3D Reconstruction for Geo-aware Tasks

Presented by Kevin Matzen

Street View Motion-from-Structure-from-Motion

Klingner, Martin, Roseborough (Google), ICCV 2013

Detecting Dynamic Objects with Multi-View Background Subtraction

Diaz, Hallman, Fowlkes (UC Irvine), ICCV 2013

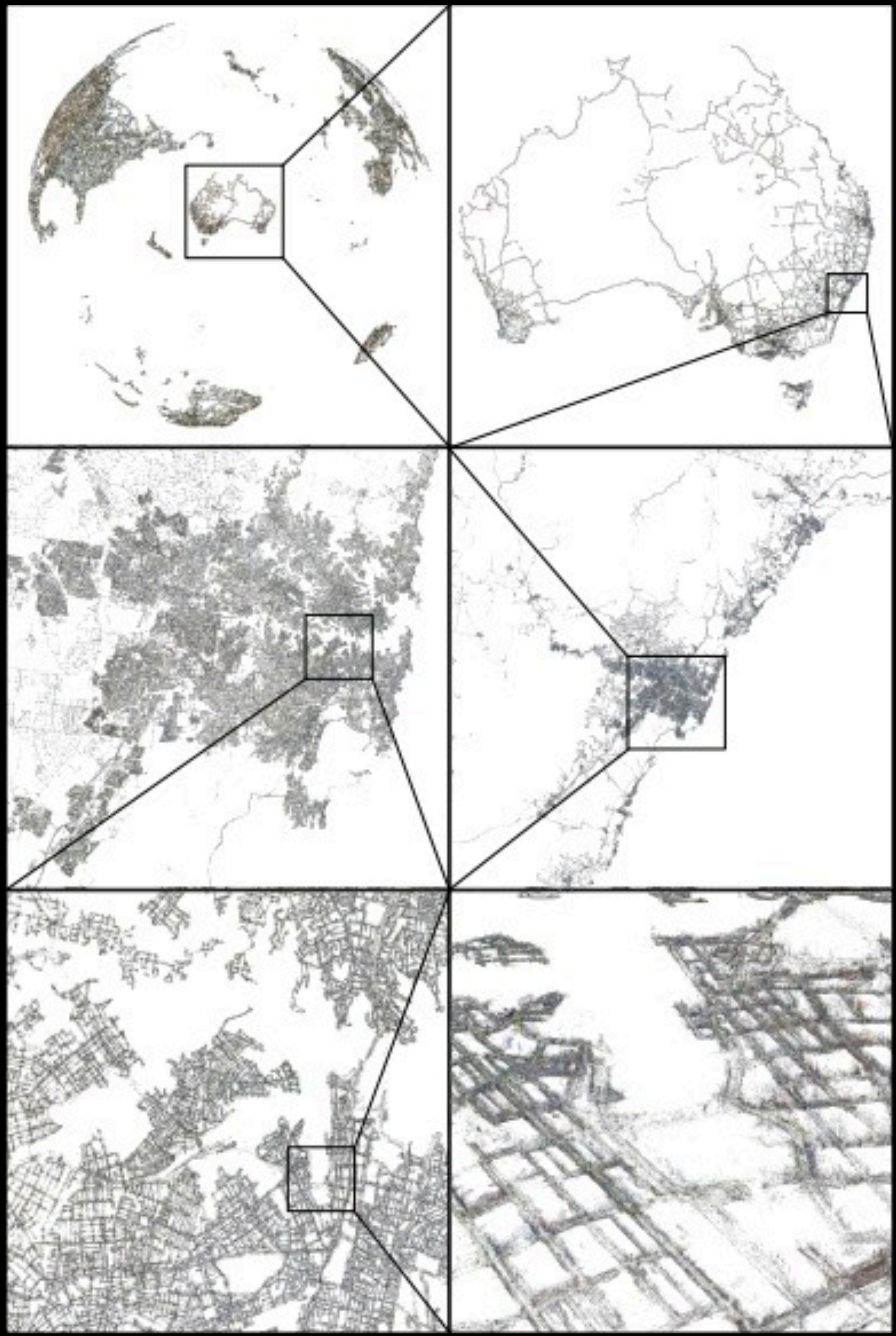
Street View Motion-from-Structure-from-Motion

Klingner, Martin, Roseborough (Google), ICCV 2013

Detecting Dynamic Objects with Multi-View Background Subtraction

Diaz, Hallman, Fowlkes (UC Irvine), ICCV 2013

Goal: Build a planet-scale 3D reconstruction from Street View images.



Reconstruction Basics

- Feature generation (SIFT, SURF, HOG)
- Feature matching/tracking
- Camera pose initialization
- Triangulation
- Bundle adjustment

Core Contributions

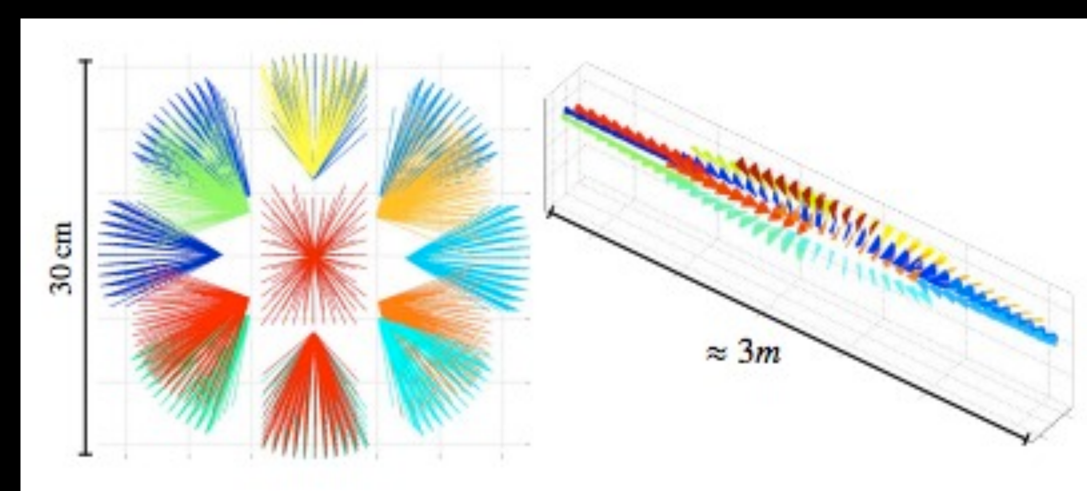
- Generalized camera model - supports rigidly attached cameras; rolling shutter
- Scalable track generation method
- Bundle adjustment procedure for generalized camera model

Generalized camera model

- Many rigidly attached cameras
- Rolling shutter

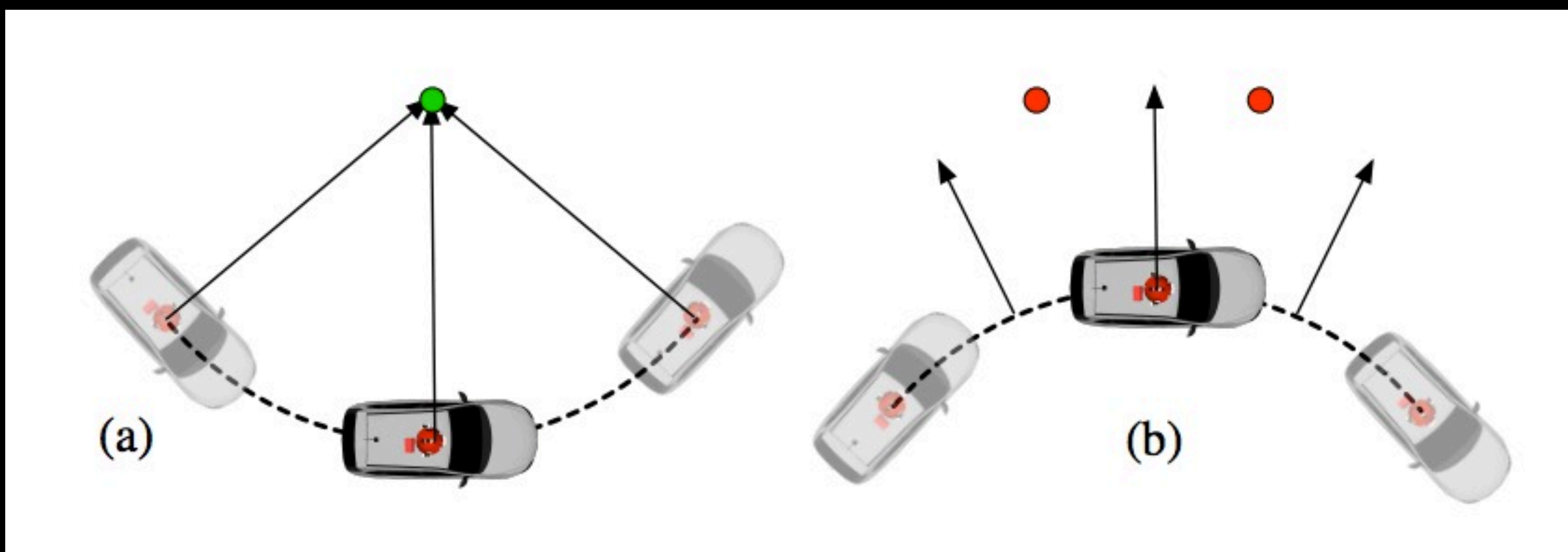
$${}_{\mathrm{im}}T_{\mathrm{w}}(t) = \overbrace{{}_{\mathrm{im}}T_{\mathrm{c}}}^{\text{Lens Model}} \cdot \underbrace{{}_{\mathrm{c}}T_{\mathrm{r}} \cdot \overbrace{{}_{\mathrm{r}}T_{\mathrm{w}}(t)}^{\text{Rosette Pose}}}_{\text{Camera Pose}}$$

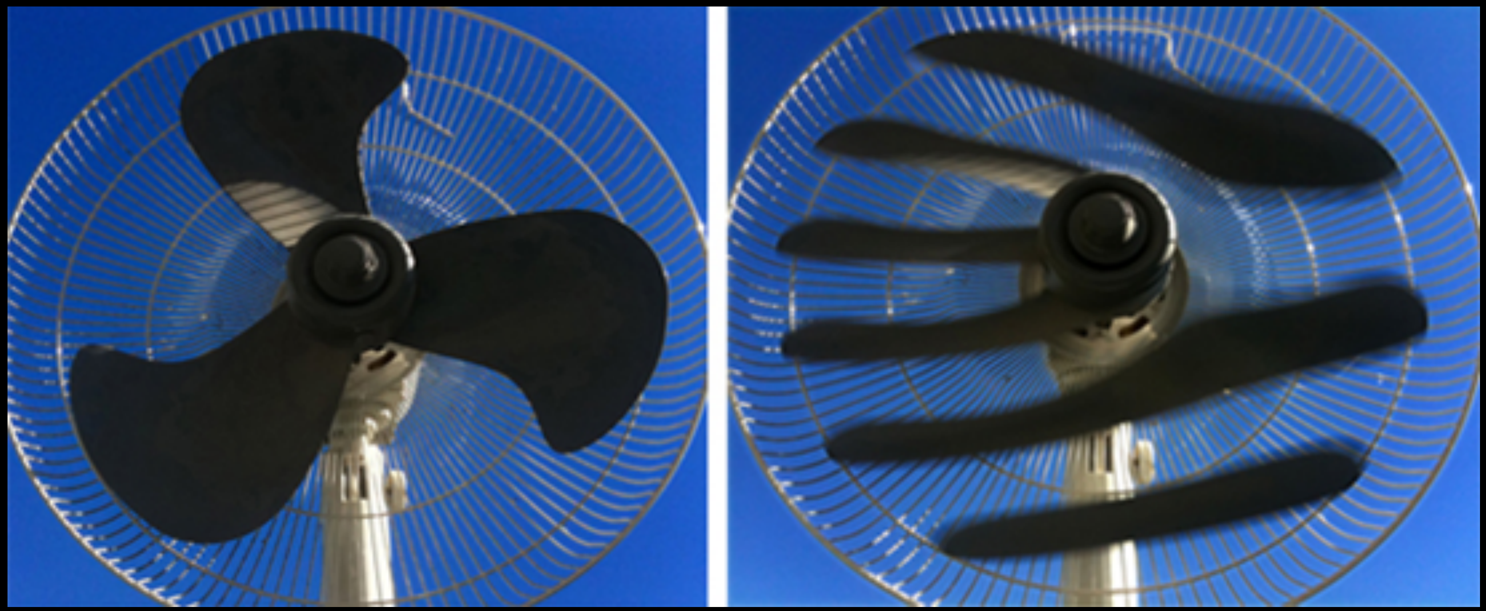
$$x_{\mathrm{w}} = {}_{\mathrm{w}}T_{\mathrm{r}}(t(x_{\mathrm{im}})) \cdot {}_{\mathrm{r}}T_{\mathrm{c}} \cdot {}_{\mathrm{c}}T_{\mathrm{im}} \cdot x_{\mathrm{im}}$$



$$x_{\mathrm{im}} = {}_{\mathrm{im}}\mathsf{T}_{\mathrm{c}} \cdot {}_{\mathrm{c}}\mathsf{T}_{\mathrm{r}} \cdot {}_{\mathrm{r}}\mathsf{T}_{\mathrm{w}}(t(x_{\mathrm{im}})) \cdot x_{\mathrm{w}}$$

Permits many or no solutions





http://www.digitalbolex.com/global-shutter/

Related Work

- Ait-Aider, et. al., Structure and kinematics triangulation with a rolling shutter stereo rig, ICCV 2009.
- Baker, et. al., Removing Rolling Shutter Wobble, CVPR 2010.
- Hedborg, et. al., Rolling Shutter Bundle Adjustment, CVPR 2012.

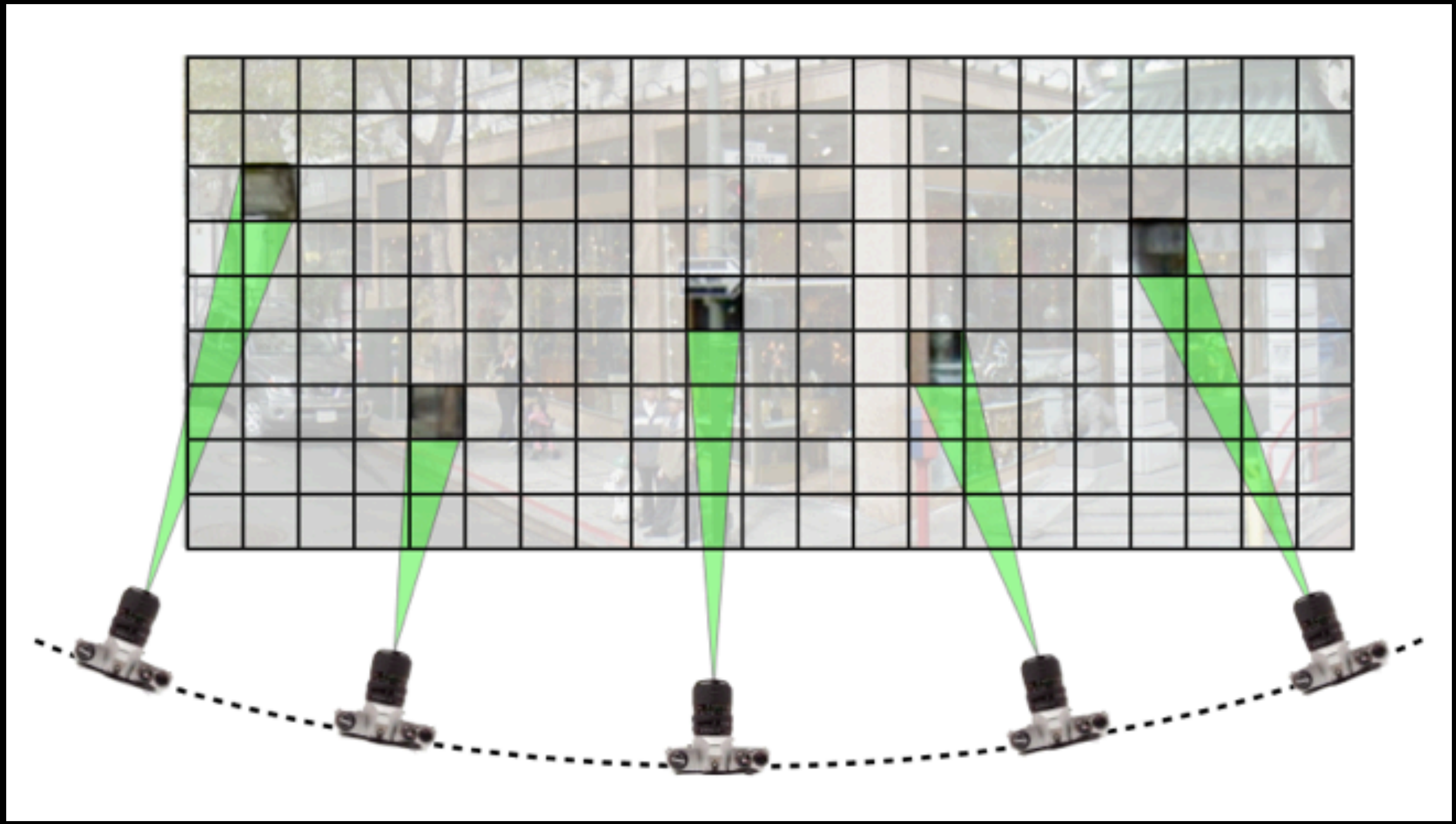
Triangulation

Observation: rolling shutters are fast

Time for true scanline $\longrightarrow t(x_{\mathrm{im}}) \approx t(\hat{x}_{\mathrm{im}}) \longleftarrow$ Time for estimated scanline

(when the reprojection error is small)

$$\underset{x_{\mathrm{w}}}{\operatorname{argmin}} \sum_{\text{views}} \|x_{\mathrm{im}} - \hat{x}_{\mathrm{im}}\|^2 \longrightarrow \underset{x_{\mathrm{w}}}{\operatorname{argmin}} \sum_{\text{views}} \| \underbrace{\overbrace{{}_{\mathrm{im}}\mathrm{T}_{\mathrm{c}}}^{\text{Lens}} \cdot \overbrace{{}_{\mathrm{c}}\mathrm{T}_{\mathrm{r}} \cdot {}_{\mathrm{r}}\mathrm{T}_{\mathrm{w}}(t(\hat{x}_{\mathrm{im}}))}^{\text{Camera Pose}} \cdot x_{\mathrm{w}}}_{x_{\mathrm{im}}} - \hat{x}_{\mathrm{im}}\|^2$$



$$x_{\mathrm{f}} = \underbrace{{}_{\mathrm{f}}\top_{\mathrm{c}} \cdot {}_{\mathrm{c}}\top_{\mathrm{r}} \cdot {}_{\mathrm{r}}\top_{\mathrm{w}}(t(\hat{x}_{\mathrm{im}}))}_{{}_{\mathrm{f}}\top_{\mathrm{w}}} \cdot x_{\mathrm{w}}$$

$$x_{\mathrm{f}} = {}_{\mathrm{f}}\top_{\mathrm{c}} \cdot {}_{\mathrm{c}}\top_{\mathrm{r}} \cdot \overbrace{{}_{\mathrm{r}}\top_{\mathrm{n}}(t(\hat{x}_{\mathrm{im}}) - t_n) \cdot {}_{\mathrm{n}}\top_{\mathrm{w}}(t_n)}^{{}_{\mathrm{r}}\top_{\mathrm{w}}(t(\hat{x}_{\mathrm{im}}))} \cdot x_{\mathrm{w}}$$

$$x_{\mathrm{f}} = \overbrace{{}_{\mathrm{f}}\top_{\mathrm{n}}(t(\hat{x}_{\mathrm{im}}) - t_n)}^{\text{Feature Camera}} \cdot \overbrace{{}_{\mathrm{n}}\top_{\mathrm{w}}(t_n)}^{\text{Rosette Pose}} \cdot x_{\mathrm{w}}$$

Generalized Bundle Adjustment

$$\underset{\{x_{\mathrm{w}},\ {}_{\mathrm{n}}\mathrm{T}_{\mathrm{w}}(t_n)\}}{\operatorname{argmin}} \sum_{\text{points}} \sum_{\text{views}} \|x_{\mathrm{f}}\|^2$$

$$x_{\mathrm{f}} = \overbrace{{}_{\mathrm{f}}\mathrm{T}_{\mathrm{n}}(t(\hat{x}_{\mathrm{im}}) - t_n)}^{\text{Feature Camera}} \cdot \overbrace{{}_{\mathrm{n}}\mathrm{T}_{\mathrm{w}}(t_n)}^{\text{Rosette Pose}} \cdot x_{\mathrm{w}}$$

High frequency - provided by IMU

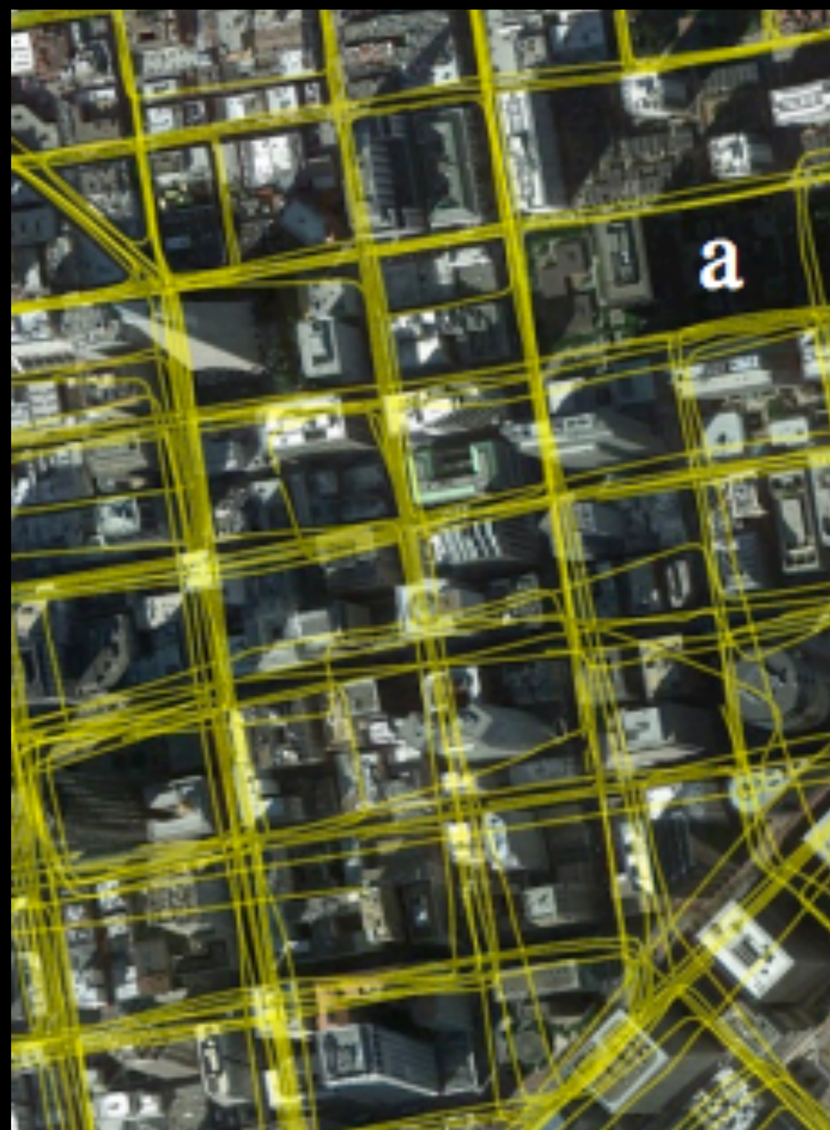
Mid frequency - bundle adjusted

Initial Pose Generation

- Accelerometers
- Gyroscopes
- Wheel encoders
- GPS receivers

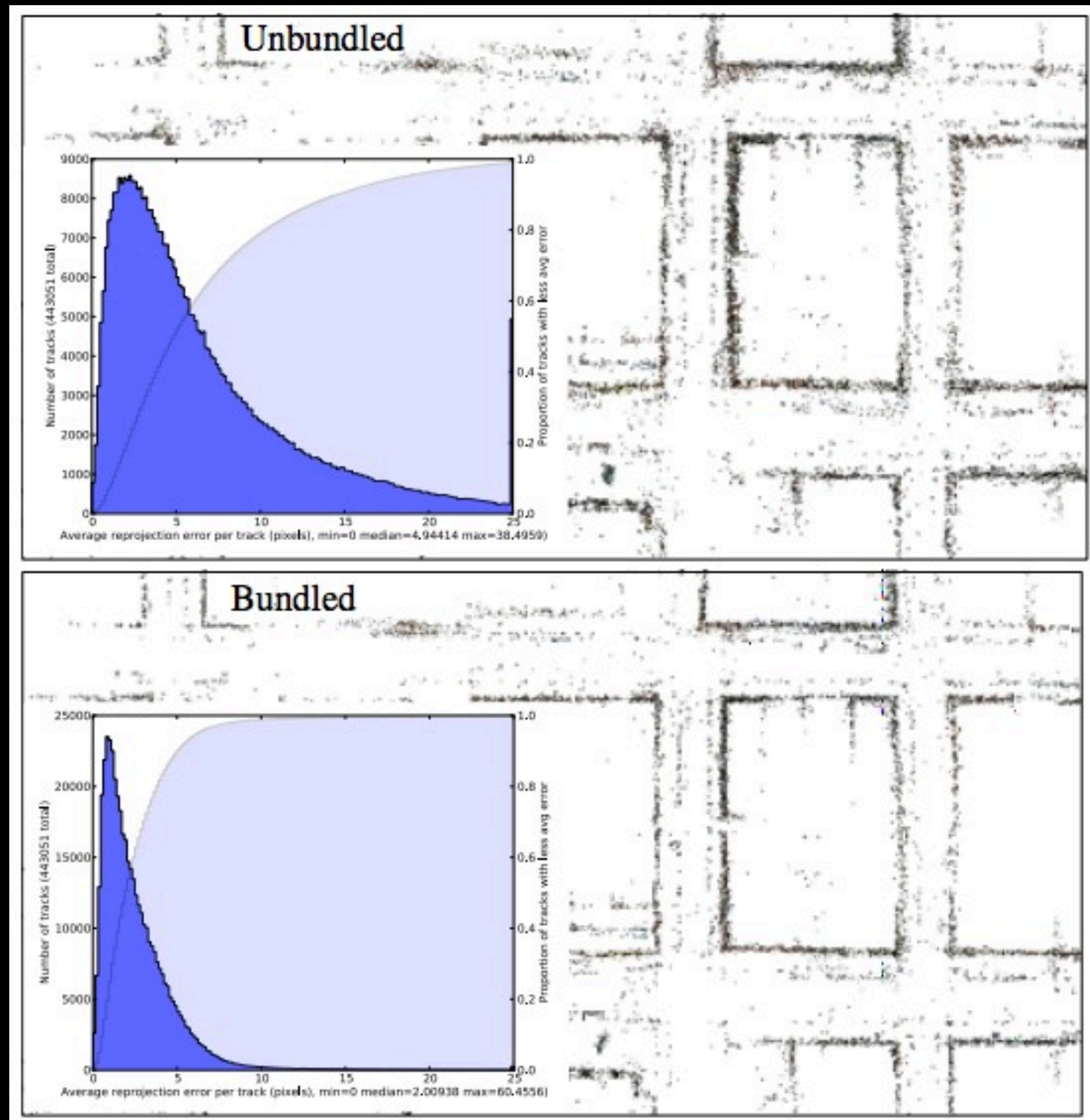
Track Generation

- Track rather than match - O(n) versus O(n^2)
- Strict bidirectional 1-NN match criterion
- Retain feature descriptors for later loop closure



Before

After



Unbundled
Number of tracks (443051 total)
Proportion of tracks with less avg error
Average reprojection error per track (pixels), min=0 median=4.94414 max=38.4959)
Bundled
Number of tracks (443051 total)
Proportion of tracks with less avg error
Average reprojection error per track (pixels), min=0 median=2.00938 max=60.4556)

Street View Motion-from-Structure-from-Motion

Klingner, Martin, Roseborough (Google), ICCV 2013

Detecting Dynamic Objects with Multi-View Background Subtraction

Diaz, Hallman, Fowlkes (UC Irvine), ICCV 2013

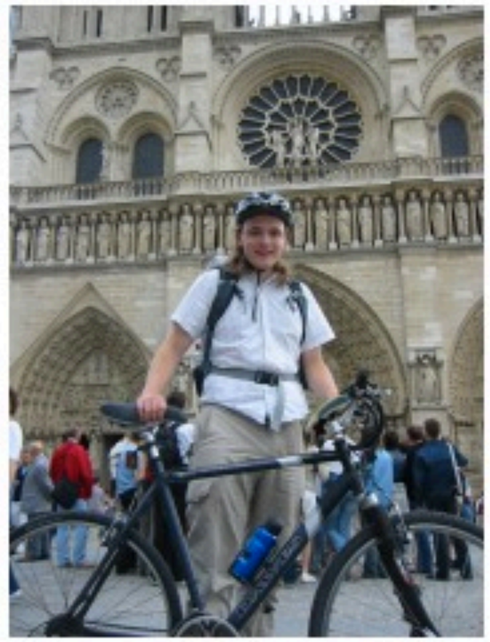
Goal: Use 3D reconstructions with appearance information to construct scene-specific object detectors.

Contributions

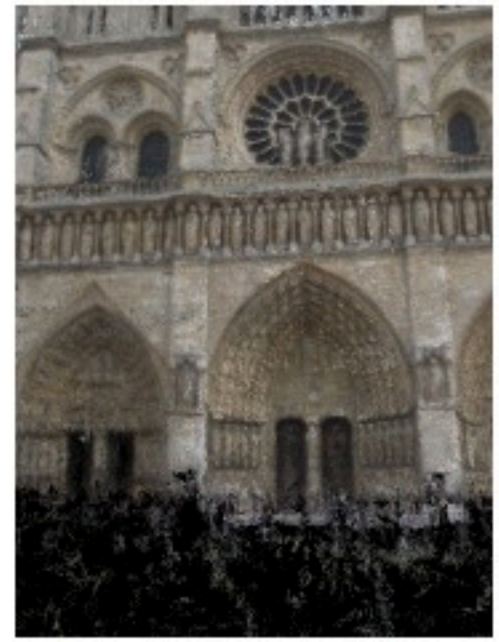
(1) Problem: Expensive to manually label training data for each location.

Idea: Use reconstruction to detect background and generate scene-specific negatives.

(2) Project reconstruction into test images to detect foreground.



(a) input image



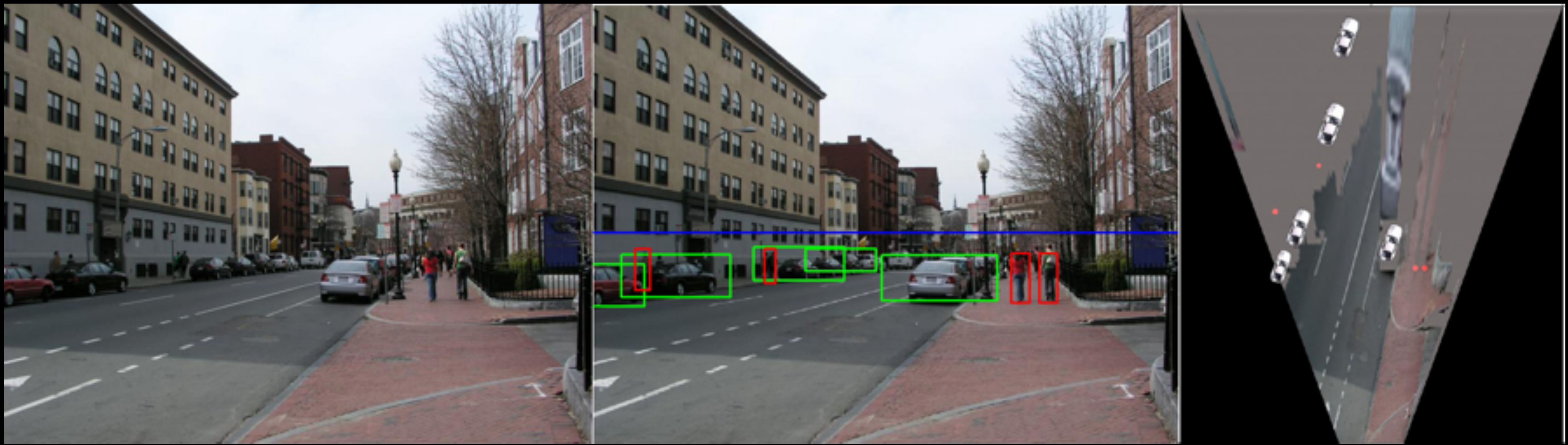
(b) scene reconstruction

(c) background mask

(d) detected foreground

Related Work

D. Hoiem, A.A. Efros, and M. Hebert, "Geometric Context from a Single Image", ICCV 2005.



D. Hoiem, A.A. Efros, and M. Hebert, "Putting Objects in Perspective", CVPR 2006.

Reconstruction Pipeline

- Sparse scene structure and camera calibration (Bundler) [Snavely 2006]
- Camera clustering (CMVS) [Furukawa 2010]
- Dense reconstruction (PMVS) [Furukawa 2007]

Background Detection

- Comparing a photo directly to a model is messy
- Instead, compare to photos used to generate the model

$$match(p) = \frac{1}{|V(p)|} \sum_{J \in V(p)} h(p, I, J)$$

h(p, I, J) - NCC for 5x5 window

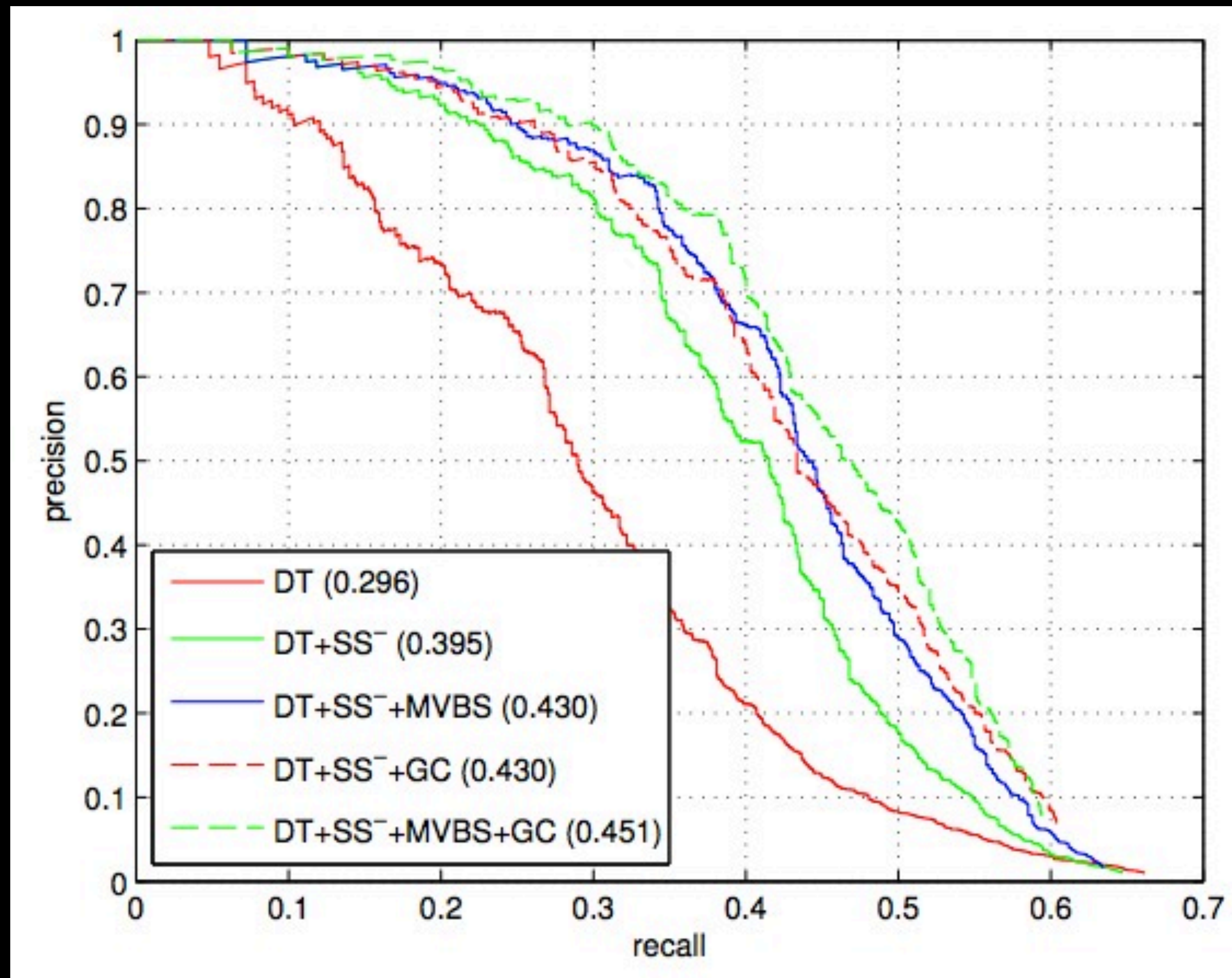
Threshold at 0.5 to generate binary mask

Background-Aware Training

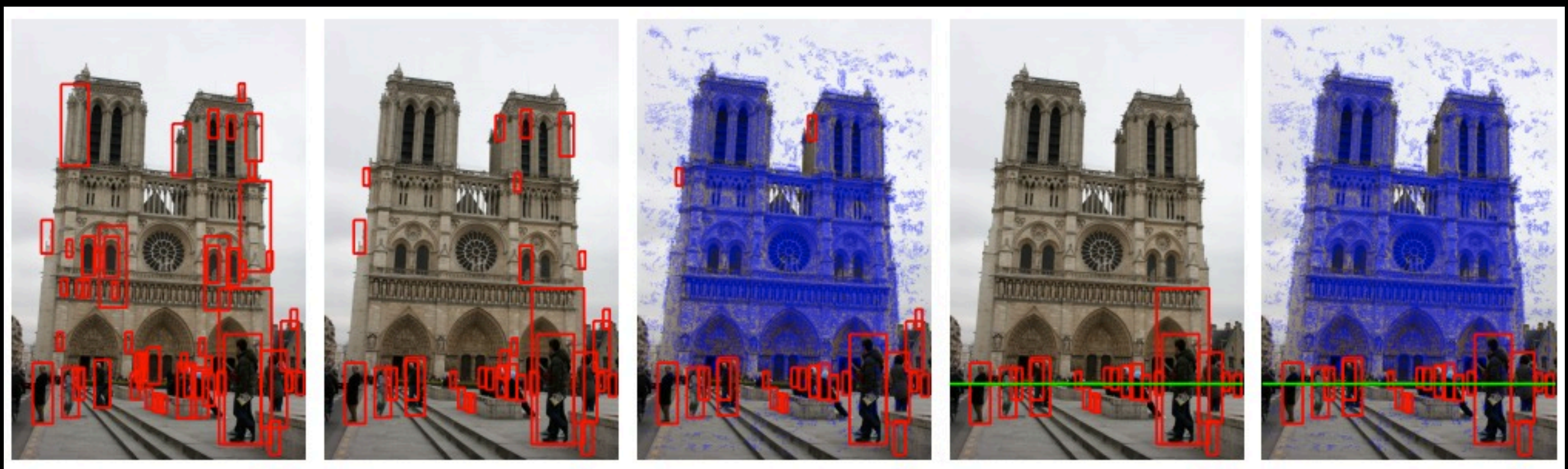
- Train generic detector
- Perform hard negative mining [Dalal and Triggs 2005] on location-specific photos
    - Consider bounding boxes with percentage background pixels > 0.2 as negatives

Multi-View Background Subtraction

- Given a novel test image, improve detection results using background mask
- Tried GrabCut, super-pixels, comparing average shape masks
- Rejecting background percentage > 0.2 worked well



DT - Dalal and Triggs 2005
SS- - Scene-specific negatives
MVBS - Background subtraction criterion
GC - Threshold by SfM horizon



DT DT+SS$^-$ DT+SS$^-$+MVBS DT+SS$^-$+GC DT+SS$^-$+MVBS+GC

	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 [13]	0.323	0.322	0.348	0.323
PoP+SfM	0.405	0.406	0.404	0.337

	DT+FS$^-$	DT+FS	DPM+FS$^-$	DPM+FS
Detection	0.41	0.43	0.55	0.63

Average Precision

DT - Dalal and Triggs 2005

SS- - Scene-specific negatives

DPM - Felzenswalb, et. al. 2008

PoP - Hoiem, et. al. 2006

SfM - SfM-based horizon prior

FS- - Fully supervised negatives

FS - Fully supervised pos+neg

Questions?