 프레임 워크에서도 유사한 개선이 이루어졌습니다.

FCN을 사용한 글로벌 평균 풀링이 의미론적 세분화 결과를 개선할 수 있음을 입증했습니다. 그러나, 우리의 실험은 이러한 글로벌 설명자가 도전적인 ADE20K 데이터에 충분히 대표적이지 않다는 것을 보여줍니다. 따라서 글로벌 풀링과 달리 피라미드 장면 구문 분석 네트워크를 통해 different-region-based 컨텍스트 집계를 통해 글로벌 컨텍스트 정보의 기능을 활용합니다.

3. PSPNet

우리는 장면 구문 분석에 FCN 방법을 적용할 때 대표적인 실패 사례를 관찰하고 분석하는 것으로 시작합니다. 그들은 이전의 효과적인 글로벌 컨텍스트로서 우리의 피라미드 풀링 모듈의 제안에 동기를 부여합니다. 그런 다음 그림 3에 설명된 피라미드 장면 구문 분석 네트워크(PSPNet)는 복잡한 장면 구문 분석에서 열린 어휘 개체 및 재료 식별의 성능을 향상시키기 위해 설명됩니다.

3.1. 주요 관찰 포인트

새로운 ADE20K 데이터 세트에는 150개의 물건/물체 범주 레이블(예: 벽, 하늘 및 나무)과 1,038개의 이미지 레벨 장면 설명자(예: 공항 터미널, 침실 및 거리)가 포함되어 있습니다. 그래서 많은 양의 레이블과 광대한 분포의 장면들이 생겨납니다. 에 제공된 FCN 기준선의 예측 결과를 검사하여 복잡한 장면 구문 분석에 대한 몇 가지 일반적인 문제를 요약합니다.

불일치 관계 상황 관계는 보편적이며 특히 복잡한 장면 이해에 중요합니다. 동시에 발생하는 시각적 패턴이 존재합니다. 예를 들어, 비행기는 도로를 건너지 않고 활주로에 있거나 하늘을 날 가능성이 있습니다. 그림 2의 첫 번째 줄 예에서, FCN은 노란색 상자에 있는 보트를 외관에 따라 "자동차"로 예측합니다. 하지만 일반 상식은 자동차가 강 위를 달리는 경우는 거의 없다는 것입니다. 상황별 정보를 수집할 수 있는 능력이 부족하면 잘못 분류될 가능성이 높아집니다.

혼동 카테고리 ADE20K 데이터 세트에는 분류에 혼란스러운 많은 클래스 레이블 쌍이 있습니다. 예를 들어 들판과 지구, 산과 언덕, 벽, 집, 건물, 마천루가 있습니다. 그들은 비슷한 외모를 가지고 있습니다. 전체 데이터 세트에 레이블을 지정한 전문가 주석자는 에 설명된 대로 여전히 17.60% 픽셀 오류를 발생시킵니다. 그림 2의 두 번째 줄에서, FCN은 상자 안에 있는 물체를 마천루와 건물의 일부로 예측합니다. 이러한 결과는 제외되어야만 전체 객체가 마천루 또는 빌딩이 될 수 있지만 둘 다 될 수는 없습니다. 이 문제는 범주 간의 관계를 활용하여 해결할 수 있습니다.

잘 보이지 않는 클래스 장면(scene)에 임의 크기의 개체/물건이 있습니다. 가로등과 간판과 같은 몇몇 작은 크기의 것들은 매우 중요할 수도 있지만 찾기 어렵습니다. 반대로, 큰 물체나 물건은 FCN의 수용 영역을 초과하여 불연속적인 예측을 일으킬 수 있습니다.

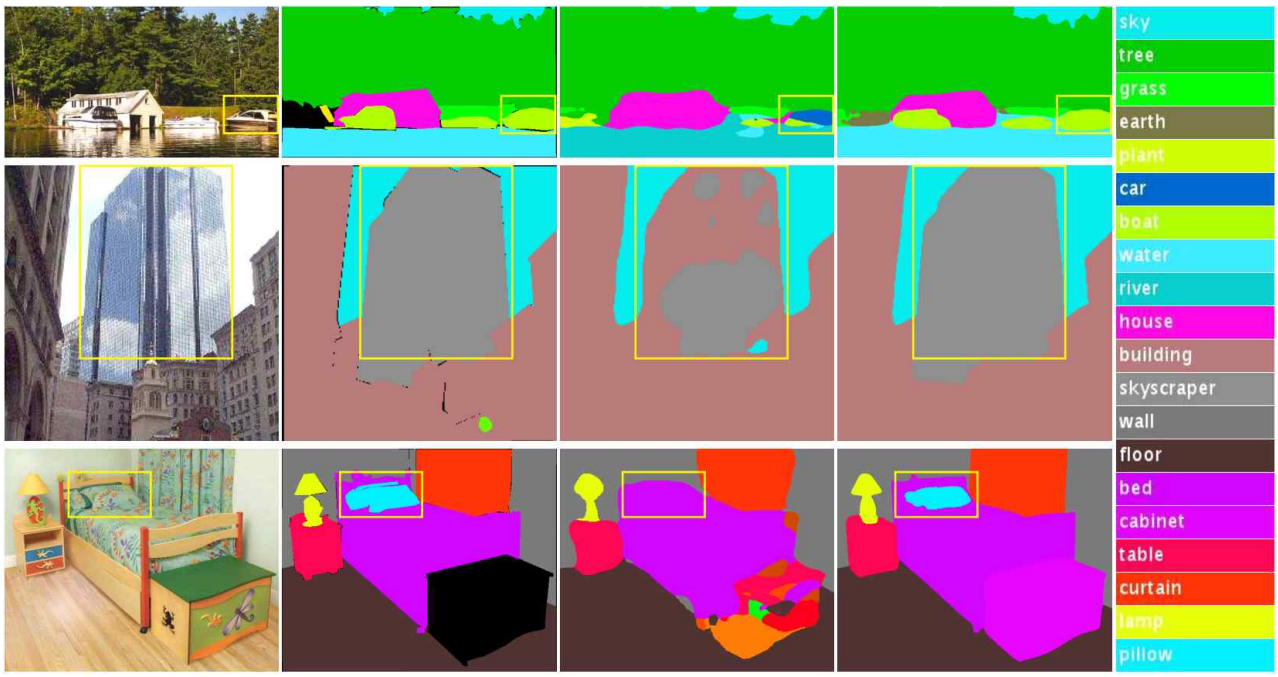


그림 2. ADE20K 데이터 세트에서 관찰되는 장면 구문 분석 문제입니다. 첫 번째 행은 불일치 관계에 대한 문제를 보여줍니다. 자동차는 보트보다 물 위에 있는 경우가 거의 없습니다. 두 번째 행은 클래스 "빌딩"이 "스카이스카이퍼"로 쉽게 혼동되는 혼란 범주를 보여줍니다. 세 번째 행은 눈에 띄지 않는 클래스를 보여줍니다. 이 예제에서 베개는 색상과 질감 면에서 침대 시트와 매우 유사합니다. 이러한 눈에 띄지 않는 물체는 FCN에 의해 쉽게 잘못 분류됩니다.

그림 2의 세 번째 줄에서 보는 바와 같이, 베개는 시트와 비슷한 외관을 가지고 있습니다. 전역 장면 범주를 간과하면 베개를 구문 분석하지 못할 수 있습니다. 현저하게 작거나 큰 객체의 성능을 향상시키려면 눈에 띄지 않는 범주가 포함된 여러 하위 영역에 많은 주의를 기울여야 합니다. 이러한 관찰을 요약하면, 많은 오류는 서로 다른 수용 필드에 대한 상황별 관계 및 전역 정보와 부분적으로 또는 완전히 관련이 있습니다. 따라서 적절한 global-scene-level가 있는 심층 네트워크는 장면 구문 분석 성능을 크게 향상시킬 수 있습니다.

3.2. 피라미드 풀링 모듈

위의 분석을 통해, 다음에서, 우리는 경험적으로 효과적인 전역 컨텍스트 사전임을 증명하는 피라미드 풀링 모듈을 소개합니다. 심층 신경망에서 수용 필드의 크기는 대략 우리가 컨텍스트 정보를 얼마나 사용하는지를 나타낼 수 있습니다. 이론적으로 ResNet의 수용 필드는 이미 입력 이미지보다 크지만, Zhou 등은 CNN의 경험적 수용 필드가 특히 높은 수준의 레이어에서 이론적 필드보다 훨씬 작다는 것을 보여줍니다. 이것은 많은 네트워크가 이전의 중요한 글로벌 풍경을 충분히 통합하지 못하게 만듭니다. 우리는 효과적인 글로벌 사전 표현을 제안함으로써 이 문제를 해결합니다.

글로벌 평균 풀링은 이미지 분류 작업에서 일반적으로 사용되는 글로벌 컨텍스트 사전의 좋은 기준 모델입니다.

비모수 장면 구문 분석, 의미론적 분할에 성공적으로 적용되었습니다. 그러나 ADE20K의 복잡한 장면 이미지와 관련하여, 이 전략은 필요한 정보를 포함하기에 충분하지 않습니다.

이러한 장면 이미지의 픽셀에는 많은 물건과 물체에 대한 주석이 달려 있습니다. 이들을 직접 융합하여 단일 벡터를 형성하면 공간적 관계가 상실되고 모호성이 발생할 수 있습니다. 하위 지역 컨텍스트와 함께 전역 컨텍스트 정보는 다양한 범주를 구별하는 데 도움이 됩니다. 보다 강력한 표현은 이러한 수용 필드를 가진 서로 다른 하위 영역의 정보를 융합할 수 있습니다. 장면/이미지 분류의 고전적인 작업에서도 유사한 결론이 도출되었습니다.

시각적 인식을 위한 심층 컨볼루션 네트워크의 공간 피라미드 풀링에서, 피라미드 풀링에 의해 생성된 다양한 수준의 특징 맵이 마침내 평평해지고 연결되어 분류를 위해 완전히 연결된 레이어에 공급되었습니다. 이 글로벌 사전은 이미지 분류를 위한 CNN의 고정 크기 제약을 제거하기 위해 설계되었습니다. 서로 다른 하위 영역 간의 컨텍스트 정보 손실을 더욱 줄이기 위해, 우리는 서로 다른 척도와 서로 다른 하위 영역을 가진 정보를 포함하는 계층적 전역 사전 작업을 제안합니다.

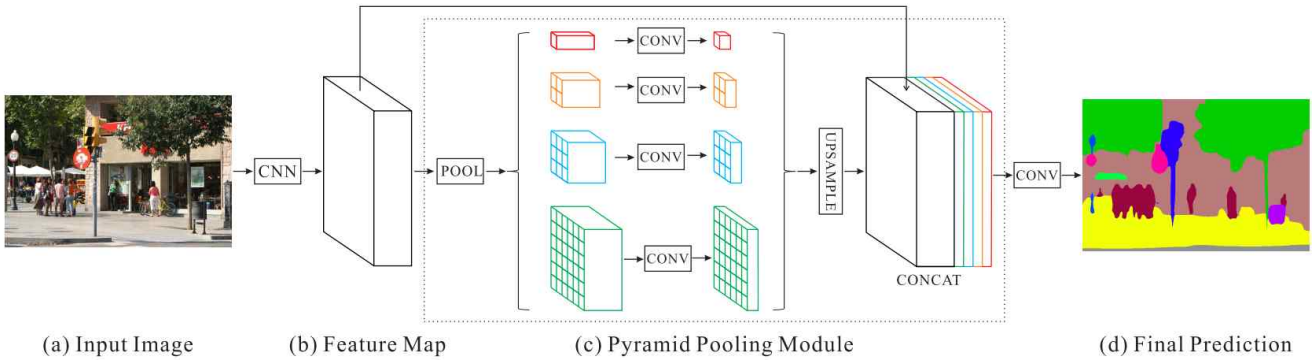


그림 3. 제안된 PSPNet에 대한 개요입니다. 입력 이미지(a)가 주어지면 먼저 CNN을 사용하여 마지막 컨볼루션 레이어(b)의 피처 맵을 가져온 다음 피라미드 구문 분석 모듈을 적용하여 다른 하위 영역 표현을 수확한 다음 업샘플링 및 연결 레이어를 형성하여 (c)의 로컬 및 전역 컨텍스트 정보를 모두 전달합니다. 마지막으로, 표현은 최종 픽셀당 예측(d)을 얻기 위해 컨볼루션 레이어에 입력됩니다.

우리는 그림 3의 (c) 부분에 표시된 것처럼 심층 신경망의 최종 레이어 기능 맵에서 사전 구축을 위한 전역 장면을 위한 피라미드 풀링 모듈이라고 부릅니다.

피라미드 풀링 모듈은 4개의 서로 다른 피라미드 축척에서 특징을 융합합니다. 빨간색으로 강조 표시된 가장 거친 수준은 단일 빈 출력을 생성하기 위한 글로벌 풀링입니다. 다음 피라미드 수준은 형상 지도를 서로 다른 하위 영역으로 분리하고 서로 다른 위치에 대해 풀링된 표현을 형성합니다. 피라미드 풀링 모듈의 다양한 레벨의 출력에는 다양한 크기의 피쳐 맵이 포함되어 있습니다. 전역 기능의 가중치를 유지하기 위해 각 피라미드 수준 뒤에 1 * 1 컨볼루션 레이어를 사용하여 피라미드의 수준 크기가 N인 경우 컨텍스트 표현의 차원을 원래 것의 1/N로 줄입니다. 그런 다음 이중 선형 보간을 통해 원래 형상 맵과 동일한 크기 형상을 얻기 위해 저차원 형상 맵을 직접 업샘플링합니다. 마지막으로, 다양한 수준의 기능이 최종 피라미드 풀링 글로벌 기능으로 연결됩니다.

각 레벨의 피라미드 수 및 크기는 수정할 수 있습니다. 피라미드 풀링 계층에 공급되는 피쳐 맵의 크기와 관련이 있습니다. 이 구조는 다양한 크기의 풀링 커널을 몇 단계로 채택하여 다양한 하위 영역을 추상화합니다. 따라서 다단계 커널은 표현상의 합리적인 간격을 유지해야 합니다.

우리의 피라미드 풀링 모듈은 빈 크기가 각각 1 * 1, 2 * 2, 3 * 3 및 6 * 6인 4단계입니다. 최대값과 평균값 사이의 풀링 작업 유형에 대해 5.2절의 차이를 보여주기 위해 광범위한 실험을 수행합니다.

3.3. 망 구조

피라미드 풀링 모듈을 사용하여, 우리는 그림 3과 같이 피라미드 중간 장면 구문 분석 네트워크(PSPNet)를 제안합니다. 그림 3(a)의 입력 이미지가 주어지면, 우리는 피처 맵을 추출하기 위해 확장된 네트워크 전략을 가진 사전 훈련된 ResNet 모델을 사용합니다. 최종 형상 맵 크기는 그림 3(b)와 같이 입력 이미지의 1/8입니다. 지도 위에 (c)에 표시된 피라미드 풀링 모듈을 사용하여 컨텍스트 정보를 수집합니다.

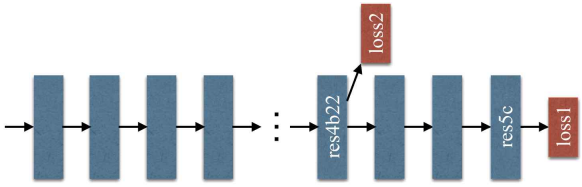


그림 4. ResNet101의 보조 손실에 대한 그림입니다. 각 파란색 상자는 잔여 블록을 나타냅니다. 보조 손실은 res4b22 잔여 블록 뒤에 추가됩니다.

4단계 피라미드를 사용하여 풀링 커널은 이미지의 전체, 절반 및 작은 부분을 커버합니다. 그들은 글로벌 선행으로 융합되었습니다. 그런 다음 (c)의 마지막 부분에서 이전 기능을 원래 기능 맵과 연결합니다. 이어서 (d)의 최종 예측 맵을 생성하기 위해 컨볼루션 레이어를 사용합니다.

우리의 구조를 설명하기 위해 PSPNet은 픽셀 수준 장면 구문 분석에 앞서 효과적인 전역 컨텍스트를 제공합니다. 피라미드 풀링 모듈은 글로벌 풀링보다 더 대표적인 수준의 정보를 수집할 수 있습니다. 계산 비용 측면에서, 우리의 PSPNet은 원래의 확장된 FCN 네트워크에 비해 그것을 크게 증가시키지 않습니다. 엔드 투 엔드 학습에서 글로벌 피라미드 풀링 모듈과 로컬 FCN 기능을 동시에 최적화할 수 있습니다.

4. ResNet 기반 FCN 심층 슈퍼 비전

사전 학습 기간을 많이 받은 네트워크는 우수한 성능을 제공합니다. 그러나 네트워크의 깊이가 증가하면 이미지 분류에 대한 추가적인 최적화 어려움이 발생할 수 있습니다. ResNet은 각 블록의 건너뛰기 연결을 통해 이 문제를 해결합니다. 심층 ResNet의 후자 계층은 주로 이전 계층을 기반으로 잔류물을 학습합니다. 우리는 반대로 추가적인 손실로 감독을 통해 초기 결과를 생성하고, 최종 손실로 이후 잔여물을 학습할 것을 제안합니다. 따라서, 심층 네트워크의 최적화는 두 가지로 분해되며, 각각은 해결하기 더 쉽습니다.

깊이 감독된 ResNet101 모델의 예는 그림 4에 나와 있습니다. 최종 분류기를 훈련시키기 위해 소프트맥스 손실을 사용하는 주요 분기를 제외하고, 4단계, 즉 res4b22 잔류 블록 이후에 또 다른 분류기가 적용됩니다. 역방향 보조 손실을 여러 개의 얕은 층으로 차단하는 릴레이 역 전파와 달리, 우리는 두 손실 함수가 이전의 모든 층을 통과하도록 합니다. 보조 손실은 학습 프로세스를 최적화하는 데 도움이 되며 마스터 분기 손실이 가장 큰 책임을 집니다. 보조 손실의 균형을 맞추기 위해 무게를 더합니다.

테스트 단계에서는 이 보조 분기를 포기하고 최종 예측을 위해 잘 최적화된 마스터 분기만 사용합니다. ResNet 기반 FCN에 대한 이러한 심층 지도 교육 전략은 다양한 실험 설정에서 광범위하게 유용하며 사전 훈련된 ResNet 모델과 함께 작동합니다. 이것은 그러한 학습 전략의 일반성을 나타냅니다. 자세한 내용은 섹션 5.2에 나와 있습니다.

5. 실험

우리가 제안한 방법은 장면 구문 분석 및 의미 분할 문제에 성공합니다. 우리는 이 섹션에서 ImageNet 장면 구문 분석 챌린지 2016, PASCAL VOC 2012 의미 분할 및 도시 장면 이해 데이터 세트 Cityscapes를 포함한 세 가지 데이터 세트에 대해 평가합니다.

5.1. 세부 구현 정보

실용적인 딥러닝 시스템의 경우, 악마는 항상 세부 사항에 있습니다. 우리의 구현은 공공 플랫폼 카페를 기반으로 합니다. 영감을 받아 현재 학습률이 $\left(1-\frac{er}{\max_i ter}\right)^{power}$ 과 동일한 "폴리" 학습 속도 정책을 사용합니다. 우리는 기본 학습률을 0.01로, 전력을 0.9로 설정합니다. ImageNet 실험의 경우 150K, PASCAL VOC의 경우 30K, Cityscape의 경우 90K로 설정된 반복 횟수를 늘리면 성능이 향상될 수 있습니다. 운동량과 무게 감쇠는 각각 0.9와 0.0001로 설정됩니다. 데이터 증강을 위해 모든 데이터 세트에 대해 0.5에서 2 사이의 랜덤 미러와 랜덤 크기를 채택하고, 추가로 -10도에서 10도 사이의 랜덤 회전과 ImageNet 및 PASCAL VOC의 랜덤 가우스 블러를 추가합니다. 이 포괄적인 데이터 확대 체계는 네트워크가 과적합에 저항하도록 만듭니다. 우리의 네트워크는 확장된 컨볼루션(convolution)을 포함합니다.실험 과정에서, 우리는 적절하게 큰 "크롭사이즈"가 좋은 성능을 낼 수 있고 배치 정규화 계층에서 "배치 크기"가 매우 중요하다는 것을 알게 되었습니다. GPU 카드의 물리적 메모리가 제한되어 있기 때문에 교육 중에 "배치 크기"를 16으로 설정합니다. 이를 위해 분기와 함께 카페를 수정하고 OpenMPI를 기반으로 여러 GPU에서 수집된 데이터에 대한 일괄 정규화를 지원하도록 합니다. 보조 손실의 경우 실험에서 무게를 0.4로 설정합니다.

Method	Mean IoU(%)	Pixel Acc.(%)
ResNet50-Baseline	37.23	78.01
ResNet50+B1+MAX	39.94	79.46
ResNet50+B1+AVE	40.07	79.52
ResNet50+B1236+MAX	40.18	79.45
ResNet50+B1236+AVE	41.07	79.97
ResNet50+B1236+MAX+DR	40.87	79.61
ResNet50+B1236+AVE+DR	**41.68**	**80.04**

표 1. 다른 설정으로 PSPNet을 조사합니다. 기준선은 확장된 네트워크를 사용하는 ResNet50 기반 FCN입니다. 'B1' 및 'B1236'은 각각 빈 크기 {1 * 1} 및 {1 * 1, 2 * 2, 3 * 3, 6 * 6}의 풀링 형상 맵을 나타냅니다. 'MAX'와 'AVE'는 각각 최대 풀링 작업과 평균 풀링 작업을 나타냅니다. 'DR'은 풀링 후 치수 축소를 취함을 의미합니다. 결과는 단일 척도 입력으로 검증 세트에서 테스트됩니다.

5.2. ImageNet Scene Parsing Challenge 2016

데이터셋 및 평가 메트릭스 ADE20K 데이터 세트는 Image Net 장면 구문 분석 챌린지 2016에 사용됩니다. 다른 데이터 세트와 달리, ADE20K는 총 1,038개의 이미지 레벨 레이블이 있는 최대 150개의 클래스 및 다양한 장면에서 더 어렵습니다. 과제 데이터는 교육, 검증 및 테스트를 위해 20K/2K/3K 이미지로 나뉩니다. 또한 장면에서 객체와 내용을 모두 구문 분석해야 하므로 다른 데이터 세트보다 어렵습니다. 평가를 위해 픽셀 단위 정확도(Pixel Acc.)와 유니언에 대한 클래스 단위 교차 평균(Mean IoU)이 모두 사용됩니다.

PSPNet에 대한 절제 연구 PSPNet을 평가하기 위해, 우리는 풀링 작업 후와 연결 전에 치수 축소를 포함하거나 포함하지 않고 최대 및 평균 풀링 유형, 하나의 전역 기능 또는 4단계 기능으로 풀링하는 것을 포함한 몇 가지 설정을 사용하여 실험을 수행합니다. 표 1에 나열된 대로 평균 풀링은 모든 설정에서 최대 풀링보다 더 잘 작동합니다. 피라미드 구문 분석을 사용한 풀링은 글로벌 풀링을 사용한 풀링보다 성능이 뛰어납니다. 치수 축소를 통해 성능이 더욱 향상됩니다. 제안된 PSPNet을 통해, 최상의 설정은 평균 IoU 및 픽셀 액센트(%) 측면에서 41.68/80.04의 결과를 산출합니다. Liu et al. [24]의 아이디어에서 40.07/79.52의 글로벌 평균 풀링을 1.61/0.52만큼 능가합니다. 그리고 기준과 비교하여 PSPNet은 절대적인 개선 측면에서 4.45/2.03 그리고 상대적인 차이 측면에서 11.95/2.60으로 그것을 능가했습니다.

보조 손실에 대한 절제 연구 도입된 보조 손실은 마스터 지점의 학습에 영향을 미치지 않으면서 학습 프로세스를 최적화하는 데 도움이 됩니다. 우리는 보조 손실 무게α를 0과 1 사이로 설정하는 실험을 하고 그 결과를 표 2에 보여줍니다. 이 기준선은 확장된 네트워크와 함께 ResNet50 기반 FCN을 사용하며, 최적화를 위해 마스터 지점의 소프트맥스 손실이 발생합니다. 보조 손실 분기를 추가하면 α = 0.4가 최상의 성능을 발휘합니다. 평균 IoU 및 픽셀 액센트(%) 측면에서 1.41/0.94의 개선으로 기준치를 능가합니다. 우리는 새로운 증강 보조 손실을 고려할 때 더 깊은 네트워크가 더 많은 이익을 얻을 것이라고 믿습니다.

Loss Weight α	Mean IoU(%)	Pixel Acc.(%)
ResNet50 (without AL)	35.82	77.07
ResNet50 (with $\alpha = 0.3$)	37.01	77.87
ResNet50 (with $\alpha = 0.4$)	**37.23**	**78.01**
ResNet50 (with $\alpha = 0.6$)	37.09	77.84
ResNet50 (with $\alpha = 0.9$)	36.99	77.87

표 2. 보조 분기에서 적절한 손실 중량α를 설정하는 것이 중요합니다. 'AL'은 보조 손실을 나타냅니다. 기준선은 확장된 네트워크를 사용하는 ResNet50 기반 FCN입니다. 경험적으로 α = 0.4가 최상의 성능을 제공합니다. 결과는 단일 척도 입력으로 검증 세트에서 테스트됩니다.

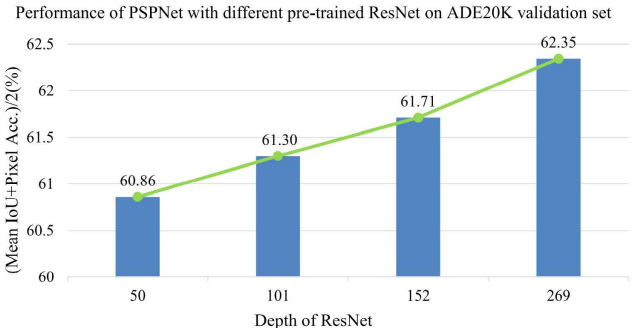


그림 5. 네트워크의 깊이에 따라 성능이 향상됩니다. 결과는 단일 척도 입력으로 검증 세트에서 얻습니다.

Method	Mean IoU(%)	Pixel Acc.(%)
PSPNet(50)	41.68	80.04
PSPNet(101)	41.96	80.64
PSPNet(152)	42.62	80.80
PSPNet(269)	43.81	80.88
PSPNet(50)+MS	42.78	80.76
PSPNet(101)+MS	43.29	81.39
PSPNet(152)+MS	43.51	81.38
PSPNet(269)+MS	44.94	81.69

표 3. 사전 훈련된 심층 모델이 더 높은 성능을 제공합니다. 괄호 안의 숫자는 ResNet의 깊이를 의미하며 'MS'는 다중 스케일 테스트를 의미합니다.

분기, α = 0.4가 최상의 성능을 제공합니다. 평균 IoU 및 픽셀 액센트(%) 측면에서 1.41/0.94의 개선으로 기준치를 능가합니다. 우리는 새로운 증강 보조 손실을 고려할 때 더 깊은 네트워크가 더 많은 이익을 얻을 것이라고 믿습니다.

사전 훈련 모델을 위한 절제 연구 심층 신경망은 이전 연구에서 대규모 데이터 분류에 도움이 되는 것으로 나타났습니다. PSPNet을 추가로 분석하기 위해 사전 훈련된 ResNet의 다양한 깊이에 대한 실험을 수행합니다. 우리는 {50, 101, 152, 269}의 네 가지 깊이를 테스트합니다. 그림 5와 같이, 동일한 설정으로 ResNet의 깊이를 50에서 269로 증가시키면 (평균 IoU + Pixel Acc.) / 2 (%)의 점수를 60.86에서 62.35로 향상시킬 수 있으며, 1.49 절대 개선됩니다. 다양한 깊이 ResNet 모델에서 사전 교육을 받은 PSPNet의 자세한 점수는 표 3에 나와 있습니다.

Method	Mean IoU(%)	Pixel Acc.(%)
FCN [26]	29.39	71.32
SegNet [2]	21.64	71.00
DilatedNet [40]	32.31	73.55
CascadeNet [43]	34.90	74.52
ResNet50-Baseline	34.28	76.35
ResNet50+DA	35.82	77.07
ResNet50+DA+AL	37.23	78.01
ResNet50+DA+AL+PSP	**41.68**	**80.04**
ResNet269+DA+AL+PSP	43.81	80.88
ResNet269+DA+AL+PSP+MS	**44.94**	**81.69**

표 4. 제안된 PSPNet에 대한 자세한 분석과 다른 PSPNet과의 비교입니다. 우리의 결과는 마지막 행을 제외한 단일 척도 입력으로 검증 세트에서 얻습니다. FCN, SegNet 및 DelatedNet의 결과가 보고됩니다. 'DA'는 우리가 수행한 데이터 증강을 의미하고, 'AL'은 우리가 추가한 보조 손실을 의미하며, 'PSP'는 제안된 PSPNet을 의미하며, 'MS'는 다중 스케일 테스트가 사용되는 것을 의미합니다.

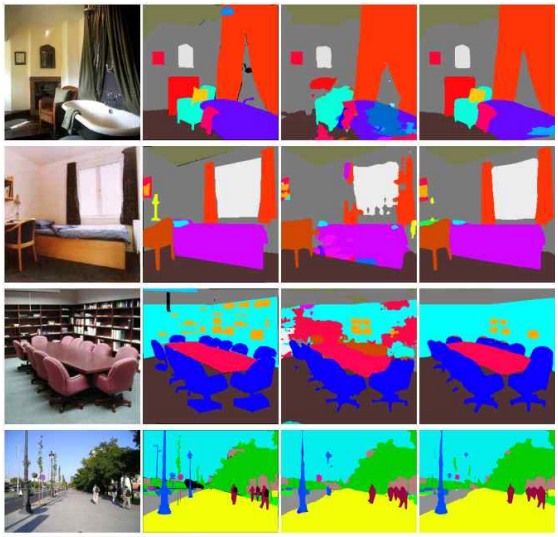
Rank	Team Name	Final Score (%)
1	**Ours**	**57.21**
2	Adelaide	56.74
3	360+MCG-ICT-CAS_SP	55.56
-	(our single model)	(55.38)
4	SegModel	54.65
5	CASIA_IVA	54.33
-	DilatedNet [40]	45.67
-	FCN [26]	44.80
-	SegNet [2]	40.79

표 5. ImageNet 장면 구문 분석 챌린지 2016의 결과입니다. 각 팀의 최고의 엔트리가 나열됩니다. 최종 점수는 평균 IoU 및 픽셀 Acc의 평균입니다. 결과는 테스트 세트에서 평가됩니다.

보다 상세한 성능 분석 우리는 표 4에서 ADE20K의 검증 세트에 대한 보다 자세한 분석을 보여줍니다. 마지막 행을 제외한 모든 결과는 단일 척도 테스트를 사용합니다. "ResNet269+입니다.DA+AL+PSP+MS"는 다중 스케일 테스트를 사용합니다. 우리의 기준선은 확장된 네트워크가 있는 ResNet50에서 조정되어 MeanIoU 34.28 및 픽셀 액세스 76.35를 산출합니다. 강력한 ResNet 덕분에 이미 다른 이전 시스템보다 성능이 뛰어납니다.

제안된 아키텍처는 기준과 비교하여 더욱 개선됩니다. 데이터 증가를 사용하여, 우리의 결과는 기준치를 1.54/0.72만큼 초과하고 35.82/77.07에 도달합니다. 보조 손실은 1.41/0.94만큼 더 개선될 수 있으며 37.23/78.01에 도달합니다. PSPNet을 사용하면 4.45/2.03의 개선 작업이 비교적 많이 진행됩니다. 결과는 41.68/80.04에 도달합니다. 기준 결과와의 차이는 절대 개선 측면에서 7.40/3.69이고 상대성 측면에서 21.59/4.83%입니다. ResNet269의 더 깊은 네트워크는 43.81/80.88까지 더 높은 성능을 제공합니다. 마지막으로, 다중 척도 검사 방식은 점수를 44.94/81.69로 이동합니다.

결과 우리는 PSPNet을 사용하여, 우리 팀은 ImageNet 장면 구문 분석 챌린지 2016에서 1위를 차지했습니다. 표 5는 이 대회에서 몇 가지 결과를 보여줍니다.



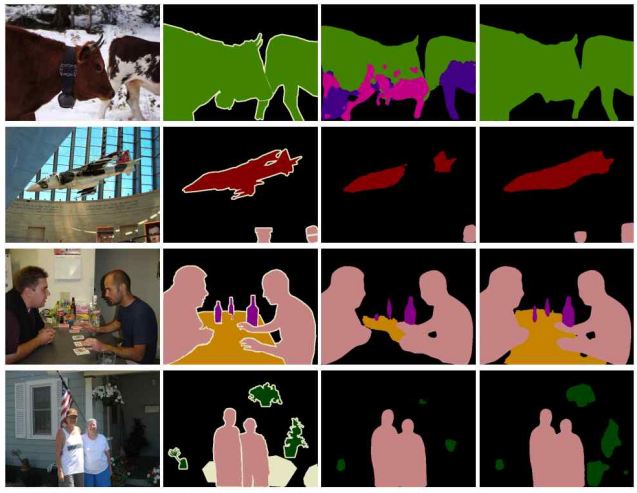
(a) Image (b) Ground Truth (c) Baseline (d) PSPNet

그림 6. ADE20K, PSPNet의 시각적 개선으로 보다 정확하고 상세한 결과를 얻을 수 있습니다.

우리의 앙상블 제출은 테스트 세트에서 57.21%의 점수를 달성합니다. 우리의 단일 모델은 55.38%의 점수를 얻으며, 이는 다른 몇 개의 다중 모델 앙상블 제출보다 더 높습니다. 이 점수는 유효성 검사 집합과 테스트 집합 간의 데이터 분포 차이로 인해 유효성 검사 집합의 점수보다 낮습니다. 그림 2의 열 (d)에 나타난 바와 같이, PSPNet은 FCN의 일반적인 문제를 해결합니다. 그림 6은 ADE20K의 검증 세트에 대한 다른 몇 가지 구문 분석 결과를 보여줍니다. 우리의 결과는 기준선에 비해 더 정확하고 상세한 구조를 포함하고 있습니다.

5.3. PASCAL VOC 2012

우리의 PSPNet은 또한 의미론적 분할에 대해 만족스럽게 작동합니다. 우리는 20개의 객체 범주와 하나의 배경 클래스를 포함하는 PASCAL VOC 2012 세분화 데이터 세트에 대한 실험을 수행합니다. 우리는 10,582, 1,449 및 1,456 이미지의 주석이 있는 증강 데이터를 교육, 검증 및 테스트에 사용합니다. 결과는 표 6에 나와 있으며, 우리는 두 가지 설정, 즉 MS-COCO 데이터 세트에 대한 사전 교육 유무에 기초하여 PSPNet을 테스트 세트의 이전 최고 성능 방법과 비교합니다. MS-COCO로 사전 교육을 받은 방법은 '†'로 표시됩니다. 장면 구문 분석/시맨틱 분할 작업에서 현재 ResNet 기반 프레임워크와 공정한 비교를 위해 CRF와 같은 후처리 없이 ResNet101을 기반으로 아키텍처를 구축합니다. 우리는 여러 척도 입력으로 PSPNet을 평가하고 다음과 같은 평균 결과를 사용합니다.



(a) Image (b) Ground Truth (c) Baseline (d) PSPNet

그림 7. PASCAL VOC 2012 데이터에 대한 시각적 개선 사항입니다. PSPNet은 보다 정확하고 자세한 결과를 생성합니다.

표 6에 나와 있는 것처럼 PSPNet은 두 설정 모두에서 이전 방법을 능가합니다. VOC 2012 데이터만으로 교육을 받은 결과 82.6%의 정확도를 달성했습니다2 - 20개 클래스 모두에서 가장 높은 정확도를 얻었습니다. PSPNet을 MS-COCO 데이터 세트로 사전 교육하면 85.4%의 정확도에 도달하며, 20개 클래스 중 19개 클래스가 가장 높은 정확도를 받습니다. 흥미롭게도, VOC 2012 데이터만으로 훈련된 PSPNet은 MS-COCO 사전 훈련된 모델로 훈련된 기존 방법을 능가합니다.

ResNet이 최근 제안된 이후 우리의 기반 분류 모델이 몇 가지 이전 방법보다 더 강력하다고 주장할 수 있습니다. 우리의 고유한 기여를 보여주기 위해, 우리는 우리의 방법이 FCRN, LRR 및 DeepLab을 포함하여 동일한 모델을 사용하는 state-of-the-art 프레임워크도 능가한다는 것을 보여줍니다. 이 과정에서는 시간이 많이 걸리지만 CRF와 같은 효과적인 후처리를 [4, 9]에서와 같이 사용하지 않습니다.

그림 7에 몇 가지 예가 나와 있습니다. 1행의 "소"의 경우, 기본 모델은 "말"과 "개"로 취급하고 PSPNet은 이러한 오류를 수정합니다. 두 번째와 세 번째 행에 있는 "항공기" 및 "표"의 경우 PSPNet은 누락된 부품을 찾습니다. 다음 행의 "사람", "병" 및 "식물"의 경우, PSP-Net은 기본 모델과 비교하여 이미지의 이러한 작은 크기의 개체 클래스에서 우수한 성능을 발휘합니다. PSPNet과 다른 방법 간의 시각적 비교는 그림 9에 포함되어 있습니다.

5.4. Cityscapes

Cityscapes는 의미론적 도시 장면 이해를 위해 최근에 출시된 데이터 세트입니다. 여기에는 계절에 따라 50개 도시에서 수집한 5,000개의 고품질 픽셀 수준의 정교한 주석이 달린 이미지가 포함되어 있습니다.

Method	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mIoU
FCN [26]	76.8	34.2	68.9	49.4	60.3	75.3	74.7	77.6	21.4	62.5	46.8	71.8	63.9	76.5	73.9	45.2	72.4	37.4	70.9	55.1	62.2
Zoom-out [28]	85.6	37.3	83.2	62.5	66.0	85.1	80.7	84.9	27.2	73.2	57.5	78.1	79.2	81.1	77.1	53.6	74.0	49.2	71.7	63.3	69.6
DeepLab [3]	84.4	54.5	81.5	63.6	65.9	85.1	79.1	83.4	30.7	74.1	59.8	79.0	76.1	83.2	80.8	59.7	82.2	50.4	73.1	63.7	71.6
CRF-RNN [41]	87.5	39.0	79.7	64.2	68.3	87.6	80.8	84.4	30.4	78.2	60.4	80.5	77.8	83.1	80.6	59.5	82.8	47.8	78.3	67.1	72.0
DeconvNet [30]	89.9	39.3	79.7	63.9	68.2	87.4	81.2	86.1	28.5	77.0	62.0	79.0	80.3	83.6	80.2	58.8	83.4	54.3	80.7	65.0	72.5
GCRF [36]	85.2	43.9	83.3	65.2	68.3	89.0	82.7	85.3	31.1	79.5	63.3	80.5	79.3	85.5	81.0	60.5	85.5	52.0	77.3	65.1	73.2
DPN [25]	87.7	59.4	78.4	64.9	70.3	89.3	83.5	86.1	31.7	79.9	62.6	81.9	80.0	83.5	82.3	60.5	83.2	53.4	77.9	65.0	74.1
Piecewise [20]	90.6	37.6	80.0	67.8	74.4	92.0	85.2	86.2	39.1	81.2	58.9	83.8	83.9	84.3	84.8	62.1	83.2	58.2	80.8	72.3	75.3
PSPNet	**91.8**	**71.9**	**94.7**	**71.2**	**75.8**	**95.2**	**89.9**	**95.9**	**39.3**	**90.7**	**71.7**	**90.5**	**94.5**	**88.8**	**89.6**	**72.8**	**89.6**	**64.0**	**85.1**	**76.3**	**82.6**
CRF-RNN† [41]	90.4	55.3	88.7	68.4	69.8	88.3	82.4	85.1	32.6	78.5	64.4	79.6	81.9	86.4	81.8	58.6	82.4	53.5	77.4	70.1	74.7
BoxSup† [7]	89.8	38.0	89.2	68.9	68.0	89.6	83.0	87.7	34.4	83.6	67.1	81.5	83.7	85.2	83.5	58.6	84.9	55.8	81.2	70.7	75.2
Dilation8† [40]	91.7	39.6	87.8	63.1	71.8	89.7	82.9	89.8	37.2	84.0	63.0	83.3	89.0	83.8	85.1	56.8	87.6	56.0	80.2	64.7	75.3
DPN† [25]	89.0	61.6	87.7	66.8	74.7	91.2	84.3	87.6	36.5	86.3	66.1	84.4	87.8	85.6	85.4	63.6	87.3	61.3	79.4	66.4	77.5
Piecewise† [20]	94.1	40.7	84.1	67.8	75.9	93.4	84.3	88.4	42.5	86.4	64.7	85.4	89.0	85.8	86.0	67.5	90.2	63.8	80.9	73.0	78.0
FCRNs† [38]	91.9	48.1	93.4	69.3	75.5	94.2	87.5	92.8	36.7	86.9	65.2	89.1	90.2	86.5	87.2	64.6	90.1	59.7	85.5	72.7	79.1
LRR† [9]	92.4	45.1	94.6	65.2	75.8	**95.1**	89.1	92.3	39.0	85.7	70.4	88.6	89.4	88.6	86.6	65.8	86.2	57.4	85.7	77.3	79.3
DeepLab† [4]	92.6	60.4	91.6	63.4	76.3	95.0	88.4	92.6	32.7	88.5	67.6	89.6	92.1	87.0	87.4	63.3	88.3	60.0	86.8	74.5	79.7
PSPNet†	**95.8**	**72.7**	**95.0**	**78.9**	**84.4**	94.7	**92.0**	**95.7**	**43.1**	**91.0**	**80.3**	**91.3**	**96.3**	**92.3**	**90.1**	**71.5**	**94.4**	**66.9**	**88.8**	**82.0**	**85.4**

표 6. PASCAL VOC 2012 테스트 세트의 클래스별 결과입니다. MS-COCO에서 사전 교육을 받은 방법은 다음과 같이 표시됩니다. '†'

Method	IoU cla.	iIoU cla.	IoU cat.	iIoU cat.
CRF-RNN [41]	62.5	34.4	82.7	66.0
FCN [26]	65.3	41.7	85.7	70.1
SiCNN [16]	66.3	44.9	85.0	71.2
DPN [25]	66.8	39.1	86.0	69.1
Dilation10 [40]	67.1	42.0	86.5	71.1
LRR [9]	69.7	48.0	88.2	74.7
DeepLab [4]	70.4	42.6	86.4	67.7
Piecewise [20]	71.6	51.7	87.3	74.1
PSPNet	**78.4**	**56.7**	**90.6**	**78.6**
LRR‡ [9]	71.8	47.9	88.4	73.9
PSPNet‡	**80.2**	**58.1**	**90.6**	**78.2**

표 7. 도시경관 테스트 세트 결과입니다. 미세 데이터와 거친 데이터를 모두 사용하여 훈련된 방법은 다음과 같이 표시됩니다.'‡'.

이미지는 교육, 유효성 검사 및 테스트를 위해 2,975, 500 및 1,525번 세트로 나뉩니다. 그것은 물건과 물건을 모두 포함하는 19개의 범주를 정의합니다. 또한 두 가지 설정, 즉 미세 데이터만 사용하거나 미세 데이터와 거친 데이터를 모두 사용하여 훈련하기 위해 대략적으로 주석이 달린 20,000개의 이미지가 제공됩니다. 미세 데이터와 거친 데이터를 모두 사용하여 학습한 방법은 '‡'로 표시됩니다. 자세한 결과는 표 7에 나와 있습니다. 우리의 기본 모델은 공정한 비교를 위해 DeepLab에서와 같이 ResNet101이며 테스트 절차는 섹션 5.3을 따릅니다.

표 7의 통계에 따르면 PSPNet은 다른 방법보다 성능이 뛰어나며 주목할 만한 이점을 가지고 있습니다. 교육에 미세 데이터와 거친 데이터를 모두 사용하면 80.2의 정확도를 얻을 수 있습니다. 그림 8에 몇 가지 예가 나와 있습니다. 테스트 세트에 대한 자세한 클래스별 결과는 표 8에 나와 있습니다.

6. 논문을 마치며

복잡한 장면 이해를 위한 효과적인 피라미드 장면 구문 분석 네트워크를 제안했습니다. 전역 피라미드 풀링 기능은 추가 상황별 정보를 제공합니다.

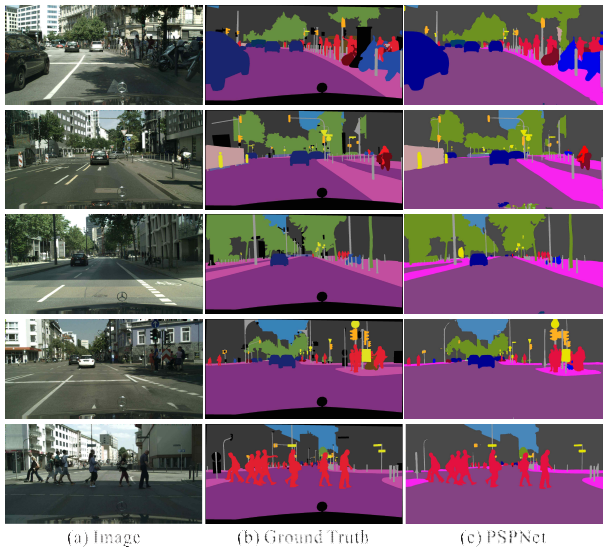


그림 8. Cityscapes 데이터 세트에 대한 PSPNet 결과의 예.

또한 ResNet 기반 FCN 네트워크에 대해 심층적으로 감독되는 최적화 전략을 제공했습니다. 우리는 공개적으로 사용할 수 있는 구현 세부 정보가 커뮤니티가 장면 구문 분석 및 의미 분할을 위한 유용한 전략을 채택하고 관련 기술을 발전시키는 데 도움이 되기를 바랍니다.

도움을 주신 분들

기본 분류 모델인 Qun Luo 기술 지원을 위한 교육에 도움을 주신 Gang Sun 및 Tong Xiao에 감사드립니다. 이 작업은 홍콩 특별행정구 연구 보조금 위원회(프로젝트 번호 2150760)의 보조금으로 지원됩니다.

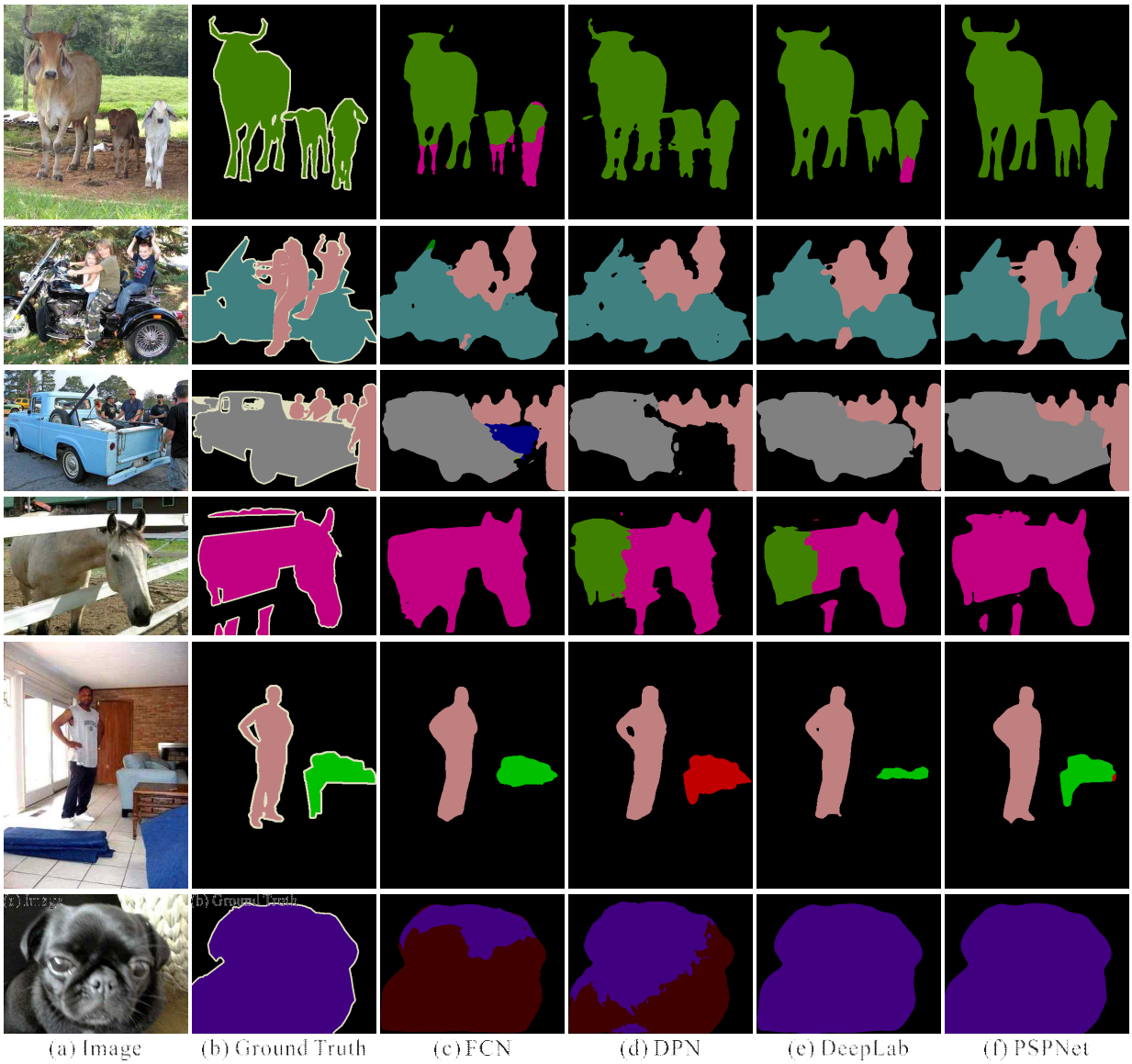


그림 9. PASCAL VOC 2012 데이터에 대한 시각적 비교. (a) 이미지. (b) 실측 자료. (c) FCN. (d) DPN. (e) DeepLab. (f) PSPNet.

Method	road	swalk	build.	wall	fence	pole	tlight	sign	veg.	terrain	sky	person	rider	car	truck	bus	train	mbike	bike	mIoU
CRF-RNN [41]	96.3	73.9	88.2	47.6	41.3	35.2	49.5	59.7	90.6	66.1	93.5	70.4	34.7	90.1	39.2	57.5	55.4	43.9	54.6	62.5
FCN [26]	97.4	78.4	89.2	34.9	44.2	47.4	60.1	65.0	91.4	69.3	93.9	77.1	51.4	92.6	35.3	48.6	46.5	51.6	66.8	65.3
SiCNN+CRF [16]	96.3	76.8	88.8	40.0	45.4	50.1	63.3	69.6	90.6	67.1	92.2	77.6	55.9	90.1	39.2	51.3	44.4	54.4	66.1	66.3
DPN [25]	97.5	78.5	89.5	40.4	45.9	51.1	56.8	65.3	91.5	69.4	94.5	77.5	54.2	92.5	44.5	53.4	49.9	52.1	64.8	66.8
Dilation10 [40]	97.6	79.2	89.9	37.3	47.6	53.2	58.6	65.2	91.8	69.4	93.7	78.9	55.0	93.3	45.5	53.4	47.7	52.2	66.0	67.1
LRR [9]	97.7	79.9	90.7	44.4	48.6	58.6	68.2	72.0	92.5	69.3	94.7	81.6	60.0	94.0	43.6	56.8	47.2	54.8	69.7	69.7
DeepLab [4]	97.9	81.3	90.3	48.8	47.4	49.6	57.9	67.3	91.9	69.4	94.2	79.8	59.8	93.7	56.5	67.5	57.5	57.7	68.8	70.4
Piecewise [20]	98.0	82.6	90.6	44.0	50.7	51.1	65.0	71.7	92.0	72.0	94.1	81.5	61.1	94.3	61.1	65.1	53.8	61.6	70.6	71.6
PSPNet	98.6	86.2	92.9	50.8	58.8	64.0	75.6	79.0	93.4	72.3	95.4	86.5	71.3	95.9	68.2	79.5	73.8	69.5	77.2	78.4
LRR‡ [9]	97.9	81.5	91.4	50.5	52.7	59.4	66.8	72.7	92.5	70.1	95.0	81.3	60.1	94.3	51.2	67.7	54.6	55.6	69.6	71.8
PSPNet‡	98.6	86.6	93.2	58.1	63.0	64.5	75.2	79.2	93.4	72.1	95.1	86.3	71.4	96.0	73.5	90.4	80.3	69.9	76.9	80.2

표 8. 도시경관 테스트 세트에 대한 클래스별 결과입니다. 미세 집합과 거친 집합을 모두 사용하여 훈련된 방법은 '‡'로 표시됩니다.

참조

[1] A. Arnab, S. Jayasumana, S. Zheng, and P. H. S. Torr. Higher order conditional random fields in deep neural networks. In *ECCV*, 2016. 2

[2] V. Badrinarayanan, A. Kendall, and R. Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *arXiv:1511.00561*, 2015. 6

[3] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv:1412.7062*, 2014. 1, 2, 4, 7, 8

[4] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *arXiv:1606.00915*, 2016. 5, 7, 8, 9

[5] L. Chen, Y. Yang, J. Wang, W. Xu, and A. L. Yuille. Atten- tion to scale: Scale-aware semantic image segmentation. In *CVPR*, 2016. 2

[6] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler,
R. Benenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016. 2, 5, 7

[7] J. Dai, K. He, and J. Sun. Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic seg- mentation. In *ICCV*, 2015. 7, 8

[8] M. Everingham, L. J. V. Gool, C. K. I. Williams, J. M. Winn, and A. Zisserman. The pascal visual object classes VOC challenge. *IJCV*, 2010. 1, 2, 5, 7

[9] G. Ghiasi and C. C. Fowlkes. Laplacian pyramid reconstruc- tion and refinement for semantic segmentation. In *ECCV*, 2016. 7, 8, 9

[10] B. Hariharan, P. Arbelaez, L. D. Bourdev, S. Maji, and J. Malik. Semantic contours from inverse detectors. In *ICCV*, 2011. 7

[11] B. Hariharan, P. A. Arbela´ez, R. B. Girshick, and J. Malik. Hypercolumns for object segmentation and fine-grained localization. In *CVPR*, 2015. 2

[12] K. He, X. Zhang, S. Ren, and J. Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. In *ECCV*, 2014. 1, 3

[13] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 2, 3, 4, 5, 6

[14] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, 2015. 5

[15] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. B. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *ACM MM*, 2014. 5

[16] I. Kreso, D. Causevic, J. Krapac, and S. Segvic. Convolutional scale invariance for semantic segmentation. In *GCPR*, 2016. 8, 9

[17] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012. 2, 4

[18]

[19] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006. 1, 2, 3

[20] C. Lee, S. Xie, P. W. Gallagher, Z. Zhang, and Z. Tu. Deeply-supervised nets. In *AISTATS*, 2015. 4

[21] G. Lin, C. Shen, I. D. Reid, and A. van den Hengel. Efficient piecewise training of deep structured models for semantic segmentation. In *CVPR*, 2016. 8, 9

[22] T. Lin, M. Maire, S. J. Belongie, J. Hays, P. Perona, D. Ra- manan, P. Dolla´r, and C. L. Zitnick. Microsoft coco: Com- mon objects in context. In *ECCV*, 2014. 7

[23] C. Liu, J. Yuen, and A. Torralba. Nonparametric scene pars- ing: Label transfer via dense scene alignment. In *CVPR*, 2009. 1

[24] C. Liu, J. Yuen, and A. Torralba. Nonparametric scene pars- ing via label transfer. *TPAMI*, 2011. 1

[25] W. Liu, A. Rabinovich, and A. C. Berg. Parsenet: Looking wider to see better. *arXiv:1506.04579*, 2015. 2, 3, 4, 5, 7, 9

[26] Z. Liu, X. Li, P. Luo, C. C. Loy, and X. Tang. Semantic im- age segmentation via deep parsing network. In *ICCV*, 2015. 2, 8, 9

[27] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015. 1, 2, 6, 7, 8, 9

[28] A. Lucchi, Y. Li, X. B. Bosch, K. Smith, and P. Fua. Are spatial and global constraints really necessary for segmentation? In *ICCV*, 2011. 2

[29] M. Mostajabi, P. Yadollahpour, and G. Shakhnarovich. Feedforward semantic segmentation with zoom-out features. In *CVPR*, 2015. 8

[30] R. Mottaghi, X. Chen, X. Liu, N. Cho, S. Lee, S. Fidler, R. Urtasun, and A. L. Yuille. The role of context for object detection and semantic segmentation in the wild. In *CVPR*, 2014. 1

[31] H. Noh, S. Hong, and B. Han. Learning deconvolution net- work for semantic segmentation. In *ICCV*, 2015. 2, 8

[32] G. Papandreou, L. Chen, K. P. Murphy, and A. L. Yuille. Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation. In *ICCV*, 2015. 7

[33] L. Shen, Z. Lin, and Q. Huang. Relay backpropagation for effective learning of deep convolutional neural networks. In *ECCV*, 2016. 4, 5

[34] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv:1409.1556*, 2014. 2, 4

[35] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. E. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *CVPR*, 2015. 2, 3

[36] C. Szegedy, S. E. Reed, D. Erhan, and D. Anguelov. Scalable, high-quality object detection. *arXiv:1412.1441*, 2014. 2

[37] R. Vemulapalli, O. Tuzel, M. Liu, and R. Chellappa. Gaus- sian conditional random field network for semantic segmen- tation. In *CVPR*, 2016. 8

[38] L. Wang, Y. Xiong, Z. Wang, and Y. Qiao. Towards good practices for very deep two-stream convnets. *arXiv:1507.02159*, 2015. 5

[39] Z. Wu, C. Shen, and A. van den Hengel. Bridging category-level and instance-level semantic image segmenta- tion. *arXiv:1605.06885*, 2016. 7, 8

[40] F. Xia, P. Wang, L. Chen, and A. L. Yuille. Zoom better to see clearer: Human and object parsing with hierarchical auto-zoom net. In *ECCV*, 2016. 2

[41] F. Yu and V. Koltun. Multi-scale context aggregation by di- lated convolutions. *arXiv:1511.07122*, 2015. 1, 2, 4, 6, 8, 9

[42] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. H. S. Torr. Conditional ran- dom fields as recurrent neural networks. In *ICCV*, 2015. 2, 8, 9

[43] B. Zhou, A. Khosla, À. Lapedriza, A. Oliva, and A. Tor- ralba. Object detectors emerge in deep scene cnns. *arXiv:1412.6856*, 2014. 3

[44] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba. Semantic understanding of scenes through the ADE20K dataset. *arXiv:1608.05442*, 2016. 1, 2, 3, 5, 6