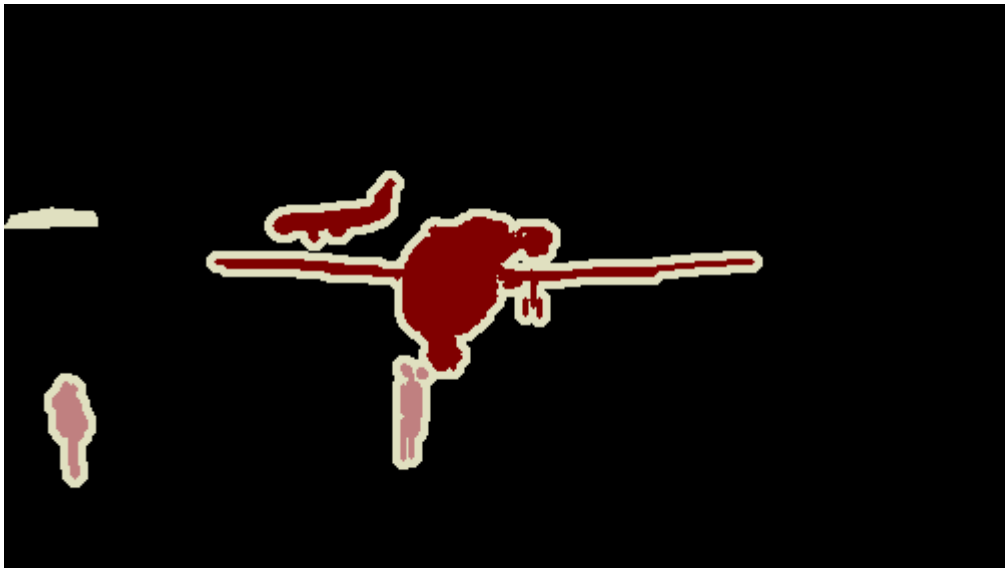
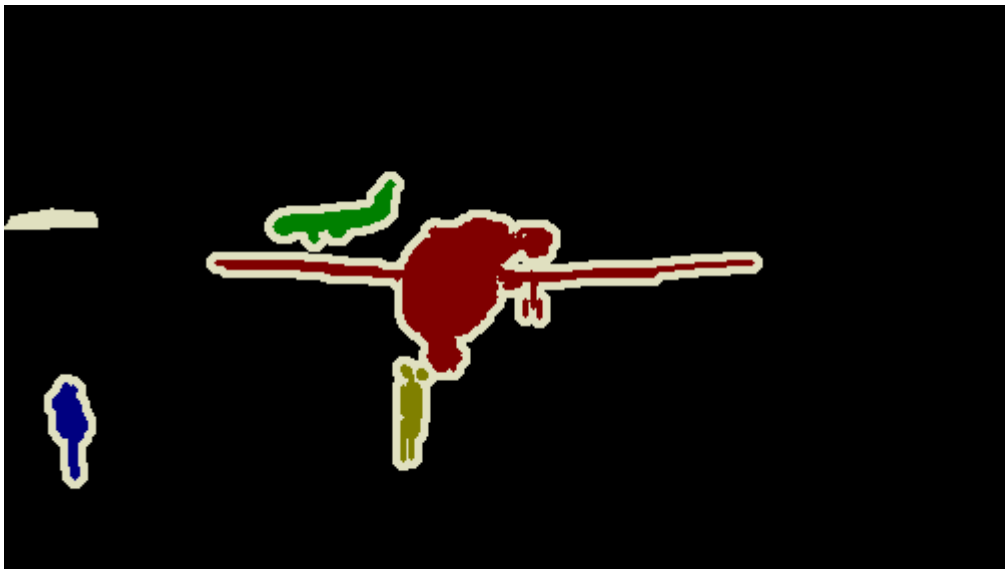


SegmentationClass



SegmentationObject

3.2. Segmentation from Extreme Points

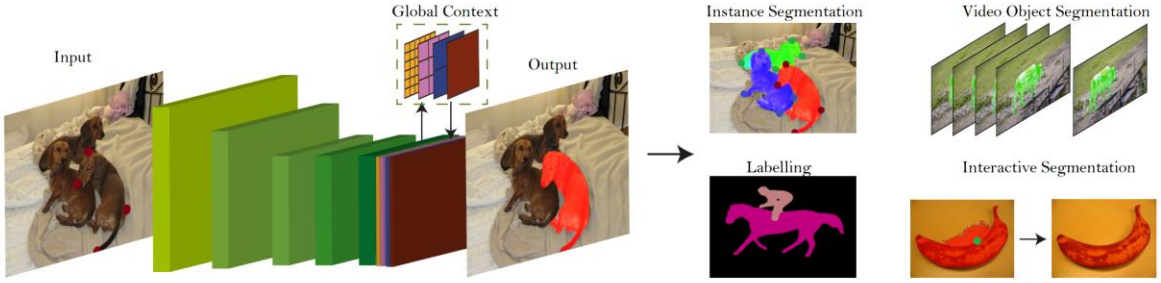


Figure 2. **Architecture of DEXTR**: Both the RGB image and the labeled extreme points are processed by the CNN to produce the segmented mask. The applicability of this method is illustrated for various tasks: Instance, Semantic, Video, and Interactive segmentation.

Annotated extreme points는 네트워크 입력에 대한 가이드 신호로 주어집니다. 이 때문에, 이를 위해, extreme points 영역에서 활성화되는 히트맵을 만듭니다. 단일 히트맵을 만들기 위해, 각 포인트의 중심에 2D Gaussian을 배치합니다. 히트맵은 입력 이미지의 RGB 채널과 연계되어 CNN 4 채널 입력을 형성합니다. Object of interest에 포커스를 맞추기 위해, 입력은 extreme point annotations로부터 형성되는 bounding box에 의해 crop됩니다. Crop된 결과에 context를 포함시키기 위해, 타이트한 bounding box를 몇 픽셀씩 relax해줍니다. Extreme clicks에서만 독점적으로 보이는 전처리 단계 이후, 입력은 object를 포함한 RGB crop과 그 extreme points로 구성됩니다.

본 연구에서는 backbone architecture로 ResNet-101 [13]을 선택했습니다. (다양한 segmentation 방법론 [6, 12]에서 성공적인 것으로 입증됨) dense prediction을 위해, 적용 가능한 출력 해상도를 보존하기 위해서, 마지막 두 단계의 max pooling layer 뿐만 아니라 FC layer 또한 제거했습니다. 그리고 같은 receptive field를 유지하기 위해 마지막 두 단계에서 atrous convolutions를 도입했습니다. 마지막 ResNet-101 단계 이후, 최종 feature map에 global context를 종합하기 위해 pyramid scene parsing module [43]을 도입했습니다. ImageNet에서 pre-training 한 것으로 네트워크 weight를 초기화한 것은 다양한 작업에 이득인 것으로 입증되었습니다. [22, 40, 12] 대부분의 실험에서, 본 연구는 ImageNet에서 pre-train된 Deeplab-v2 모델을 제공받아 사용하고 semantic segmentation을 위해 PASCAL에서 fine-tune합니다.

CNN의 출력은 픽셀이 우리가 segmentation하고자 하는 object에 속하는지 아닌지를 나타내는 probability map 입니다. CNN은 dataset에서 서로 다른 빈도로 서로 다른 클래스가 발생한다는 점을 고려해서, standard cross entropy loss를 최소화하도록 학습되었습니다.

$$\mathcal{L} = \sum_{j \in Y} w_{y_j} C\left(y_j, \hat{y}_j\right), \qquad j \in 1, \ldots, |Y|$$

w_{y_j}는 픽셀 j의 라벨 y_j에 의존합니다. $y_j \in \{0, 1\}$ 인 w_{y_j} 를 minibatch 내 라벨의 역 정규화 주파수로 정의합니다. $C(.)$는 라벨과 prediction y_j 사이의 standard cross-entropy loss를 나타냅니다. Balanced loss는 샘플의 대다수가 background class에 속하는 경계 검출 [40, 23] 에서 매우 잘 수행되는 것으로 입증되었습니다. 본 연구는, 이 방법이 네트워크로의 guiding signal로서 extreme points를 이용하여, 강력한 mask-level supervision으로부터, 공개적으로 이용 가능한 dataset에 대해 학습된다는 점에 주목했습니다.

Object를 segmentation하기 위해, 우리의 방법은 object centered crop을 사용하므로, 따라서 background보다 foreground에 속하는 샘플의 수가 훨씬 많고 balanced loss의 사용이 유익하다는 것을 증명합니다.

최종 모델에 사용된 각 컴포넌트에 대한 대안은 ablation analysis에서 연구되었으며, 자세한 비교는 Section 4.2에서 확인할 수 있습니다.