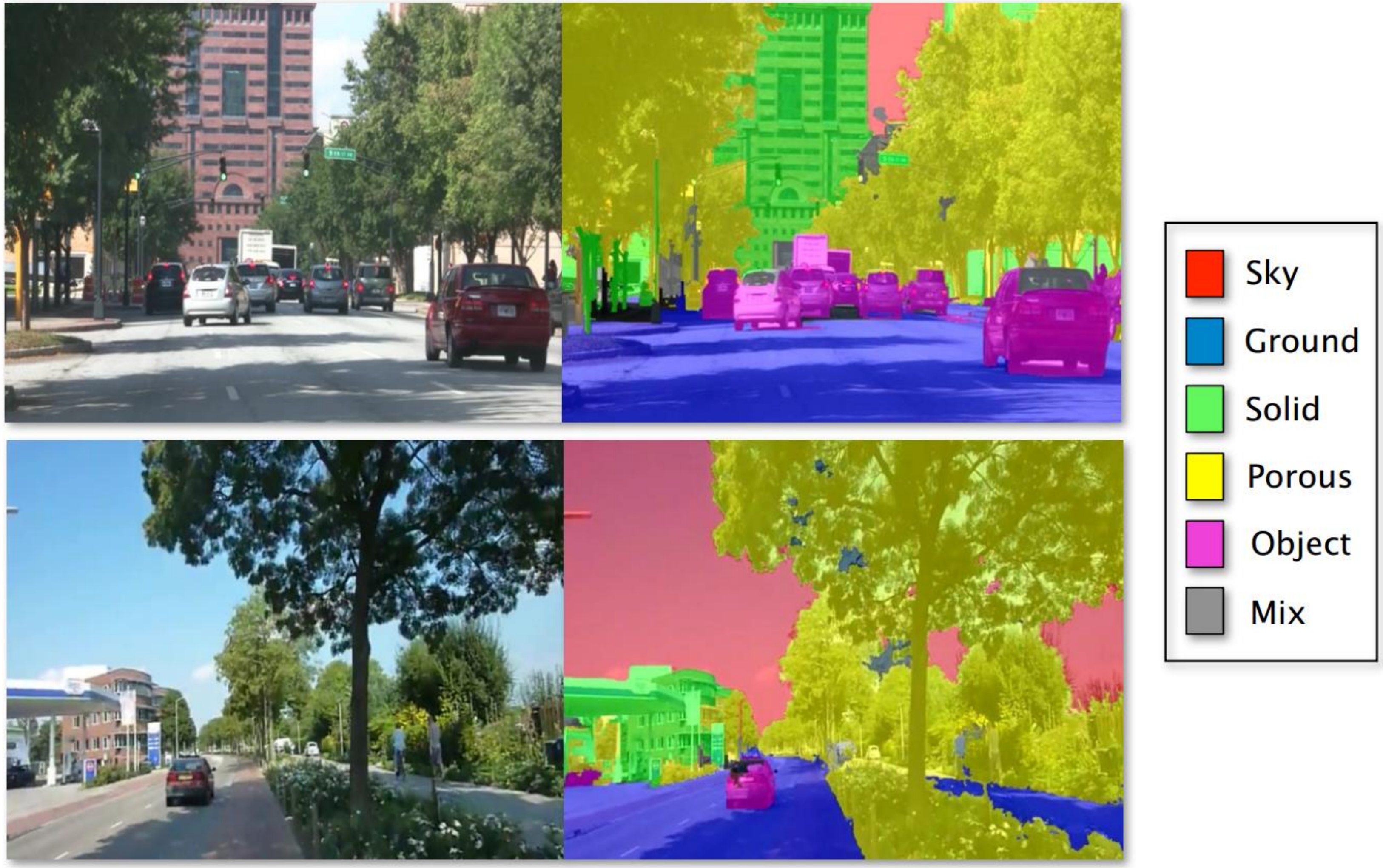
Geometric Context from Video

S. Hussain Raza and Matthias Grundmann and Irfan Essa

CVPR 2013

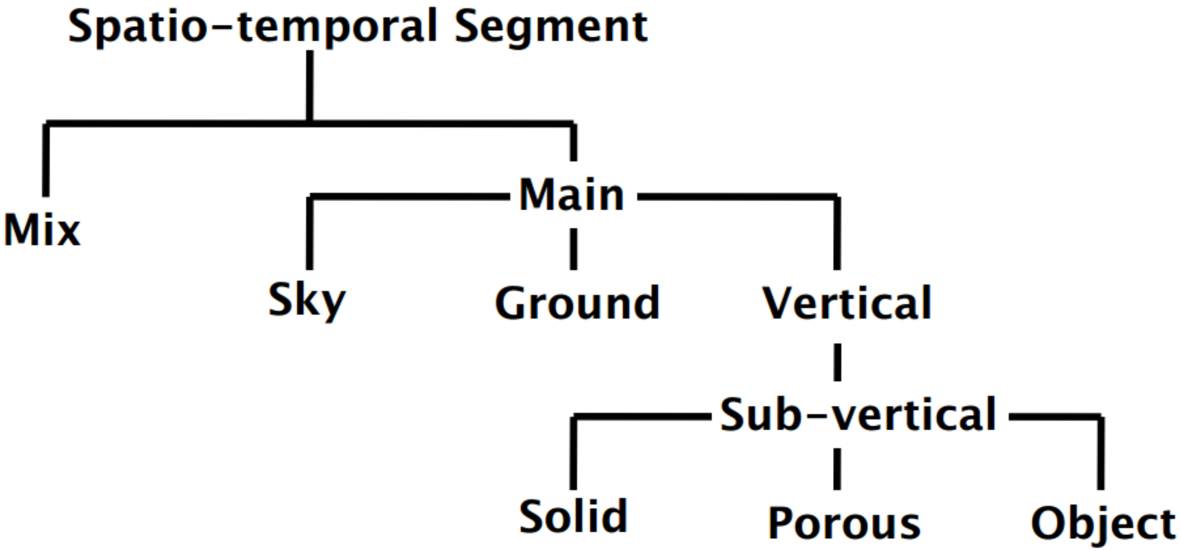
- Purpose: Estimating the broad 3D spatio-temporal structure of outdoor video scenes by labeling regions.



Dataset

- Total: 160 outdoor videos.
  - 100 pixel - level annotated videos (20K frames)
    - training & test
  - 60 unannotated videos (14K frames)
    - semi-supervised learning

Contents of Videos



Spatio-temporal Segment

Mix

Main

Sky

Ground

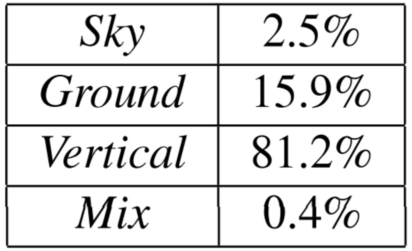
Vertical

Sub-vertical

Solid

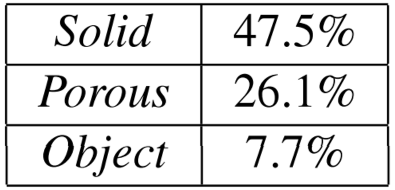
Porous

Object



Sky	2.5%
Ground	15.9%
Vertical	81.2%
Mix	0.4%

(a) Main Classes

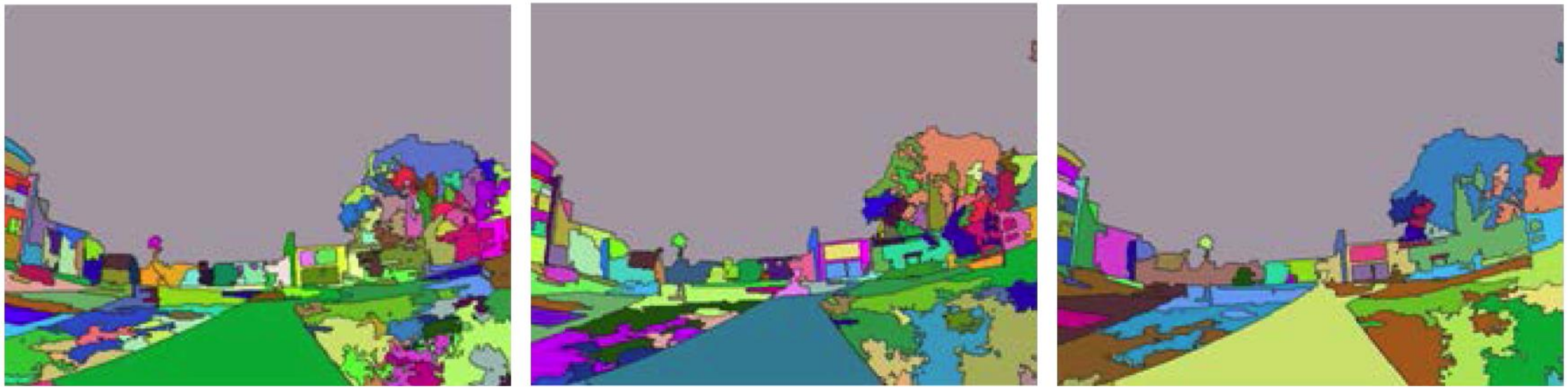


Solid	47.5%
Porous	26.1%
Object	7.7%

(b) Sub-vertical Classes

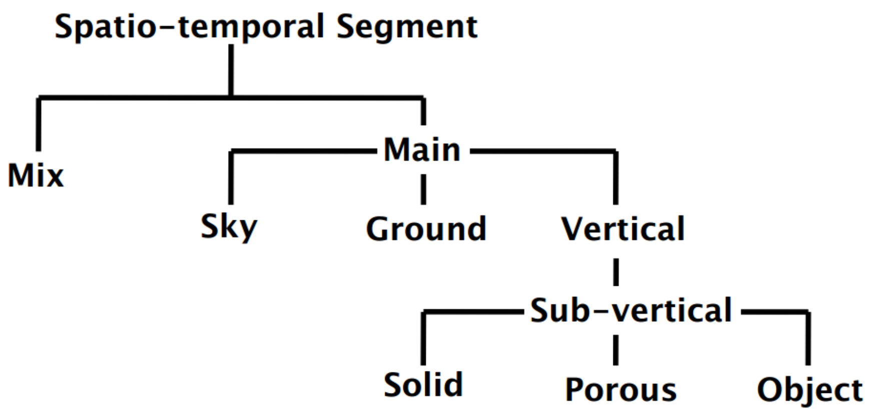
Video Segmentation

- Purpose: Group similar pixels into spatio-temporal regions that are coherent in both appearance and motion.
  - Method: *M. Grundmann, V. Kwatra, M. Han, and I. Essa. Efficient hierarchical graph-based video segmentation. In IEEE CVPR, 2010.*
    - Graph-based segmentation in spatio-temporal domain → Over-segmented video volume
    - Over-segmented video volume → Hierarchy of super-regions (based on a graph which is constructed using region descriptors)



Video Annotation

- Over-segmented regions (super-voxels) are labeled.
- Labels of regions in upper levels of hierarchy are determined via majority voting.



Spatio-temporal Segment

Mix

Main

Sky

Ground

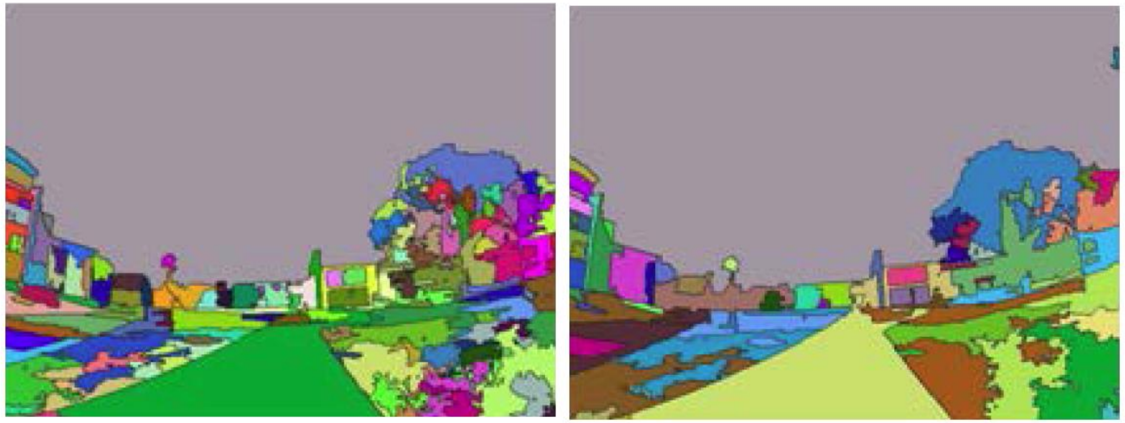
Vertical

Sub-vertical

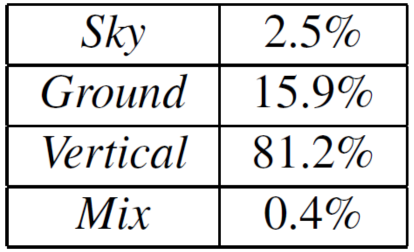
Solid

Porous

Object

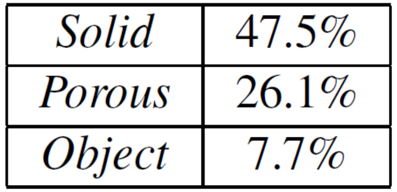


Video Annotation



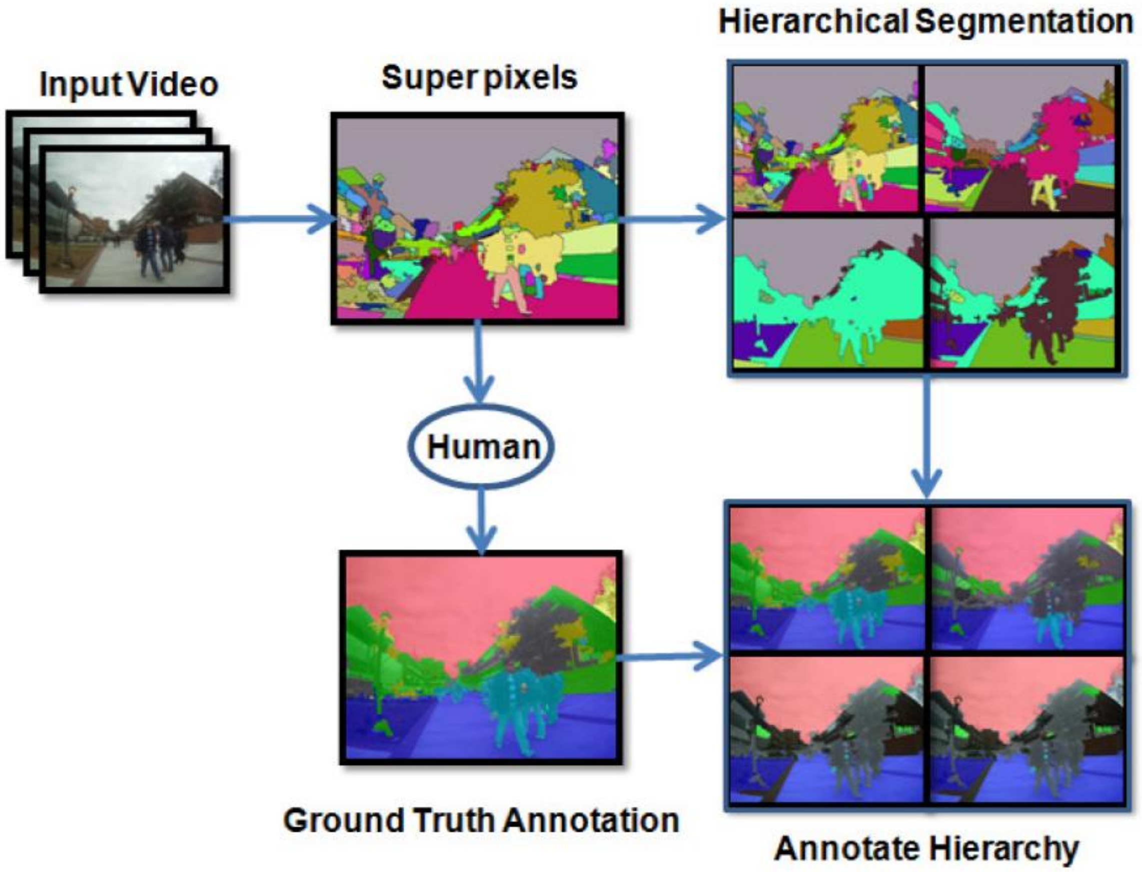
Sky	2.5%
Ground	15.9%
Vertical	81.2%
Mix	0.4%

(a) Main Classes



Solid	47.5%
Porous	26.1%
Object	7.7%

(b) Sub-vertical Classes



Features

- Features are extracted from 2D segments.
    - Appearance-based features:
        - Color
        - Texture
        - Location
        - Perspective
    - Motion-based features
        - Histogram of dense optical flow
        - Mean motion of a segment
        - Spatial flow differentials for the dense optical flow field
        - Mean location change
        - …

Multiple Segmentations

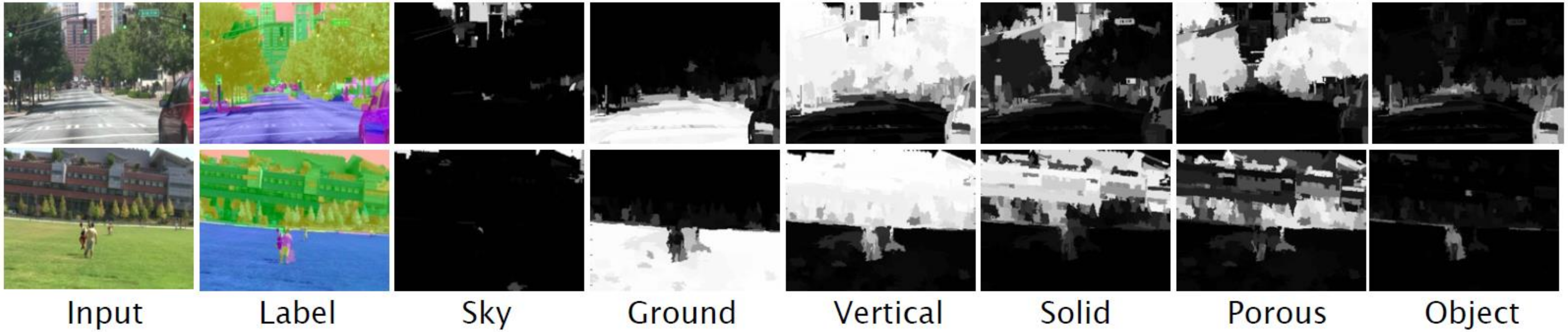
- Features are extracted for different hierarchy levels (10%, 20%, 30%, 40%, 50%) and their predicted labels will be combined based on homogeneity.

Classification

- Method: Boosted decision trees based on a logistic regression version of Adaboost (5-fold cross validation for each segment in different hierarchical levels).
- Output: Class probability.
- 3 classifiers are trained:
  - Main classes (multi-class classifier)
  - Vertical class (multi-class classifier)
  - Homogeneity classifier for the “mix” class (binary classifier)

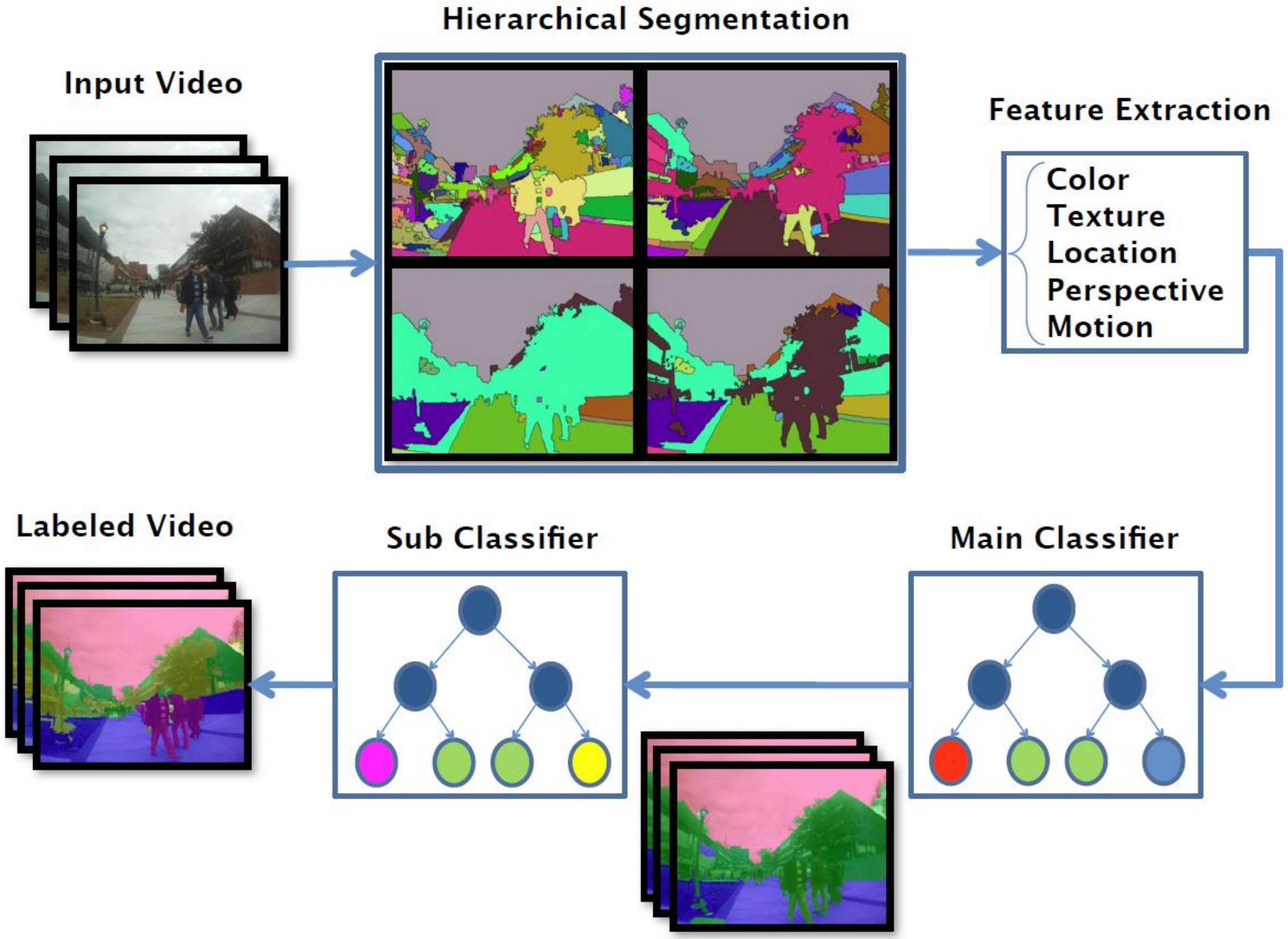
Prediction

$$P(y_i = k|\mathbf{x}_i) = \sum_{j}^{n_s} P(y_j = k|\mathbf{x}_j, s_j)P(s_j|\mathbf{x}_j),$$

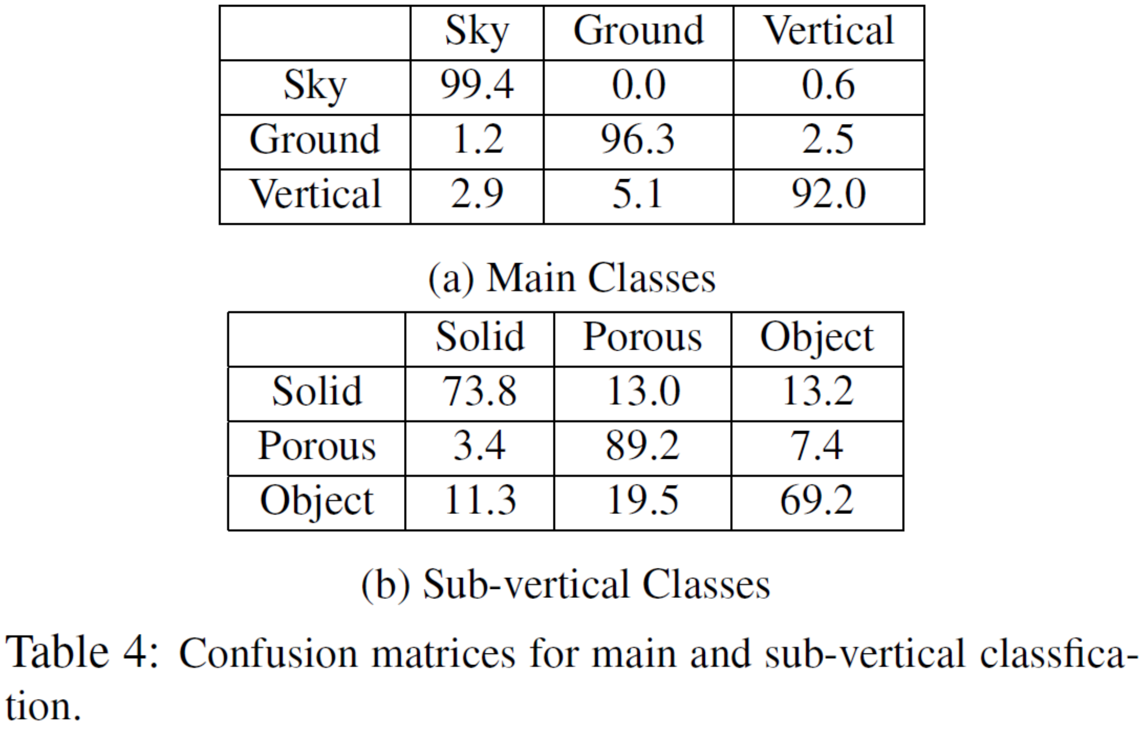


*sub-vertical classifier only applied to segments that are labeled as vertical by the main classifier

Overview


Hierarchical Segmentation
Input Video
Feature Extraction
Color
Texture
Location
Perspective
Motion
Labeled Video
Sub Classifier
Main Classifier

Results – Overall Accuracy



	Sky	Ground	Vertical
Sky	99.4	0.0	0.6
Ground	1.2	96.3	2.5
Vertical	2.9	5.1	92.0

(a) Main Classes

	Solid	Porous	Object
Solid	73.8	13.0	13.2
Porous	3.4	89.2	7.4
Object	11.3	19.5	69.2

(b) Sub-vertical Classes

Table 4: Confusion matrices for main and sub-vertical classfication.

Results – Overall Accuracy

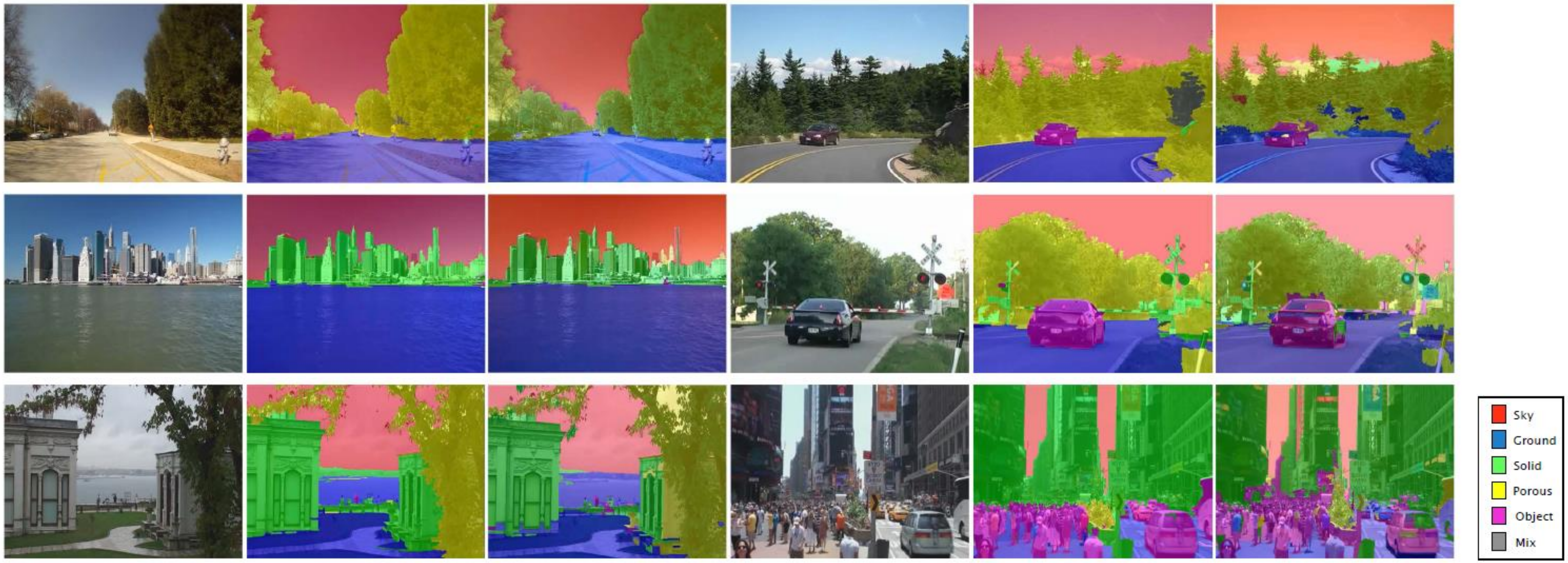
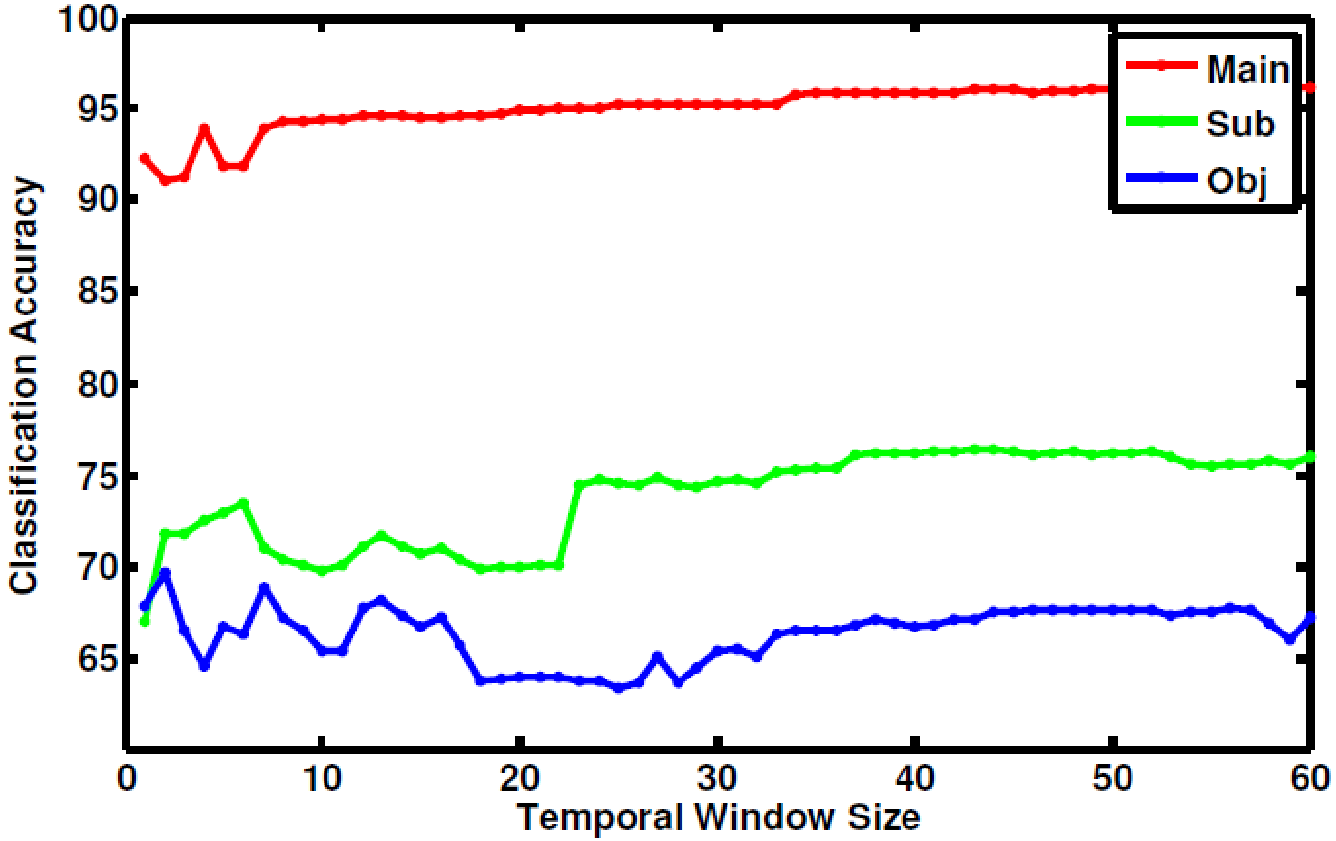
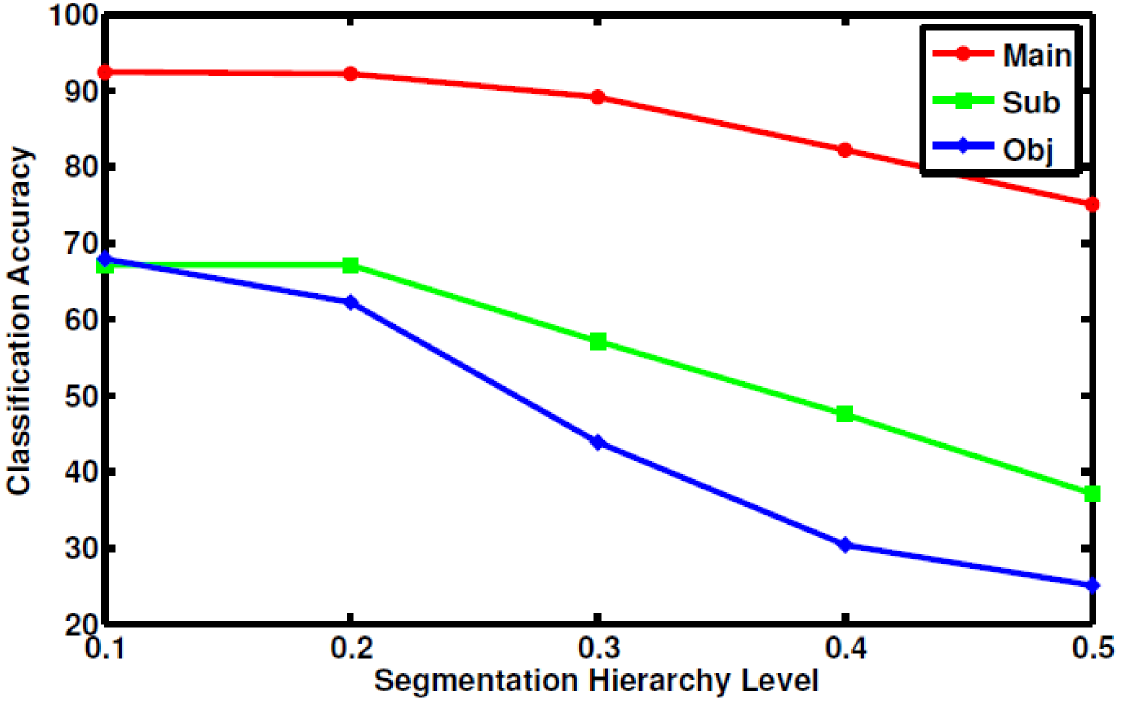


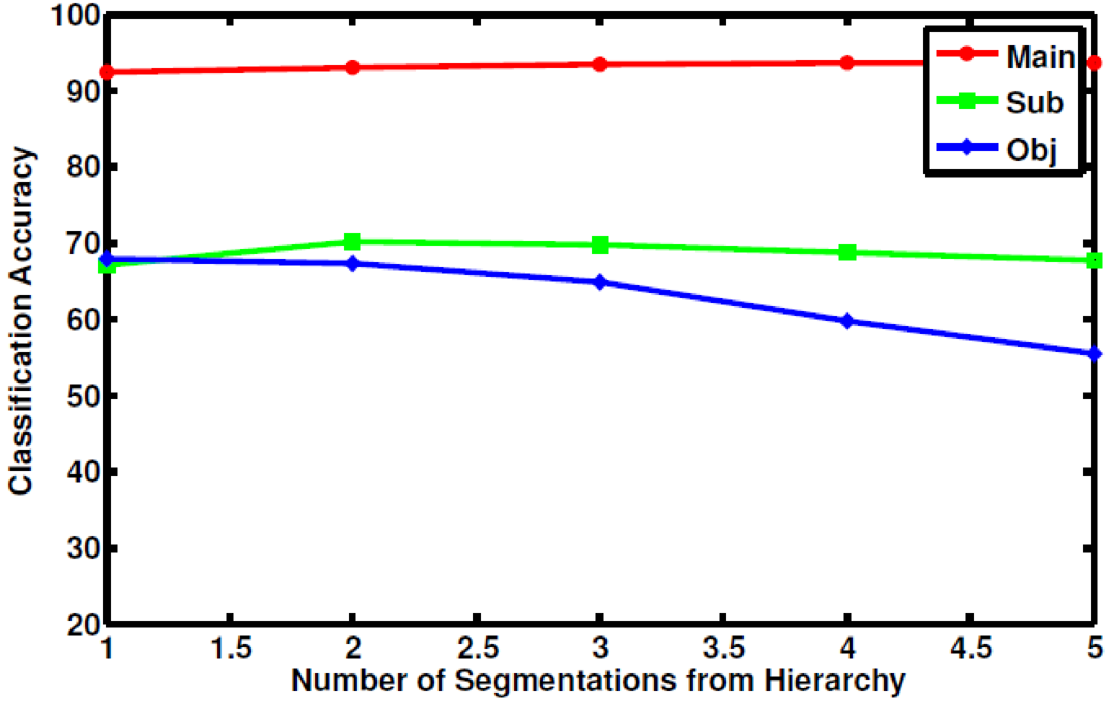
Figure 7: Qualitative results: From left to right: Input video frames, ground truth labels and predicted geometric labels. Our system performs well in challenging settings accurately predicting crowds, objects and foliage.

Results – Effect of Temporal Redundancy

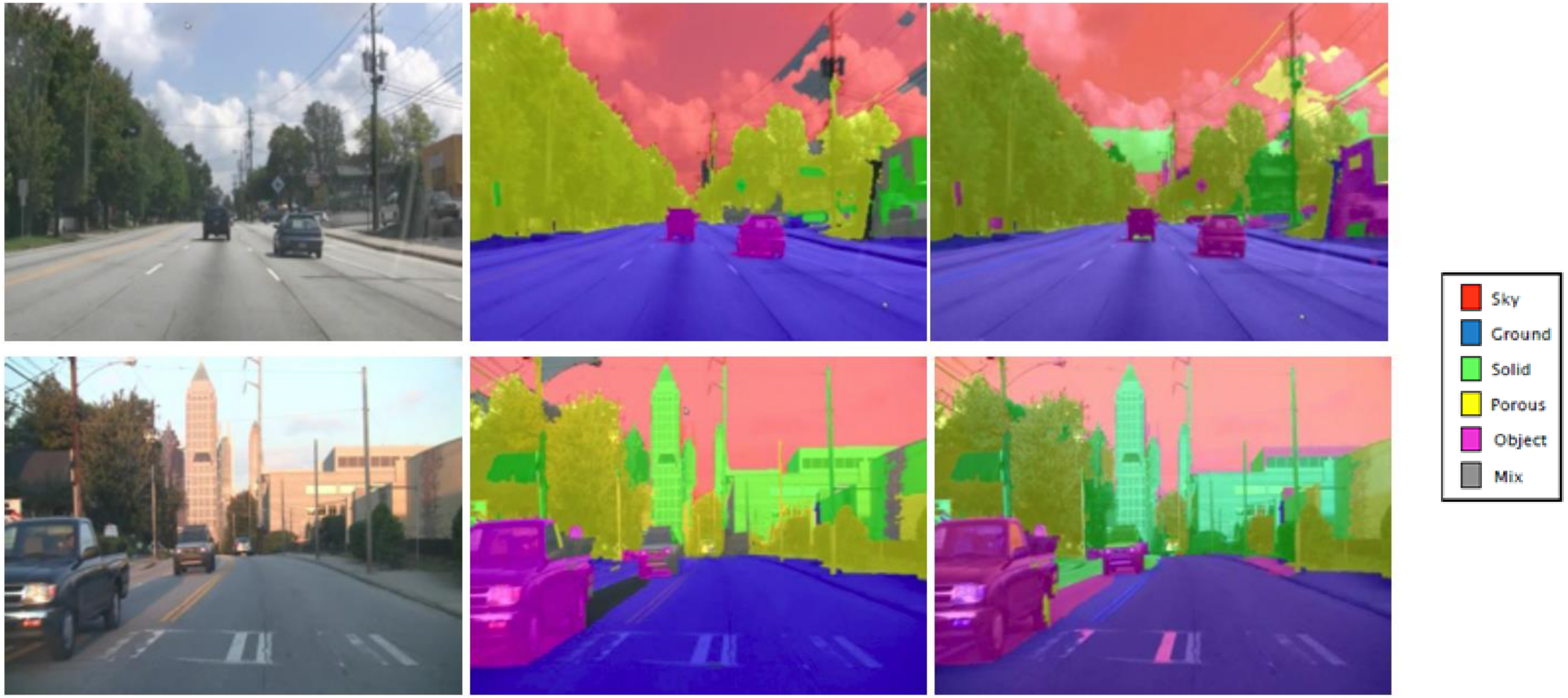


Results – Effect of Hierarchy



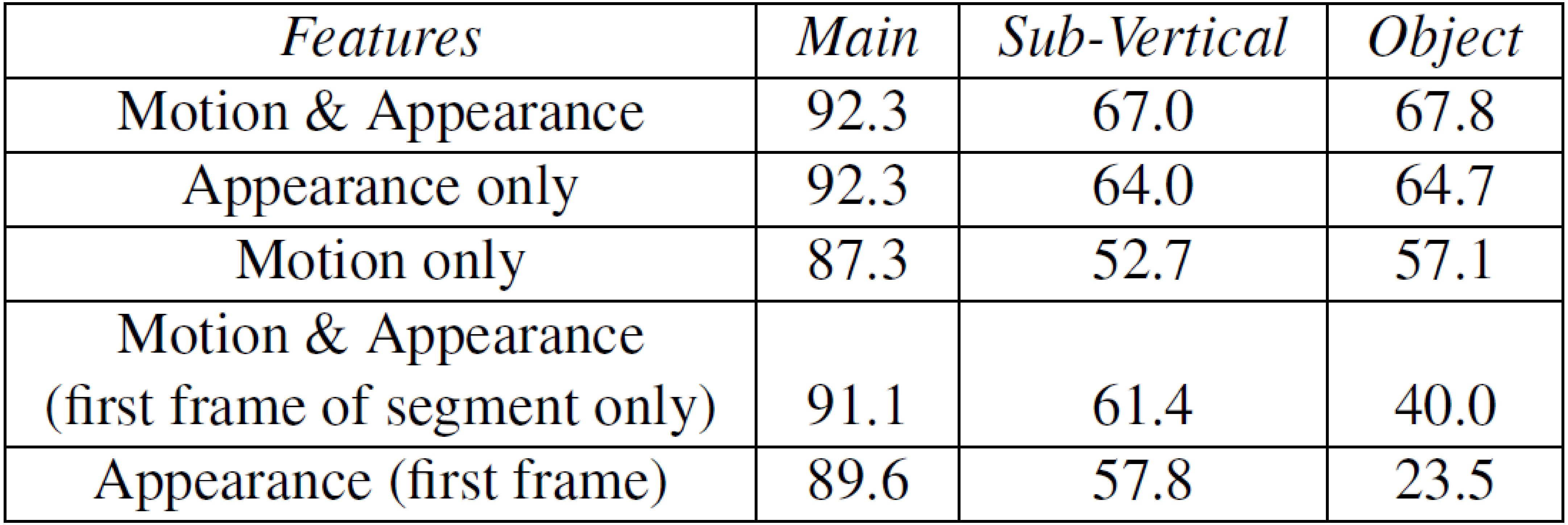


Misclassifications



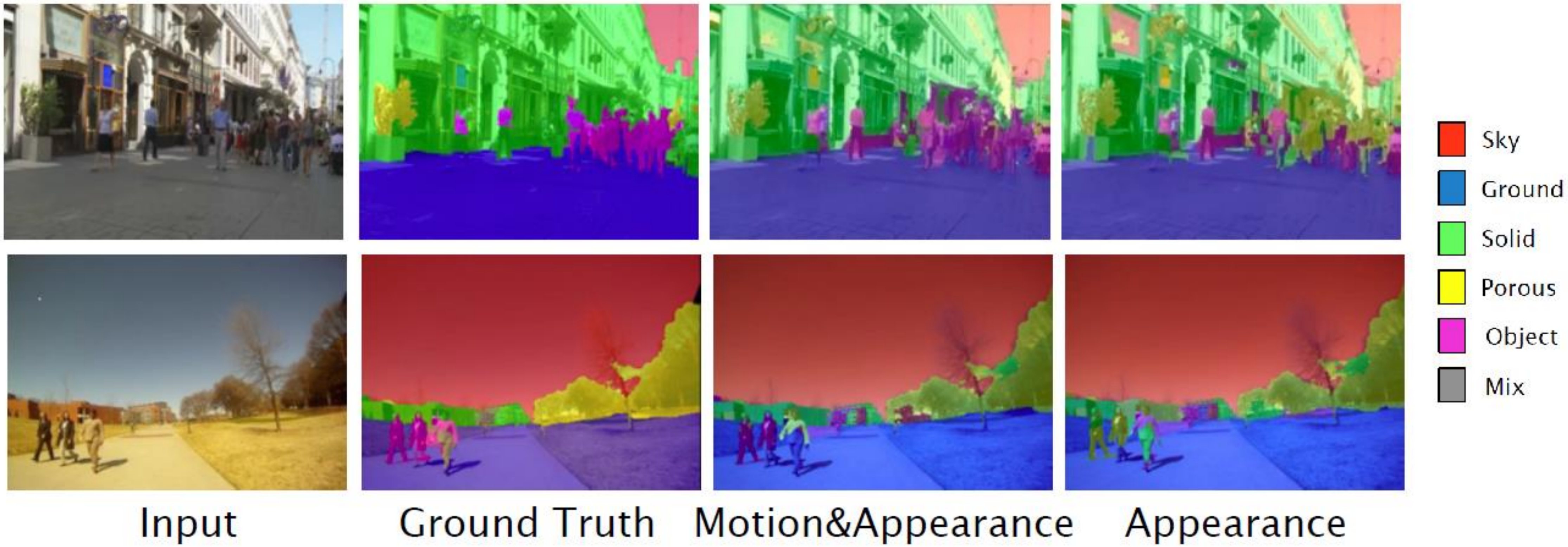
Input Ground Truth Labels

Results – Importance of Features

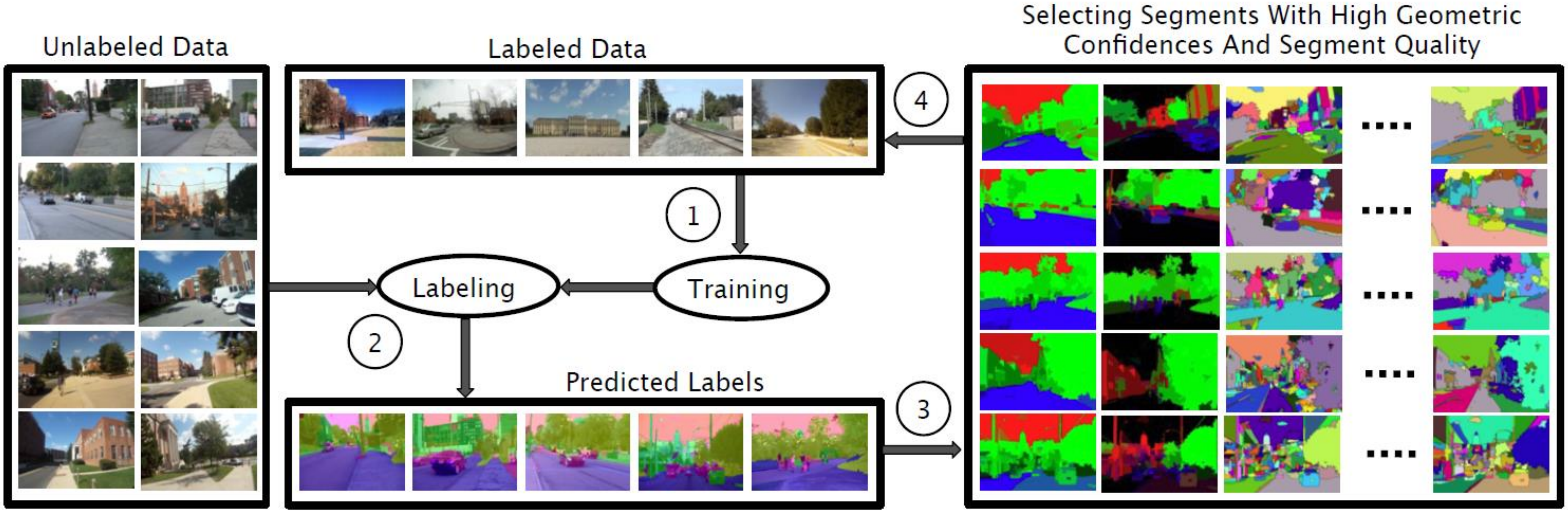


Features	*Main*	*Sub-Vertical*	*Object*
Motion & Appearance	92.3	67.0	67.8
Appearance only	92.3	64.0	64.7
Motion only	87.3	52.7	57.1
Motion & Appearance (first frame of segment only)	91.1	61.4	40.0
Appearance (first frame)	89.6	57.8	23.5

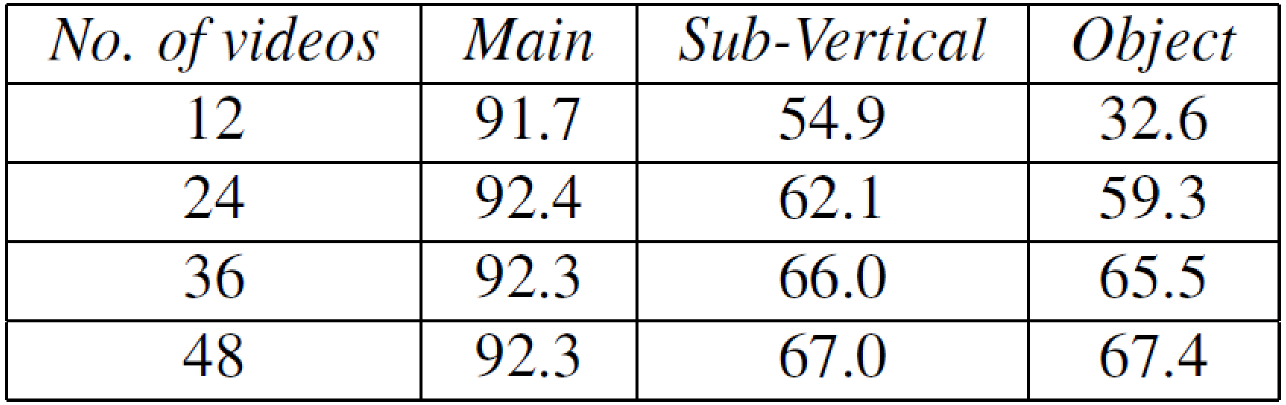
Results – Importance of Features



Semi-supervised Learning

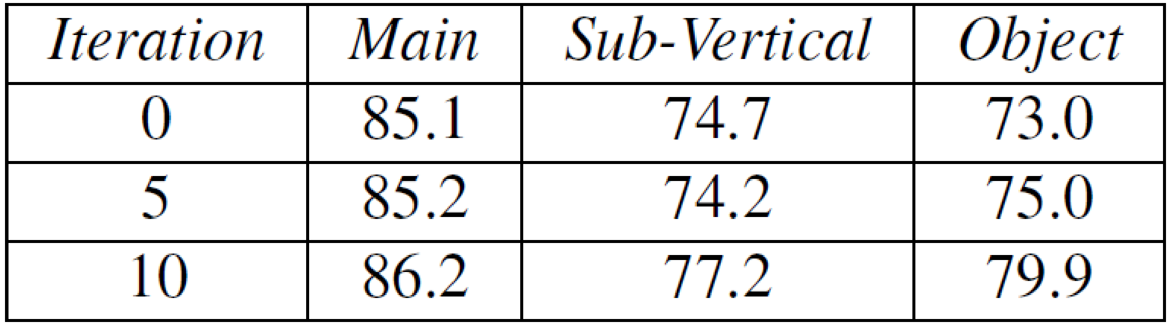

Unlabeled Data
Labeled Data
Selecting Segments With High Geometric Confidences And Segment Quality
4
1
Labeling
Training
2
Predicted Labels
3
....

Semi-supervised Learning



No. of videos	*Main*	*Sub-Vertical*	*Object*
12	91.7	54.9	32.6
24	92.4	62.1	59.3
36	92.3	66.0	65.5
48	92.3	67.0	67.4

(a) Data-size dependency in supervised learning



Iteration	*Main*	*Sub-Vertical*	*Object*
0	85.1	74.7	73.0
5	85.2	74.2	75.0
10	86.2	77.2	79.9

(b) Semi-supervised bootstrap learning

Thanks