Deep Extreme Cut: From Extreme Points to Object Segmentation

K.-K. Maninis* S. Caelles* J. Pont-Tuset L. Van Gool

Computer Vision Lab, ETH Zurich, Switzerland

10/31/2018
Presented by: Xi Fang

Extreme Points

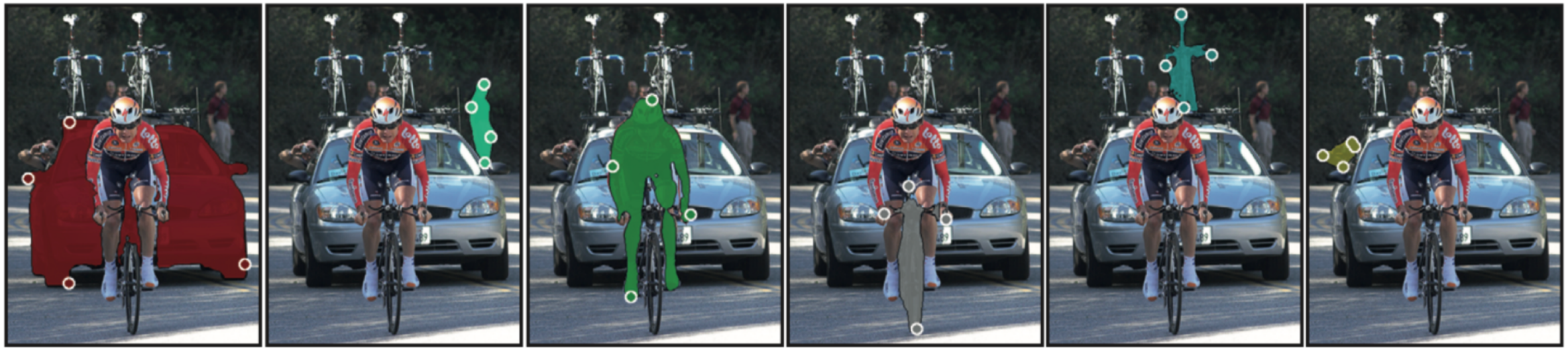


Figure 1. **Example results of DEXTR**: The user provides the extreme clicks for an object, and the CNN produces the segmented masks.

Content

- Motivation
- Contribution
- Related Work
- Method
- Cases and Datasets
- Experiment
- Performance

Motivation

- Supervised techniques are performed best in many benchmarks and challenges
- Weakly-supervised techniques are significantly behind of the state of the art
- Self-automatic techniques can circumventing the expensive annotations

Contribution

- Tackles all these scenarios in a unified way and show state of the art results
- Presents Deep Extreme Cut(DEXTR), that obtains an object segmentation from four extreme points: the left-most, right-most, top and bottom pixels
- Performs well for Self-automatic object segmentation, video object segmentation, annotation

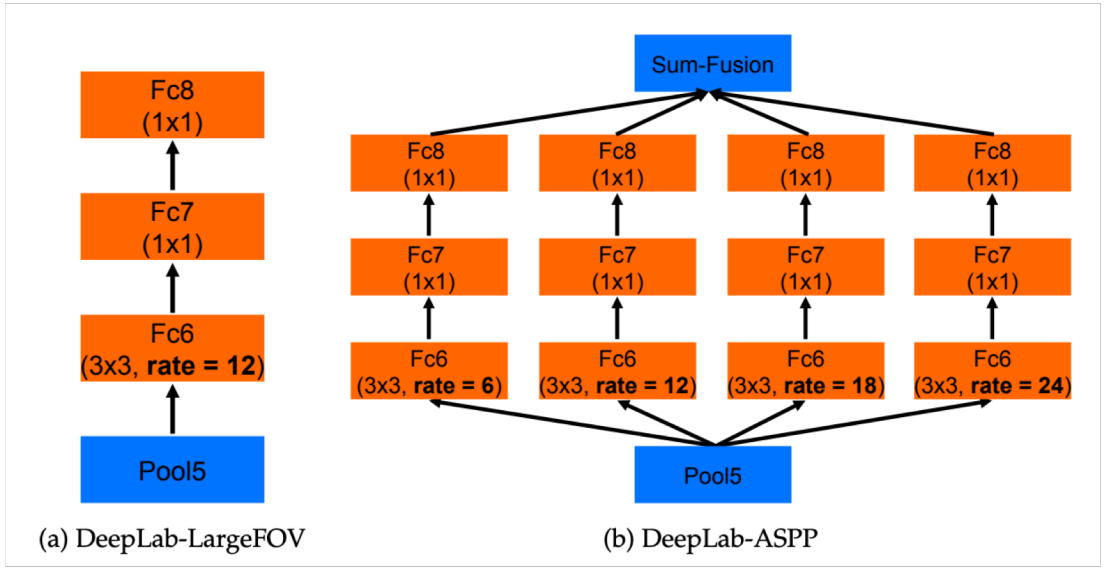
Related Work

- Weakly Supervised Signals for segmentation
  - Point-level supervision
  - Box-regression
  - Extreme clicks as additional information
- Instance Segmentation
  - Automatically segment object proposals
  - Guiding signals in form of a bounding box
- Interactive Segmentation from points
  - Grabcut: gradually updates appearance model
  - iFCN: updates results by user-defined clicks
  - RIS-Net: adding local context

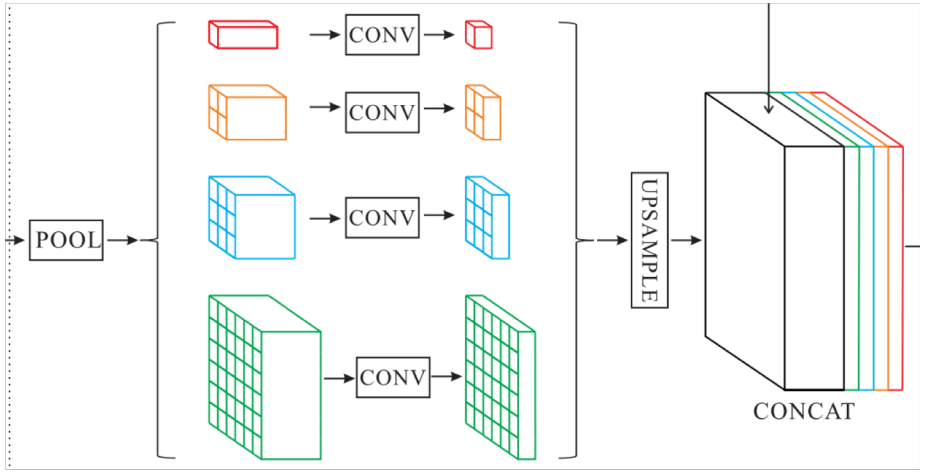
Method

- Network

layer name	output size	18-layer	34-layer	50-layer	101-layer	152-layer
conv1	112×112	7×7, 64, stride 2				
conv2_x	56×56	3×3 max pool, stride 2				
		$\begin{bmatrix} 3\times3, 64 \\ 3\times3, 64 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 64 \\ 3\times3, 64 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}\times3$
conv3_x	28×28	$\begin{bmatrix} 3\times3, 128 \\ 3\times3, 128 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 128 \\ 3\times3, 128 \end{bmatrix}\times4$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix}\times4$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix}\times4$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix}\times8$
conv4_x	14×14	$\begin{bmatrix} 3\times3, 256 \\ 3\times3, 256 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 256 \\ 3\times3, 256 \end{bmatrix}\times6$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}\times6$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}\times23$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}\times36$
conv5_x	7×7	$\begin{bmatrix} 3\times3, 512 \\ 3\times3, 512 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 512 \\ 3\times3, 512 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix}\times3$
	1×1	average pool, 1000-d fc, softmax				
FLOPs		1.8×10^9	3.6×10^9	3.8×10^9	7.6×10^9	11.3×10^9



(a) DeepLab-LargeFOV (b) DeepLab-ASPP



Method

- Procedure

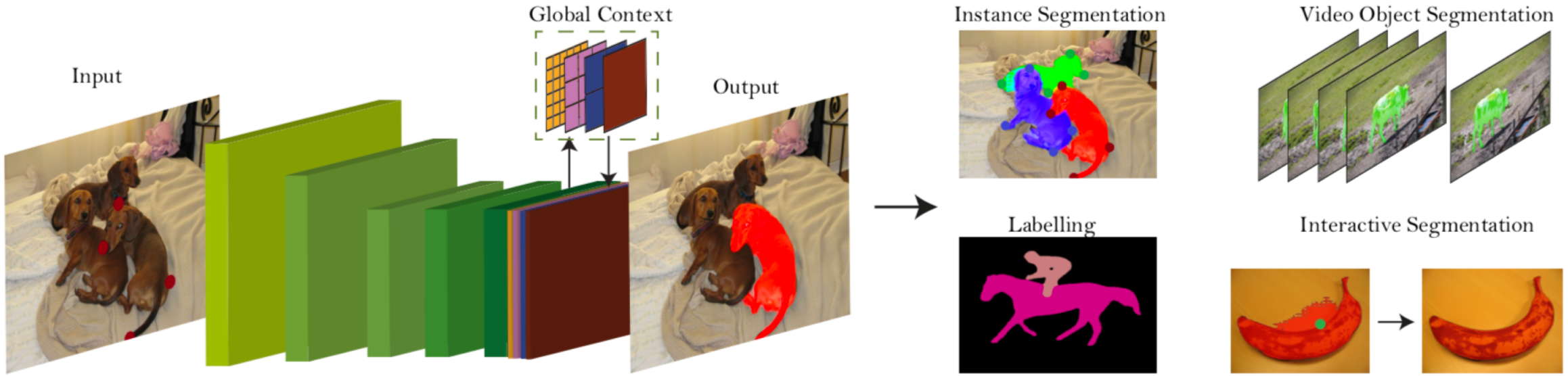


Figure 2. **Architecture of DEXTR**: Both the RGB image and the labeled extreme points are processed by the CNN to produce the segmented mask. The applicability of this method is illustrated for various tasks: Instance, Semantic, Video, and Interactive segmentation.

Method

- Extreme points
  - Bounding box Drawing
    - Click points outside the object, drag the box diagonally, which takes a long labeling time
    - Extreme clicking leads to high quality bounding boxes that are on par with the ones obtained from traditional methods
  - Provides more information

Method

- Segmentation from Extreme Points
  - Guiding signal to input network
  - Input: image cropped by bounding box
  - Architecture:
    - ResNet101 as backbone
    - Atrous convolutions in the last two stages
    - Pyramid scene parsing modules
  - Loss: Weighted cross entropy

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

  - Trained from strong mask-level supervision

Cases and Datasets

- Class-agnostic Instance Segmentation
  - Data PASCAL and Grab-cut
- Annotation
- Video Object Segmentation
  - DAVIS-2016 and DAVIS-2017
- Interactive Object Segmentation

Experiments

- Ablation Study
  - Deeplab-v2 vs. ResNet-101-c4 variant
  - Bounding boxes vs. extreme points
  - Standard cross entropy vs. class-balanced entropy
  - Full image vs. crops
  - Atrous spatial pyramid(ASPP) vs pyramid scene parsing(PSP) module
  - Manual vs. simulated points
  - Distance-map vs. fixed points

Experiments

- Summary

Variant	IoU (%)	Gain
Full Image (Deeplab-v2 + PSP + Extreme Points)	82.6	
Crop on Object (Deeplab-v2)	85.1	+2.5%
+ PSP	87.4	+2.3%
+ Extreme Points	90.5	+3.1%
+ SBD data (Ours)	**91.5**	+1.0%

Performance

- Class-agnostic Instance Segmentation
  - Comparison to the state of the art in PASCAL

Method	IoU
Sharpmask [30] from bounding box	69.3%
Sharpmask [30] upper bound	78.0%
[25] from extreme points	73.6%
Ours from extreme points	**80.1%**

Table 4. **Comparison on PASCAL$_{EXT}$**: IoU of our results against class-agnostic instance segmentation methods, on the objects annotated by [25] to be able to compare to them.

Performance

- Class-agnostic Instance Segmentation
  - Comparison to the state of the art in Grabcut dataset

Method	Error Rate (%)
GrabCut [33]	8.1
KernelCut [37]	7.1
OneCut [38]	6.7
[25] from extreme points	5.5
BoxPrior [18]	3.7
MILCut [39]	3.6
DeepGC [41]	3.4
Ours from extreme points	**2.3**

Table 5. **Comparison on the Grabcut dataset**: Error rates compared to the state-of-the-art techniques.

Performance

- Class-agnostic Instance Segmentation
  - Generalization to unseen categories and across datasets

	Train	Test	IoU
Unseen categories	PASCAL	COCO MVal w/o PASCAL classes	80.3%
	PASCAL	COCO MVal only PASCAL classes	79.9%
Dataset generalization	PASCAL	COCO MVal	80.1%
	COCO	COCO MVal	82.1%
	COCO	PASCAL	87.8%
	PASCAL	PASCAL	89.8%

Table 6. **Generalization to unseen classes and across datasets**: Intersection over union results of training in one setup and testing on another one. MVal stands for mini-val.

Performance

- Class-agnostic Instance Segmentation
  - Generalization to unseen categories and across datasets

	Train	Test	IoU
Unseen categories	PASCAL	COCO MVal w/o PASCAL classes	80.3%
	PASCAL	COCO MVal only PASCAL classes	79.9%
Dataset generalization	PASCAL	COCO MVal	80.1%
	COCO	COCO MVal	82.1%
	COCO	PASCAL	87.8%
	PASCAL	PASCAL	89.8%

Table 6. **Generalization to unseen classes and across datasets**: Intersection over union results of training in one setup and testing on another one. MVal stands for mini-val.

  - Generalization to background (stuff) categories

Performance

- Annotation
  - DEXTR's masks and those from ground truth

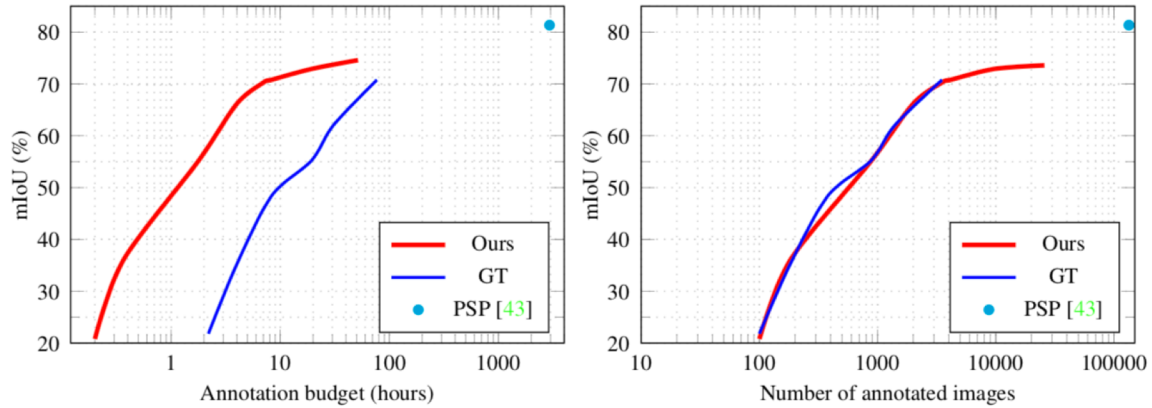


Figure 4. **Quality vs. annotation budget**: mIoU for semantic segmentation on PASCAL val set trained on our masks or the input, as a function of annotation budget (left) and the number of annotated images (right).

Performance

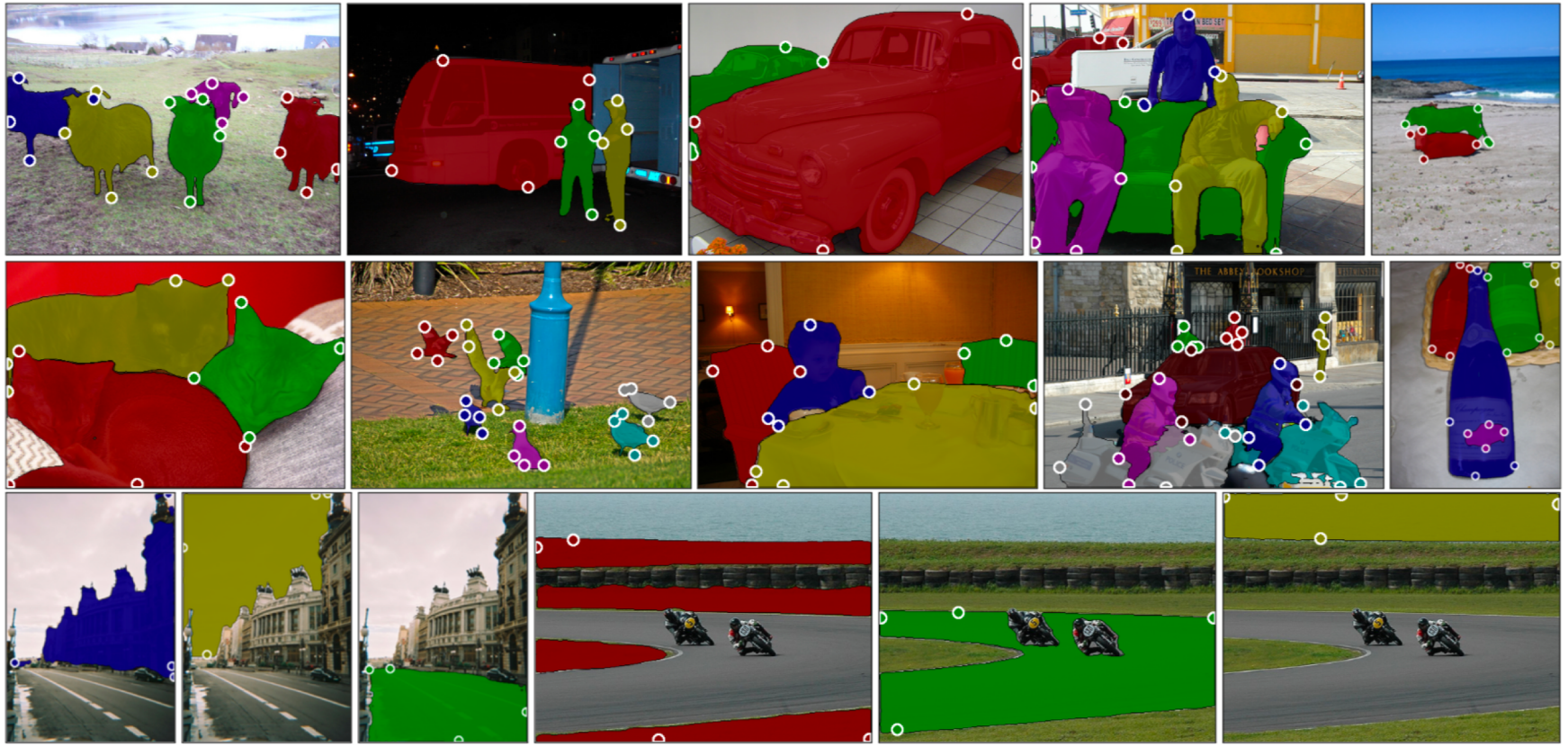


Figure 3. **Qualitative results by DEXTR on PASCAL:** Each instance with the simulated extreme points used as input and the resulting mask overlayed. The bottom row shows results on PASCAL Context stuff categories.

Performance

- Video Object Segmentation

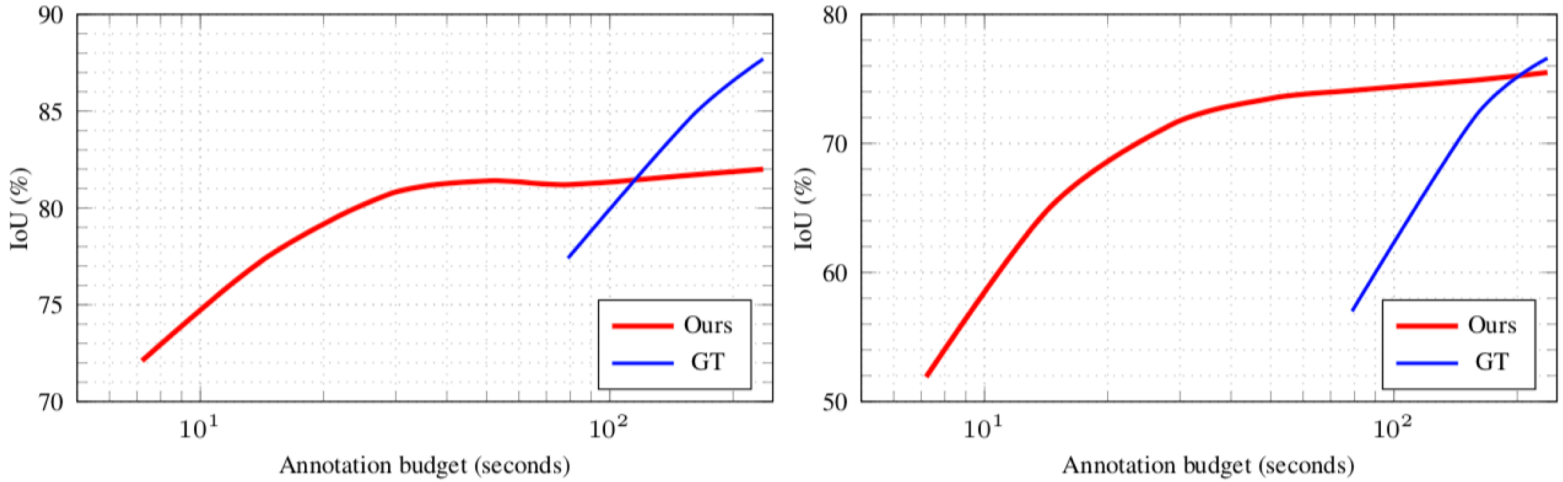


Figure 5. **Quality vs. annotation budget in video object segmentation**: OSVOS' performance when trained from the masks of DEXTR or the ground truth, on DAVIS 2016 (left) and on DAVIS 2017 (right).

Performance

- Interactive Segmentation
  - Interactive Object Segmentation Evaluation

Trained on	4 points	4 points-all	5 points	5 points + OHEM
IoU	59.6%	69.0%	69.2%	**73.2%**

Table 7. **Interactive Object Segmentation Evaluation**: Average IoU on difficult cases of PASCAL VOC 2012 validation dataset.

Performance

- Interactive Segmentation
  - Comparison to the state of the art

Method	Number of Clicks PASCAL@85%	Number of Clicks Grabcut@90%	IoU (%) @ 4 clicks PASCAL	IoU (%) @ 4 clicks Grabcut
GraphCut [4]	> 20	> 20	41.1	59.3
Geodesic matting [2]	> 20	> 20	45.9	55.6
Random walker [9]	16.1	15	55.1	56 .9
iFCN [42]	8.7	7.5	75.2	84.0
RIS-Net [10]	5.7	6.0	80.7	85.0
Ours	**4.0**	**4.0**	**91.5**	**94.4**

Table 8. **PASCAL and Grabcut Dataset evaluation:** Comparison to interactive segmentation methods in terms of number of clicks to reach a certain quality and in terms of quality at 4 clicks.