Appendix

Hyperparameters

Operation	Kernel size	Stride	Feature maps	Padding
Network – $513 \times 513 \times 3$ input				
ResNet-v1-101				
Convolutional LSTM	1	1	512	SAME
Convolutional LSTM	1	1	256	SAME
Convolutional LSTM	1	1	$\lvert\text{classes}\rvert$	SAME
Canvas (add:0)			$\lvert\text{classes}\rvert$	
Bilinear upsampling			$\lvert\text{classes}\rvert$	
Padding mode	Zeros			
Normalization	Batch normalization after every ResNet convolution			
Optimizer	SGD with Momentum (momentum $= 0.95$)			
Parameter updates	30,000			
Learning rate schedule	$(1e^{-3} - \epsilon) \cdot \left(1 - \frac{\text{step}}{\text{total steps}}\right)^{0.9} + \epsilon$ where $\epsilon = 1e^{-6}$			
Batch size	16			
Weight initialization	Glorot normal [25]			

Table 4. Details of the recurrent network architecture for image segmentation. $\lvert\text{classes}\rvert$ is 21 for the PASCAL VOC 2012 semantic segmentation dataset, and 19 for the Cityscapes dataset. The final block4 of the ResNet-v1-101 was augmented with dilation rates of (2, 4, 8) in the three units of block4, following [9].

Supplemental methods

Convolutional LSTM

For the recurrent network architecture, we use stacked convolutional LSTM layers [70] defined as

$$
\begin{aligned}
\mathbf{i}_t &= \sigma\left(\mathbf{W^{ih}} * \mathbf{h}_{t-1} + \mathbf{W^{ix}} * \mathbf{x}_t + \mathbf{b^i}\right) && (1)\\
\mathbf{f}_t &= \sigma\left(\mathbf{W^{fh}} * \mathbf{h}_{t-1} + \mathbf{W^{fx}} * \mathbf{x}_t + \mathbf{b^f} + b^{fg}\right) && (2)\\
\mathbf{c}_t^{in} &= \tanh\left(\mathbf{W^{ch}} * \mathbf{h}_{t-1} + \mathbf{W^{cx}} * \mathbf{x}_t + \mathbf{b^c}\right) && (3)\\
\mathbf{c}_t &= \mathbf{f}_t \cdot \mathbf{c}_{t-1} + \mathbf{i}_t \cdot \mathbf{c}_t^{in} && (4)\\
\mathbf{o}_t &= \sigma\left(\mathbf{W^{oh}} * \mathbf{h}_{t-1} + \mathbf{W^{ox}} * \mathbf{x}_t + \mathbf{b^o}\right) && (5)\\
\mathbf{h}_t &= \mathbf{o}_t \cdot \tanh(\mathbf{c}_t), && (6)
\end{aligned}
$$

where $\sigma(\cdot)$ is the logistic function, $*$ is the convolution operator, where i, f, o represent the input, forget and output gates, respectively, and c and h are the state of the LSTM. Forget gate offset bias $b^{fg} = 1.0$.

Training details

We found that the best performance was achieved by having a batch-size of 12 - 16 and a relatively large ResNet output stride of 16.

We also found that the crop-size had a significant effect on performance, with the highest performance achieved with keeping the crop-size of the input image as large as the native resolution - 513×513 for PASCAL VOC images and 1025×2049 for Cityscapes images. In each case the crop size is an integer divisible by 32, plus one, in order to avoid edge effects with the ResNet output stride.

Estimating computational cost

We used the Tensorflow profiler (tf.profiler.Profiler) to estimate FLOPS during evaluation of the models. We also constructed Tensorflow models of the Pyramid Scene Parsing network [72] and Deeplab V3 [9], following the methods reported as closely as possible. From these models we used Tensorflow profiling as before to estimate the FLOPS for these models. Note however that all performance numbers for both models are taken from the values reported in the original papers.

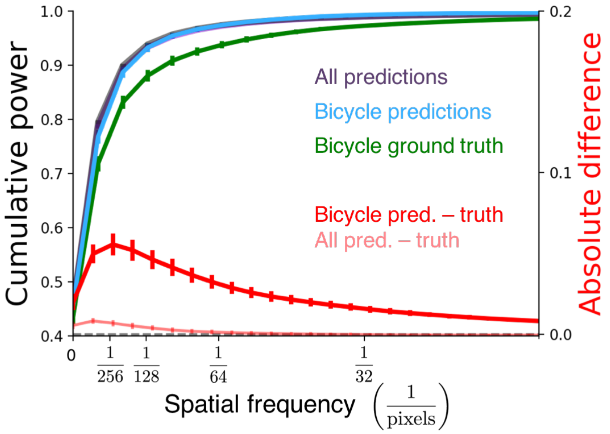

Figure 8. Spatial frequency analysis of segmentation errors. Cumulative distribution of power spectral densities of RNN predictions on all classes (purple), images containing bicycles (blue), and ground truth bicycle segmentations (green). The difference between spectral density distributions for the RNN and ground truth labels are shown in red for the bicycle class (dark red) and all semantic classes (light red).