Supplementary material for CVPR submission

Paper ID 1665

Fine-grained recognition of thousands of object categories with single-example training

Description of supplementary data

The supplementary data for the paper consists of three parts: (i) test videos from datasets, discussed in the paper, annotated with the output of our proposed algorithm; (ii) Precision-Recall graphs of the proposed algorithm on test data, (iii) visualization of the data augmentation for CNN training, and (iv) the appendix with the proof of proposition stated in Equation (2) of the paper.

Test video files

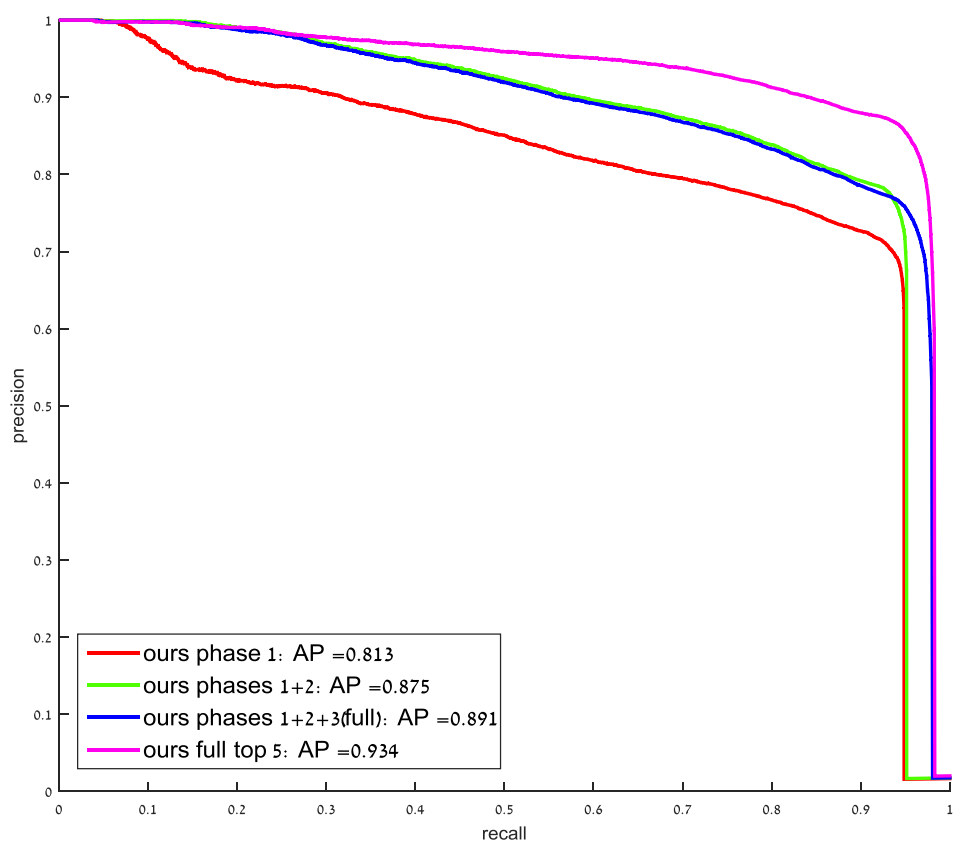
File	Description
car_PCPE_compressed.mp4 engine_PCPE_compressed.mp4 phone_PCPE_compressed.mp4 room_PCPE.mp4 sprinklers_PCPE_compressed.mp4	In these videos, various objects are captured with a moving camera. In the training stage, the objects were scanned to obtain a 3D point cloud. In the test videos (taken by a different camera at a different time, and for car & phone in different environment), the 3D point cloud presence and pose are automatically detected by the proposed approach and overlaid on the original footage. The pose estimation is done in each frame separately, with no temporal smoothing in order to illustrate the method performance on still images (the individual frames). In the overlay, the test images are displayed in greyscale, while the 3D point cloud in the pose detected by the proposed method is displayed in color on top of the test image. In videos of car and of the engine, the screen is split in two, displaying in the upper half the original image without the overlay – this is done for a better visibility of the original scene. In the video sprinklers_PCPE_compressed.mp4, the training data consists of two separate point clouds which both are detected and overlaid in the video. The video files were downsized and compressed to meet the 100Mb limitation of the submission file; we apologize for the decreased visibility.
GameStop_015_compressed.mp4 GameStop_018_compressed.mp4 GameStop_023_compressed.mp4	Videos taken in retail stores of video games, the branches of the GameStop company (they are used with the kind permission of GameStop®). The overlay consists of bounding boxes on the detected objects, and of the database (studio) images associated with the detections (in the lower left corner of each bounding box). The number in the upper left corner of each bounding box represents the score produce by our method (the higher the score, the greater is the certainty of the detection).

	The displayed results were produced by the full pipeline of the proposed method – stages of detection, classification and temporal tracking.
Retail121_1_compressed.mp4 Retail121_2_compressed.mp4	Videos of our in-house shelf of retail supermarket products, displaying detections of the products in the same fashion as was done for the GameStop videos (see the row above).

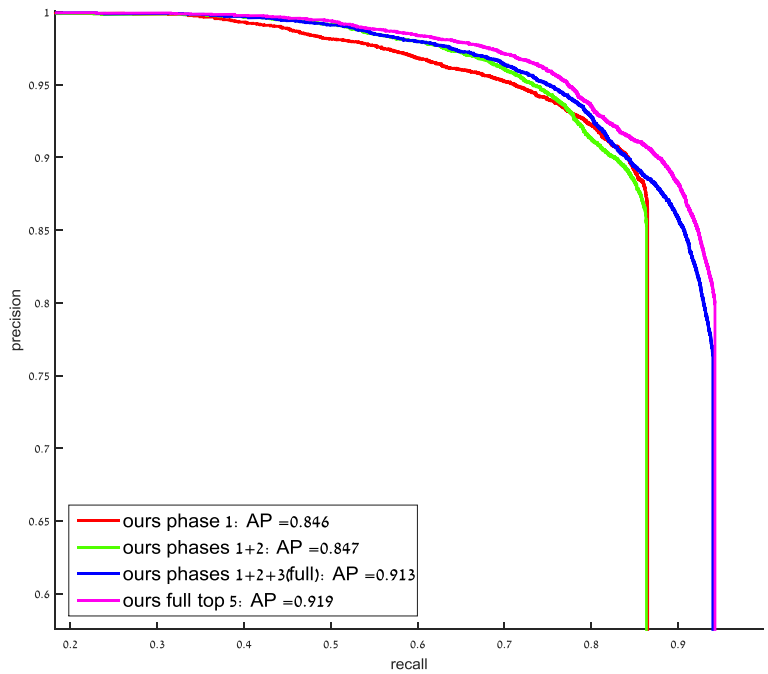
Precision-Recall curves for detection results in test datasets

We present precision-recall curves for four test datasets, which corresponding Average Precision values (areas under the curves) are reported in Table 1 of the submitted paper and in the figure legend.

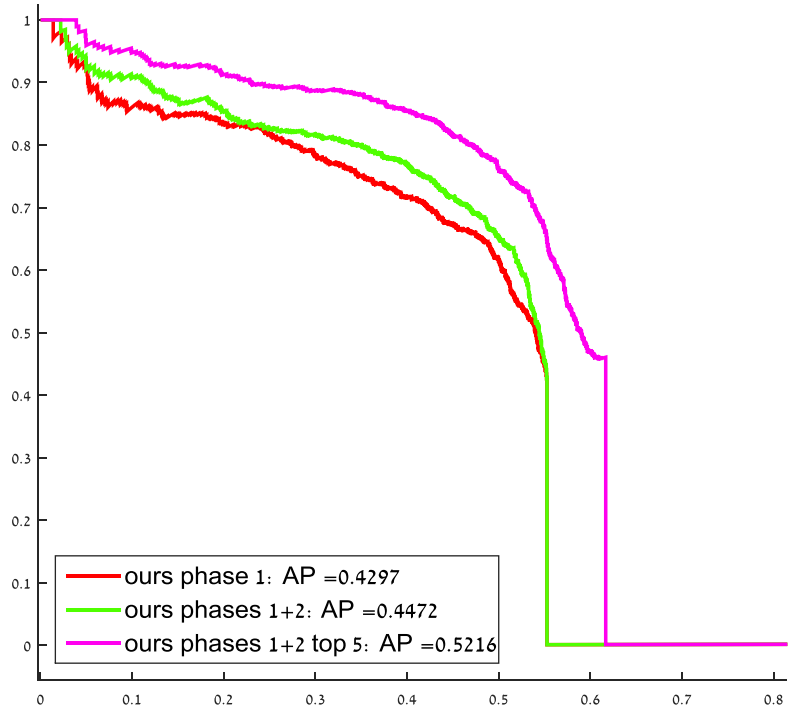
1. **GameStop dataset:**



2. **Retail121 dataset:**

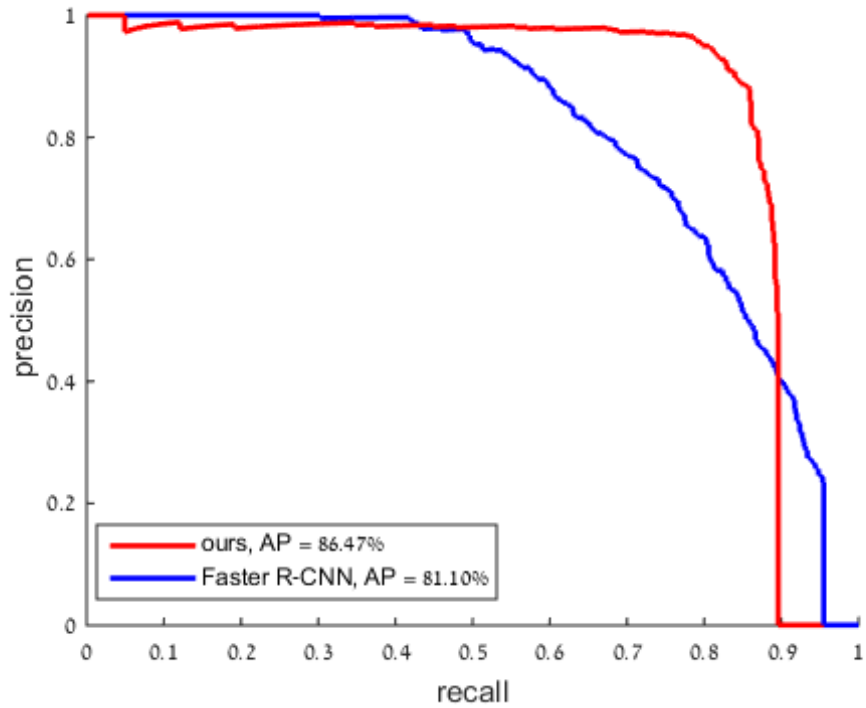


3. **Grozi-3.2K dataset:**

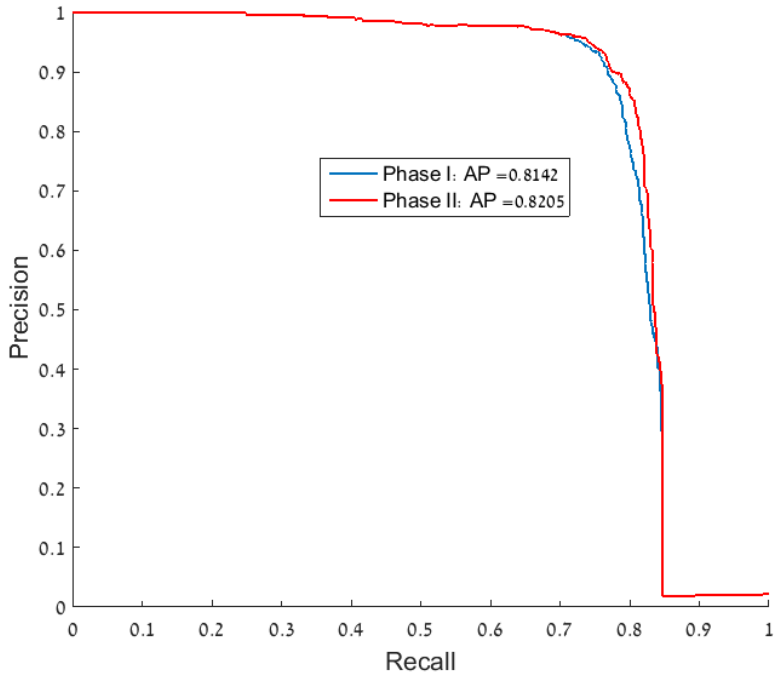


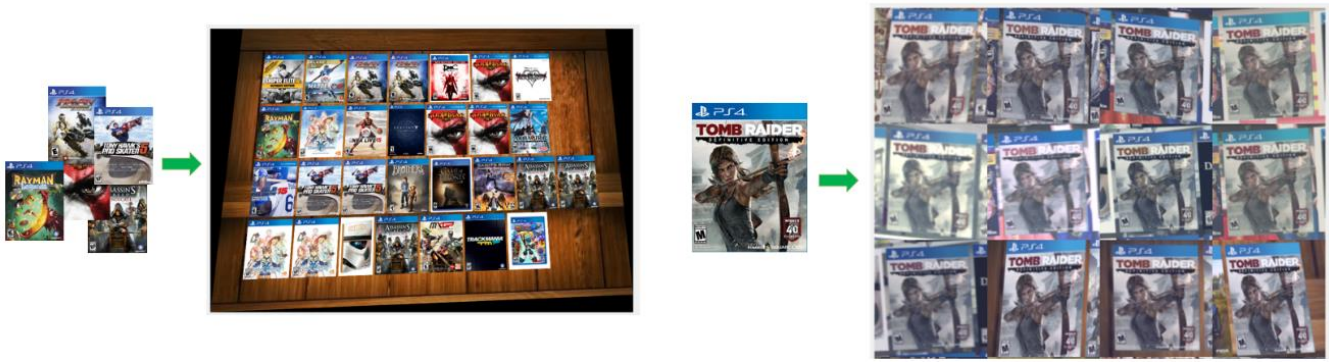
4. Precision-Recall curves for the set of large 27 retail categories in Grozi-3.2K dataset (see the discussion in the paper on lines 534-568). The red curve is produced by our method,

while the blue one is the result of applying the Faster-RCNN.



5. Precision-Recall graphs for FlickrLogos-32 dataset performance of our method. Blue – phase 1, red – phase 1+2.





Synthesis process for data augmentation in DNN training. (i) build a lattice of product images on a random background; (ii) apply random homography; (iii) apply a random photometric transformation.

Appendix. Proof of the proposition in Equation (2) of the paper.

Claim:

Let $P(F_q^j|U,R^j) = Q(F_q^j|U)$ when $R^j = 1$ and $P(F_q^j|U,R^j) = Q(F_q^j)$ when $R^j = 0$, and assume that $P(R^j = 0)$ and $P(R^j = 1)$ are fixed (independent of j constant priors for any patch to be generated from the background or foreground respectively), and that $Q(F_q^j) \gg 0$, and that $P(R^j = 0) \gg P(R^j = 1)$ then:

$$\sum_j \log P(F_q^j|U) \approx const + const \cdot \sum_j \frac{Q(F_q^j|U)}{Q(F_q^j)}$$

Proof:

$$\log P(F_q^j|U) = \log[P(R^j = 0) \cdot P(F_q^j|U,R^j = 0) + P(R^j = 1) \cdot P(F_q^j|U,R^j = 1)] =$$
$$\log[P(R^j = 0) \cdot Q(F_q^j) + P(R^j = 1) \cdot Q(F_q^j|U)] =$$
$$\log\left[1 + \frac{P(R^j=1)\cdot Q(F_q^j|U)}{P(R^j=0)\cdot Q(F_q^j)}\right] + \log P(R^j = 0) + \log Q(F_q^j) \approx \qquad (*)$$
$$\frac{P(R^j=1)\cdot Q(F_q^j|U)}{P(R^j=0)\cdot Q(F_q^j)} + \log P(R^j = 0) + \log Q(F_q^j) = \qquad (**)$$
$$const \cdot \frac{Q(F_q^j|U)}{Q(F_q^j)} + const$$

Here (*) follows from the fact that we have assumed that $P(R^j = 0) \gg P(R^j = 1)$ (i.e. the background model is “richer” and it is much more likely any patch is generated from

it) and that $Q\left(F_q^j\right) \gg 0$ (i.e. any patch has a "significant" non-zero probability to be generated from the background) and hence $\frac{P\left(R^j=1\right)\cdot Q\left(F_q^j \middle| U\right)}{P\left(R^j=0\right)\cdot Q\left(F_q^j\right)} = \varepsilon \ll 1$ and since $\log(1+\varepsilon) \approx \varepsilon$ we have (*). **<u>Note:</u>** it is also reasonable to assume that $Q\left(F_q^j \middle| U\right)$ is smaller then $Q\left(F_q^j\right)$ also supporting the claim above. Finally, (**) follows from $P\left(R^j = 0\right)$ and $P\left(R^j = 1\right)$ being fixed and $\log Q\left(F_q^j\right)$ being independent of the assignment to the unobserved variables U (i.e. constant with respect to the given query image and its given set of sampled descriptors $\left\{F_q^j\right\}$).