Multi-View Background Subtraction for Object Detection

UCIRVINE

Raúl Díaz, Sam Hallman, Charless C. Fowlkes
School of Information and Computer Sciences
The Henry Samueli School of Engineering

Sponsored by The Balsells Fellowship Program

Introduction

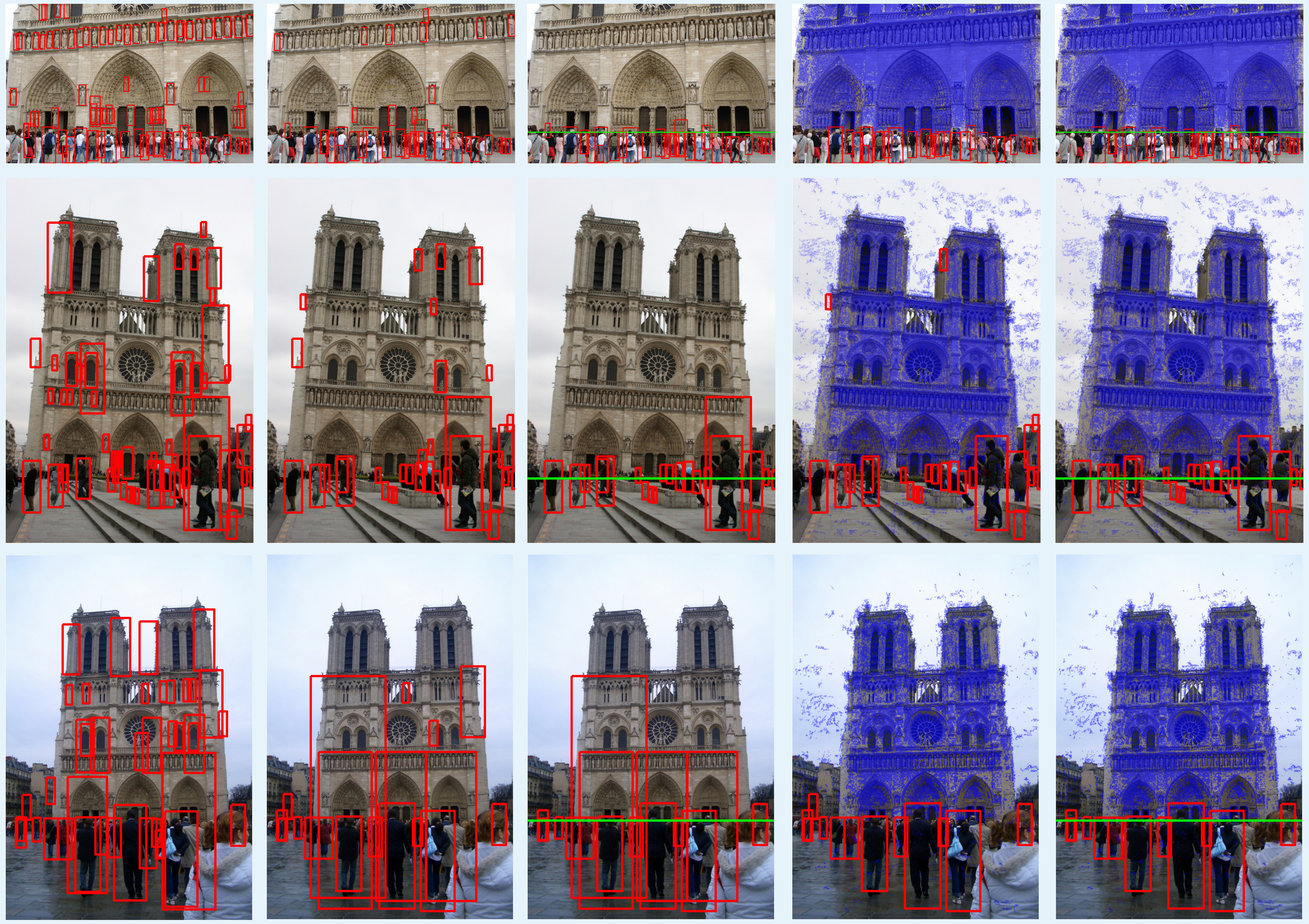
The confluence of robust algorithms for structure from motion (SfM) and multi-view stereo (MVS) suggests that it will soon be feasible to accurately estimate camera pose of pictures taken in outdoor, urban environments. **How can we exploit camera localization and large sets of photographs available online to better understand the contents of a particular scene?** It is useful to divide scene components into static, rigid background (buildings, streets...) and dynamic objects (people, bikes, cars...).

SfM and MVS are useful techniques to build up an explicit model of the static background geometry and appearance. We want to investigate how such information can be used to improve the detection of dynamic objects like pedestrians and cars. We evaluate these ideas using a dataset of tourist photos with estimated camera pose.

Methodology

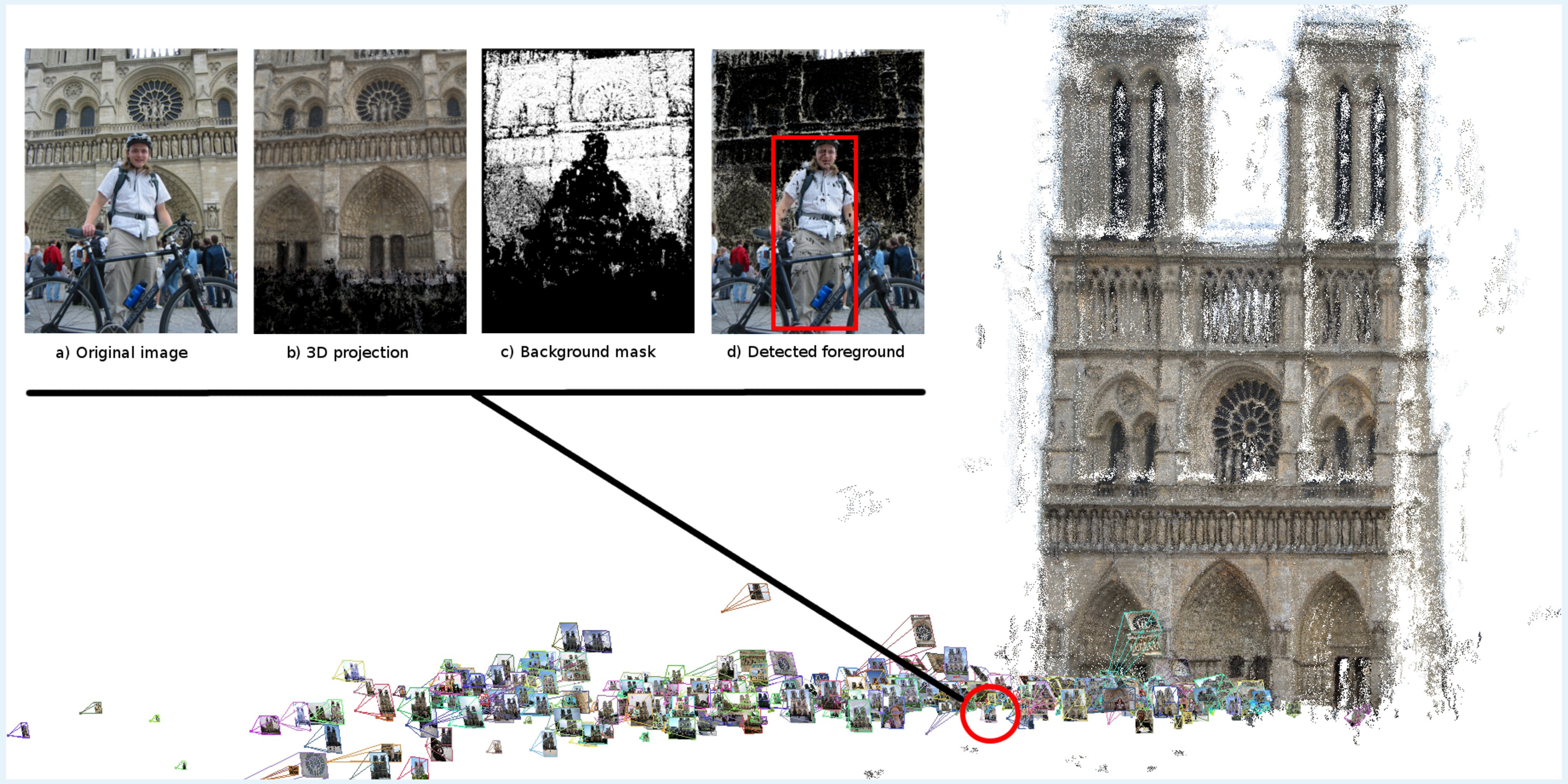
Generate 3D static scene model using internet photo collections and SfM + MVS. Use it for two approaches:

1.**Scene-specific detectors** utilize a stronger model of background statistics to improve accuracy
 ➔**Unsupervised hard-negative mining** via internet photo collections
 ➔Only need approximate camera localization at test time
2.**Multi-View Background Subtraction** builds a view-specific background model from nearby images
 ➔Estimate background mask via multi-view strereo matching
 ➔Suppress false-positive detections by masking out the estimated background from model



a) DT baseline b) DT+SS^- c) DT+SS^-+GC d) DT+SS^-+MVBS e) DT+SS^-+MVBS+GC

Figure 1: example detector outputs at 50% recall. Scene-specific detectors (SS^-) ignore background clutter (e.g. statues) while MVBS and geometric consistency (GC) can prune additional false positives.



a) Original image b) 3D projection c) Background mask d) Detected foreground

Figure 2: MVS matching estimates which pixels from image (a) belong to static background (b) extracted from other images from the same scene. (c) shows patches for which match scores are above a given threshold, while (d) shows a sample detection after removing background areas. SfM + MVS reconstruction builds a model of the static background which can be used along with camera pose estimation to improve object detection in a novel test image.

Results

Scene-specific information and geometric concistency can boost the baseline average precision up to 50% using our **unsupervised negative-mining**, improving Dalal-Triggs (DT) from 0.30 to 0.40 and Deformable Part Models (DPM) from 0.46 to 0.55. Using **fully-supervised** scene-specific negatives yielded an AP of 0.41 for DT and 0.55 for DPM, suggesting that our unsupervised negative mining is capturing most of the useful negative examples. Additionally, estimating the horizon line from SfM significantly improves the performance of related work *Putting Objects in Perspective* (PoP).

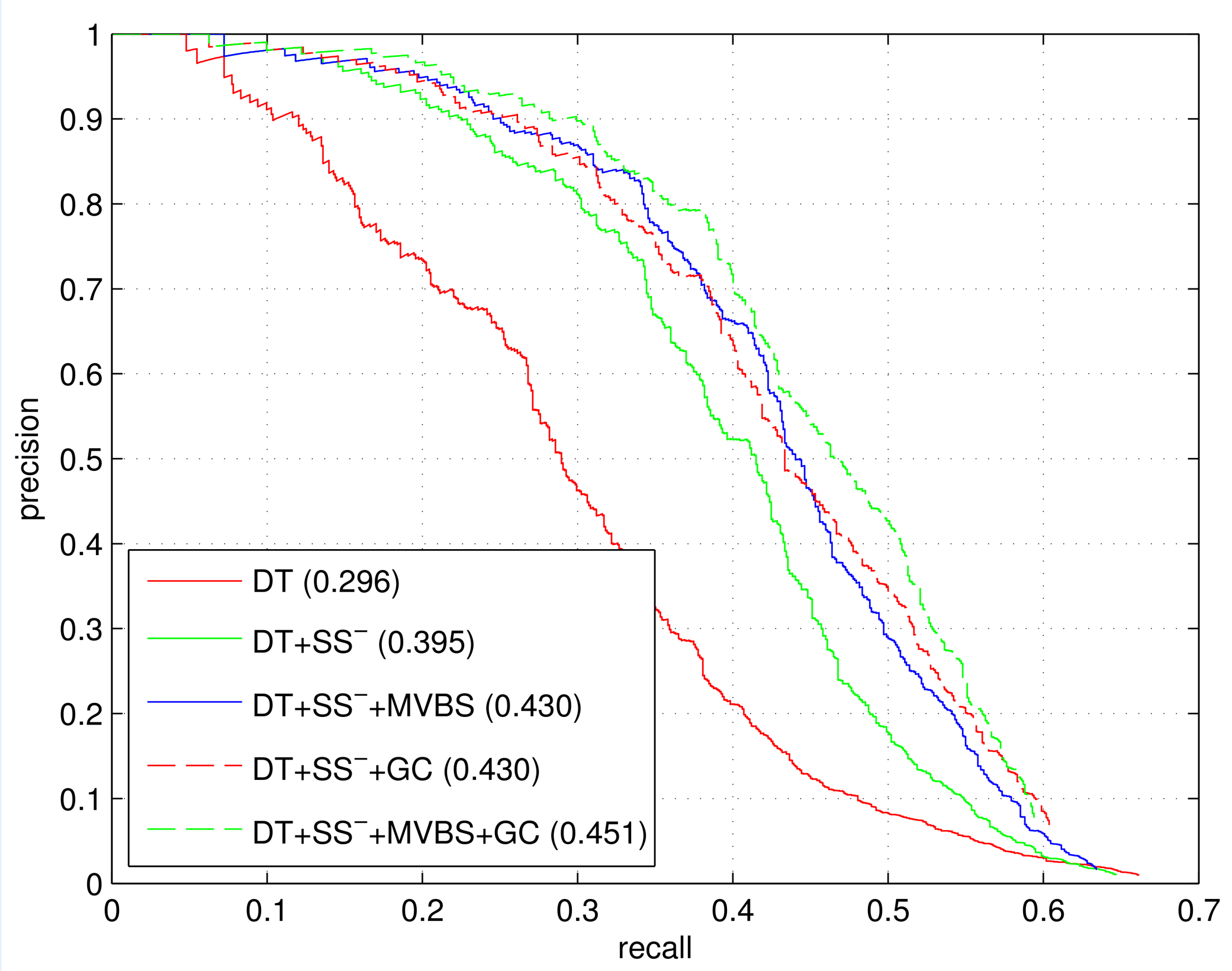


Figure 3: precision/recall results for DT

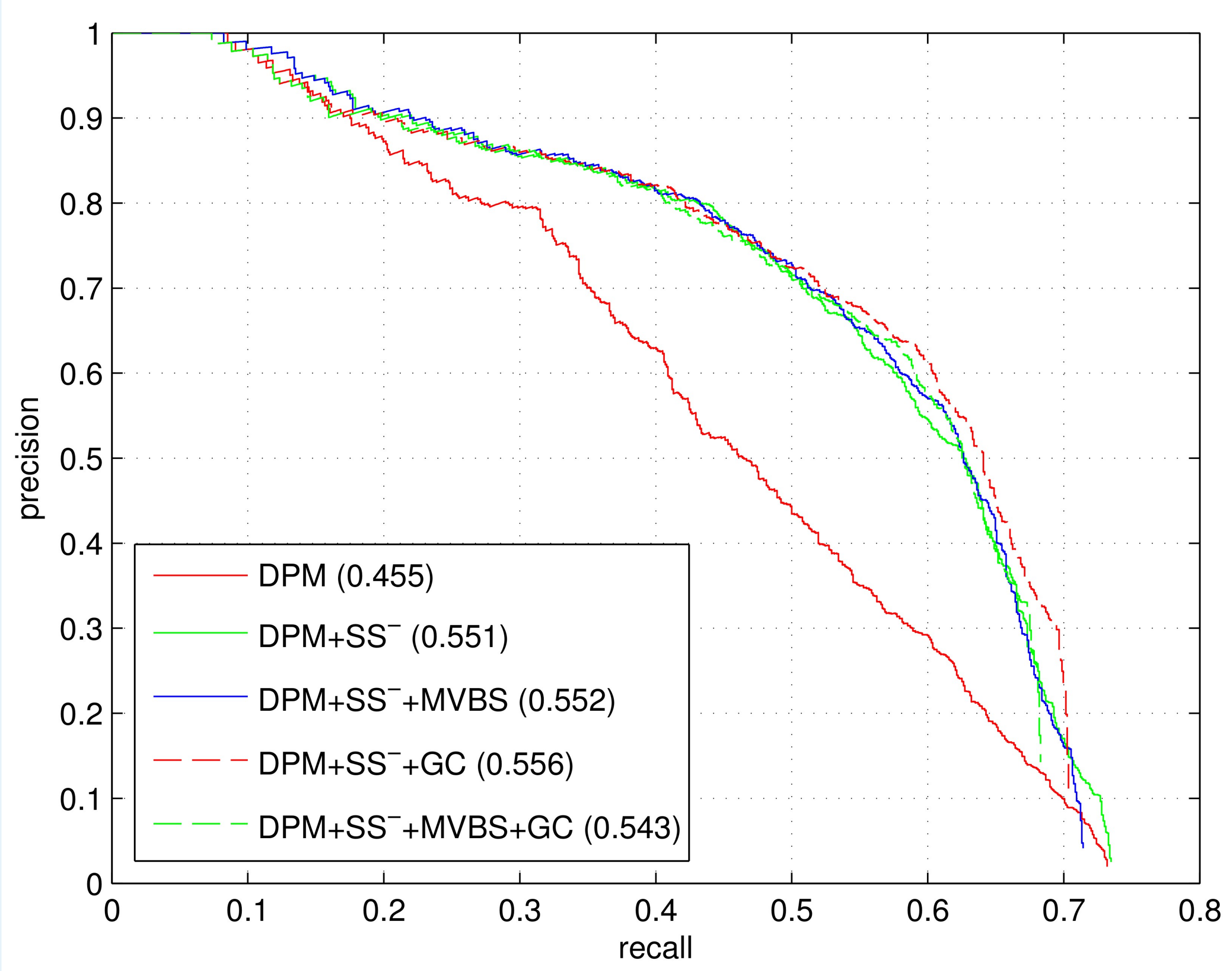


Figure 4: precision/recall results for DPM

	DT	DT+SS^-	DPM	DPM+SS^-
Detection	0.296	0.395	0.455	0.551
+MVBS	0.412	0.430	0.558	0.552
PoP	0.323	0.322	0.348	0.323
PoP+SfM	0.405	0.406	0.404	0.337

Table 1: average precision results for DT and DPM. +SS^- indicates scene-specific negatives, while +MVBS includes background pruning. +SfM substitutes the estimation of horizon line in PoP by the estimation in SfM assuming true ground is horizontal.

References

[1] Hoiem *et al*. Putting Objects in Perspective. CVPR, 2006.
[2] Agarwal *et al*. Building Rome in a day. ICCV, 2009.
[3] Furukawa *et al*. Towards Internet-scale multi-view stereo. CVPR, 2010.
[4] Dalal *et al*. Histograms of oriented gradients for human detection. CVPR, 2005
[5] Felzenszwalb *et al*. Object detection with discriminatively trained part-based models. TPAMI, 2010