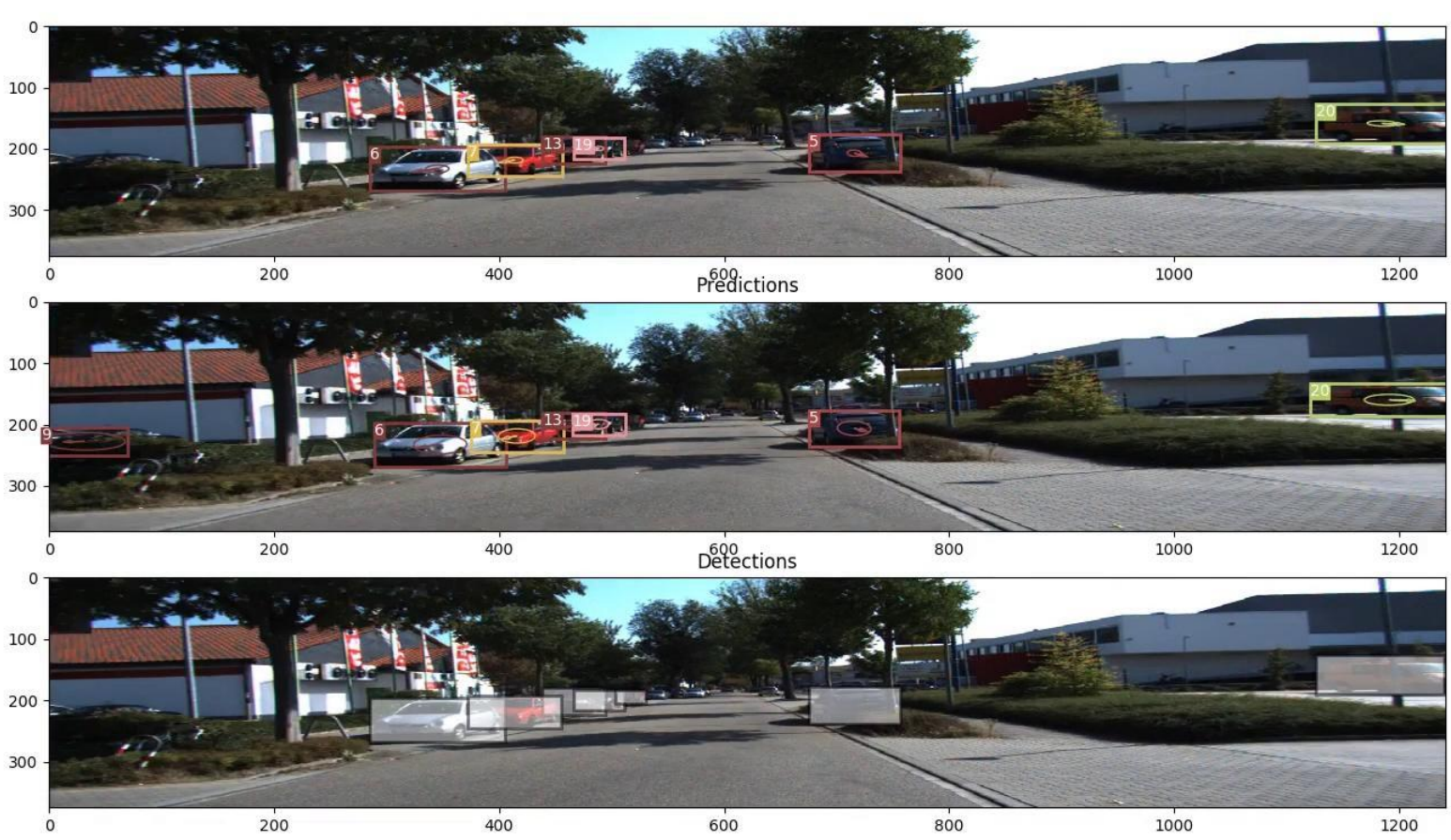
Multi-Object Tracking

Mihir Patel

Formulating the Problem

- Goal: Track objects in a video
- Input: Observations per frame ($\mathbf{O}_i$)
  - Image
- Output: State per object ($\mathbf{S}_i$)
  - Location, speed, color, ...
- Formally, maximize $P(\mathbf{S}_{1:t} \mid \mathbf{O}_{1:t})$
  - Find optimal sequence of states
- Evaluate on False Positives, False Negatives, Switches
- Upstream task for navigation


/cvgl2/u/mihirp/depth_tracking/results/new_aligned_reid/v10_2_final_train
frame 27 detections 7 confirmed 6 unconfirmed 1
6
13
19
5
20
Predictions
9
6
13
19
5
20
Detections
0
100
200
300
0
200
400
600
800
1000
1200

Types of Trackers

- Online vs. offline: can we see the future?
  - We formulate online tracking
- Detection based vs. detection free: what is part of tracking?
- 2D vs. 3D: how do we model the world?

Key Operations

- Predict: $P(\mathbf{S}_{t+1} \mid \mathbf{S}_t)$
  - What is the expected next step?
- Update: $P(\mathbf{S}_t \mid \mathbf{O}_t)$
  - How do observations update our model?
- Initiation: When have we seen a new object?
- Termination: When has an object left?

Predict

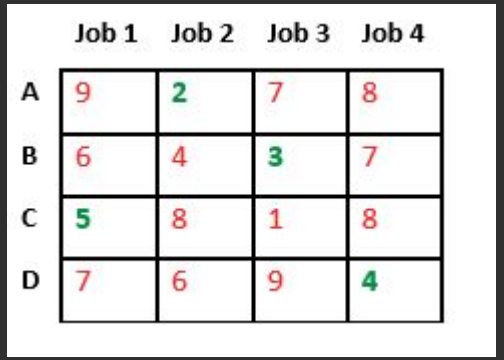
- Want to model state of object
- Track position, velocity, acceleration, appearance and predict future states
- Filters (ie Kalman, particle) allow tracking with uncertainty

Update

- Update motion models given object observation
- Which object is which?
- Worker assignment problem with Hungarian algorithm
- Add dummy node for unmatched detections and tracks

	Job 1	Job 2	Job 3	Job 4
A	9	2	7	8
B	6	4	3	7
C	5	8	1	8
D	7	6	9	4

Cost Matrices



	Job 1	Job 2	Job 3	Job 4
A	9	2	7	8
B	6	4	3	7
C	5	8	1	8
D	7	6	9	4

- How good is a track-detection pair?
- Intersection over Union (IOU) of states
- Visual appearance similarity
  - Color histogram: well studied but loses spatial info
  - Optical flow: powerful but susceptible to occlusion
  - Learned approach: embed into some vector space
- How do we combine?
  - Concatenate, sum, product, cascade, learned

Initiation

- High recall: look for n matches in a row before confirming object
- High precision detector: always initiate on detection
- Estimate start state
  - Optical flow with velocity

Termination

- High recall: miss for 1 frame
- High precision: miss for n frames
- State based: uncertainty limit based approach

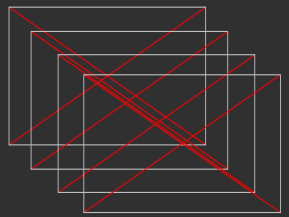
Open Ended Questions

- Hard Constraints: Cars can’t exist in a wall
- Social Cues: Objects respond to each other
- Occlusion Handling: How are states and cost matrices affected?

Tracking on JR

- Cute robot
- 2D stereo camera, 3D LIDAR data
- Drives around people and tries to not hit them

Images



AlignREID

2D Features

MLP

Feature Cost

Detector

2D Dets

IOU Cost

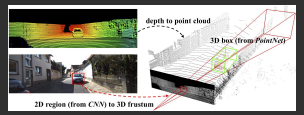
Entropy

Point Cloud

FPointNet

3D Dets

3D Features



Total Cost

Hungarian

Output

Tracks

JPDA

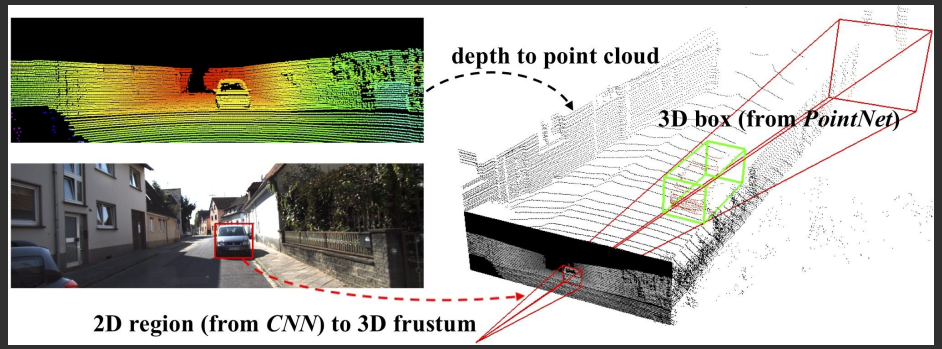
Tracks

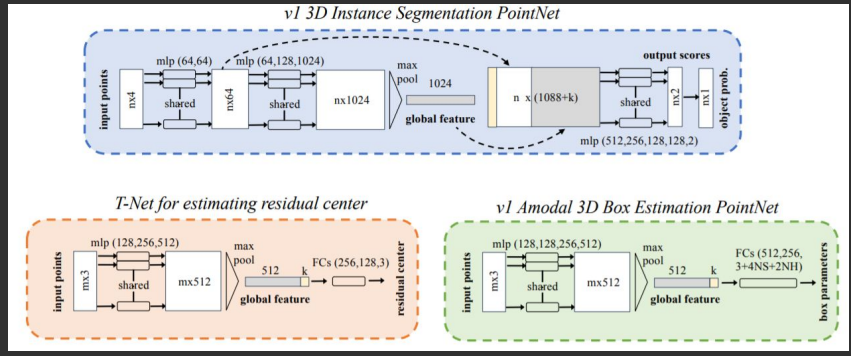
Predict with IMM

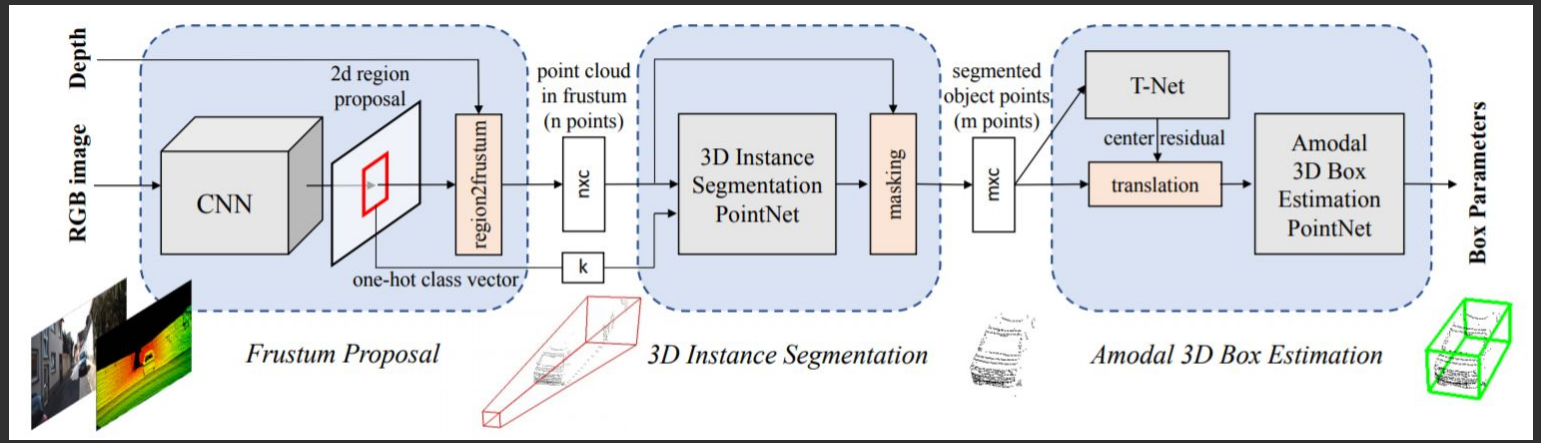
Potential Work

- How useful is this machinery?
  - How much is academic selling vs. progress?
- How can we combine two cost matrices?
  - Orderless convolutions?
- How can we handle occlusion?
  - How do appearance features change?
- How do we fuse 2D and 3D sources?

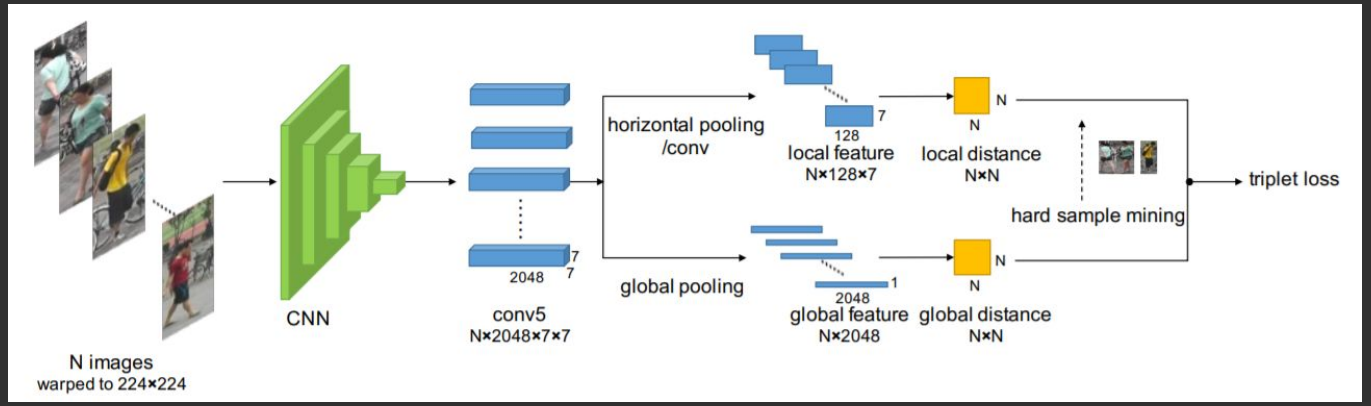
Frustum-PointNet

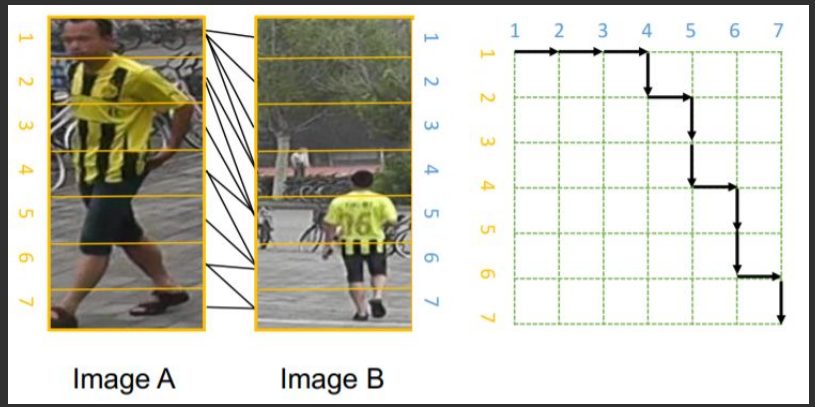






AlignREID


N images
warped to 224×224
CNN
2048
7
7
conv5
N×2048×7×7
horizontal pooling
/conv
global pooling
128
7
local feature
N×128×7
N
N
local distance
N×N
2048
1
global feature
N×2048
N
N
global distance
N×N
hard sample mining
triplet loss


1 2 3 4 5 6 7
1 2 3 4 5 6 7
1 2 3 4 5 6 7
1 2 3 4 5 6 7
Image A
Image B