25th INTERNATIONAL CONFERENCE ON PATTERN RECOGNITION
Milan, Italy 10 | 15 January 2021

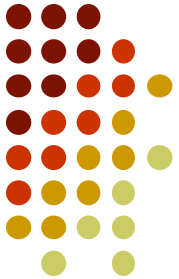
Semantic Segmentation for Pedestrian Detection from Motion in Temporal Domain

Guo Cheng, Jiang Yu Zheng

Center of VISC
Department of Computer and Information Science
Indiana University – Purdue University Indianapolis, USA

Outline

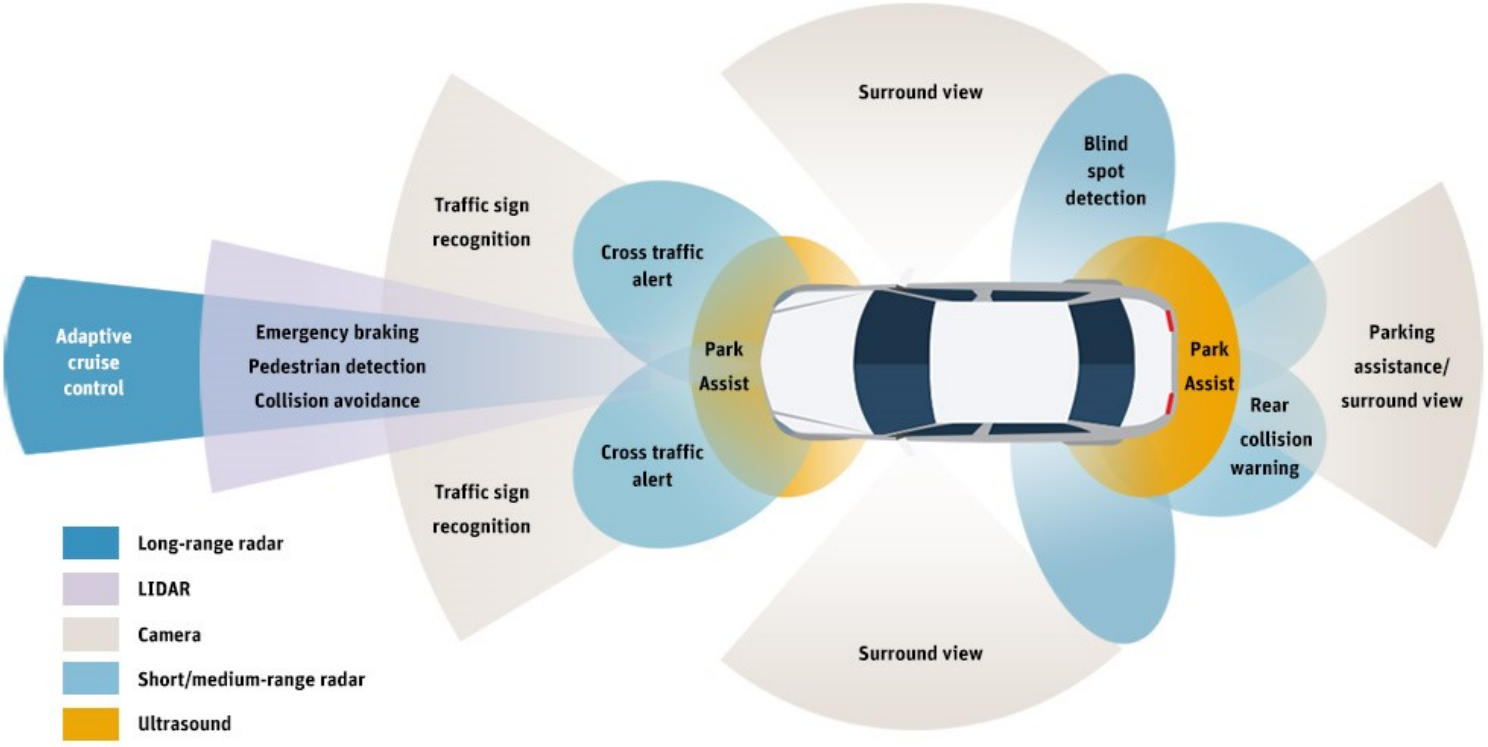
- Introduction
- Video Data Reduction through Temporal-to-Spatial
- Pedestrian Motion in Motion Profile
- Semantic Segmentation of Pedestrian Motion
- Temporal-Shift Memory
- Experiment
- Conclusion
- Appendix



I. Introduction

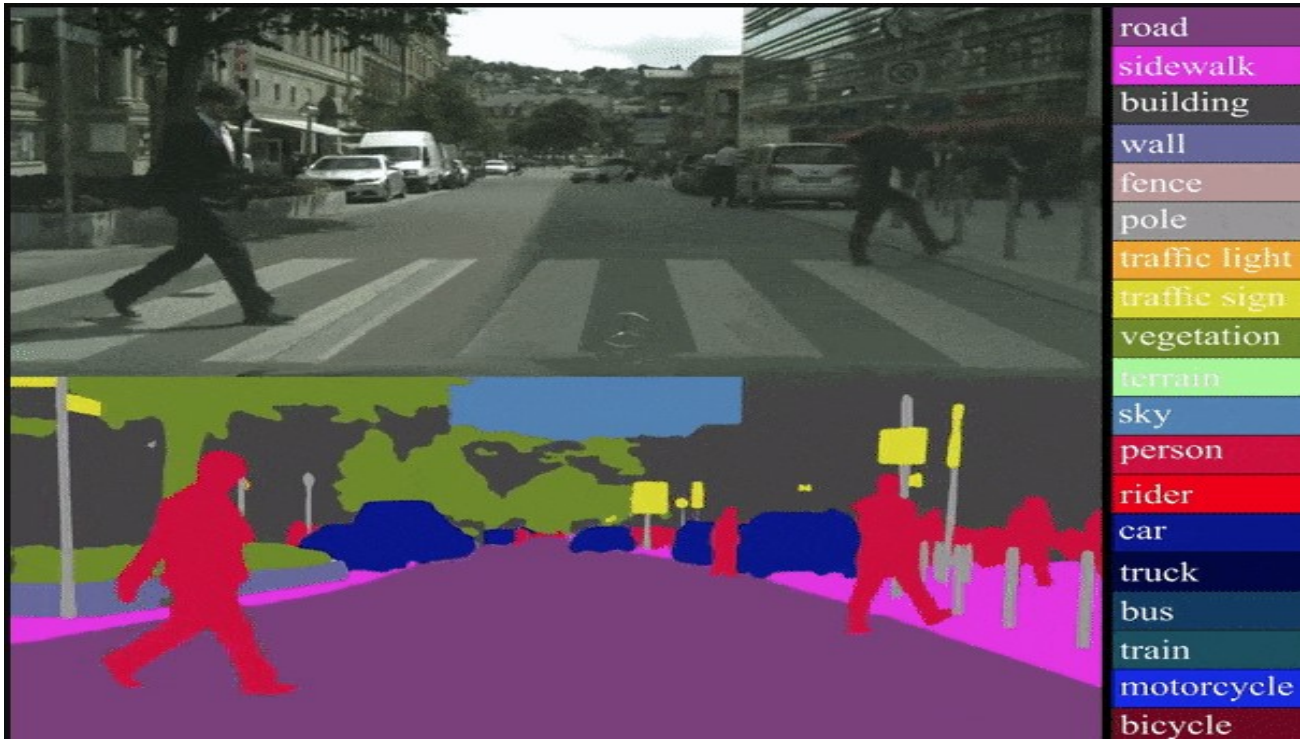
Real-Time Autonomous Driving

- Real-time autonomous driving requires fast processing of sensor-fused data from all kinds of devices embedded in the vehicle.



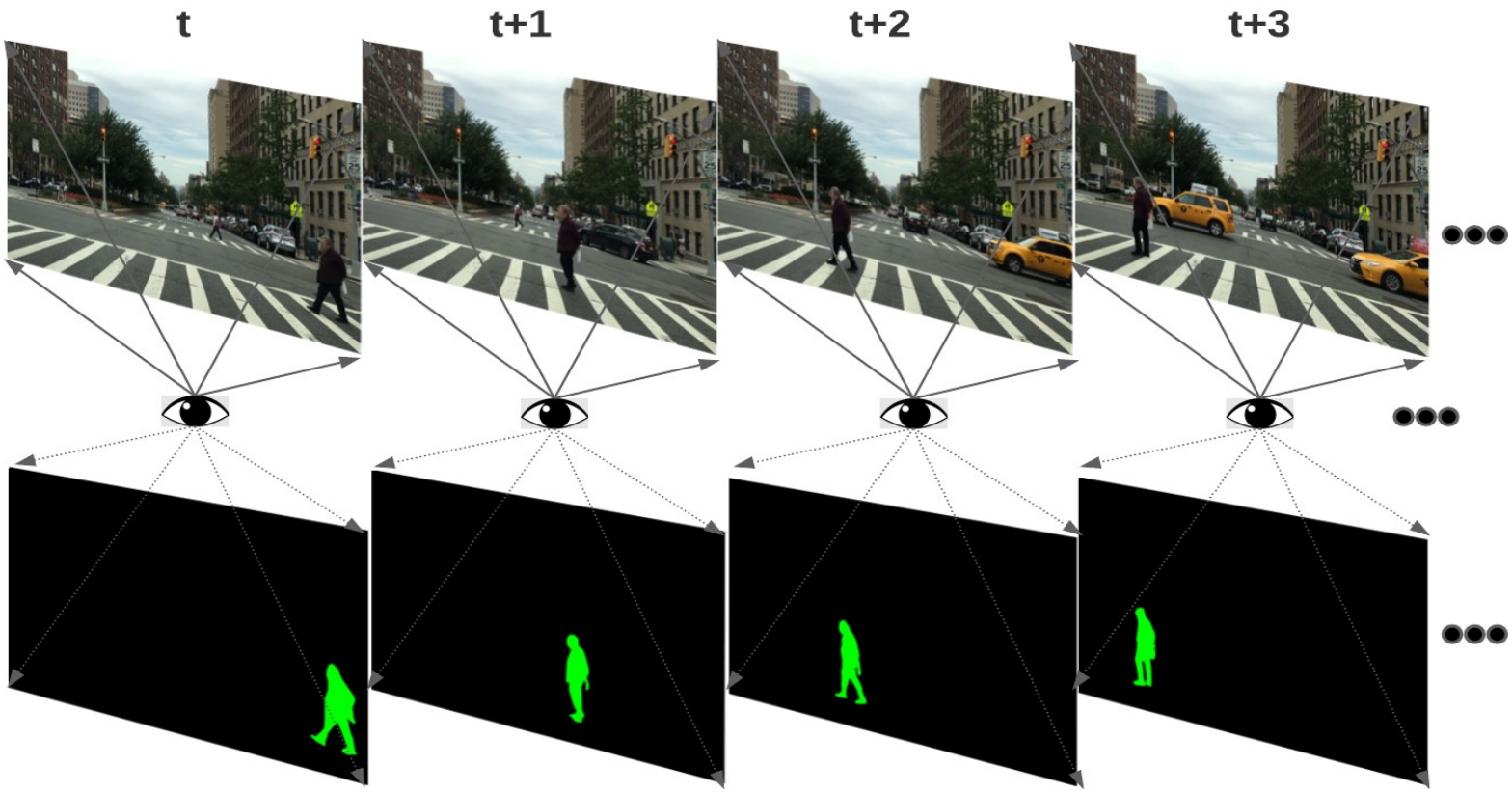
Scene Understanding

- Safety driving for autonomous vehicle requires many vision tasks such as road segmentation, pedestrian detection, and vehicle recognition by sensors, including frontal cameras and LiDAR.



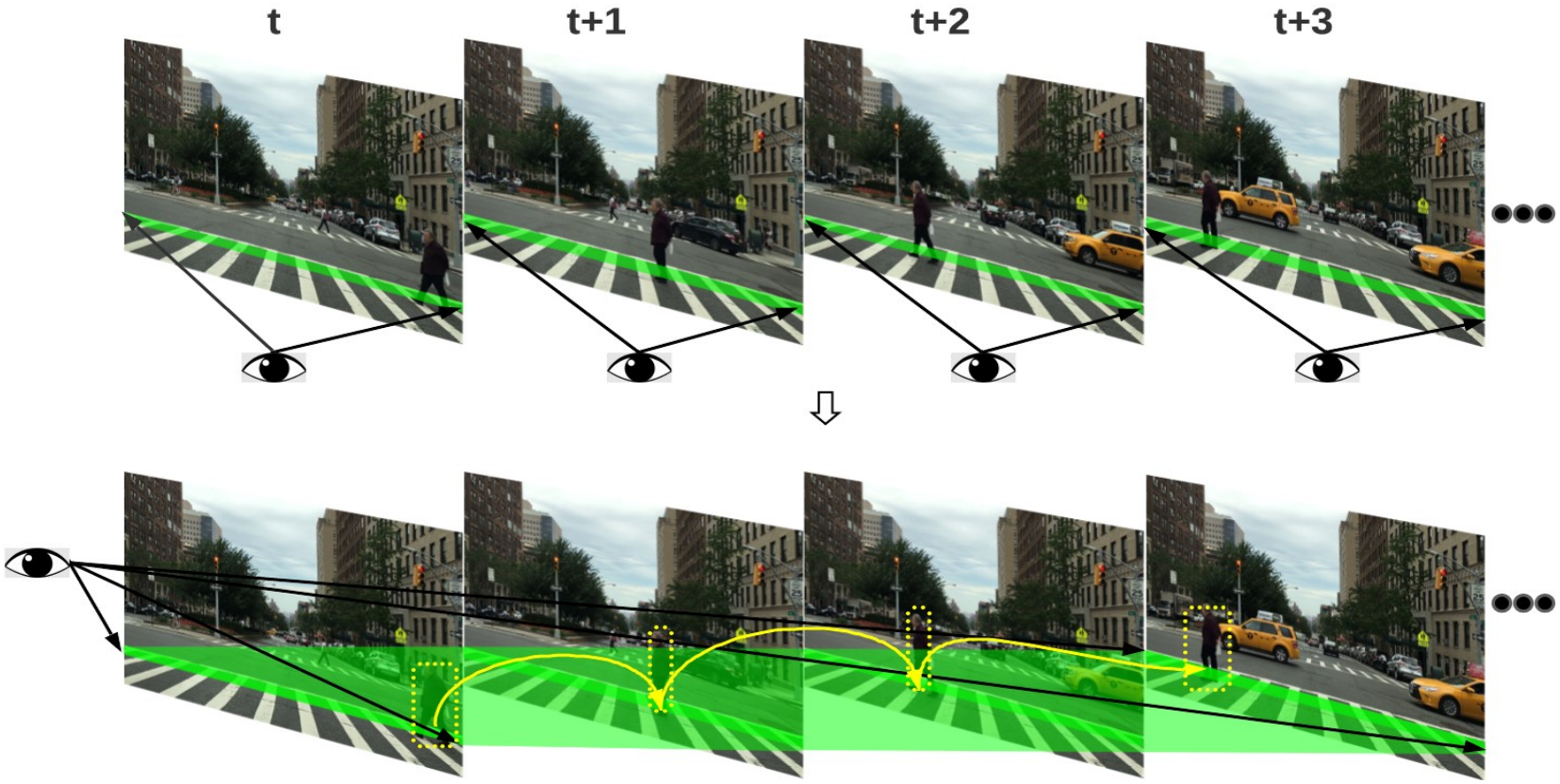
Problems in Traditional Method

- Current state-of-the-art methods detect the pedestrians and objects in spatial space, which is time-consuming.



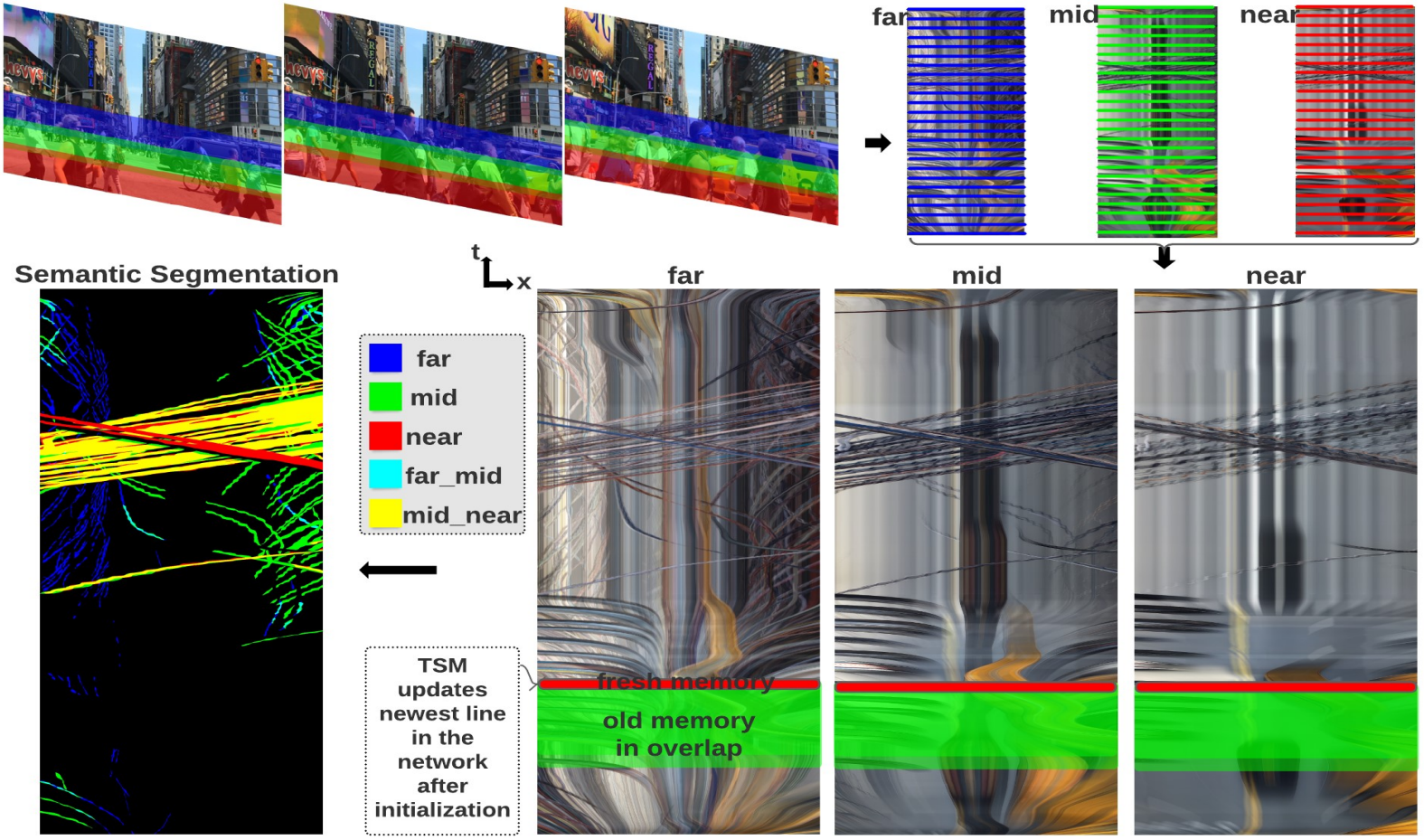
Temporal-to-Spatial

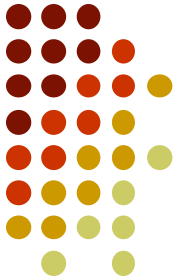
- we apply a thorough pedestrian motion segmentation and detection in a temporal-to-spatial approach.



Sequential Semantic Segmentation

- We perform semantic segmentation sequentially along the time axis in 2D temporal layout.

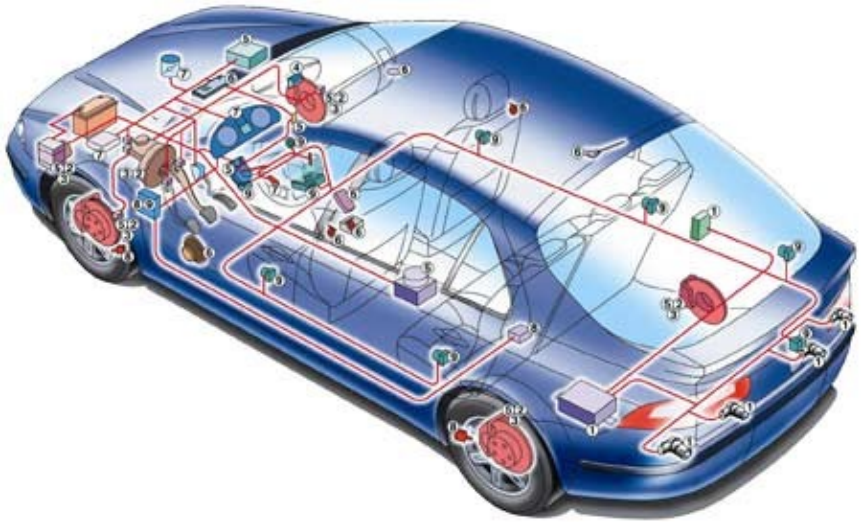




II. Video Reduction through Temporal-to-Spatial

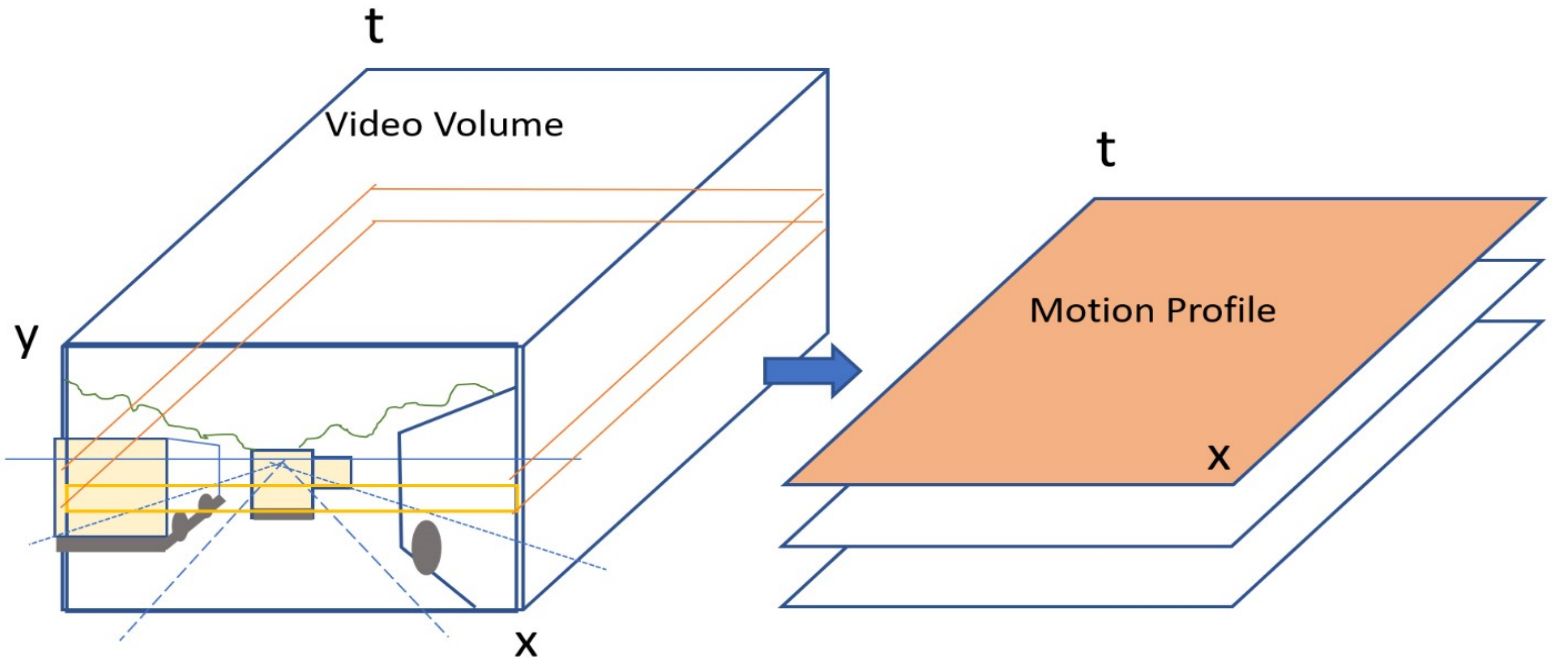
Why we adopt Temporal-to-Spatial?

- Traditional data processing based on full 2D video frames is time-consuming.
- Embedded hardware in vehicle has limited capability of computation and memory.



Motion Profile

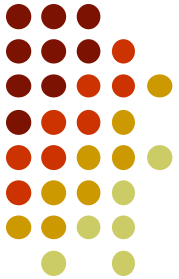
- We sample in each frame at a belt pre-defined below the calibrated horizon to catch the temporal scenes ahead of vehicle;
- We condense each belt into 1D array by averaging pixels vertically;
- the 1D arrays from consecutive frames are copied into a spatial-temporal image, i.e., Motion Profile.



Principle of Motion Profile

- The length of MP is the total number of frames in the video.
- The width of MP is exactly as the same of the frame image.
- The position of sample belt is set freely to cover a depth range.
- Multiple belts at different heights with small overlaps can cover close, middle, and far depths, respectively.

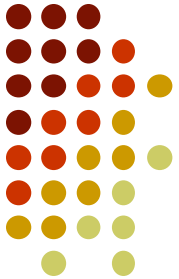
Property of Motion Profile



- Although only 1D data are sampled from each frame, their temporal concatenation in MP shows motion characteristics.

- Compared to pedestrian detection in video volume, the data sheet of MP is the smallest and the pedestrian leg motion is less varied than their pose, shape, appearance, and illumination.

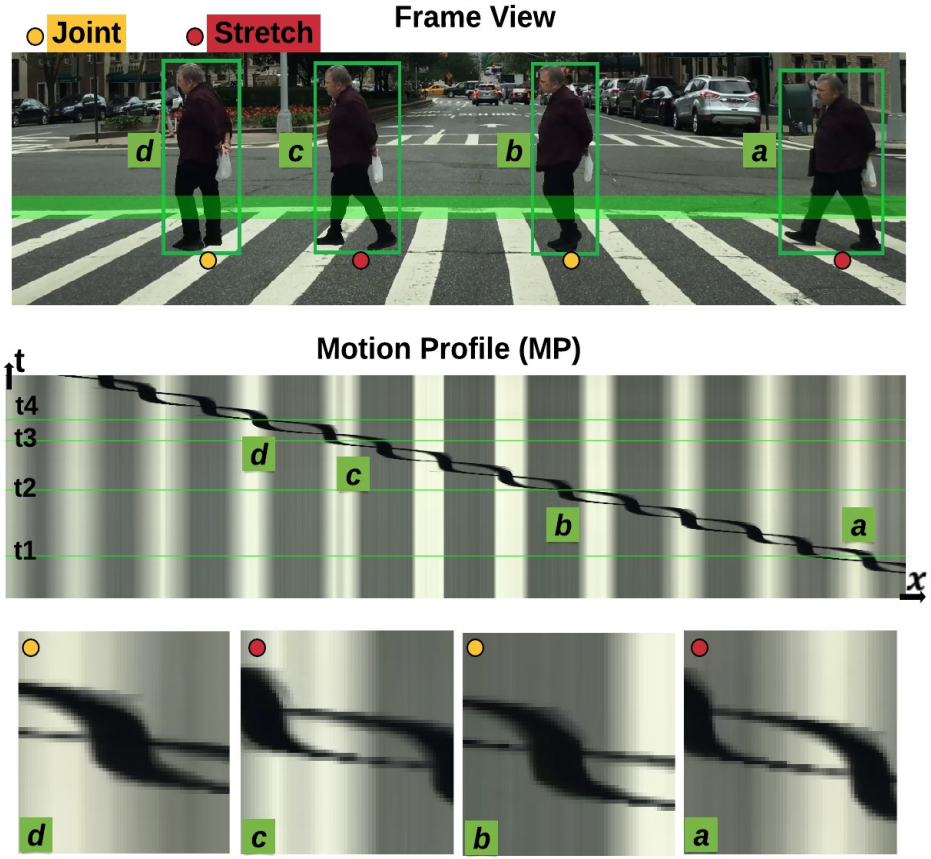
- The motion traces of objects in driving video are reserved in the temporal continuity of MP, which directly provides information about driving direction and speed.



III. Pedestrian Motion in Motion Profile

Pedestrian Motion in MP

- Pedestrian motion trace forms a crossing chain in MP.

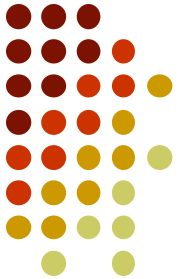


- Pedestrian motion in driving video with a green horizontal belt on legs.
- Sequentially projecting the average pixel values in the belt over consecutive frames to generate a temporal image.
- Amplified views of legs at stopping and stepping moments.

Cyclic crossing-chain

- When a pedestrian is walking, we can observe leg trajectories as a crossing chain.
- Legs step alternatively in a cycle along the trajectory.
- Because of the vertical averaging of pixels, we obtained more robust features than pixels, which are strong/long vertical edges in the frame.

MP from multiple distances



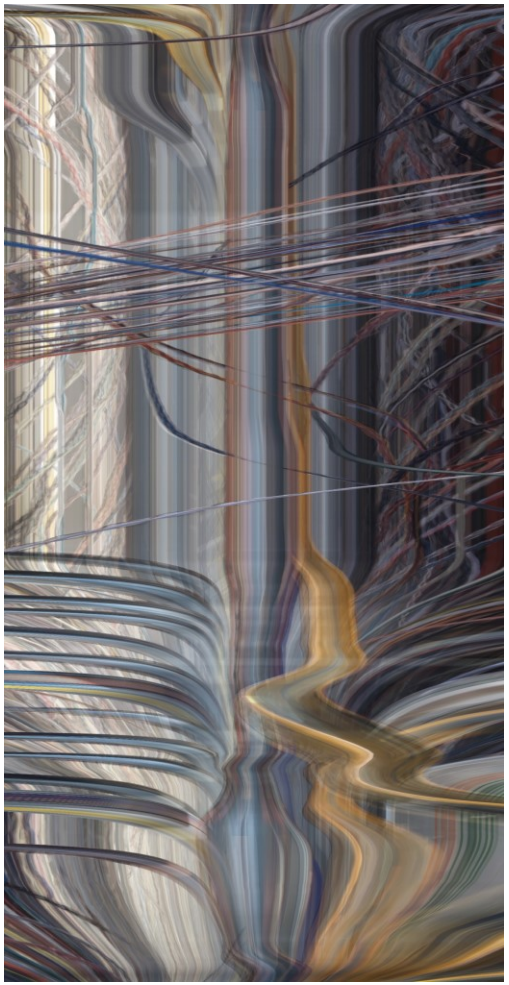
- To avoid collision in driving, we extract three MPs according to far, middle, and near distances.

- Pedestrians near to vehicle are more dangerous and thus need to be detected accurately and promptly.
- Pedestrian at the middle distance from ego-vehicle is useful for path-planning in real autonomous driving.
- Pedestrians at far distance have a lower resolution in MP but have chance to be re-identified when ego-vehicle gets closer.

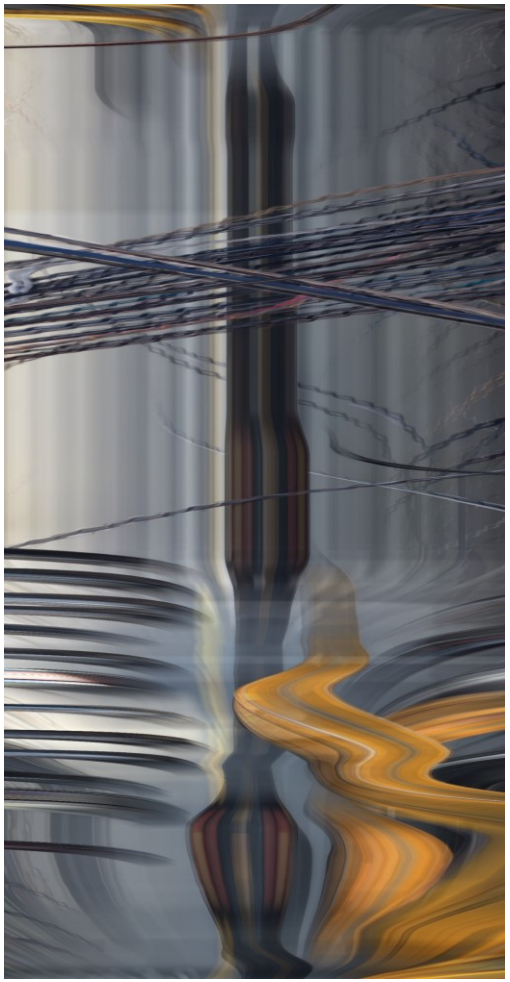
- Adjacent belts have an overlap of certain pixels, which helps leg detection timely in one of MPs and ensures pedestrian motion observed continuously.

MP from multiple distances

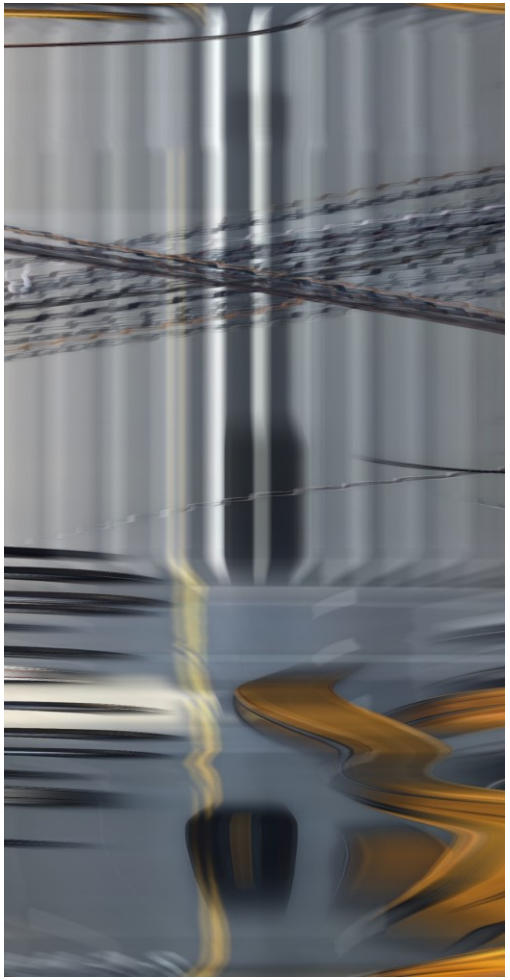
- far



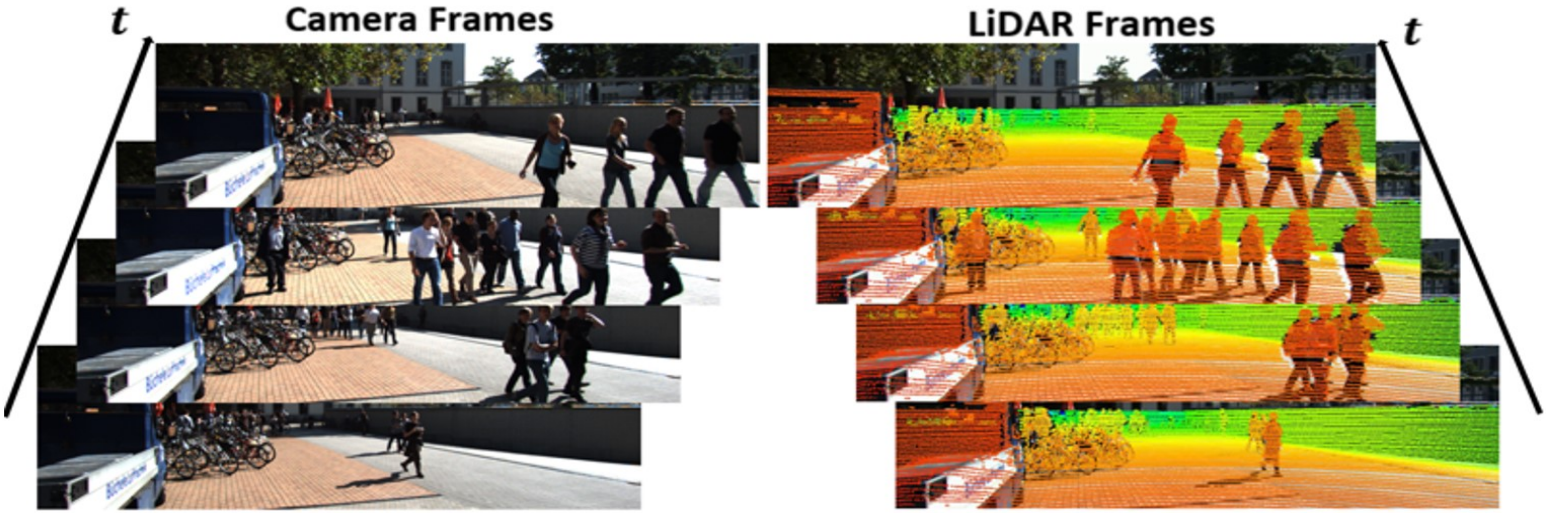
- mid

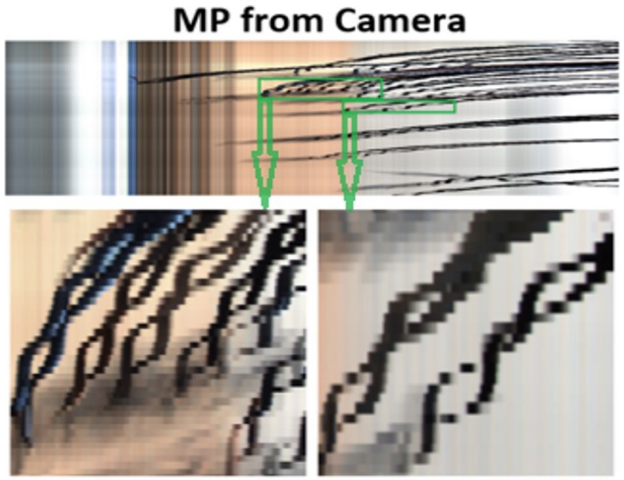


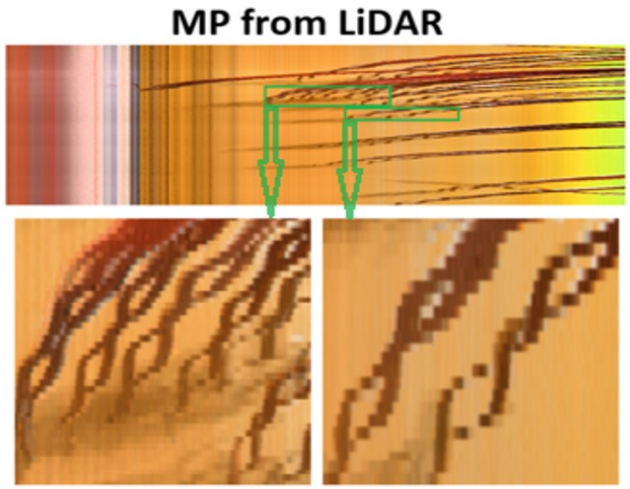
- near



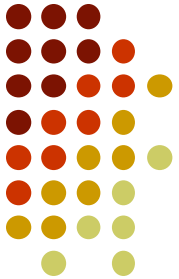
MP from different sensors







Patterns of Pedestrian Motion in MP



- Pedestrians walking on sidewalk or crossing street.
- Vertical object occlusion generates crossing traces in MP as well, which will be excluded through deep learning as well.
- Motion traces captured from a turning vehicle are skewed in MP.
- Pedestrian leg trace mixed with arm and body traces.
- Humans standing-still can be followed as a trace, rather than stepping legs, which can still be distinguished using his/her width, position, and color by deep learning.

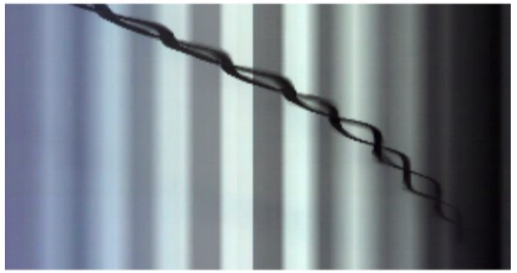
Patterns of Pedestrian Motion in MP

walking in non-static background

skewed leg trace when vehicle is making a turn

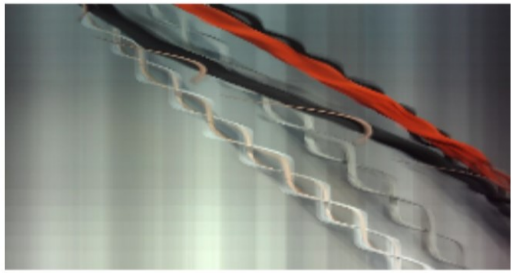
walking along sidewalk when vehicle is moving fast

walking in static background



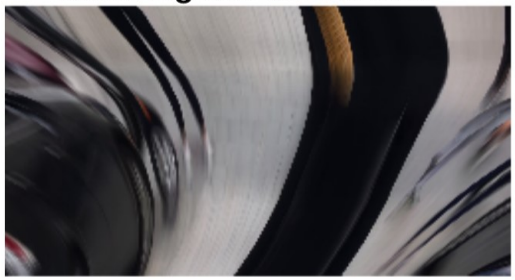
pedestrian motion in dark light

leg trace mixed with arm trace

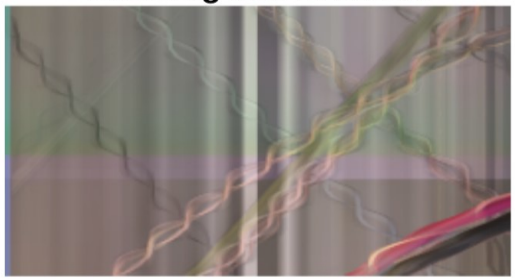


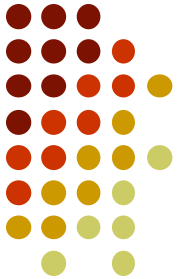
standing-still after motion

standing-still all the time



standing-still in crowd

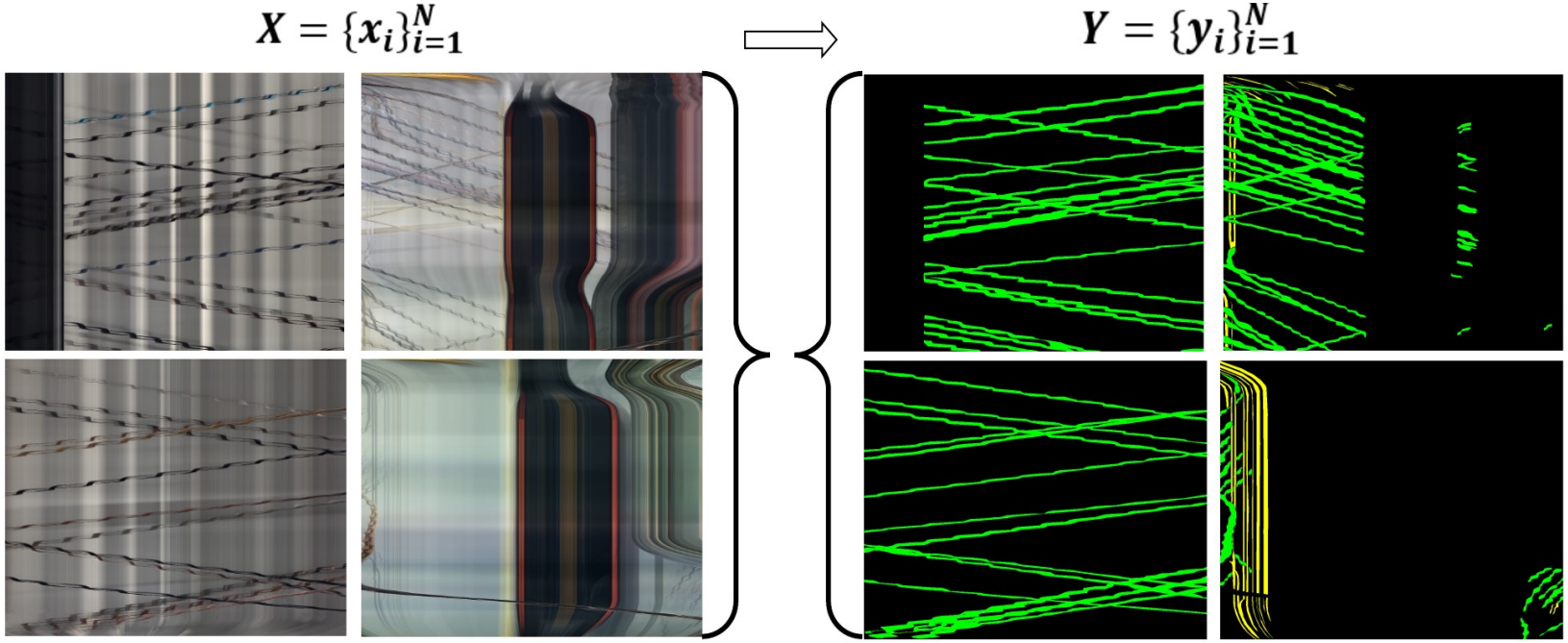




IV. Semantic Segmentation of Pedestrian Motion in Motion Profile

Data Pre1: Semantic Classes

- Paired MP patches and labels in our dataset. Three semantic classes in the labels: Pedestrian in motion (green), human standing-still (yellow) and background (black).



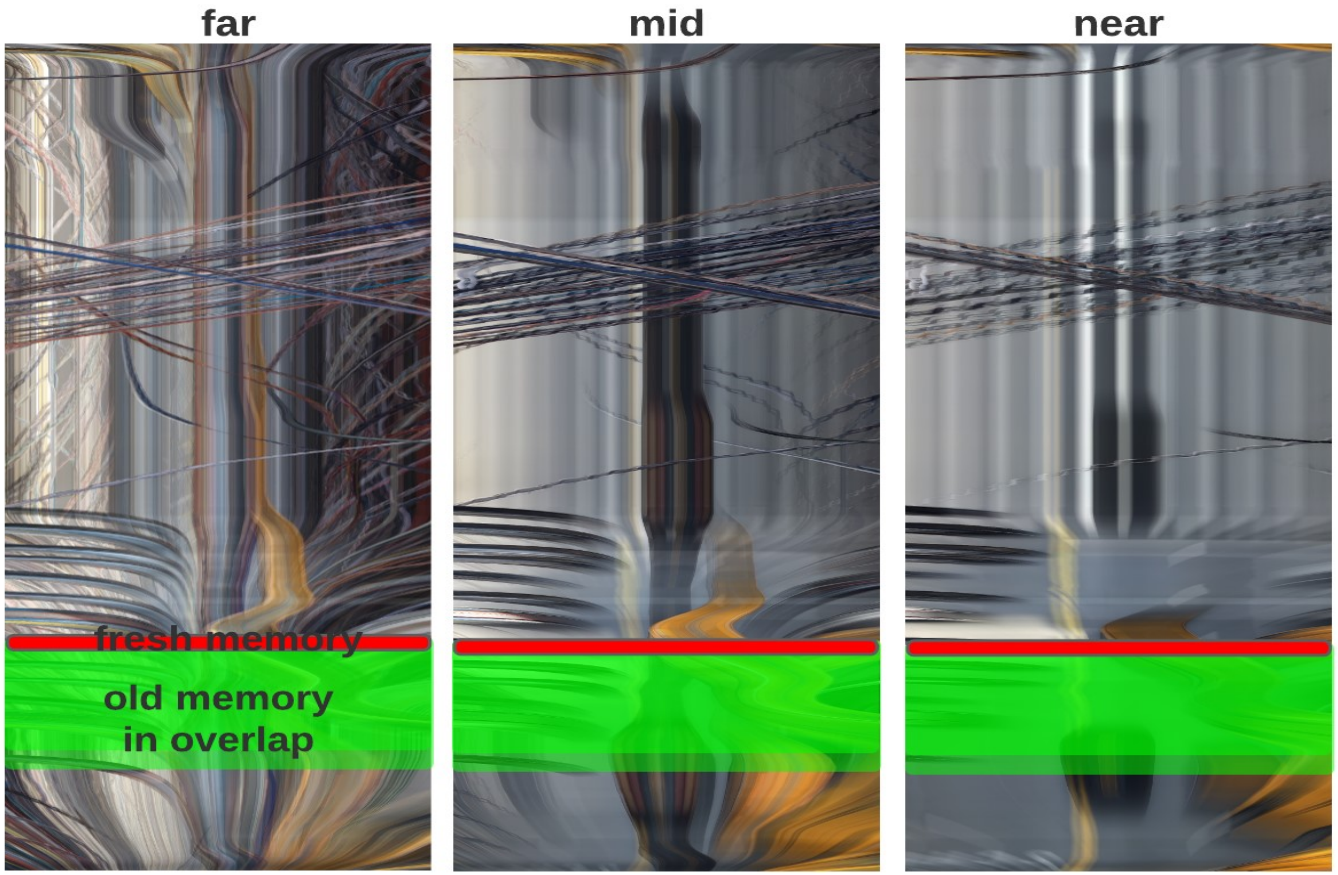
Data Pre2: Sequential Patch-Size

- To achieve a fast response in real-time driving, our semantic algorithm uses a section of motion profile at each time as input incrementally, i.e., patch based.

- A patch has a fixed width (256pixels) and a time duration T (32frames) to preserve motion continuity and context for trace identification.

- The patch is temporally shifted every frame along the time axis to achieve the finest temporal resolution and earliest response to input. Thus, our network outputs one segmented line immediately after input without time latency.

Data Pre2: Sequential Patch-Size



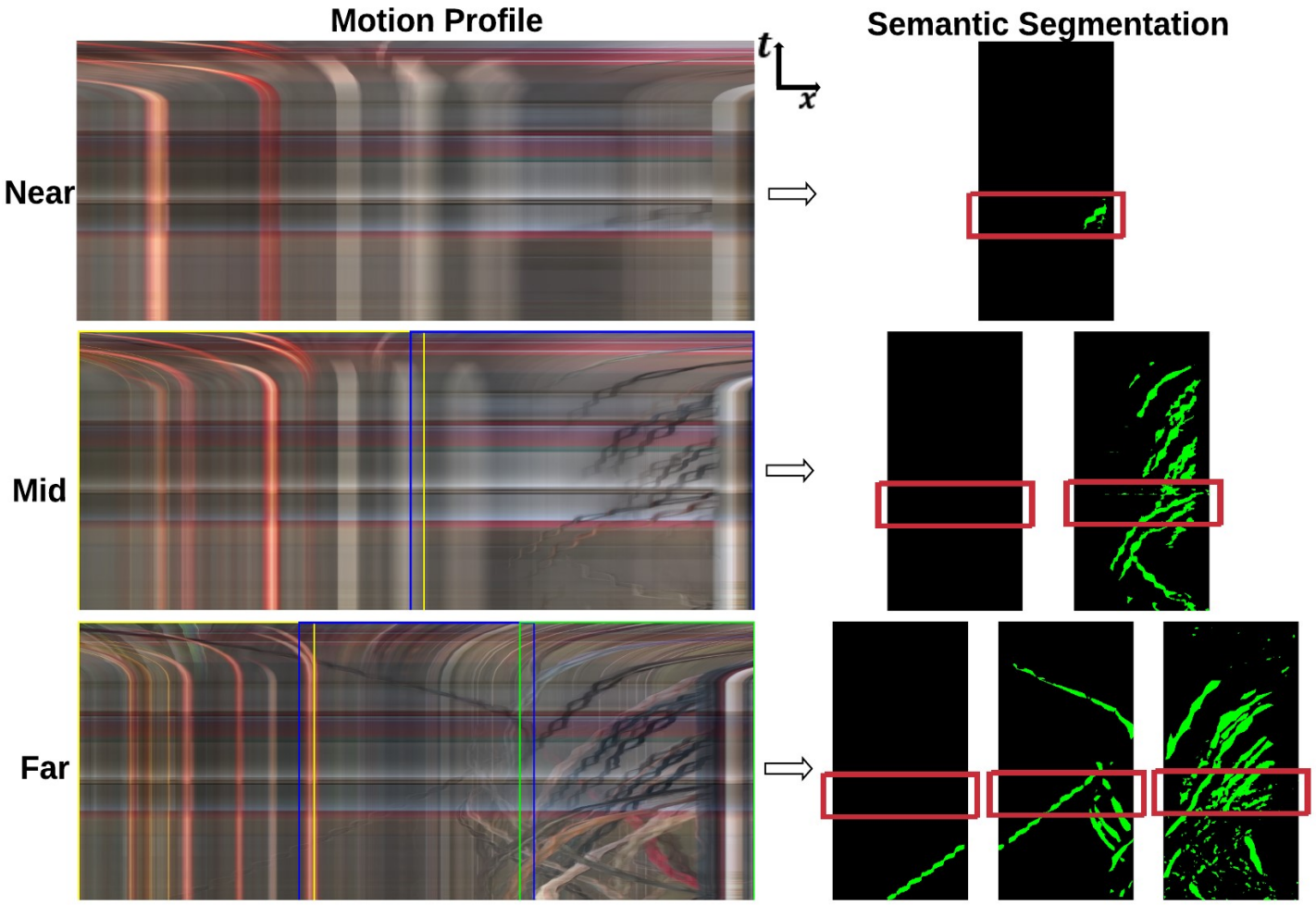
Data Pre3: Depth-Invariant

- Pedestrian motion at different depths has varied directions in MPs, but the shape of crossing-chain is clear at joints and stretches.

- To normalize the orientation for consistent training and testing by the deep learning network, three MPs from close to far are horizontally divided to 1, 2, and 3 columns, respectively.

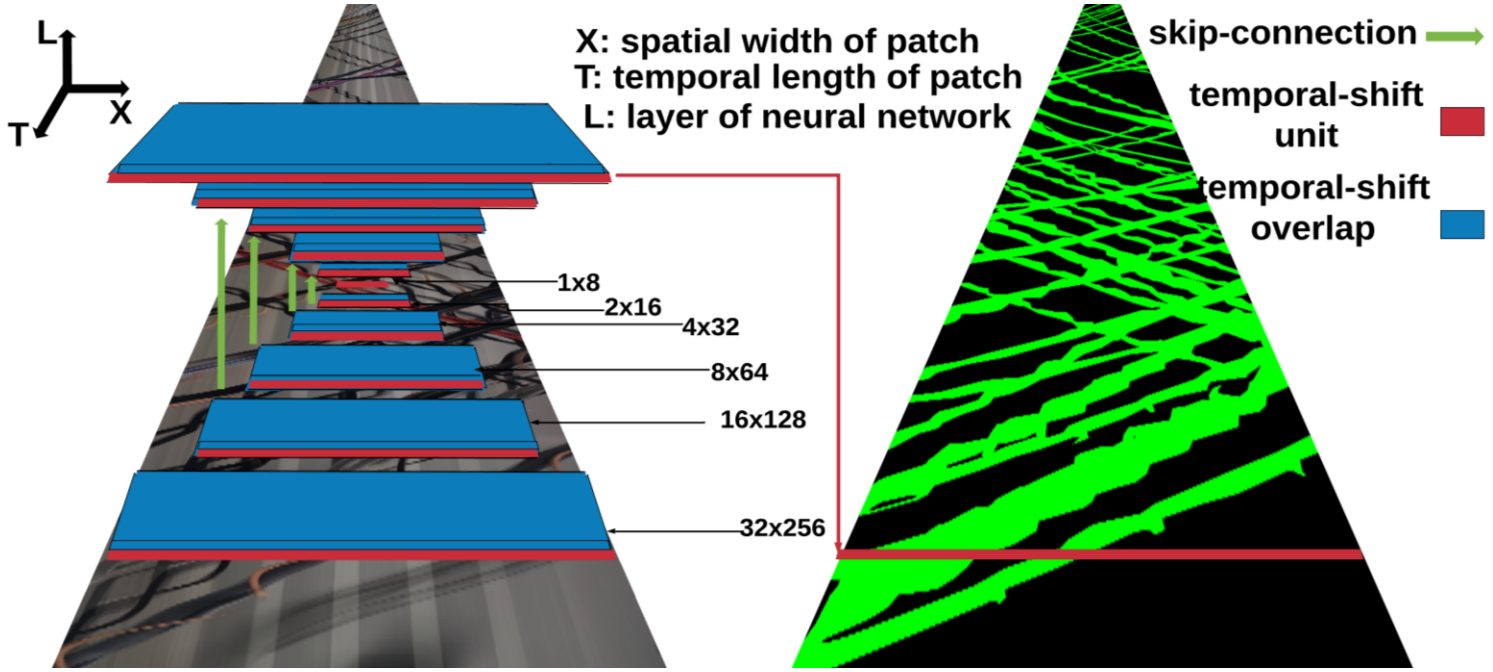
- They are horizontally scaled to a given width as input to the neural network. The time scale of them are absolute as mentioned above.

Data Pre3: Depth-Invariant



Sequential Semantic Segmentation

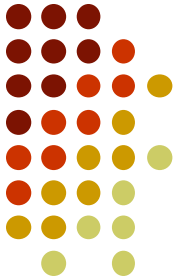
- The network performs semantic segmentation in patch size of 32×256 at each time, which consists of an encoder-decoder module embedded with skip connections depicted in layer direction vertically.



Network Structure

- **Encoder-decoder module**: The baseline of network structure of semantic segmentation for single patch is an encoder-decoder module embedded with skip-connections. It consists of five encoding and five decoding blocks symmetrically; each small block contains convolutional layer and a 2×2 pooling layer.

- **Input/Output**: Since the neural network steps 1 line (frame) for the minimum latency, it adopts small kernel size in convolutional and pooling filters of 1x1 and 2x2. We output latest line from the patch of 32×256 pixels after decoding as the segmentation result.



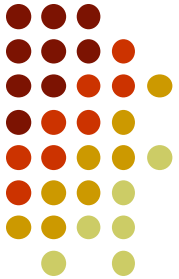
V. Temporal-Shift Memory

Overlap in Sequential Semantic Segmentation

- In testing semantic segmentation on MP, however, the deep network uses the temporal context in the pedestrian motion detection, which involves *T-1* lines computed already for the latest line.

- Thus, the pyramid structure of encoding-decoding network has a large overlap on the data when sequentially scanning the Motion Profile along time axis.

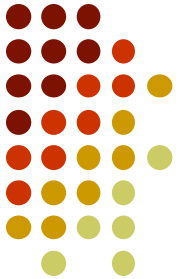
Temporal-Shift Memory



- **Referencing Model**: we propose a sequential model to avoid redundant data processing in online testing phase for real time driving.

- **Accuracy lossless**: This mechanism keeps the same mechanism of the encoding-decoding structure without accuracy loss but ensures the network to process only the newest lines at all network layers, which is in a true scanning mode.

Temporal-Shift Memory



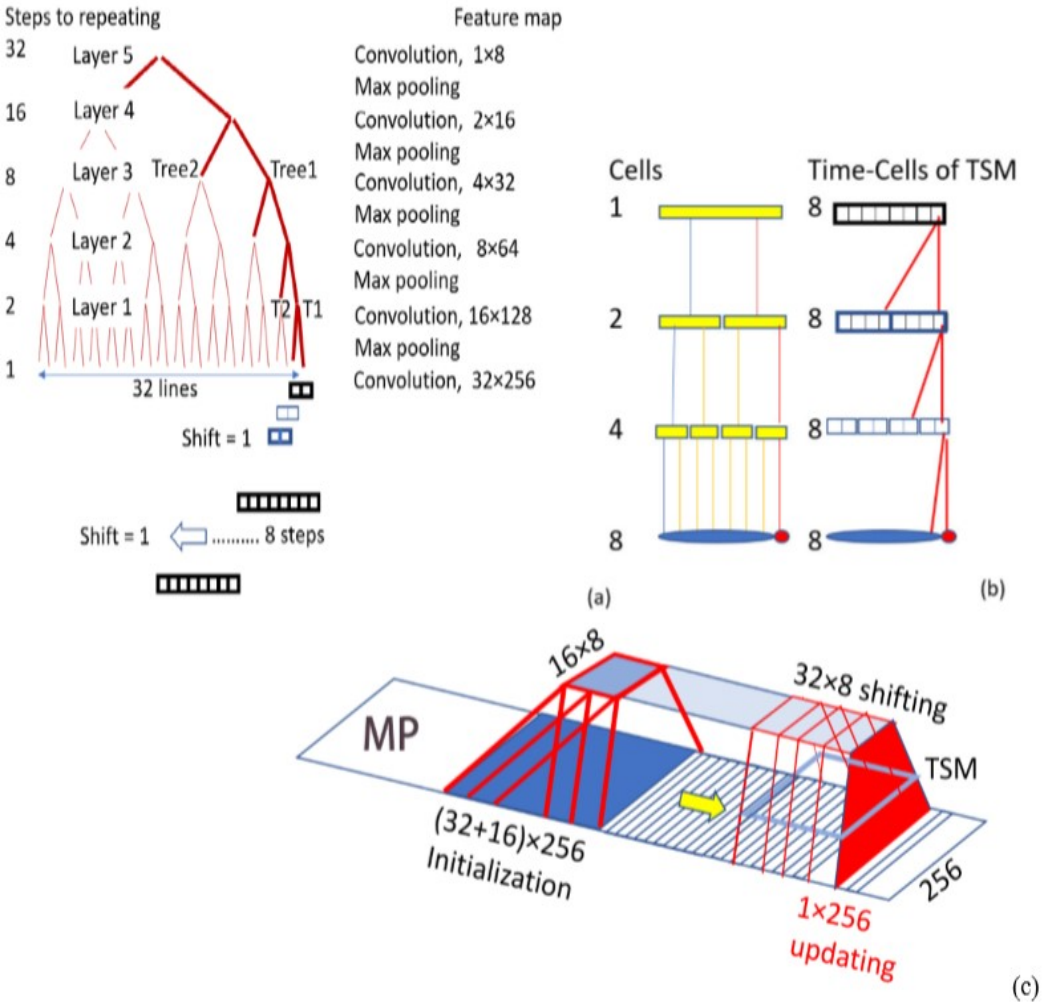
- TSM avoids repeated computation by storing previously computed nodes hierarchically along neural layers.

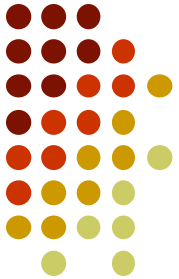
- At each layer, only the latest nodes are updated through filtering and maximum pooling from the newly input line and the previously computed node involved for this operation.

- The temporal-shift memory model requires an initial calculation of the network feature maps for the later updating of new input. After initialization, the network only works on a line for each new frame in a streaming style.

Temporal-Shift Memory

- (a) Pyramid structure of network with 2×2 pooling and data overlaps. Tree2 repeats calculation by Tree1, T2 repeats T1's.
- (b) TSM of 3 layers for illustration. Cells include temporal status. Update is done only on the newest nodes (marked in red). Time cells shifts to one older state at all layers.
- (c) TSM avoids repeating calculation on overlapped data for sequential input after an initialization.





VI. Experiment

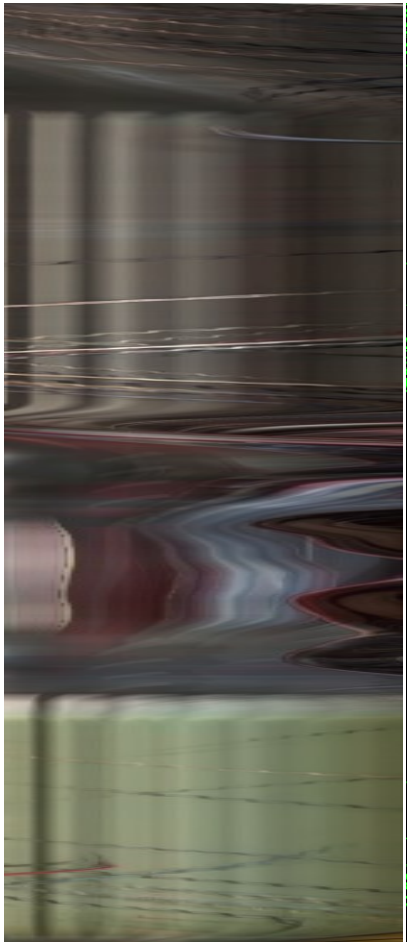
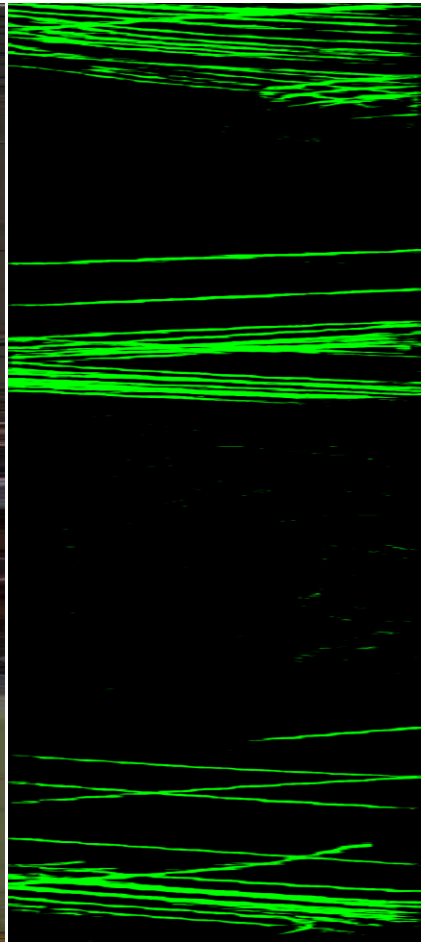
Dataset

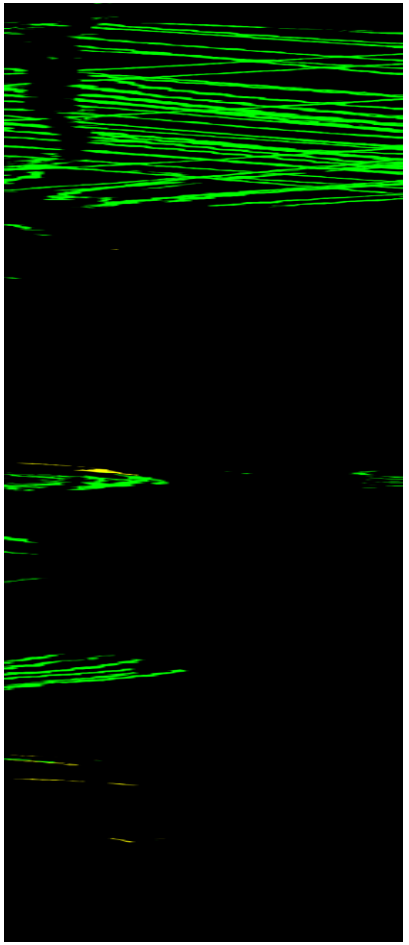
- We selected naturalistic driving videos and LiDAR data for pedestrian detection.

Sensor	Camera	Camera	LiDAR
Dataset	Village (TASI)	City (NYC)	KITTI
Training	44 clips of 5 sec. 6600 patches	32 clips of 5min with 6 colums each. Total 110,210/4 patches	36 clips of 6 sec. Total 1544 paches
Testing	10 clips of 5 sec. 1500 frames	6 clips of 5min with 6 columns each. Total 185,416 frames	3 clips of 6 sec. Total 447 frames

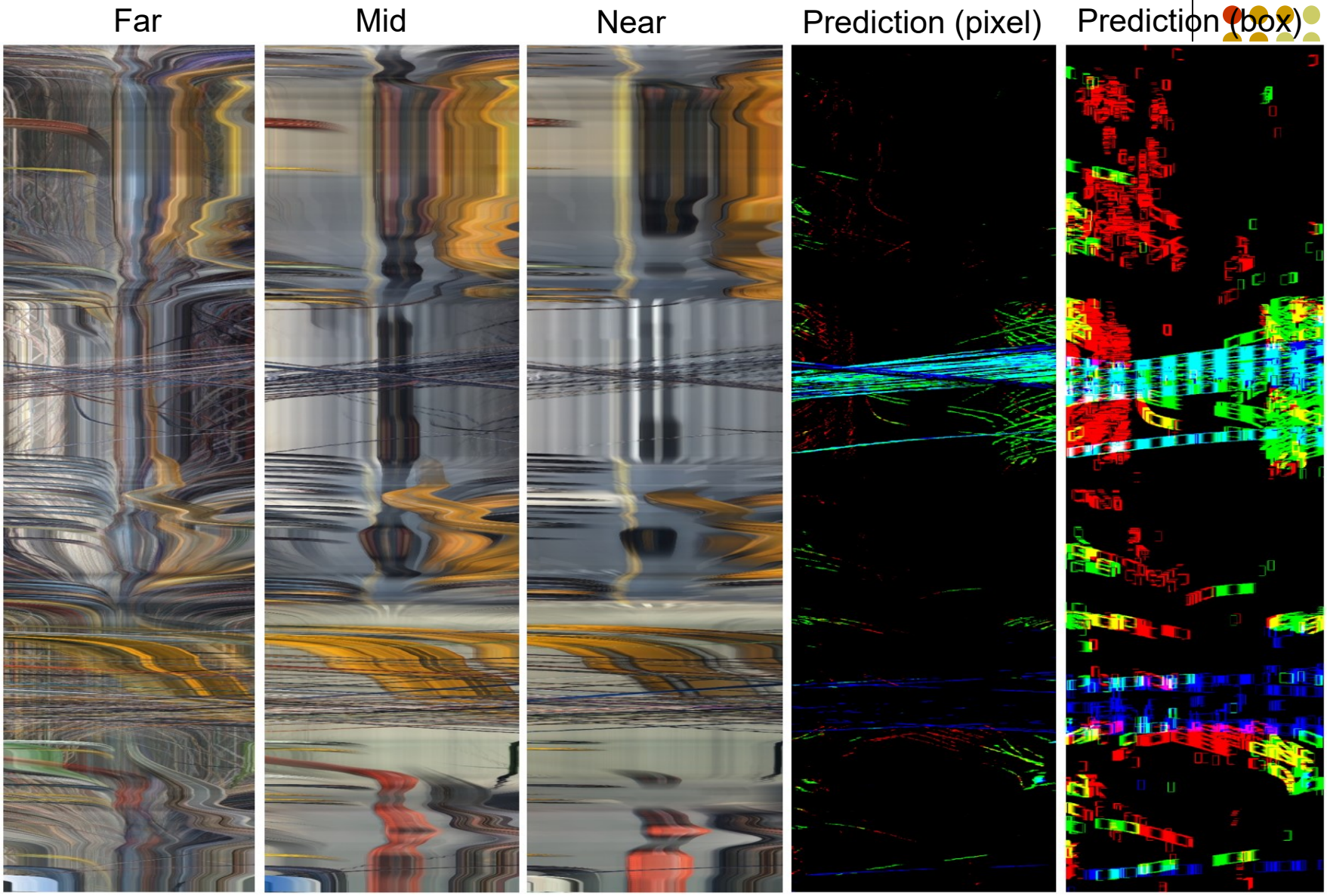
Quantitative Results

- MP
- SS
- MP
- SS

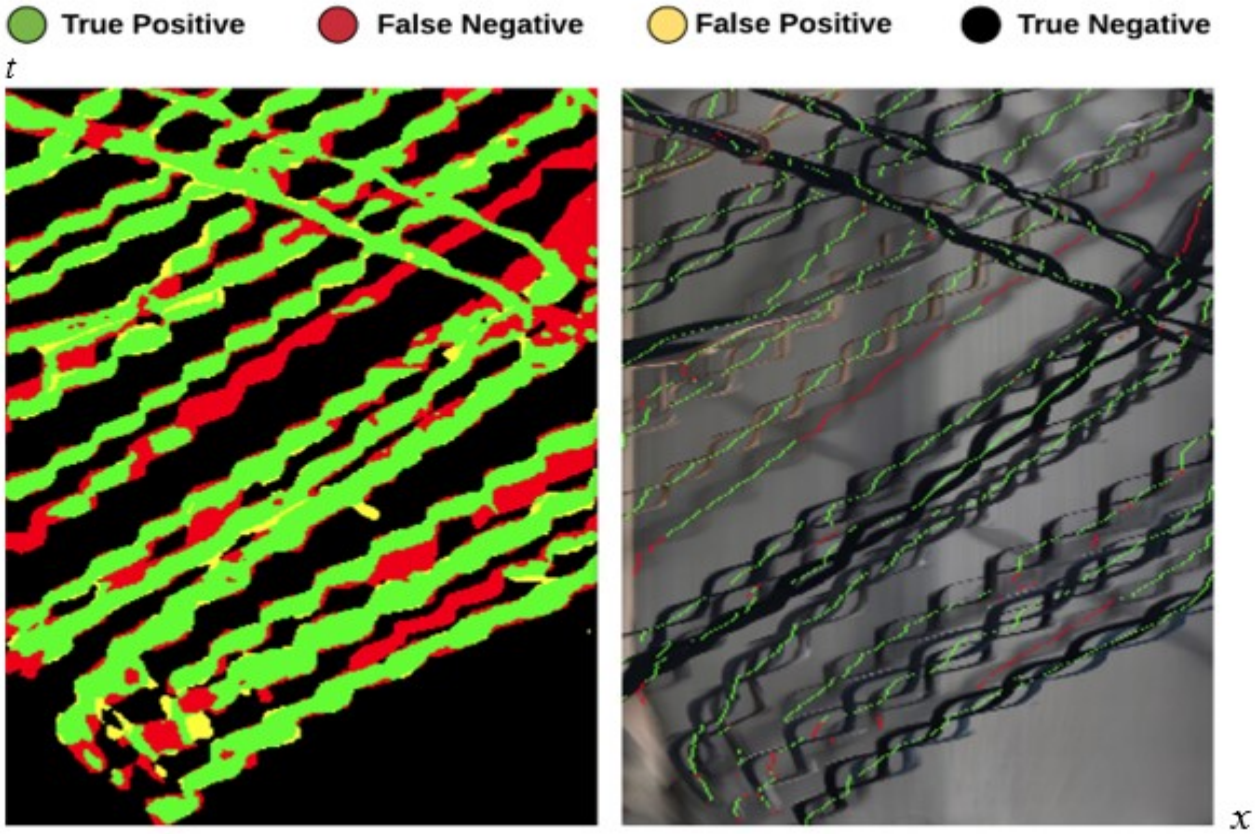





Quantitative Results



Evaluation

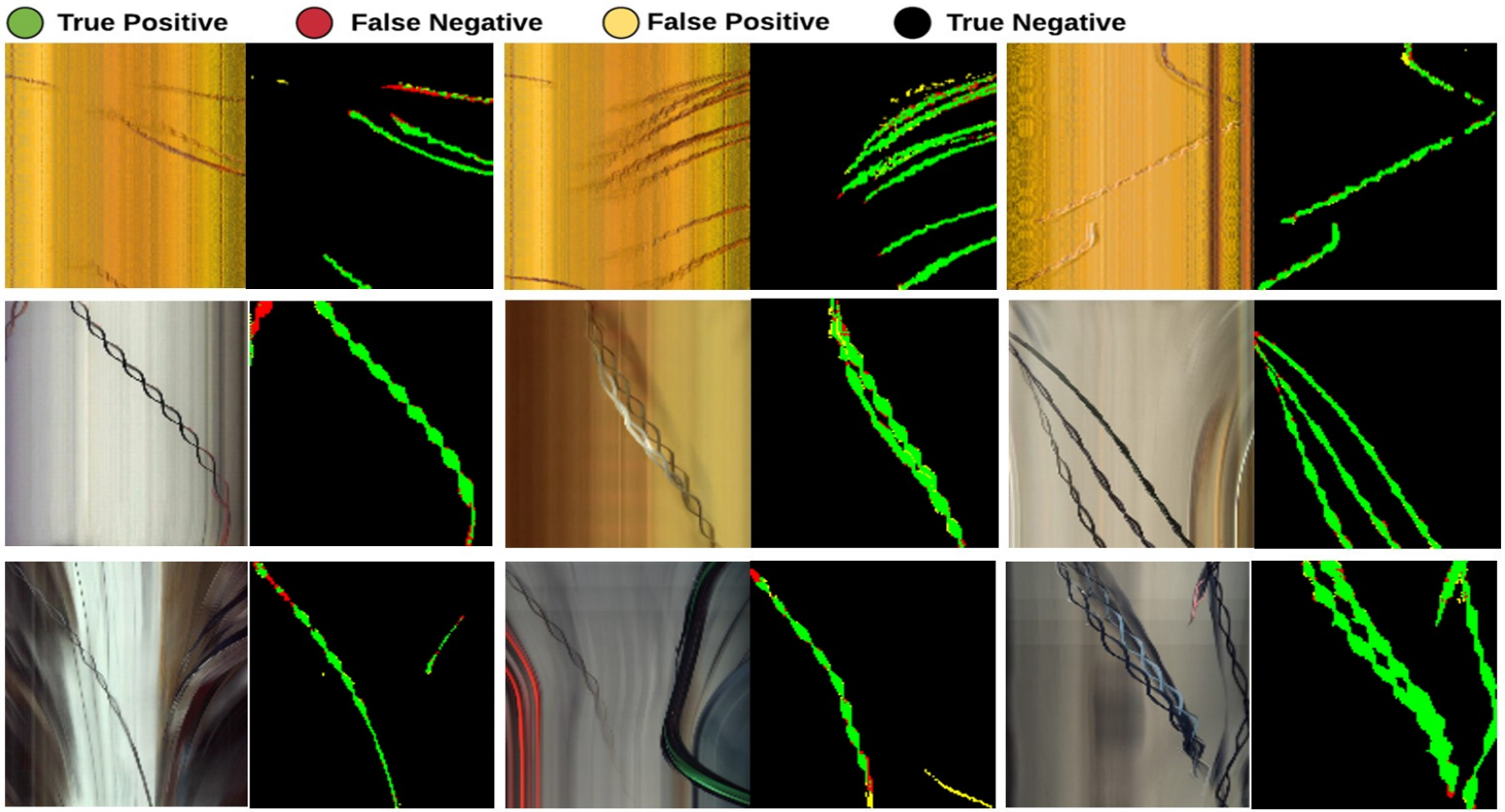


(a) Pixel-wise result in color;

(b) Skeletons of pedestrian traces from labeled ground truth for calculating frame wise pedestrian detection rate.

Evaluation

- LiDAR-based MP (top row) and camera-based MP with pixel wise semantic segmentation results.



Evaluation

- Accuracy of pedestrian trace at pixel level

Sensor	Dataset	Precision	Recall	F1	IoU	PA
LiDar	KITTI	0.743	0.650	0.691	0.538	0.964
Camera	Village(TASI)	0.820	0.906	0.861	0.759	0.982
Camera	City(far)	0.658	0.318	0.401	0.269	0.964
Camera	City(mid)	0.794	0.707	0.741	0.600	0.984
Camera	City(near)	0.892	0.928	0.910	0.834	0.996

- Accuracy of pedestrian trace at pixel level

Sensor	Dataset	Precision	Recall	F1	Det.R	PA
LiDar	KITTI	0.929	0.717	0.808	0.685	0.998
Camera	Village(TASI)	0.906	0.957	0.929	0.877	0.999
Camera	City(far)	0.655	0.511	0.553	0.398	0.996
Camera	City(mid)	0.829	0.819	0.819	0.700	0.999
Camera	City(near)	0.946	0.971	0.958	0.920	0.999

Comparison to state-of-the-art

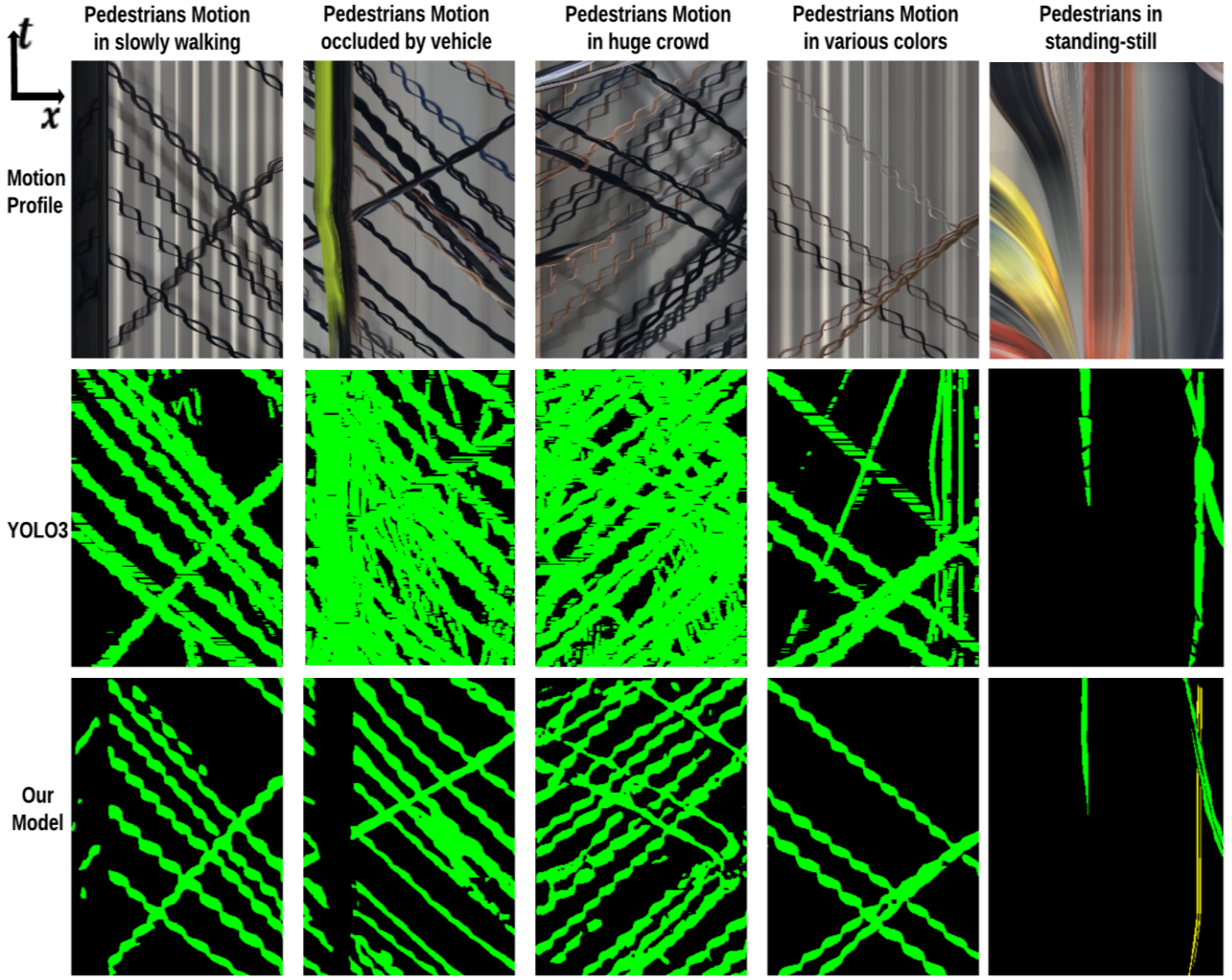
- Comparison of our method with YOLO3 and Motion Filter:

Model	DL	Motion	detection level	fps	latency (sec.)
YOLO3	✓	✗	bounding box	2.7	0.37
Motion Filter	✗	✓	bounding box	30	0.5
TSM	✓	✓	pixel	30	0.002

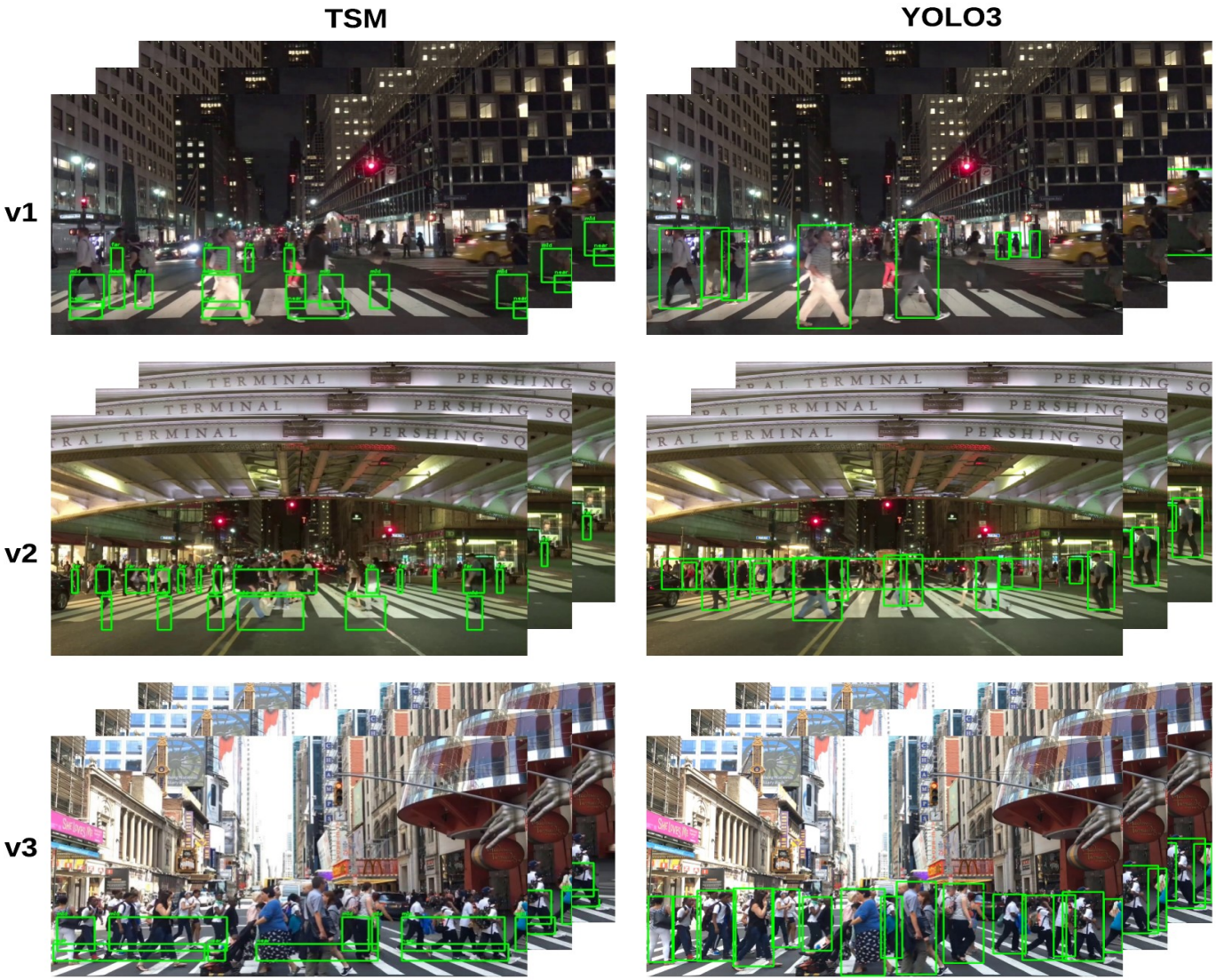
Comparison to state-of-the-art

Comparison of detecting results for the cases of:

- sparse pedestrians;
- occlusion, crowds;
- color similar to background;
- standing-still person.

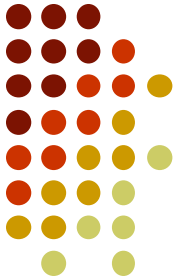


Comparison to YOLO3



Comparison to YOLO3

- (i) less data for computation.
- (ii) fast computing time (2ms) in frame advancing, much shorter than YOLO3 (370ms/fr.) on the same machine;
- (iii) preserving a better motion continuity, while YOLO3 leaves some gaps along walking chains.
- (iv) higher precision in body width and leg span than bounding boxes.



VII. Conclusion

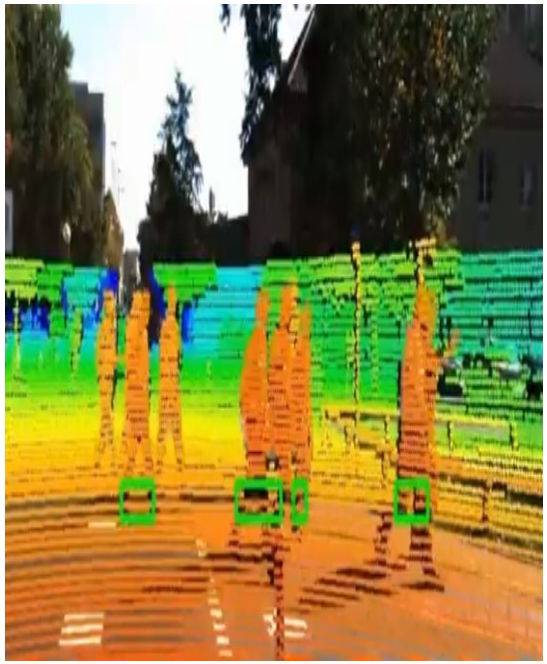
This paper presents pedestrian detection in driving video with a high efficiency based on the **minimum data size** and a **low variation of motion patterns**.

The **algorithm complexity** of pedestrian detection is significantly reduced, and the **detection speed** is drastically faster than current pedestrian detection based on spatial analysis of their shape.

VIII. Appendix

Appendix: Video Results

- https://www.youtube.com/watch?v=a-ePUpbIZKw&feature=youtu.be
- https://www.youtube.com/watch?v=BAOZOHXFq8o&feature=youtu.be
- https://www.youtube.com/watch?v=liFiHIdnsoI&feature=youtu.be



Thank you!

- Questions?