IEEE International Conference on Image Processing (ICIP 2017), Beijing, China

Improving the Discrimination Between Foreground and Background for Semantic Segmentation

Yu Liu and Michael S. Lew

Leiden Institute of Advanced Computer Science, Leiden University

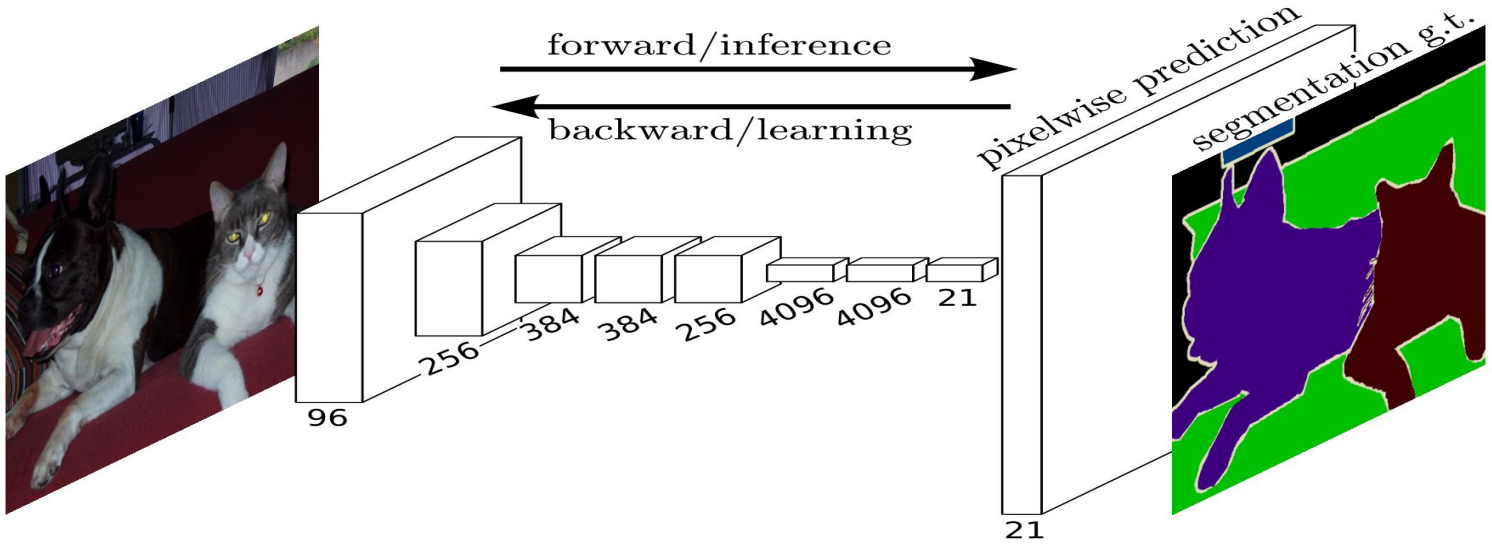
Universiteit Leiden
The Netherlands

Introduction

- Semantic segmentation aims to classify image pixels with pre-defined class labels.
- Inspired by the success from convolutional neural networks (CNN) , a great many works have applied CNNs to semantic segmentation, and yielded state-of-the-art performance.
- Particularly, fully convolutional networks (FCNs) have become the most widely-used segmentation architecture.

Introduction

- A plain FCN for semantic segmentation

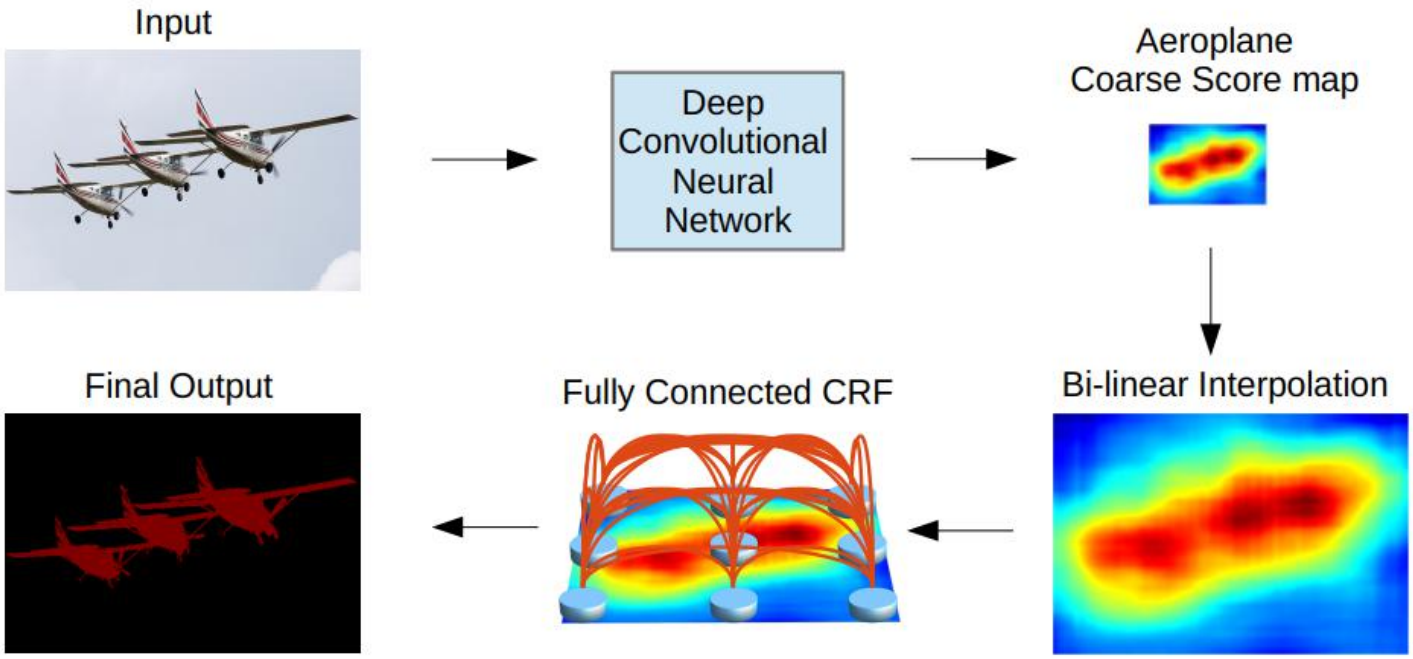


- ✓ Replace the fully-connected layers with more convolutional layers
- ✓ Upsample the convolutional layers to the original image size
- ✓ Pixel-level classification
- ✓ Image-to-image trainable network
- ✓ Multi-layer fusion: FCN-32s->FCN-16s->FCN-8s

Jonathan Long, et al. Fully Convolutional Networks for Semantic Segmentation. CVPR, 2015.

Introduction

- Conditional Random Fields (CRFs)



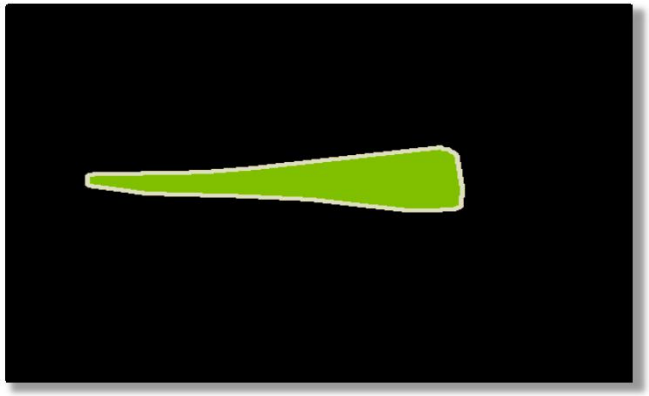
- ✓ Detailed boundary recovery
- ✓ Per-pixel probability vector (e.g. 21 classes in Pascal VOC) is fed into the unary potential of CRFs.

Liang-Chieh Chen, et al. SEMANTIC IMAGE SEGMENTATION WITH DEEP CONVOLUTIONAL NETS AND FULLY CONNECTED CRFS. ICLR, 2015.

Motivation

Input Image

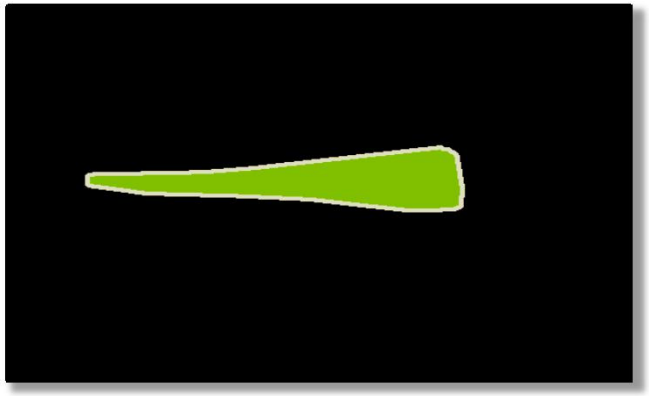
Ground-truth



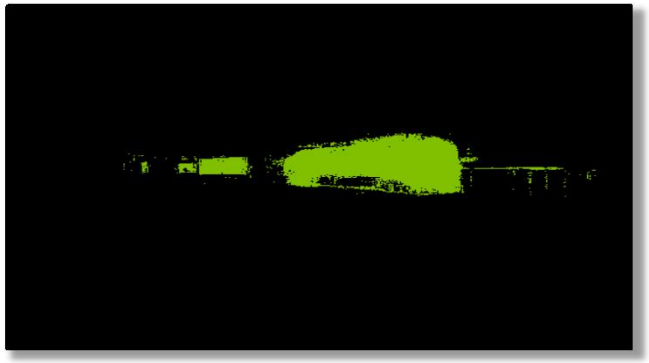
Motivation

Input Image

Ground-truth



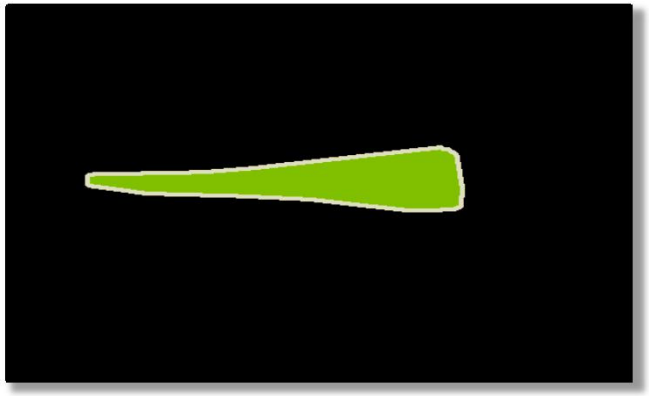
FCN+CRF



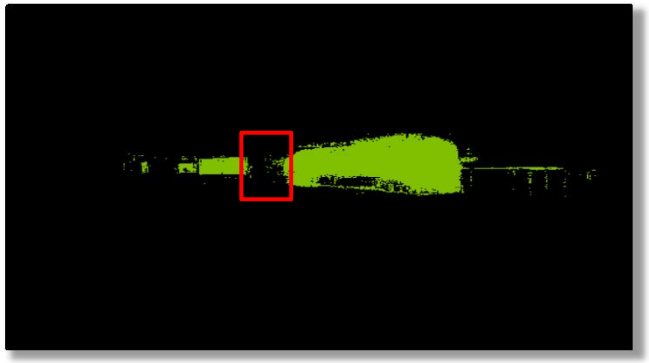
Motivation

Input Image

Ground-truth



FCN+CRF



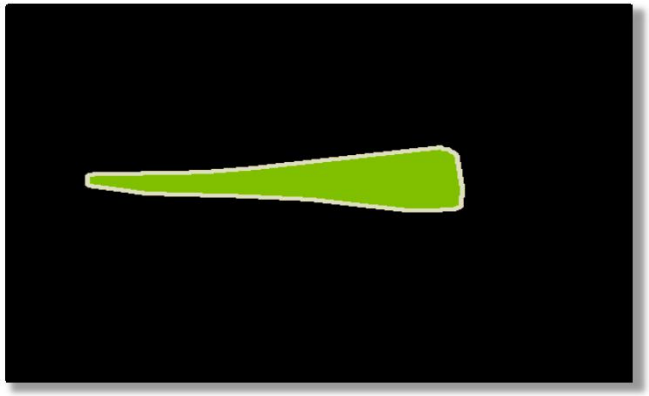
Problem:

Some object pixels (foreground) are wrongly classified as background ！

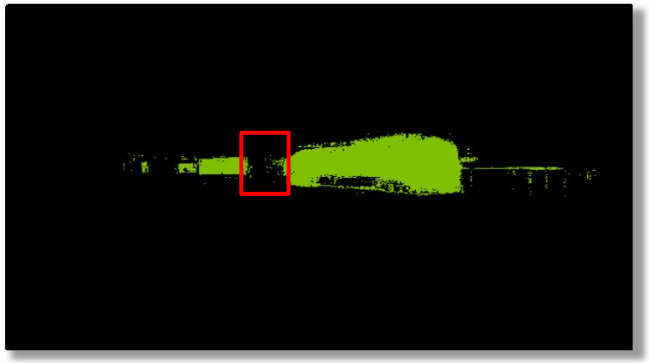
Motivation

Input Image

Ground-truth



FCN+CRF



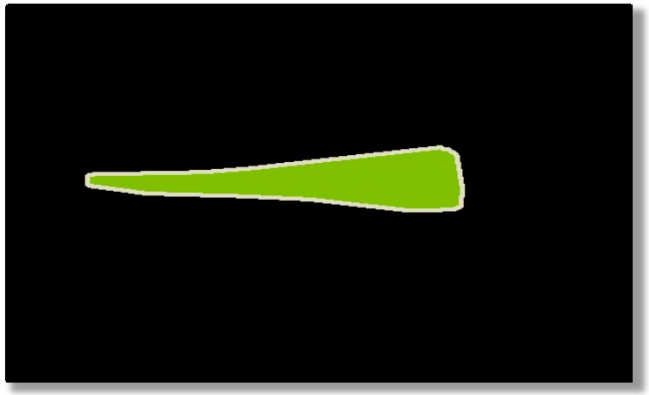
Our purpose:

- Improve the discrimination/distinction between foreground and background.
- Recover some foreground pixels from background.

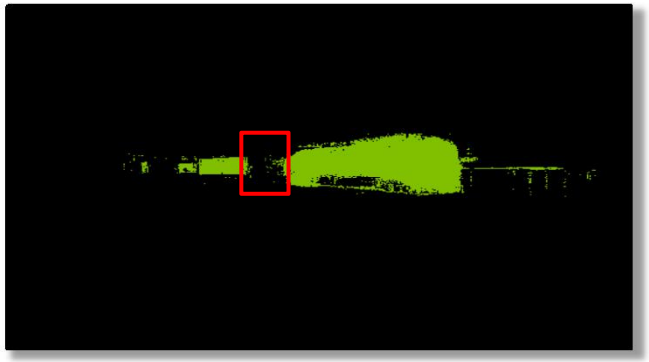
Motivation

Input Image

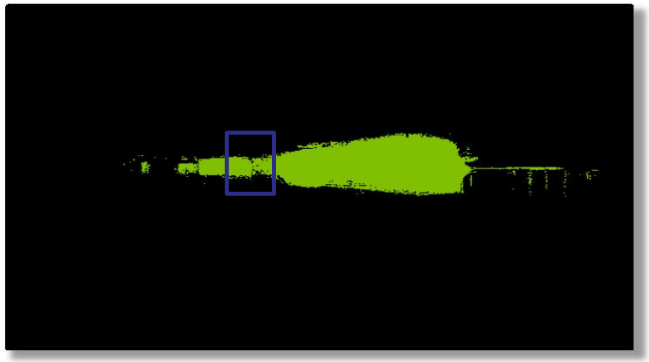
Ground-truth



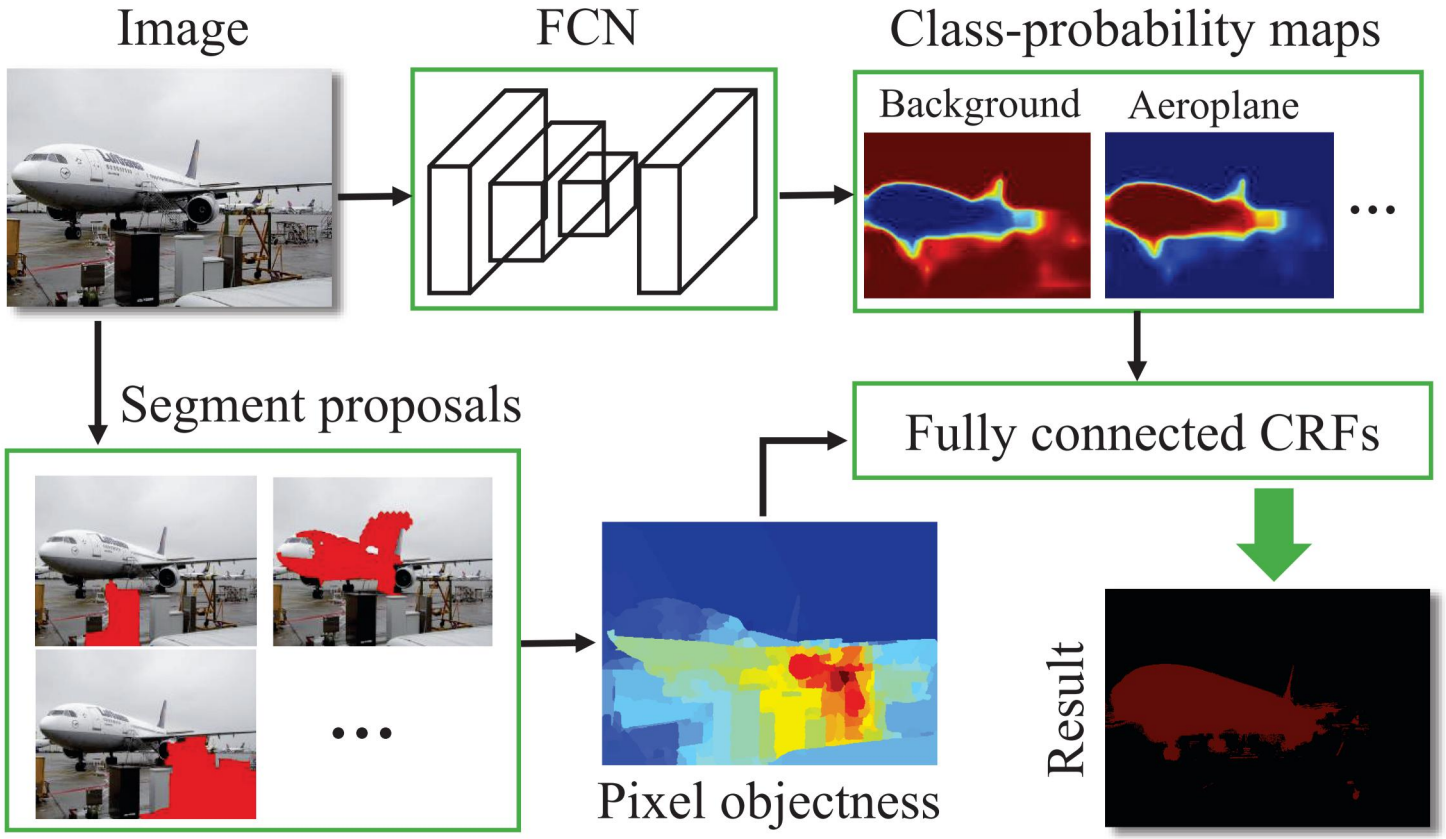
FCN+CRF



Our approach

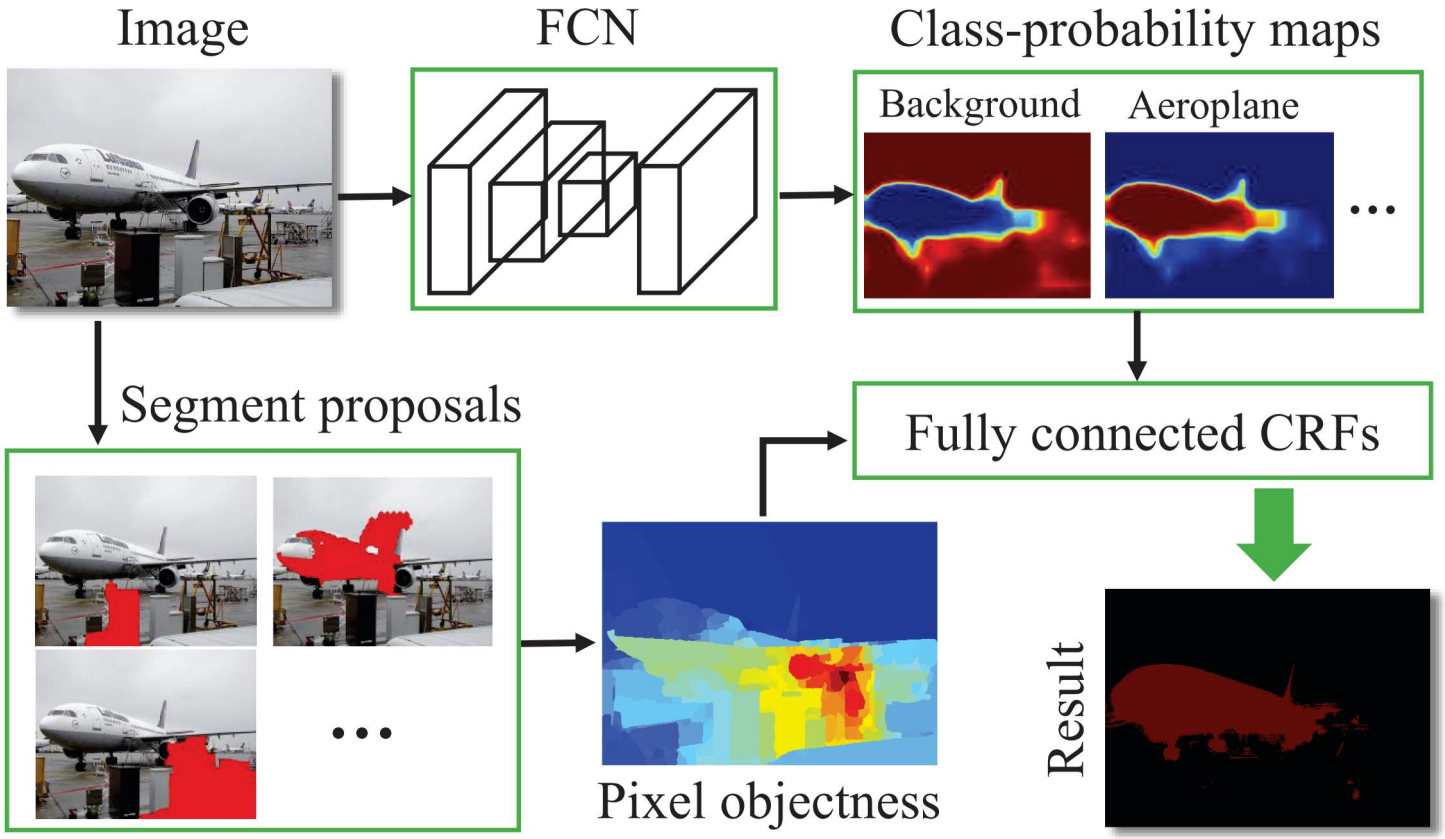


Our approach



(1) Fused loss function to train the FCN
(2) Pixel objectness to compute the CRFs

Our approach



(1) Fused loss function to train the FCN
(2) Pixel objectness to compute the CRFs

Fused loss function

(1) Softmax loss function for segmentation

$$\mathcal{L}_0 = -\frac{1}{N}\left[\sum_{i=1}^{N}\sum_{j=1}^{M_i}\sum_{k=0}^{C}\mathbf{h}(y_j^{(i)}=k)\log p_{j,k}^{(i)}\right]$$

$$p_{j,k}^{(i)} = \frac{\exp(s_{j,k}^{(i)})}{\sum_{l=0}^{C}\exp(s_{j,l}^{(i)})}$$

S: the input of the softmax layer
P: the predicted probability
N: mini-batch size
M: image size (height*width)
C: the number of object classes
y: ground-truth pixel lable

$\mathbf{h}(i=j)$ is the Kronecker delta response $\delta_{i,j}$

Fused loss function

(1) Softmax loss function for segmentation

$$\mathcal{L}_0 = -\frac{1}{N}\left[\sum_{i=1}^{N}\sum_{j=1}^{M_i}\sum_{k=0}^{C}\mathbf{h}(y_j^{(i)}=k)\log p_{j,k}^{(i)}\right]$$

$$p_{j,k}^{(i)} = \frac{\exp(s_{j,k}^{(i)})}{\sum_{l=0}^{C}\exp(s_{j,l}^{(i)})}$$

- This loss function equally computes the loss cost for all object classes and background. However, much error in semantic segmentation is attributed to the incorrect predictions between foreground and background.

Fused loss function

(2) Positive-sharing loss function for segmentation

- All object classes (foreground) are integrated as a positive class; the background is a negative class.
- This loss function is used to classifiy the foreground / background.

$$\mathcal{L}_1 = -\frac{1}{N}\left[\sum_{i=1}^{N}\sum_{j=1}^{M_i}\left(\mathbf{h}(y_j^{(i)}=0)\log p_{j,0}^{(i)} + \sum_{k=1}^{C}\mathbf{h}(y_j^{(i)}=k)\log(1-p_{j,0}^{(i)})\right)\right]$$

Background Foreground

Fused loss function

(2) Positive-sharing loss function for segmentation

- All object classes (foreground) are integrated as a positive class; the background is a negative class.
- This loss function is used to classifiy the foreground / background.

$$\mathcal{L}_1 = -\frac{1}{N}\left[\sum_{i=1}^{N}\sum_{j=1}^{M_i}\left(\mathbf{h}(y_j^{(i)}=0)\log p_{j,0}^{(i)} + \sum_{k=1}^{C}\mathbf{h}(y_j^{(i)}=k)\log(1-p_{j,0}^{(i)})\right)\right]$$

Background

Foreground

sum up the predicted probabilities of all object classes.

Fused loss function

(2) Positive-sharing loss function for segmentation

- All object classes (foreground) are integrated as a positive class; the background is a negative class.
- This loss function is used to classifiy the foreground / background.

- DeepContour:
  two-class contour detection -> multi-class classification task

- Our approach:
  multi-class semantic segmentation -> two-class classification task

Wei Shen, et al. DeepContour: A deep convolutional feature learned by positive-sharing loss for contour detection. CVPR, 2015.

Fused loss function

The final loss fuses the softmax loss function and positive-sharing loss function by

$$\mathcal{L} = W_s \cdot \mathcal{L}_0 + W_p \cdot \mathcal{L}_1$$

W_s W_p are used to balance the two loss functions.

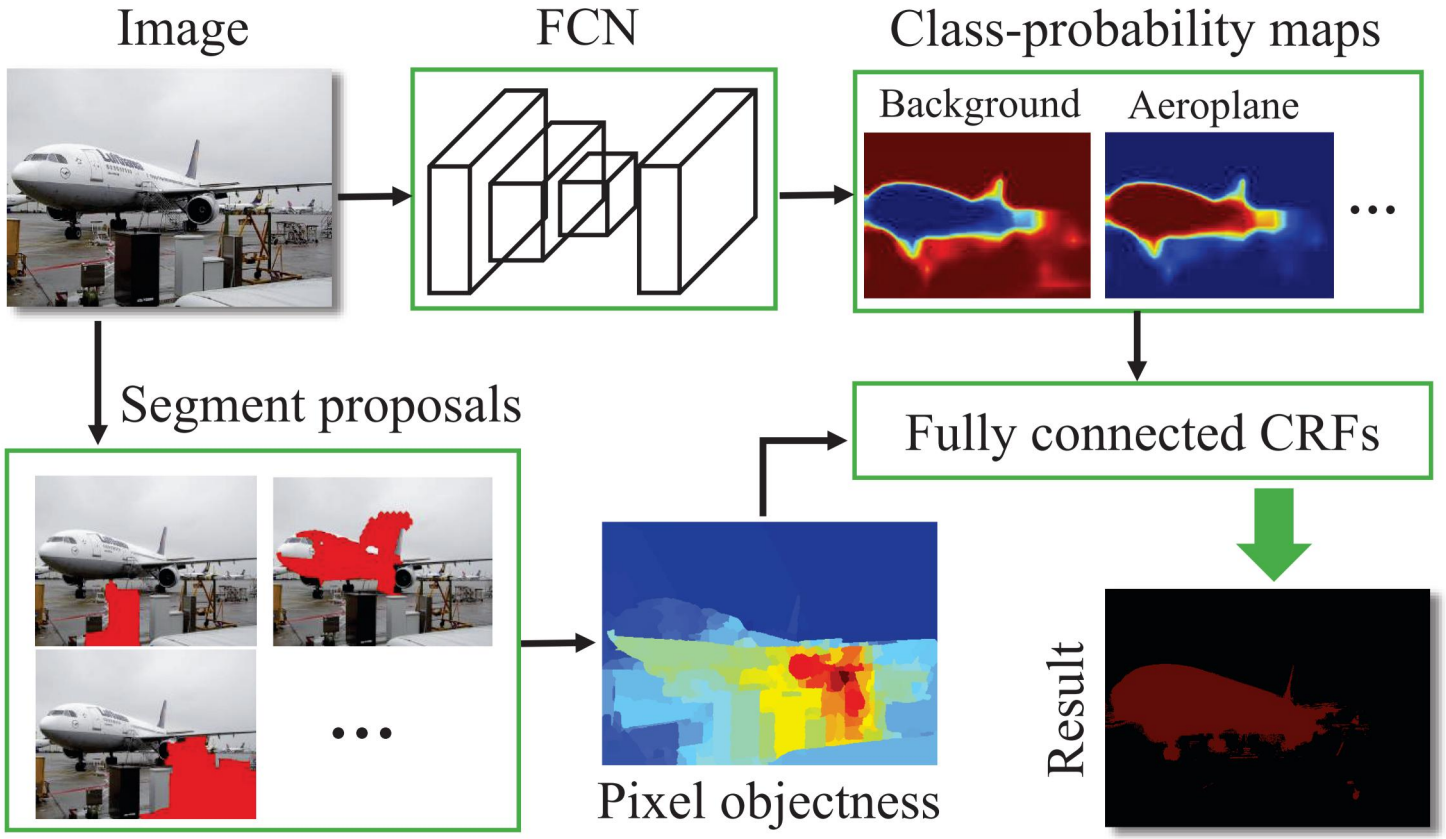
Fused loss function

For example, we show the partial derivatives of the fused loss *w.r.t.* s (*i.e.* the input of the softmax layer)

$$\frac{\partial \mathcal{L}}{\partial s_{j,0}^{(i)}} = \frac{1}{N}\Big[(W_s + W_p)\mathbf{h}(y_j^{(i)} = 0)(p_{j,0}^{(i)} - 1) + (W_s + W_p)\sum_{k=1}^{C}\mathbf{h}(y_j^{(i)} = k)p_{j,0}^{(i)}\Big],$$

$$\frac{\partial \mathcal{L}}{\partial s_{j,l}^{(i)}} = \frac{1}{N}\left[\Big(W_s + W_p\mathbf{h}(y_j^{(i)} = 0)\Big)p_{j,l}^{(i)} - W_s\mathbf{h}(y_j^{(i)} = l) - W_p\sum_{k=1}^{C}\mathbf{h}(y_j^{(i)} = k)\left(\frac{p_{j,0}^{(i)}p_{j,l}^{(i)}}{1 - p_{j,0}^{(i)}}\right)\right].$$

Our approach



(1) Fused loss function to train the FCN
(2) Pixel objectness to compute the CRFs

Pixel objectness (POS)

- POS measures the probability of a pixel locating within a salient object.
- Our hypothesis is that if there are more object proposals containing one pixel, then this pixel should be assigned with a larger weight (or objectness).
- We use the geodesic object proposals (GOP) [Philipp Krahenbuhl, et al, ECCV2014] to extract object proposals.

Pixel objectness (POS)

- POS measures the probability of a pixel locating within a salient object.
- Our hypothesis is that if there are more object proposals containing one pixel, then this pixel should be assigned with a larger weight (or objectness).
- We use the geodesic object proposals (GOP) [Philipp Krahenbuhl, et al, ECCV2014] to extract segment proposals.

$$pos_j^{(i)} = \frac{g_j^{(i)}}{G^{(i)}}, pos_j^{(i)} \in [0, 1].$$

$g_j^{(i)}$ counts how many proposals containing the *j*-th pixel.

$G^{(i)}$ is the total number of segment proposals in the *i*-th image .

Pixel objectness for CRFs

- The energy function of CRFs is represented by

$$E(x) = \underbrace{\sum_{j}^{M^{(i)}} \theta(x_j^{(i)})}_{\text{unary potential}} + \underbrace{\sum_{j_1}^{M^{(i)}} \sum_{j_2}^{M^{(i)}} \theta(x_{j_1}^{(i)}, x_{j_2}^{(i)})}_{\text{pairwise potential}}$$

(1) The unary potential is computed with FCN and POS.
(2) The pairwise potential is computed with bilateral position and color intensities.

Philipp Krahenbuhl, et al. Efficient inference in fully connected crfs with gaussian edge potentials. NIPS, 2011.

Pixel objectness for CRFs

- The unary potential is computed separately for foreground and background.

$$\theta_k(x_j^{(i)}) = \begin{cases} -\log p_{j,0}^{(i)} & k = 0 \\ -\log\left(p_{j,k}^{(i)} \cdot \exp(pos_j^{(i)})\right), & 1 \le k \le C \end{cases}$$

Background

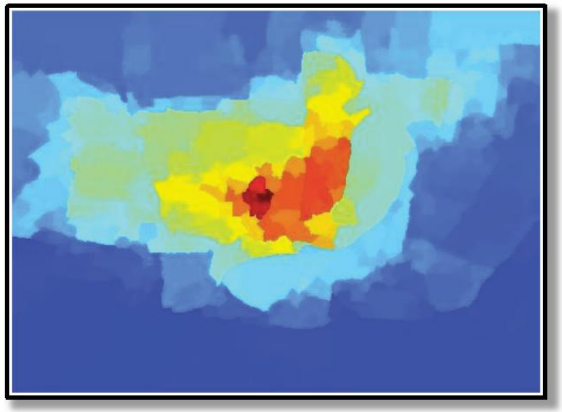
Foreground

$p_{j,k}^{(i)}$ is the probability vector predicted by FCN.

We add POS to the unary potential of foreground pixels, to improve their importance. Therefore, POS allows to avoid some important object pixels to be classified as background.

Pixel objectness for CRFs

Input Image



POS Map

without POS

with POS

Ground Truth

Results

Table. Intersection-over-union (IoU) accuracy on the Pascal VOC 2012 val set.

Method	FCN-32s	FCN-16s	FCN-8s
SoftmaxLoss	59.61	62.52	62.91
FusedLoss	60.22	63.05	63.35
FusedLoss+CRFs	63.21	66.05	66.42
FusedLoss+POS-CRFs	63.55	66.42	66.71

Results

Table. Intersection-over-union (IoU) accuracy on the Pascal VOC 2012 val set.

Method	FCN-32s	FCN-16s	FCN-8s
SoftmaxLoss	59.61	62.52	62.91
FusedLoss	60.22	63.05	63.35
FusedLoss+CRFs	63.21	66.05	66.42
FusedLoss+POS-CRFs	63.55	66.42	66.71

- The fused loss increases about 0.4-0.5% accuracy, compared with the softmax loss.

Results

Table. Intersection-over-union (IoU) accuracy on the Pascal VOC 2012 val set.

Method	FCN-32s	FCN-16s	FCN-8s
SoftmaxLoss	59.61	62.52	62.91
FusedLoss	60.22	63.05	63.35
FusedLoss+CRFs	63.21	66.05	66.42
FusedLoss+POS-CRFs	63.55	66.42	66.71

- The fused loss increases about 0.4-0.5% accuracy, compared with the softmax loss.
- Using the CRFs can boost the accuracy with remarkable improvements.

Results

Table. Intersection-over-union (IoU) accuracy on the Pascal VOC 2012 val set.

Method	FCN-32s	FCN-16s	FCN-8s
SoftmaxLoss	59.61	62.52	62.91
FusedLoss	60.22	63.05	63.35
FusedLoss+CRFs	63.21	66.05	66.42
FusedLoss+POS-CRFs	63.55	66.42	66.71

- The fused loss increases about 0.4-0.5% accuracy, compared with the softmax loss.
- Using the CRFs can boost the accuracy with remarkable improvements.
- When adding the POS to CRFs, the model can get about 0.3% IoU gain.

Results

Table. Intersection-over-union (IoU) accuracy on the Pascal VOC 2012 val set.

Method	FCN-32s	FCN-16s	FCN-8s
Baseline: SoftmaxLoss + CRFs	62.64	65.45	65.85
Ours: FusedLoss + POS-CRFs	63.55	66.42	66.71

Table. Recall measurement results on the Pascal VOC 2012 val set.

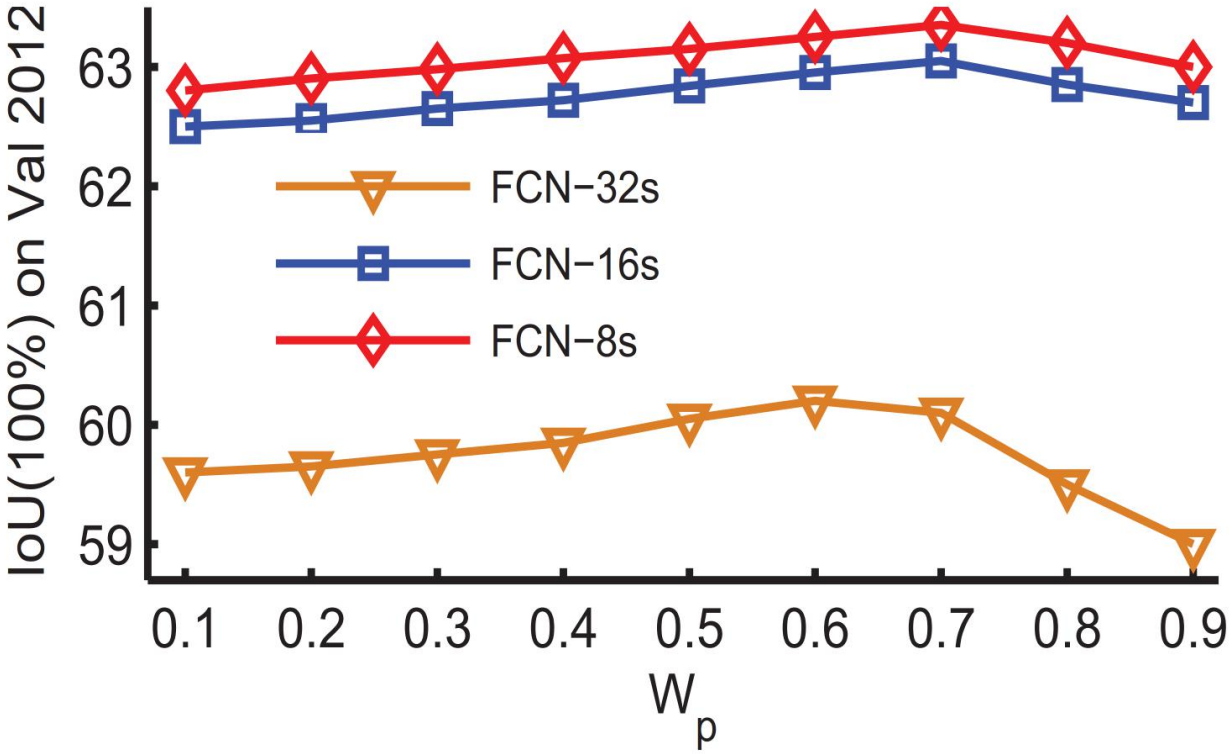
Method	FCN-32s	FCN-16s	FCN-8s
Baseline: SoftmaxLoss + CRFs	68.65	72.58	74.98
Ours: FusedLoss + POS-CRFs	70.84	74.71	77.15

$$\text{Recall measurement} = \frac{\#correct}{\#total}$$

#total is the number of object pixels in one image.
#correct indicates how many object pixels are detected correctly.

Effect of Weights

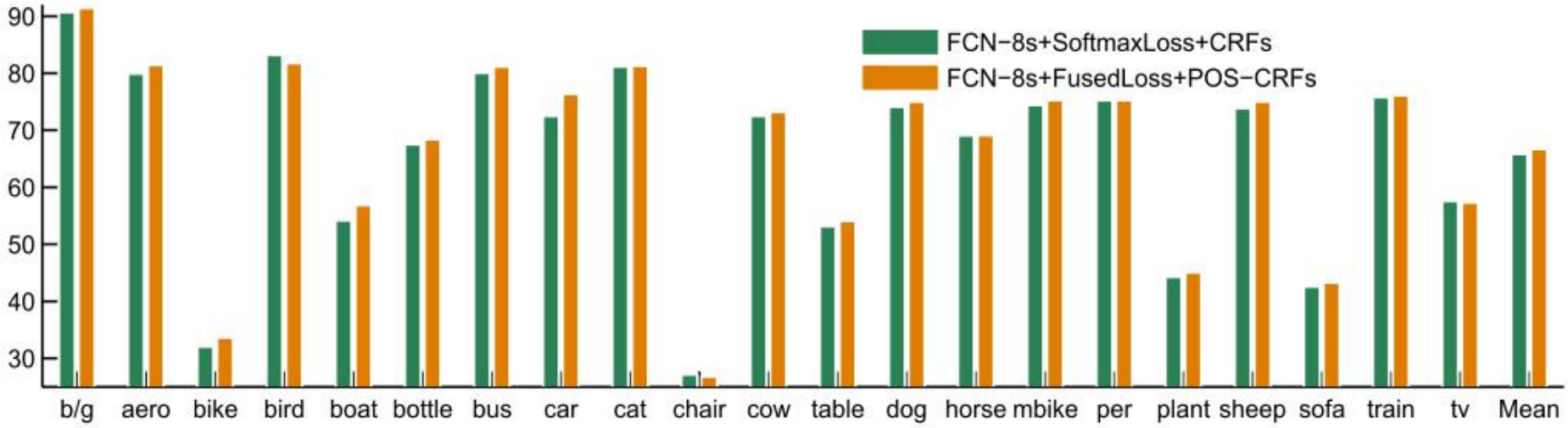
$$\mathcal{L} = W_s \cdot \mathcal{L}_0 + W_p \cdot \mathcal{L}_1$$
$$W_p + W_s = 1$$



FCN-32s: Wp = 0.6;
FCN-16s and FCN-8s: Wp = 0.7

Results

20 object classes results on the PASCAL VOC 2012 val set

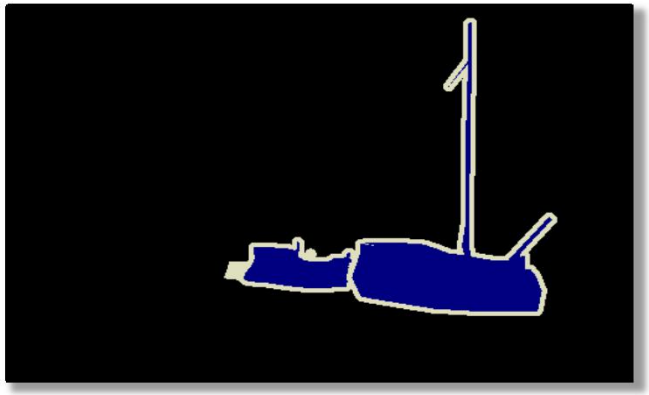


For most classes, our method (FCN-8s+FusedLoss+POS-CRFs) is better than the baseline (FCN-8s+SoftmaxLoss+CRFs).

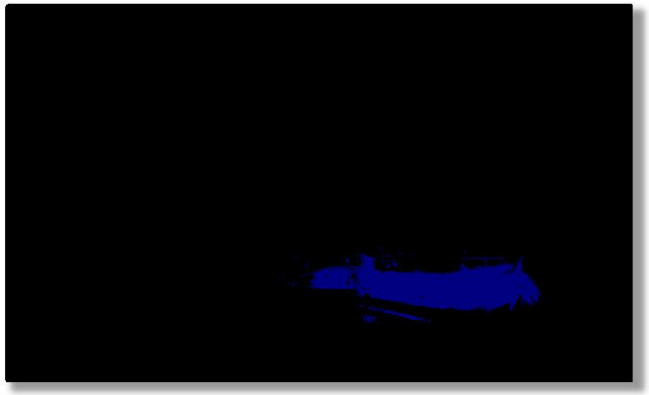
Results

Input Image

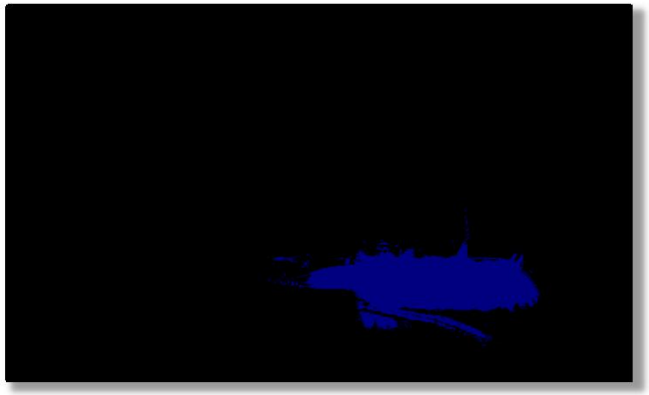
Ground-truth



SoftmaxLoss+CRFs



FusedLoss+POS-CRFs



Conclusions

- We develop two improvements to boost the distinction between foreground and background.
- Compared with the FCN baseline, our approach obtains considerable improvements in both IoU and recall performance measurements, especially small details of objects.
- The proposed method is feasible to be adapted to other state-of-the-art segmentation approaches.

Thanks for your attention!

Questions please?