Learning Detection with Diverse Proposals
Supplementary Material

Samaneh Azadi[1], Jiashi Feng[2], and Trevor Darrell[1]

[1]University of California, Berkeley, [2]National University of Singapore

1. LDDP Back-Propagation

Given the set of representative proposals, Y, and set of probable background proposals, B, we can compute the log-likelihood loss function in Eq. (1) as well as its gradient with respect to the confidence scores.

$$\mathcal{L}(\alpha) = \quad \log P_\alpha(Y|X) - \log P_\alpha(B|X) \qquad (1)$$

where α refers to the parameters of the deep network, and conditional probabilities are defined based on the DPP model [1] and our set of parameters:

$$\begin{aligned} P_\alpha(\mathbf{Y}=Y|X) &= \frac{1}{\det(L+I)}\det L_Y, \\ L_{i,j} &= \Phi_i^{1/2} S_{ij} \Phi_j^{1/2}. \end{aligned} \qquad (2)$$

where L denotes the L-ensemble matrix, S the similarity matrix, Φ the quality measure, and $L_Y := [L_{ij}]_{i,j\in Y}$ denotes the restriction of L to the entries indexed by elements of Y. Expanding the above probability distribution:

$$\begin{aligned} P_\alpha(Y|X) &= \Big(\prod_{i\in Y}\Phi_i\Big)\frac{\det S_Y}{\det(L+I)}, \\ \log P_\alpha(Y|X) &= \sum_{i\in Y}\log\Phi_i \\ &+ \log\det S_Y - \log\det(L+I) \end{aligned} \qquad (3)$$

where $\det(L+I)$ is the normalizing factor as $\sum_{Y'\subseteq\mathcal{Y}} L_{Y'}$ with $\mathcal{Y}$ as all possible sets of proposals selections.

Now, we take the gradient of each term of the above loss function with respect to the outputs of the inner product layer before softmax. As explained in the paper, for the first term, $p_1 = \log P_\alpha(Y|X)$, we have:

$$\Phi_i = \begin{cases} \text{IoU}_{i,gt^i} \times \exp\{W_{gt}^T f_i\}, & \text{if } i \in Y \\ \text{IoU}_{i,gt^i} \times \sum_{c\neq 0}\exp\{W_c^T f_i\} & \text{if } i \notin Y \end{cases} \qquad (4)$$

where W_{gt} denotes the weight vector for the corresponding ground-truth label of proposal i, and $c=0$ shows the background category. According to Eq. (3), (4), the first conditional probability distribution would be as:

$$\begin{aligned} \log P_\alpha(Y|X) &= \sum_{i\in Y}\log \text{IoU}_{i,gt^i} + \sum_{i\in Y} b_i^{gt} \\ &+ \log\det S_Y - \log\det(L+I) \end{aligned} \qquad (5)$$

where $b_i^{gt} = W_{gt}^T f_i$. The proposals indexed by $i \notin Y$ and labeled as background are not involved in this log probability resulting in a zero gradient. The same result will be applied on the proposals indexed by $i \in Y$ and labeled by category $c \neq gt$. On the other hand:

$$\log\det(L+I) = \log\sum_{Y'}\Big(\prod_{j\in Y'}\Phi_j\Big)\det S_{Y'} \qquad (6)$$

Therefore, according to Eq. (4), (6) for the proposals indexed by $i \in Y$ and labeled $c = gt$:

$$\begin{aligned} &\frac{\partial \log\det(L+I)}{\partial b_i^c} \\ &= \sum_{Y'}\mathrm{I}\{i\in Y'\}\frac{\partial\Phi_i}{\partial b_i^c}\Big(\prod_{\substack{j\in Y'\\ j\neq i}}\Phi_j\Big)\frac{\det S_{Y'}}{\det(L+I)} \\ &= \sum_{Y'}\mathrm{I}\{i\in Y'\}\Big(\prod_{j\in Y'}\Phi_j\Big)\frac{\det S_{Y'}}{\det(L+I)} = K_{ii} \end{aligned} \qquad (7)$$

Here, $K_{ii} = L_{ii}/\det(L+I)$, and $\mathrm{I}\{.\}$ is the indicator function. Combining Eq. (5), (7):

$$\frac{\partial log p_1}{\partial b_i^c} = 1 - K_{ii} \quad \forall i \in Y, c = gt \qquad (8)$$

Similarly for the proposals indexed by $i \notin Y$ and labeled as

Table 1: Ablation study on semantic similarity matrix used in LDDP inference. MS COCO minival detection average precision and average recall(%) (trained on COCO train set). All methods use VGG_CNN_M_1024 deep convolutional network.

Similarity Matrix	Avg Precision @ IoU:			Avg Precision @ Area:			Avg Recall, #Dets:			Avg Recall @ Area:		
	0.5-0.95	0.5	0.75	S	M	L	1	10	100	S	M	L
$S_{ij} = \text{IoU}_{ij} \times \text{sim}_{ij}$	15.4	32.0	13.0	4.0	16.3	25.0	17.1	24.9	25.4	7.1	27.5	41.6
$S_{ij} = \text{IoU}_{ij} \times \text{sim}_{ij}^4$	15.4	32.3	13.1	4.0	16.5	25.2	17.4	25.6	26.1	7.5	28.4	42.9
$S_{ij} = \text{IoU}_{ij}$	14.5	29.9	12.5	3.6	15.3	23.6	15.4	21.5	21.9	5.7	23.3	35.1

$c \neq 0$:

$$\begin{aligned}
&\frac{\partial \log\det(L+I)}{\partial b_i^c} \\
&= \sum_{Y'} \mathrm{I}\{i \in Y'\} \frac{\partial \Phi_i}{\partial b_i^c} \Big(\prod_{\substack{j \in Y' \\ j \neq i}} \Phi_j \Big) \frac{\det S_{Y'}}{\det(L+I)} \\
&= \sum_{Y'} \mathrm{I}\{i \in Y'\} \frac{\exp\{b_i^c\}}{\sum_{c' \neq 0} \exp\{b_i^{c'}\}} \Big(\prod_{j \in Y'} \Phi_j \Big) \frac{\det S_{Y'}}{\det(L+I)} \\
&= K_{ii} \frac{\exp\{b_i^c\}}{\sum_{c' \neq 0} \exp\{b_i^{c'}\}} \qquad (9)
\end{aligned}$$

Again, using Eq. (5), (9) results in:

$$\frac{\partial log p_1}{\partial b_i^c} = -K_{ii} \frac{\exp\{b_i^c\}}{\sum_{c' \neq 0} \exp\{b_i^{c'}\}} \quad \text{if } i \notin Y, c \neq 0 \qquad (10)$$

Thus based on Eq. (8), (10), the gradient of the first log likelihood can be summarized as:

$$\frac{\partial \log p_1}{\partial b_i^c} = \begin{cases} 1 - K_{ii}, & \text{if } i \in Y, c = \text{gt} \\ \frac{-K_{ii} \exp\{b_i^c\}}{\sum_{c' \neq 0} \exp\{b_i^{c'}\}}, & \text{if } i \notin Y, c \neq 0 \\ 0 & \text{otherwise} \end{cases} \qquad (11)$$

For the second log probability, $\log p_2 = \log P_\alpha(B|X)$, we change the quality measures as discussed in the paper. We skip the derivation of gradient of $\log p_2$ with respect to each b_i^c, which can be achieved by following a similar scheme:

$$\frac{\partial \log p_2}{\partial b_i^c} = \begin{cases} -K_{ii}, & \text{if } i \notin B, c = \text{gt} \\ \frac{-(K_{ii}-1) \exp\{b_i^c\}}{\sum_{c' \neq 0} \exp\{b_i^{c'}\}}, & \text{if } i \in B, c \neq 0 \\ 0 & \text{otherwise} \end{cases} \qquad (12)$$

2. Additional Experiments

2.1. Smaller Number of Proposals

As explained in the paper, to approve the generation of high-confidence non-redundant proposals through our proposed LDDP network, we evaluate bounding box detection performance when we restrict the number of generated proposals to different values, as shown in Figure 1. Limiting the number of proposals generated by our LDDP model to 100 drops mean AP on VOC2007 test set from 62.2% to 60.4% which is similar to the mean AP achieved by 300 proposals in Faster R-CNN network (60.5%). Thus, our LDDP model is much more efficient than the state-of-the-art Faster R-CNN approach for the task of object detection.

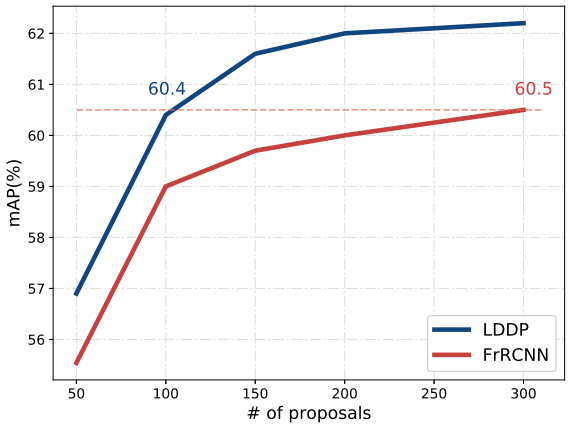

Figure 1: Detection mAP(%) vs. number of proposals generated by our end-to-end LDDP model and Faster R-CNN. Both methods use ZF deep convolutional network and are trained on VOC2007 trainval.

2.2. Ablation Study on Microsoft COCO

To understand how the semantic similarity matrix used in the kernel matrix L affects the performance of our LDDP model, we use its different powers during inference and evaluate the detection performance on the minival subset of MS COCO data set with 5K images. According to the results reported in Table 1, the semantic similarity matrix plays a crucial role in achieving accurate boxes.

2.3. Visualization

We visualize the output of our end-to-end LDDP model as well as Faster R-CNN followed by NMS both on Pascal VOC2007 and MS COCO data sets [2] in Figures 2 and 3, respectively. We use the ZF model architecture

for training the models on Pascal VOC2007 data set and the VGG_CNN_M_1024 deep network for training on MS COCO. The non-repetitive and accurate detections by the LDDP model reveal the superiority of our model against Faster R-CNN.

References

[1] A. Kulesza and B. Taskar. Determinantal point processes for machine learning. *arXiv preprint arXiv:1207.6083*, 2012. 1

[2] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pages 740–755. Springer, 2014. 2

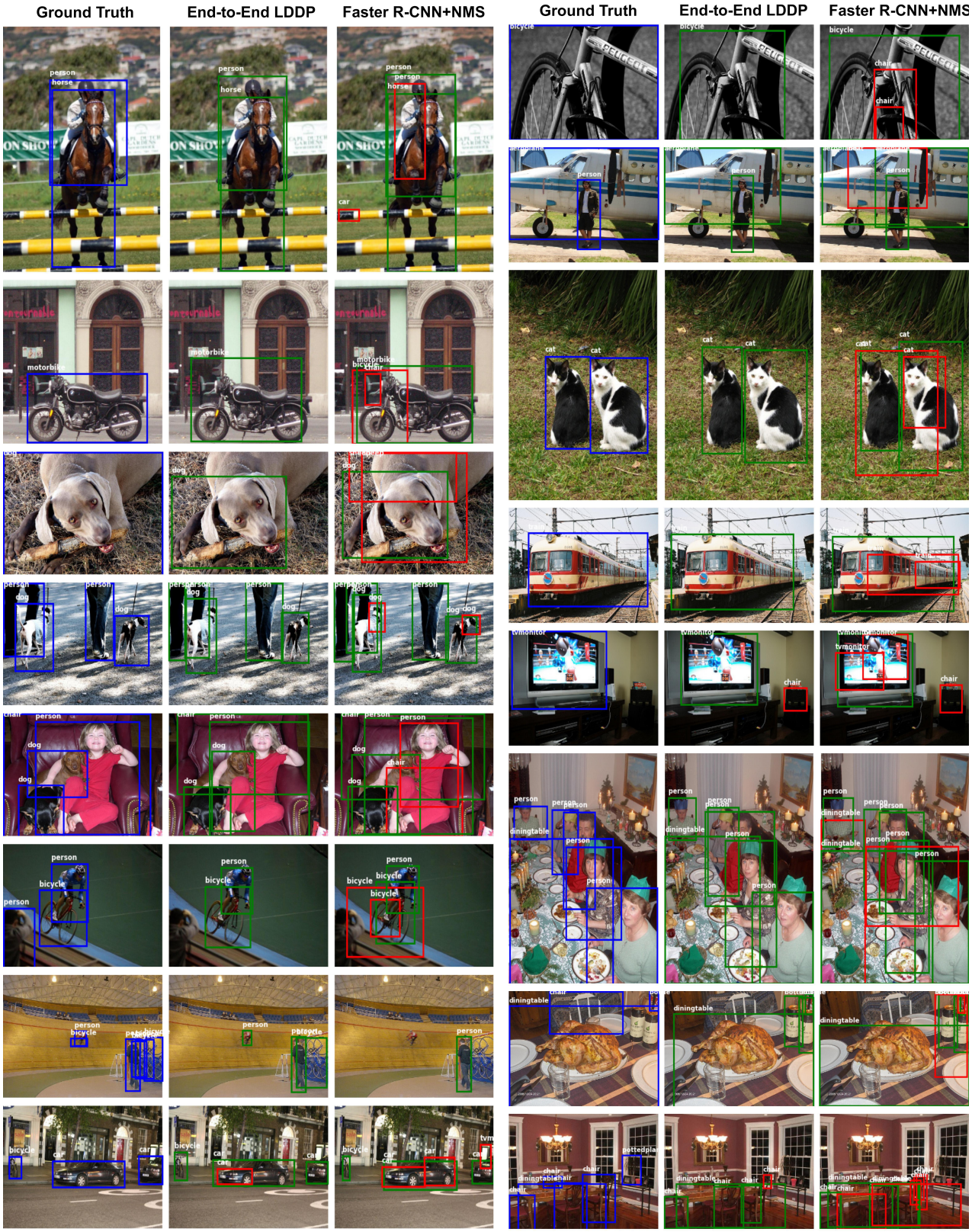

Figure 2: Example images from Pascal VOC2007 data set illustrating our end-to-end LDDP and Faster R-CNN followed by NMS. A score threshold of 0.6 is used to display images.

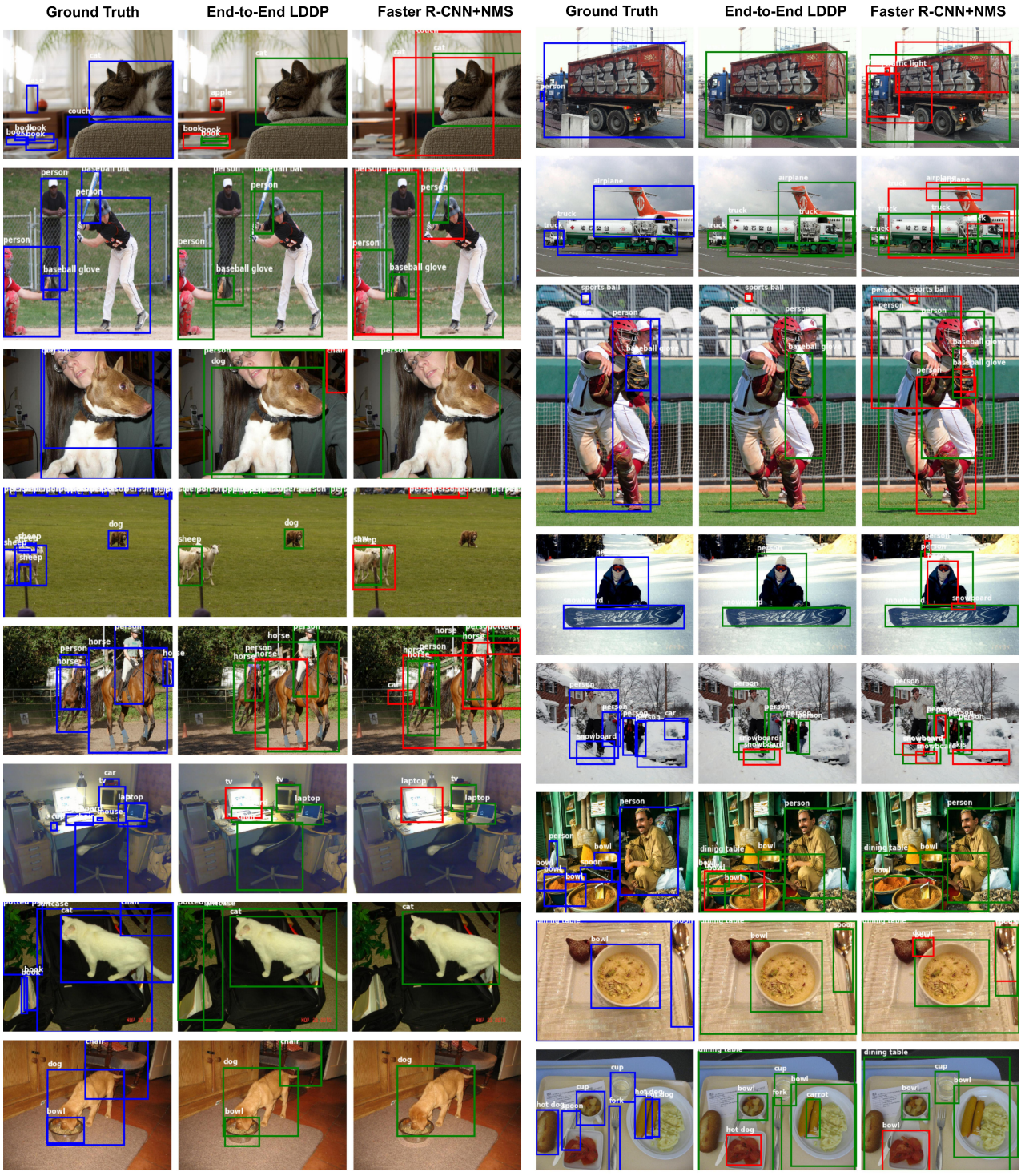

Figure 3: Example images from MS COCO data set illustrating our end-to-end LDDP and Faster R-CNN followed by NMS. A score threshold of 0.6 is used to display images.