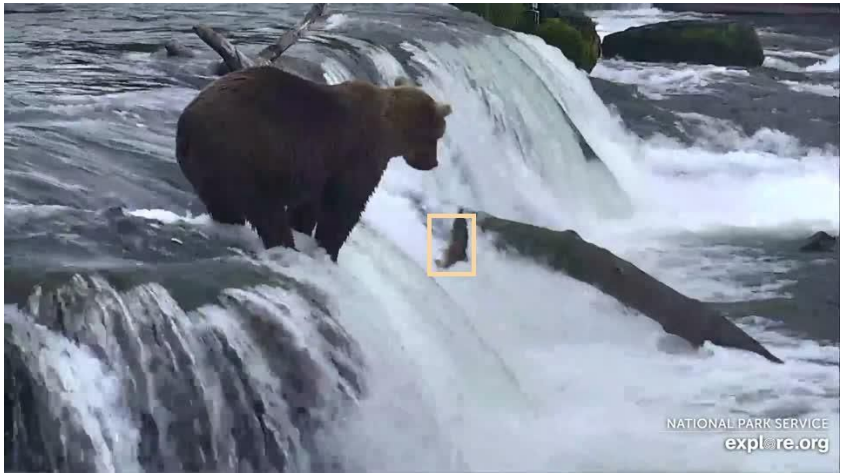
Self-Supervised Multi-Object Tracking with Cross-Input Consistency

Favyen Bastani, Songtao He, Sam Madden (MIT CSAIL)


expl re.org

Labeling tracks is expensive

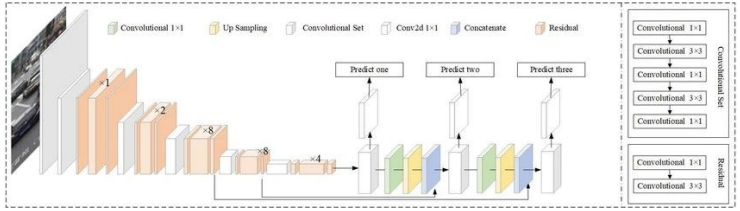


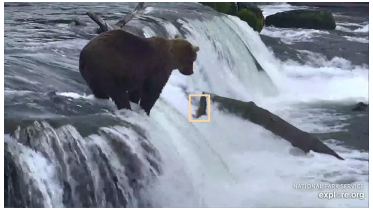
Object Detection
Training Example

Object Tracking
Training Example

Can we take advantage of unlabeled video?

1. Train an object detector (use relatively cheap image-level supervision)





2. Leverage large amounts of unlabeled video for training a robust tracker?

Our Approach: Training a Robust Tracker on Unlabeled Video

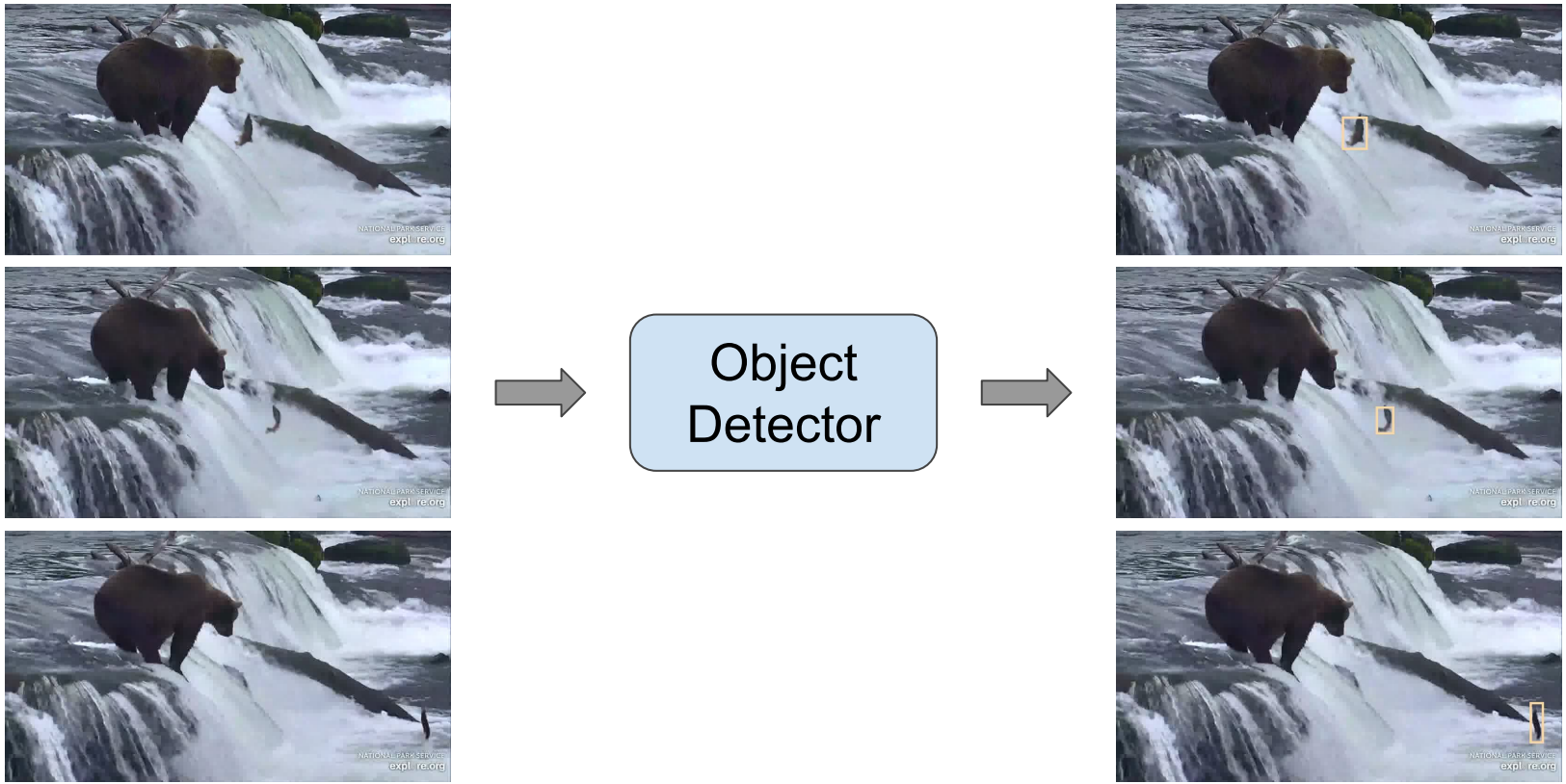
Object Detector

+

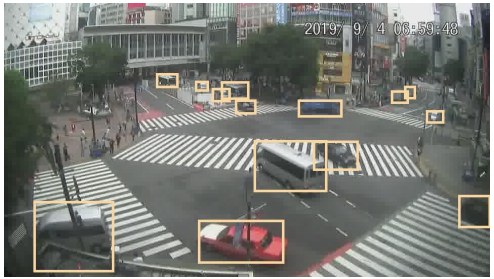
Robust Multi-Object Tracking Model

Despite training only on unlabeled video, our unsupervised approach performs competitively with several fully supervised methods that train on video-level labels in MOT17 and Kitti!

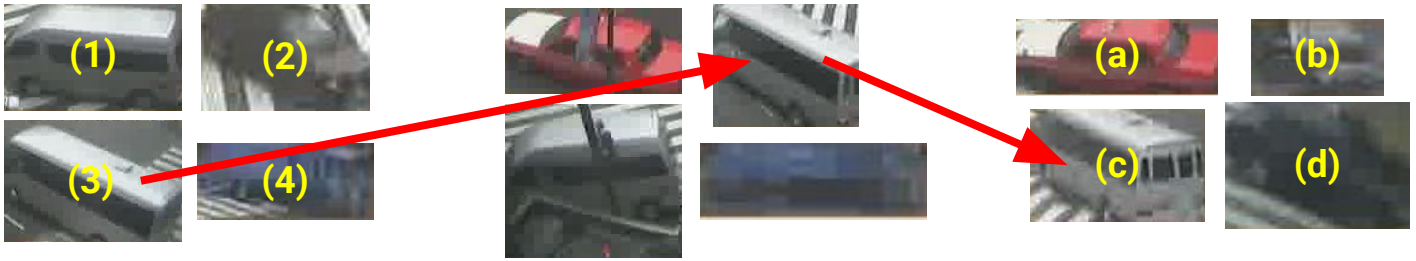
Input: Unlabeled Video + Predicted Object Bounding Boxes



Input-Hiding Scheme — Visual-Spatial Hiding

	Frame 1	Frame 2	Frame 3
Original Video Sequence			
Input 1: Visual Information			
Input 2: Spatial Information	x=93 y=681 w=176 h=103 x=950 y=459 w=143 h=75 x=652 y=333 w=78 h=87 x=899 y=337 w=62 h=38	x=490 y=429 w=211 h=140 x=740 y=575 w=235 h=127 x=821 y=335 w=63 h=47 x=848 y=325 w=51 h=38	x=458 y=313 w=62 h=49 x=847 y=325 w=53 h=37 x=819 y=336 w=62 h=47 x=609 y=282 w=75 h=48

Input 1:
Visual Information



	(a)	(b)	(c)	(d)
(1)	0.3	0.2	0.0	0.1
(2)	0.0	0.4	0.0	0.2
(3)	0.1	0.0	0.9	0.0
(4)	0.1	0.2	0.1	0.1

Use inconsistency between matrices as a learning signal!

	(a)	(b)	(c)	(d)
(1)	0.0	0.1	0.2	0.2
(2)	0.1	0.1	0.1	0.1
(3)	0.0	0.3	0.7	0.0
(4)	0.4	0.0	0.1	0.1

Input 2:
Spatial Information

(1)	(2)	(3)	(4)
x=93	x=950	x=652	x=899
y=681	y=459	y=333	y=337
w=176	w=143	w=78	w=62
h=103	h=75	h=87	h=38

x=490	x=740	x=821	x=848
y=429	y=575	y=335	y=325
w=211	w=235	w=63	w=51
h=140	h=127	h=47	h=38

(a)	(b)	(c)	(d)
x=458	x=847	x=819	x=609
y=313	y=325	y=336	y=282
w=62	w=53	w=62	w=75
h=49	h=37	h=47	h=48

Why wouldn’t model create arbitrary tracking outputs?

Saturation: 16

Saturation: 18

???

Frame 1

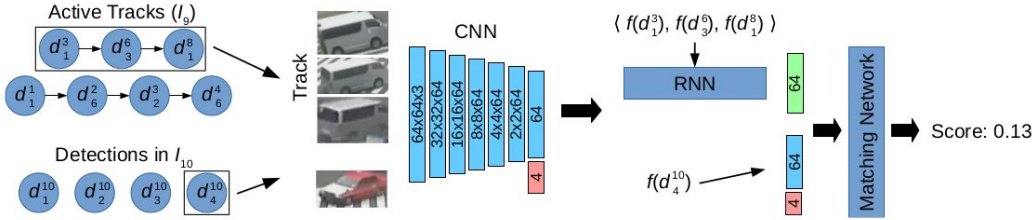
x=93	x=950	x=652	x=899
y=681	y=459	y=333	y=337
w=176	w=143	w=78	w=62
h=103	h=75	h=87	h=38

Frame 2

x=490	x=740	x=821	x=848
y=429	y=575	y=335	y=325
w=211	w=235	w=63	w=51
h=140	h=127	h=47	h=38

Frame 3

x=458	x=847	x=819	x=609
y=313	y=325	y=336	y=282
w=62	w=53	w=62	w=75
h=49	h=37	h=47	h=48



Model Architecture

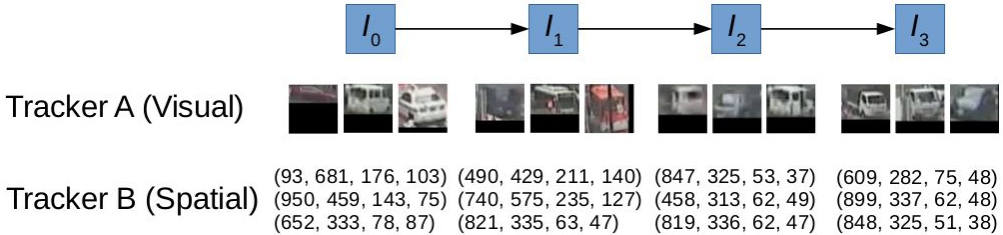
$$\begin{bmatrix} 0.3 & 0.2 & 0.0 & 0.1 \\ 0.0 & 0.4 & 0.0 & 0.2 \\ 0.1 & 0.0 & 0.9 & 0.0 \\ 0.1 & 0.2 & 0.1 & 0.1 \end{bmatrix}$$

Transition Matrix

```
array([[0.40525553, 0.58487146, 0.35028426],
       [0.17897008, 0.99813136, 0.65189838],
       [0.53625954, 0.51995857, 0.71777389]])

array([[0.42718929, 0.52618008, 0.84496057],
       [0.8125701 , 0.29344281, 0.09825419],
       [0.19191158, 0.11864856, 0.82009976]])
```

Consistency Loss Function

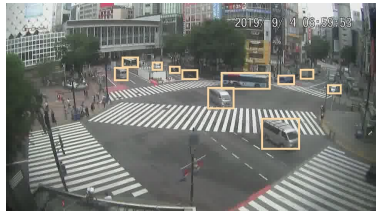


Input-Hiding Scheme

Frame i - 2

Frame i - 1

Frame i

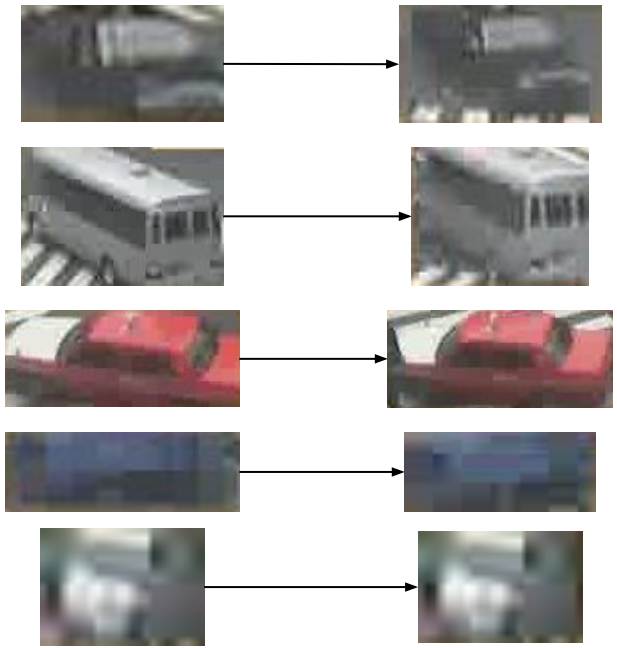


Frame i + 1

Frame i + 2

Frame i - 2 | Frame i - 1 | Frame i | Frame i + 1 | Frame i + 2

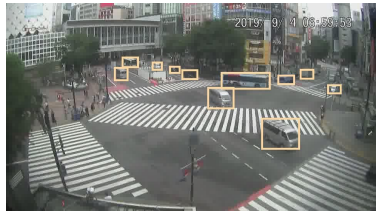
Track Prefixes



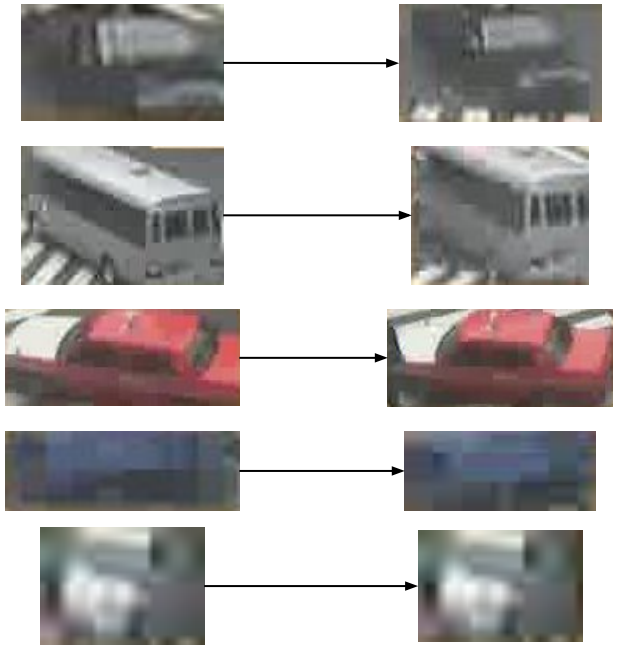
Frame i - 2

Frame i - 1

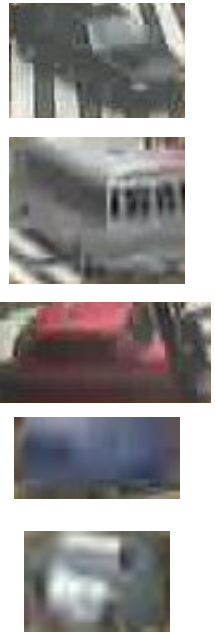
Frame i



Active Track Prefixes



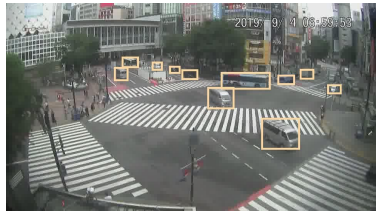
Current Detections



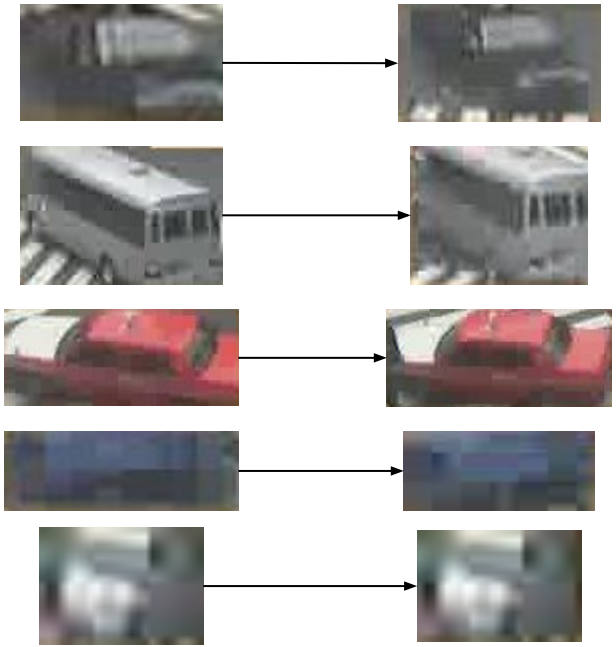
Frame i - 2

Frame i - 1

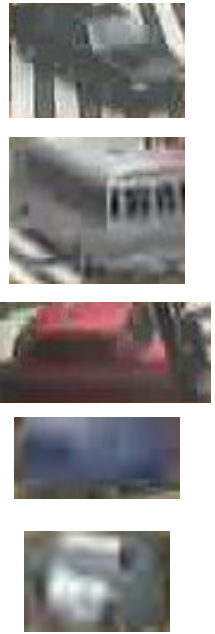
Frame i

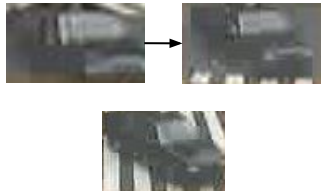


Active Track Prefixes



Current Detections





Model

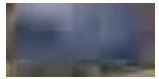
Same Object

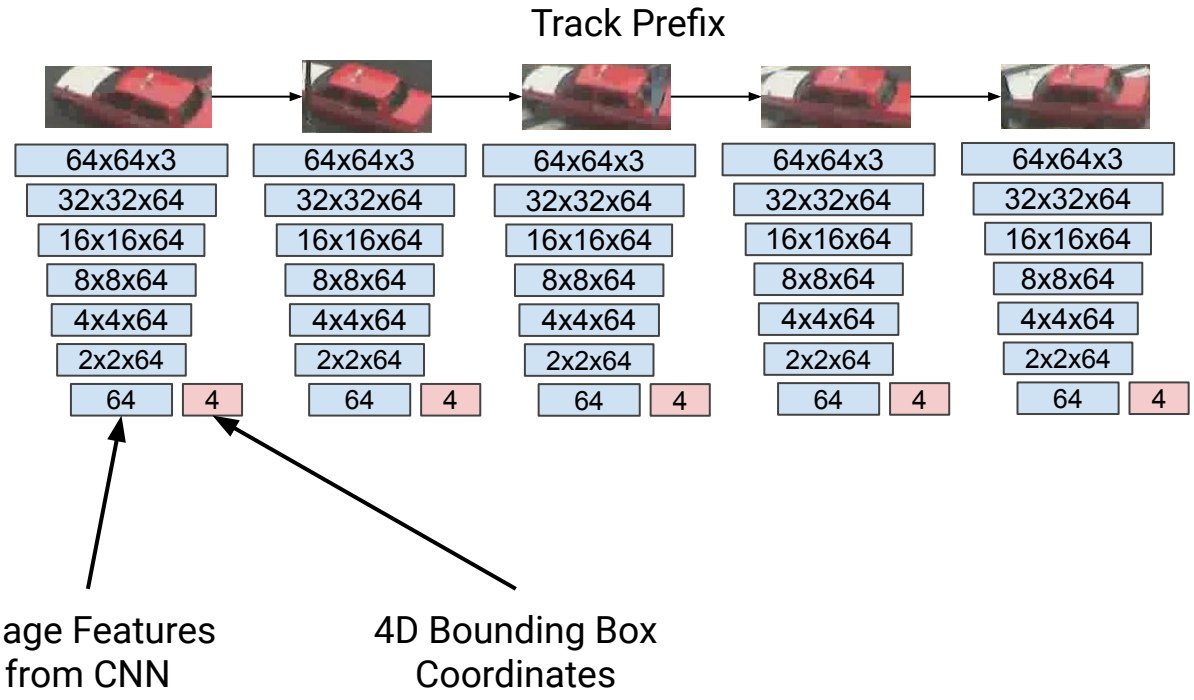
Model

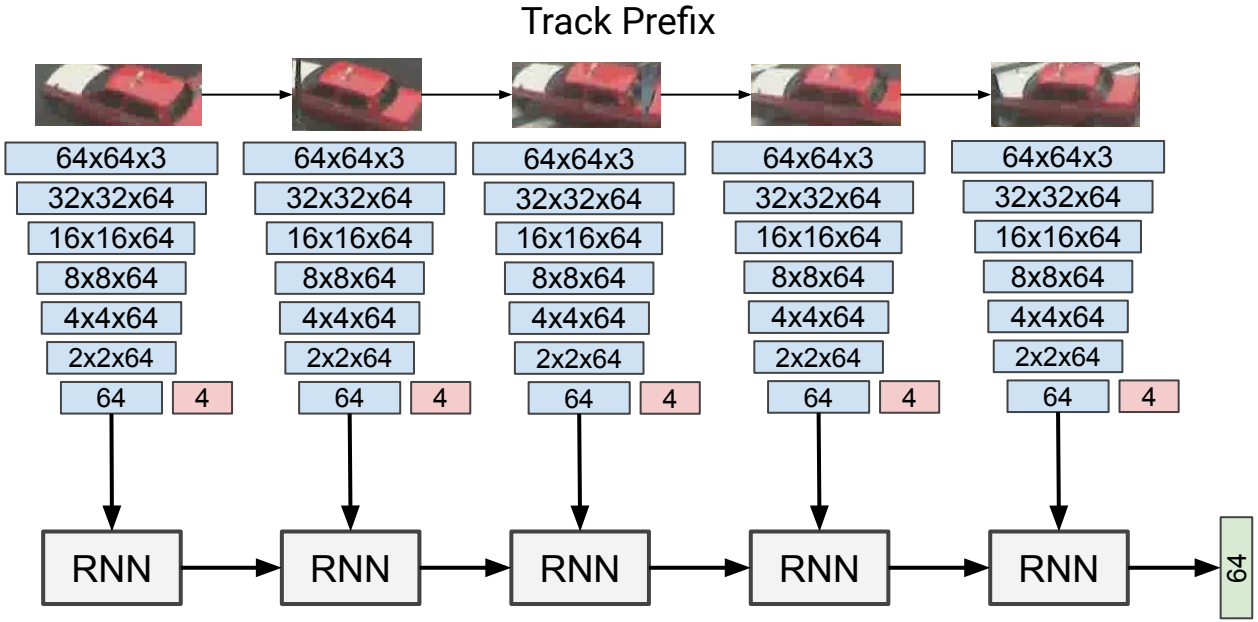
Different Objects

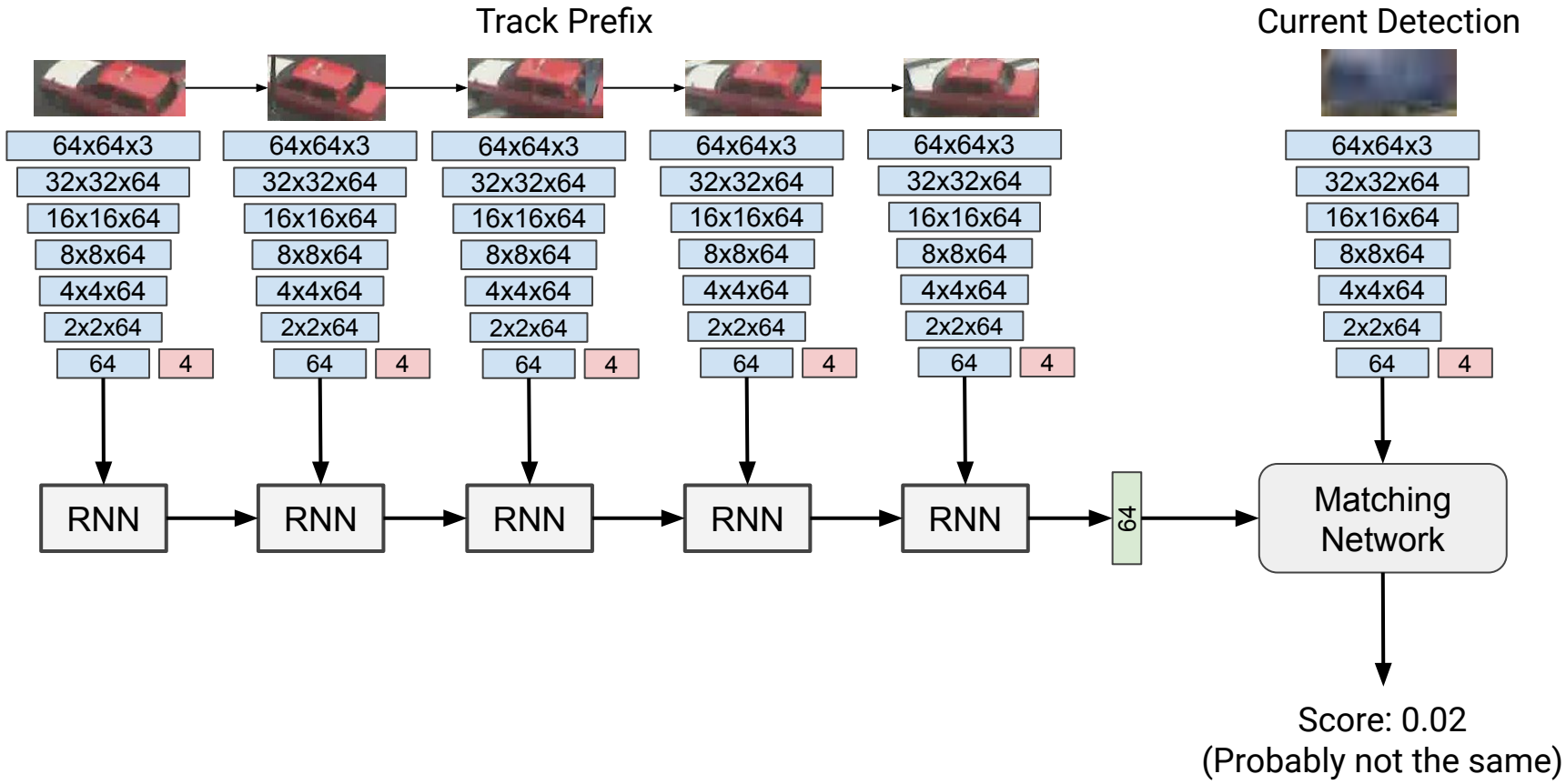
Track Prefix

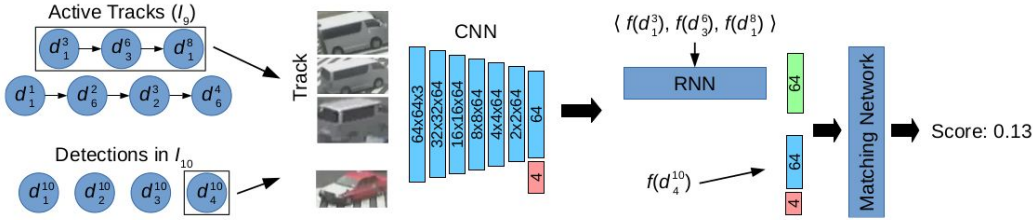
Current Detection




Track Prefix
Current Detection
64x64x3
32x32x64
16x16x64
8x8x64
4x4x64
2x2x64
64
4
Image Features from CNN
4D Bounding Box Coordinates


Track Prefix
Current Detection
64x64x3
32x32x64
16x16x64
8x8x64
4x4x64
2x2x64
64
4
RNN
64


Track Prefix
Current Detection
64x64x3
32x32x64
16x16x64
8x8x64
4x4x64
2x2x64
64
4
RNN
64
Matching Network
Score: 0.02
(Probably not the same)



Model Architecture

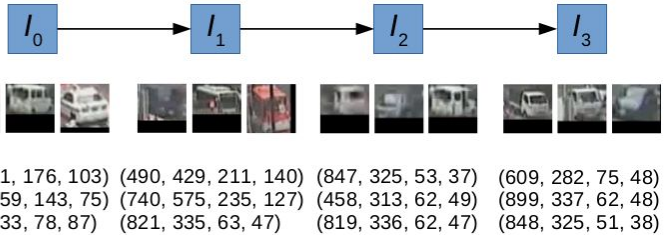
$$\begin{bmatrix} 0.3 & 0.2 & 0.0 & 0.1 \\ 0.0 & 0.4 & 0.0 & 0.2 \\ 0.1 & 0.0 & 0.9 & 0.0 \\ 0.1 & 0.2 & 0.1 & 0.1 \end{bmatrix}$$

Transition Matrix

```
array([[0.40525553, 0.58487146, 0.35028426],
       [0.17897008, 0.99813136, 0.65189838],
       [0.53625954, 0.51995857, 0.71777389]])
```

```
array([[0.42718929, 0.52618008, 0.84496057],
       [0.8125701 , 0.29344281, 0.09825419],
       [0.19191158, 0.11864856, 0.82009976]])
```

Consistency Loss Function



Input-Hiding Scheme

Training Example

Frame 1

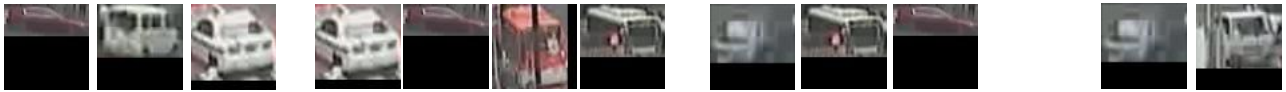
Frame 2

Frame 3

Frame 4

Training Example

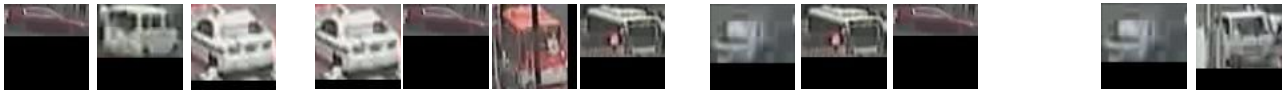
Input 1



Input 2

1A	1B	1C	2A	2B	2C	2D	3A	3B	3C	4A	4B
x=93	x=950	x=652	x=490	x=740	x=821	x=848	x=458	x=847	x=819	x=609	x=899
y=681	y=459	y=333	y=429	y=575	y=335	y=325	y=313	y=325	y=336	y=282	y=337
w=176	w=143	w=78	w=211	w=235	w=63	w=51	w=62	w=53	w=62	w=75	w=62
h=103	h=75	h=87	h=140	h=127	h=47	h=38	h=49	h=37	h=47	h=48	h=38

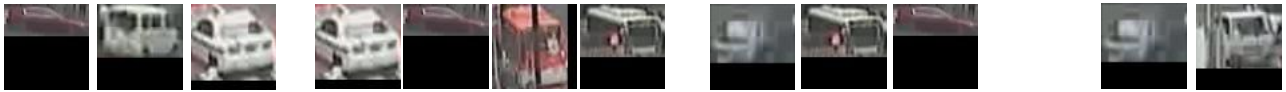
Input 1



Input 2

1A	1B	1C	2A	2B	2C	2D	3A	3B	3C	4A	4B
x=93	x=950	x=652	x=490	x=740	x=821	x=848	x=458	x=847	x=819	x=609	x=899
y=681	y=459	y=333	y=429	y=575	y=335	y=325	y=313	y=325	y=336	y=282	y=337
w=176	w=143	w=78	w=211	w=235	w=63	w=51	w=62	w=53	w=62	w=75	w=62
h=103	h=75	h=87	h=140	h=127	h=47	h=38	h=49	h=37	h=47	h=48	h=38

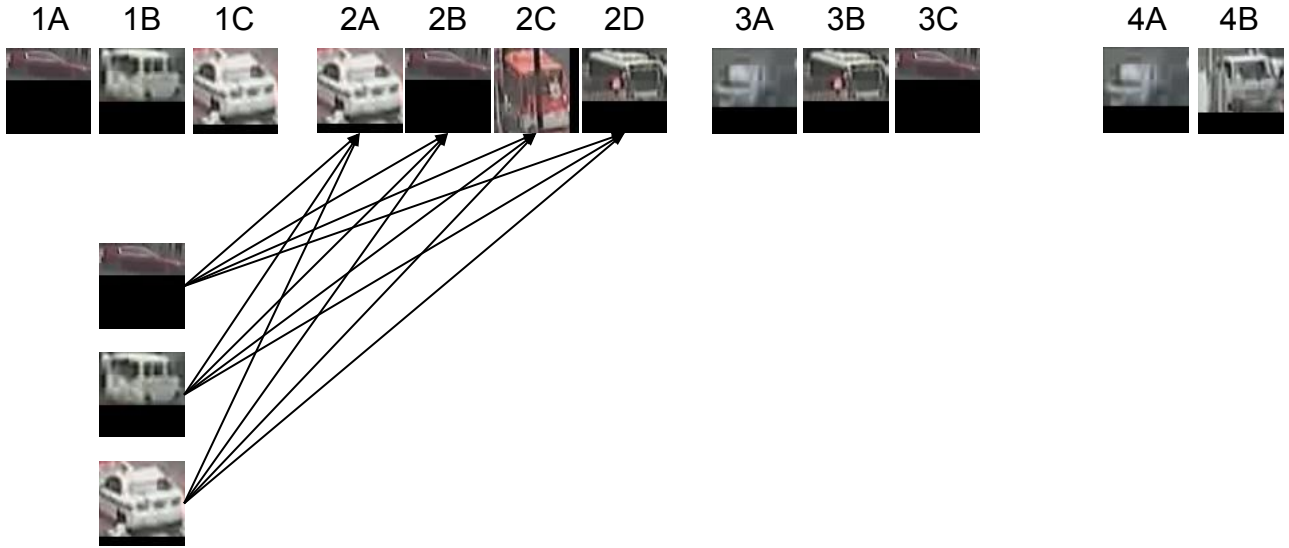
Input 1



Input 2

1A	1B	1C	2A	2B	2C	2D	3A	3B	3C	4A	4B
x=93	x=950	x=652	x=490	x=740	x=821	x=848	x=458	x=847	x=819	x=609	x=899
y=681	y=459	y=333	y=429	y=575	y=335	y=325	y=313	y=325	y=336	y=282	y=337
w=176	w=143	w=78	w=211	w=235	w=63	w=51	w=62	w=53	w=62	w=75	w=62
h=103	h=75	h=87	h=140	h=127	h=47	h=38	h=49	h=37	h=47	h=48	h=38

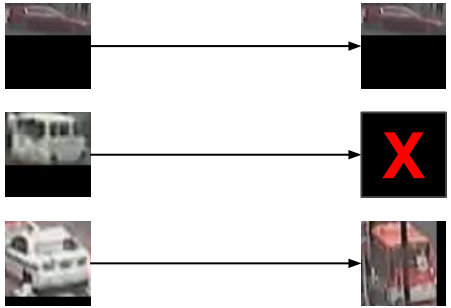
Input 1



Input 2

1A	1B	1C	2A	2B	2C	2D	3A	3B	3C	4A	4B
x=93	x=950	x=652	x=490	x=740	x=821	x=848	x=458	x=847	x=819	x=609	x=899
y=681	y=459	y=333	y=429	y=575	y=335	y=325	y=313	y=325	y=336	y=282	y=337
w=176	w=143	w=78	w=211	w=235	w=63	w=51	w=62	w=53	w=62	w=75	w=62
h=103	h=75	h=87	h=140	h=127	h=47	h=38	h=49	h=37	h=47	h=48	h=38

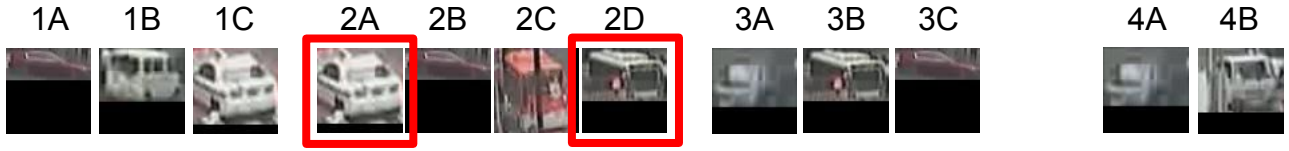
Input 1

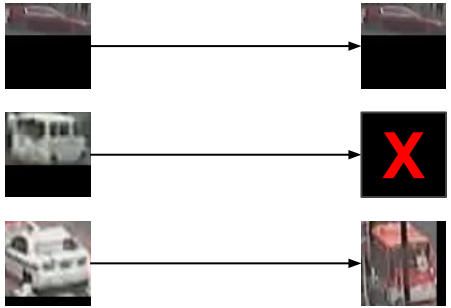


Input 2

x=93	x=950	x=652	x=490	x=740	x=821	x=848	x=458	x=847	x=819	x=609	x=899
y=681	y=459	y=333	y=429	y=575	y=335	y=325	y=313	y=325	y=336	y=282	y=337
w=176	w=143	w=78	w=211	w=235	w=63	w=51	w=62	w=53	w=62	w=75	w=62
h=103	h=75	h=87	h=140	h=127	h=47	h=38	h=49	h=37	h=47	h=48	h=38

Input 1

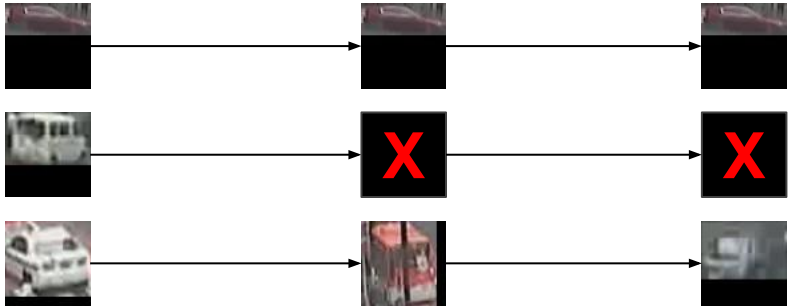




Input 2

x=93	x=950	x=652	x=490	x=740	x=821	x=848	x=458	x=847	x=819	x=609	x=899
y=681	y=459	y=333	y=429	y=575	y=335	y=325	y=313	y=325	y=336	y=282	y=337
w=176	w=143	w=78	w=211	w=235	w=63	w=51	w=62	w=53	w=62	w=75	w=62
h=103	h=75	h=87	h=140	h=127	h=47	h=38	h=49	h=37	h=47	h=48	h=38

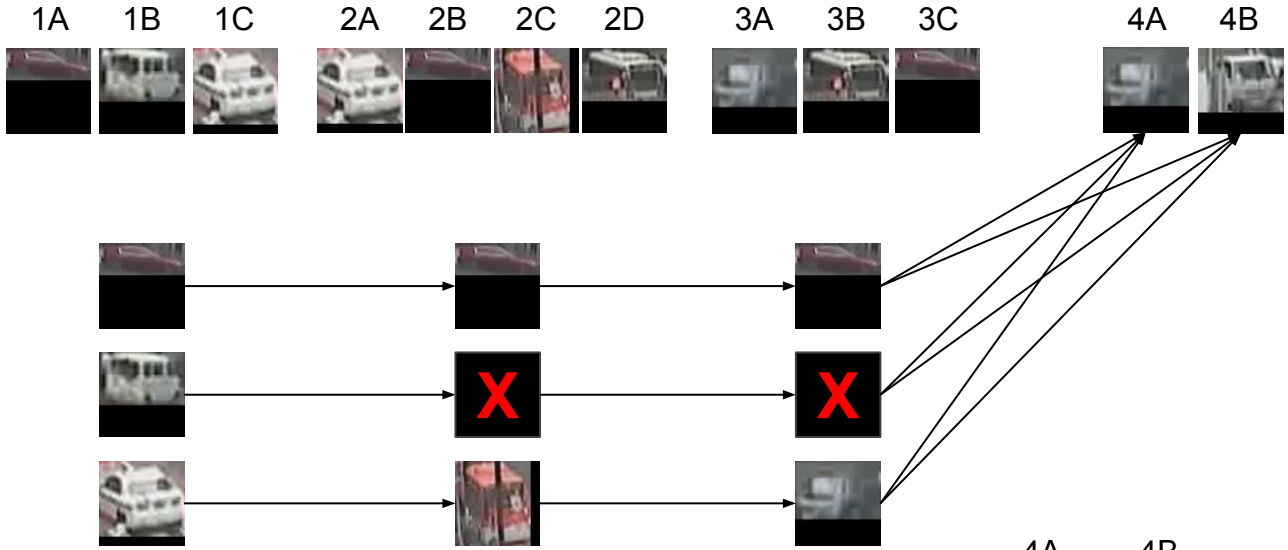
Input 1



Input 2

x=93	x=950	x=652	x=490	x=740	x=821	x=848	x=458	x=847	x=819	x=609	x=899
y=681	y=459	y=333	y=429	y=575	y=335	y=325	y=313	y=325	y=336	y=282	y=337
w=176	w=143	w=78	w=211	w=235	w=63	w=51	w=62	w=53	w=62	w=75	w=62
h=103	h=75	h=87	h=140	h=127	h=47	h=38	h=49	h=37	h=47	h=48	h=38

Input 1



	4A	4B
Track 1A	-23	-2
Track 1B	-1	-1
Track 1C	16	-15

Input 2

1A	1B	1C	2A	2B	2C	2D	3A	3B	3C	4A	4B
x=93	x=950	x=652	x=490	x=740	x=821	x=848	x=458	x=847	x=819	x=609	x=899
y=681	y=459	y=333	y=429	y=575	y=335	y=325	y=313	y=325	y=336	y=282	y=337
w=176	w=143	w=78	w=211	w=235	w=63	w=51	w=62	w=53	w=62	w=75	w=62
h=103	h=75	h=87	h=140	h=127	h=47	h=38	h=49	h=37	h=47	h=48	h=38

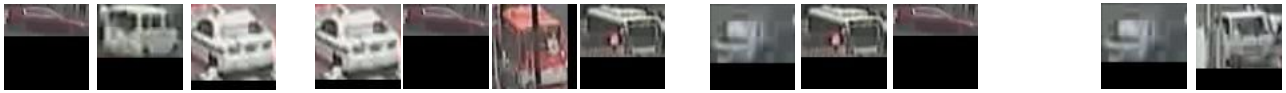
Input 1

	4A	4B
Track 1A	-23	-2
Track 1B	-1	-1
Track 1C	16	-15

Input 2

	1A	1B	1C	2A	2B	2C	2D	3A	3B	3C	4A	4B
x	93	950	652	490	740	821	848	458	847	819	609	899
y	681	459	333	429	575	335	325	313	325	336	282	337
w	176	143	78	211	235	63	51	62	53	62	75	62
h	103	75	87	140	127	47	38	49	37	47	48	38

Input 1



	4A	4B
Track 1A	-23	-2
Track 1B	-1	-1
Track 1C	16	-15

Row-wise Softmax

	0.0	1.0
	0.5	0.5
	1.0	0.0

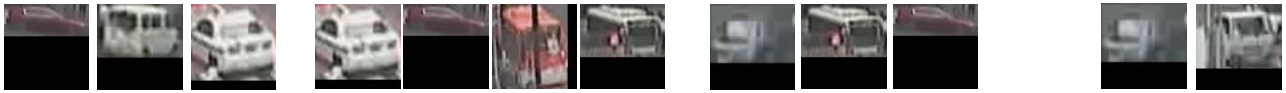
Column-wise Softmax

	0.0	0.3
	0.0	0.7
	1.0	0.0

Input 2

x=93	x=950	x=652	x=490	x=740	x=821	x=848	x=458	x=847	x=819	x=609	x=899
y=681	y=459	y=333	y=429	y=575	y=335	y=325	y=313	y=325	y=336	y=282	y=337
w=176	w=143	w=78	w=211	w=235	w=63	w=51	w=62	w=53	w=62	w=75	w=62
h=103	h=75	h=87	h=140	h=127	h=47	h=38	h=49	h=37	h=47	h=48	h=38

Input 1



	4A	4B
Track 1A	-23	-2
Track 1B	-1	-1
Track 1C	16	-15

Row-wise Softmax

	0.0	1.0
	0.5	0.5
	1.0	0.0

Column-wise Softmax

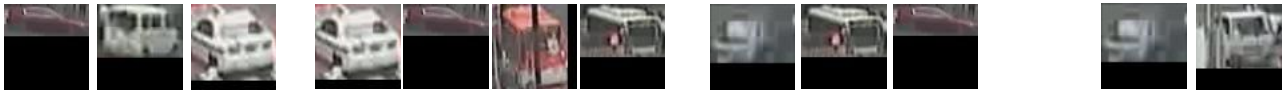
	0.0	0.3
	0.0	0.7
	1.0	0.0

0.0	0.3	
0.0	0.5	
1.0	0.0	

Input 2

x=93	x=950	x=652	x=490	x=740	x=821	x=848	x=458	x=847	x=819	x=609	x=899
y=681	y=459	y=333	y=429	y=575	y=335	y=325	y=313	y=325	y=336	y=282	y=337
w=176	w=143	w=78	w=211	w=235	w=63	w=51	w=62	w=53	w=62	w=75	w=62
h=103	h=75	h=87	h=140	h=127	h=47	h=38	h=49	h=37	h=47	h=48	h=38

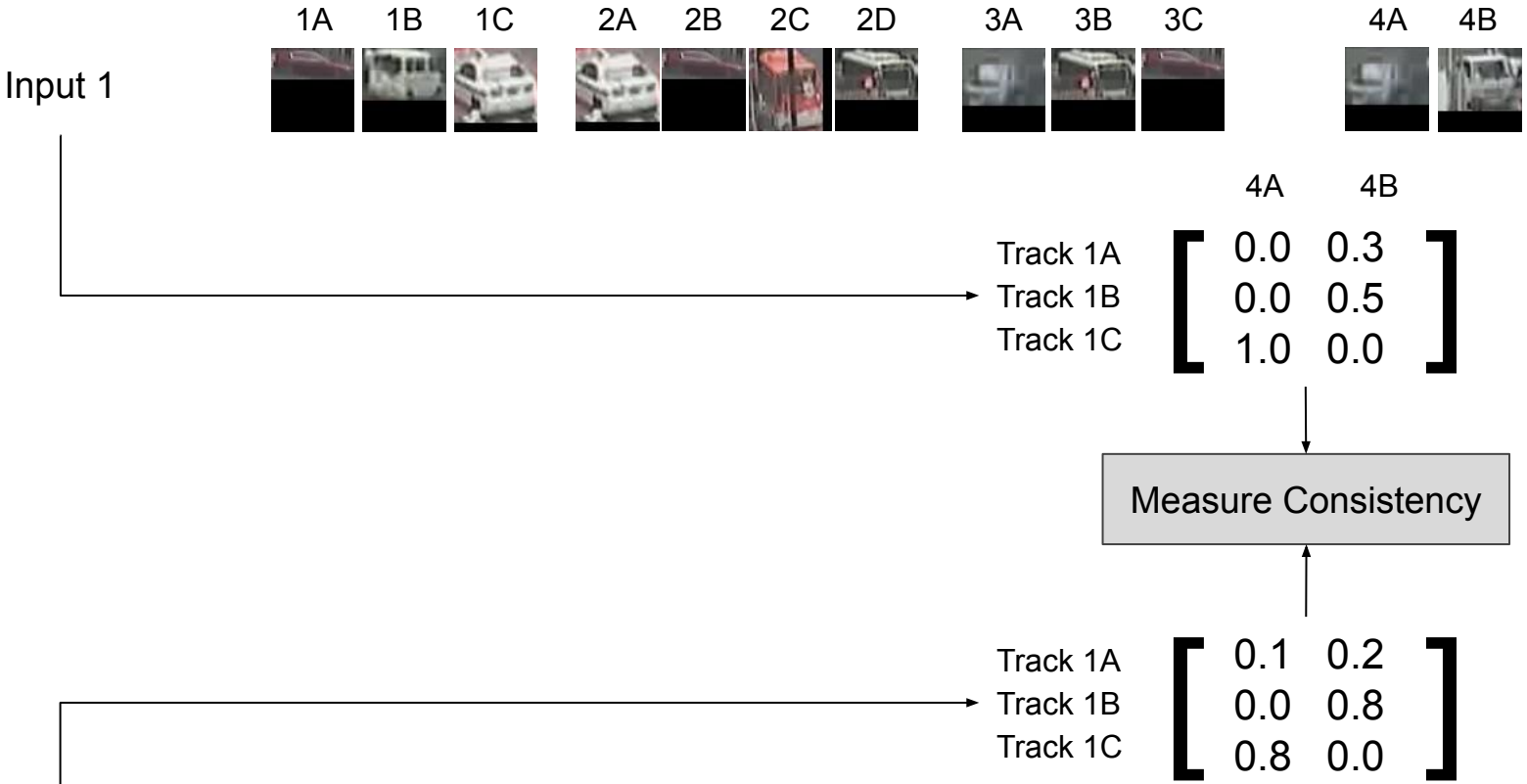
Input 1

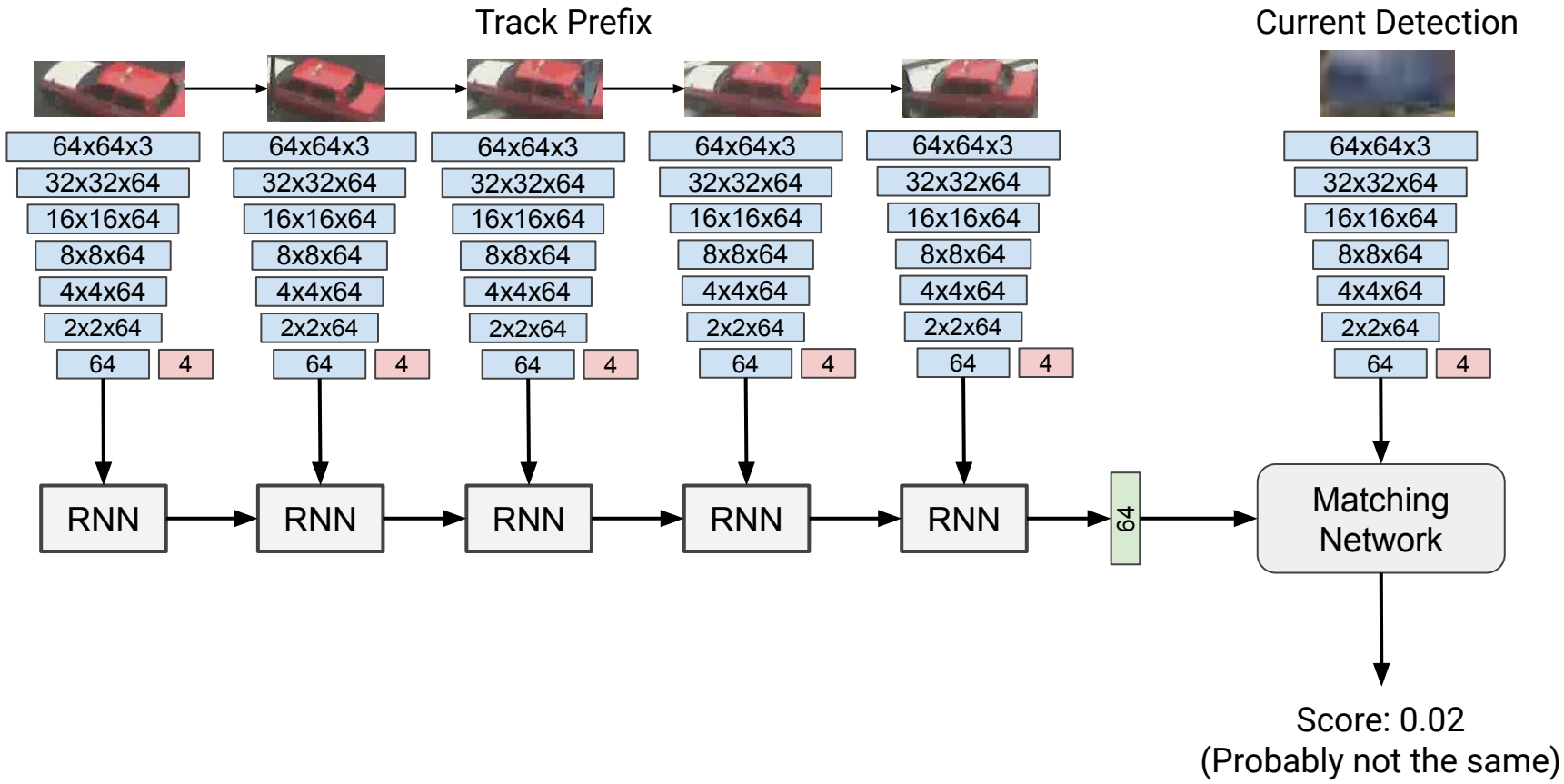


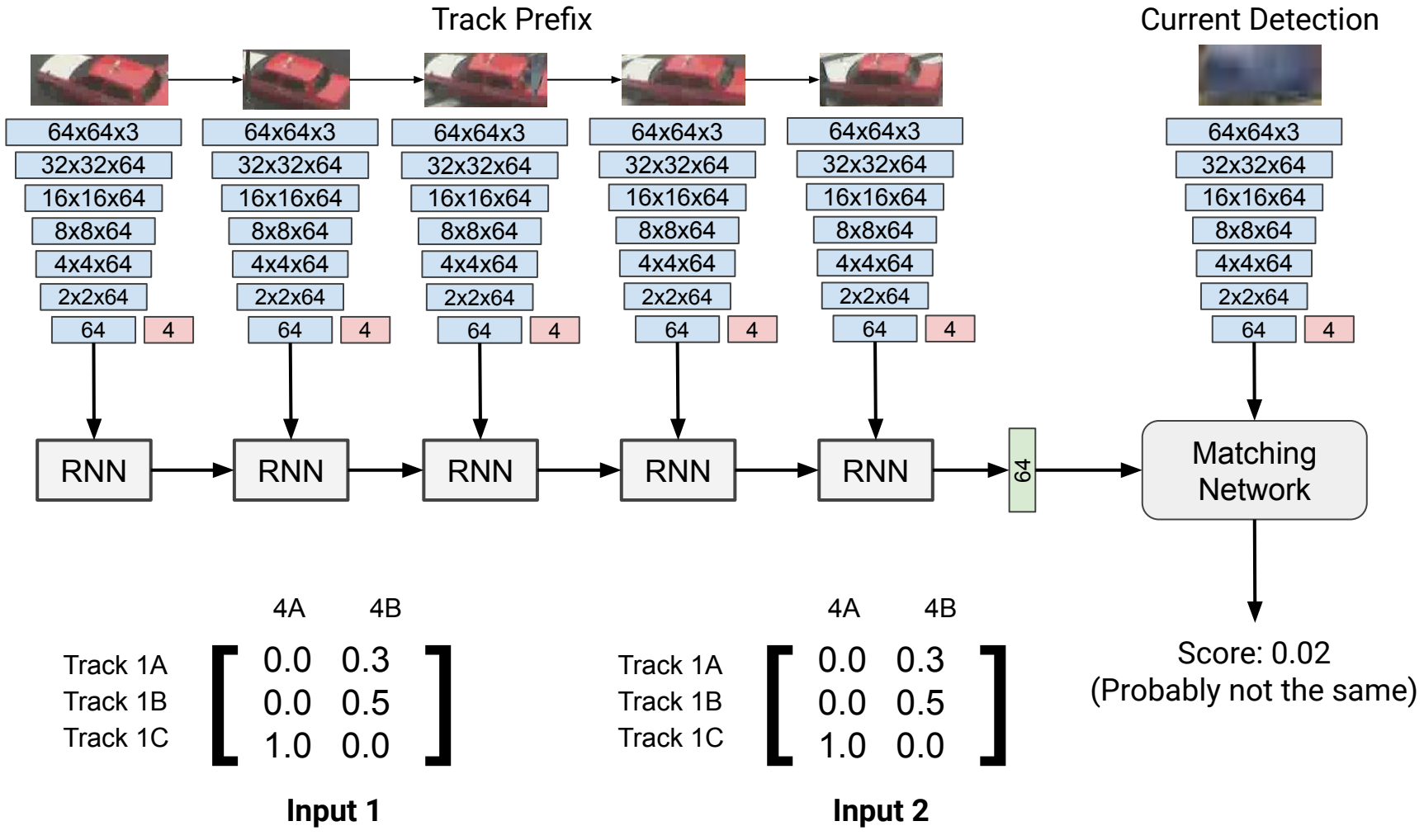
	4A	4B
Track 1A	0.0	0.3
Track 1B	0.0	0.5
Track 1C	1.0	0.0

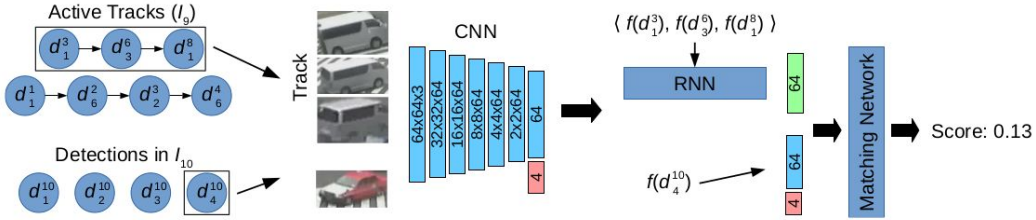
Input 2

1A	1B	1C	2A	2B	2C	2D	3A	3B	3C	4A	4B
x=93	x=950	x=652	x=490	x=740	x=821	x=848	x=458	x=847	x=819	x=609	x=899
y=681	y=459	y=333	y=429	y=575	y=335	y=325	y=313	y=325	y=336	y=282	y=337
w=176	w=143	w=78	w=211	w=235	w=63	w=51	w=62	w=53	w=62	w=75	w=62
h=103	h=75	h=87	h=140	h=127	h=47	h=38	h=49	h=37	h=47	h=48	h=38


1A 1B 1C 2A 2B 2C 2D 3A 3B 3C 4A 4B
Input 1
4A 4B
Track 1A
Track 1B
Track 1C
0.0 0.3
0.0 0.5
1.0 0.0
Measure Consistency
Track 1A
Track 1B
Track 1C
0.1 0.2
0.0 0.8
0.8 0.0
Input 2
x=93 y=681 w=176 h=103
x=950 y=459 w=143 h=75
x=652 y=333 w=78 h=87
x=490 y=429 w=211 h=140
x=740 y=575 w=235 h=127
x=821 y=335 w=63 h=47
x=848 y=325 w=51 h=38
x=458 y=313 w=62 h=49
x=847 y=325 w=53 h=37
x=819 y=336 w=62 h=47
x=609 y=282 w=75 h=48
x=899 y=337 w=62 h=38


Track Prefix
Current Detection
64x64x3
32x32x64
16x16x64
8x8x64
4x4x64
2x2x64
64
4
RNN
64
Matching Network
Score: 0.02
(Probably not the same)


Track Prefix
Current Detection
64x64x3
32x32x64
16x16x64
8x8x64
4x4x64
2x2x64
64
4
RNN
64
Matching Network
Score: 0.02
(Probably not the same)
4A
4B
Track 1A
Track 1B
Track 1C
0.0 0.3
0.0 0.5
1.0 0.0
Input 1
Input 2



Model Architecture

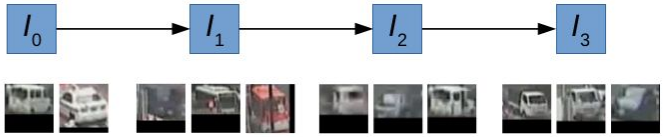
$$\begin{bmatrix} 0.3 & 0.2 & 0.0 & 0.1 \\ 0.0 & 0.4 & 0.0 & 0.2 \\ 0.1 & 0.0 & 0.9 & 0.0 \\ 0.1 & 0.2 & 0.1 & 0.1 \end{bmatrix}$$

Transition Matrix

```
array([[0.40525553, 0.58487146, 0.35028426],
       [0.17897008, 0.99813136, 0.65189838],
       [0.53625954, 0.51995857, 0.71777389]])
```

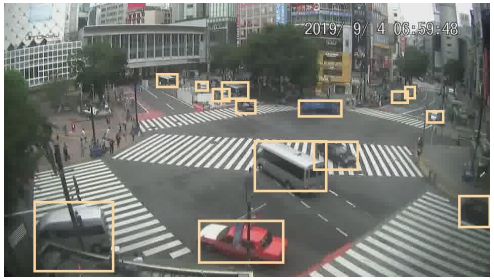
```
array([[0.42718929, 0.52618008, 0.84496057],
       [0.8125701 , 0.29344281, 0.09825419],
       [0.19191158, 0.11864856, 0.82009976]])
```

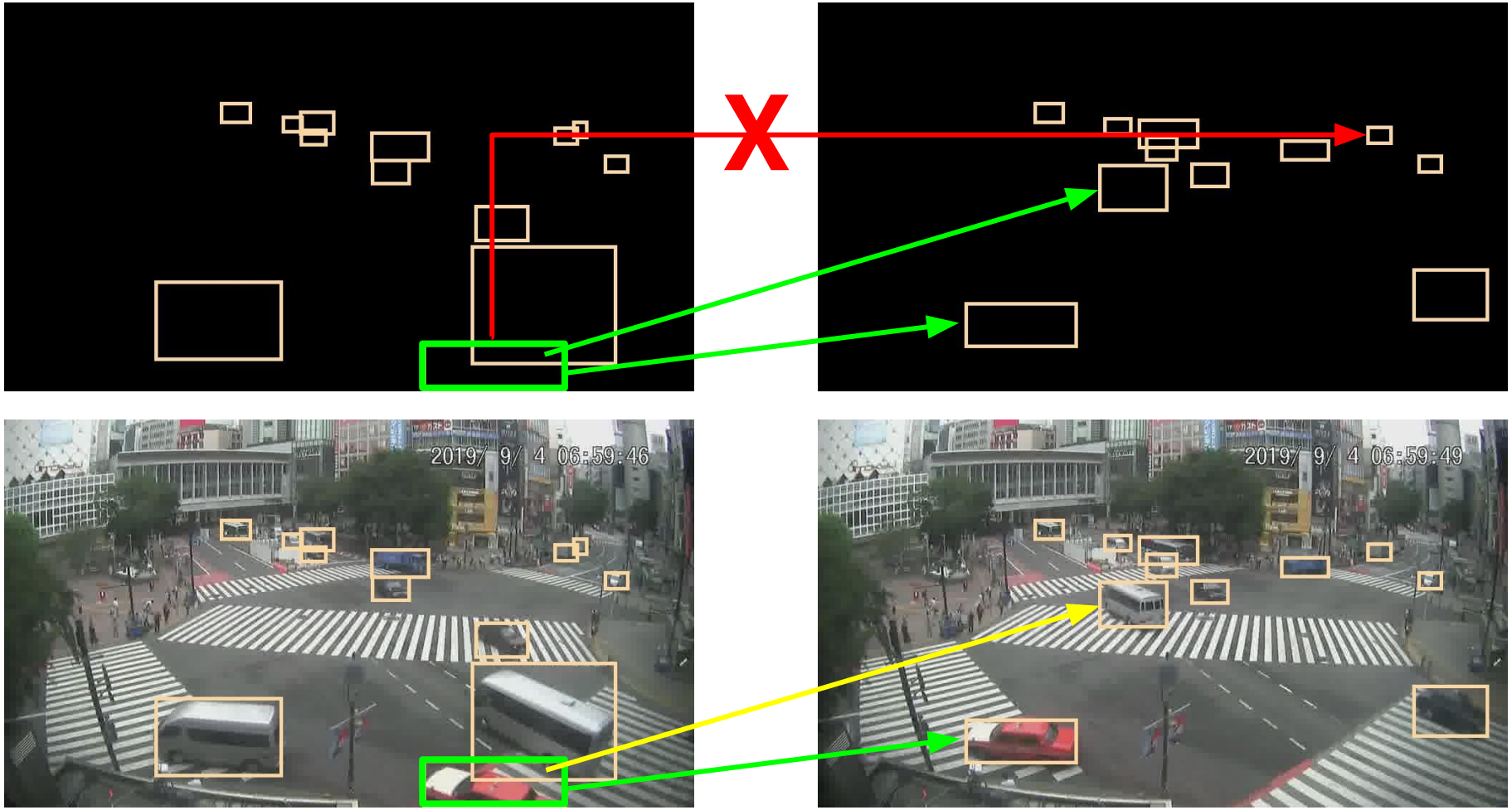
Consistency Loss Function



Input-Hiding Scheme

Input-Hiding Scheme — Visual-Spatial Hiding

	Frame 1	Frame 2	Frame 3
Original Video Sequence			
Input 1: Visual Information			
Input 2: Spatial Information	x=93 y=681 w=176 h=103 x=950 y=459 w=143 h=75 x=652 y=333 w=78 h=87 x=899 y=337 w=62 h=38	x=490 y=429 w=211 h=140 x=740 y=575 w=235 h=127 x=821 y=335 w=63 h=47 x=848 y=325 w=51 h=38	x=458 y=313 w=62 h=49 x=847 y=325 w=53 h=37 x=819 y=336 w=62 h=47 x=609 y=282 w=75 h=48


2019/ 9/ 4 06:59:46
2019/ 9/ 4 06:59:49

Results on MOT17 Benchmark

	Method	IDF1	MOTA	MT	ML	FP	FN	Idsw	Frag
Unsupervised Methods	Visual-Spatial (ours)	**58.3**	**56.8**	538	880	12K	231K	1K	2K
	SORT [2]	39.8	43.1	295	997	28K	288K	5K	7K
	IOU [3]	39.4	45.5	369	953	20K	282K	6K	7K
Supervised Methods	Tracktor++ [1]	52.3	53.5	459	861	12K	248K	2K	5K
	MHT-BLSTM [12]	51.9	47.5	429	981	26K	268K	2K	3K
	FAMNet [5]	48.7	52.0	450	787	14K	254K	3K	5K
	LSST [8]	62.3	54.7	480	944	26K	228K	1K	4K
	GSM [18]	57.8	56.4	523	813	14K	230K	1K	3K
	CenterTrack [30]	59.6	61.5	621	752	14K	201K	3K	5K

Table 2: Performance on the MOT17 test set. We compare methods in terms of IDF1 and MOTA, but include other non-comprehensive metrics from MOT17 as well for completeness.

Results on MOT17 Benchmark

	Method	IDF1	MOTA	MT	ML	FP	FN	Idsw	Frag
Unsupervised Methods	Visual-Spatial (ours)	**58.3**	**56.8**	538	880	12K	231K	1K	2K
	SORT [2]	39.8	43.1	295	997	28K	288K	5K	7K
	IOU [3]	39.4	45.5	369	953	20K	282K	6K	7K
Supervised Methods	Tracktor++ [1]	52.3	53.5	459	861	12K	248K	2K	5K
	MHT-BLSTM [12]	51.9	47.5	429	981	26K	268K	2K	3K
	FAMNet [5]	48.7	52.0	450	787	14K	254K	3K	5K
	LSST [8]	62.3	54.7	480	944	26K	228K	1K	4K
	GSM [18]	57.8	56.4	523	813	14K	230K	1K	3K
	CenterTrack [30]	59.6	61.5	621	752	14K	201K	3K	5K

Table 2: Performance on the MOT17 test set. We compare methods in terms of IDF1 and MOTA, but include other non-comprehensive metrics from MOT17 as well for completeness.

Results on MOT17 Benchmark

	Method	IDF1	MOTA	MT	ML	FP	FN	Idsw	Frag
Unsupervised Methods	Visual-Spatial (ours)	**58.3**	**56.8**	538	880	12K	231K	1K	2K
	SORT [2]	39.8	43.1	295	997	28K	288K	5K	7K
	IOU [3]	39.4	45.5	369	953	20K	282K	6K	7K
Supervised Methods	Tracktor++ [1]	52.3	53.5	459	861	12K	248K	2K	5K
	MHT-BLSTM [12]	51.9	47.5	429	981	26K	268K	2K	3K
	FAMNet [5]	48.7	52.0	450	787	14K	254K	3K	5K
	LSST [8]	62.3	54.7	480	944	26K	228K	1K	4K
	GSM [18]	57.8	56.4	523	813	14K	230K	1K	3K
	CenterTrack [30]	59.6	61.5	621	752	14K	201K	3K	5K

Table 2: Performance on the MOT17 test set. We compare methods in terms of IDF1 and MOTA, but include other non-comprehensive metrics from MOT17 as well for completeness.

Conclusion

- Labeling data for MOT is tedious and costly
- Huge amounts of unlabeled video are almost always available!
- By leveraging unlabeled video, cross-input consistency ***outperforms several fully-supervised MOT methods*** from within the last 1-2 years
- For code and more info: https://favyen.com/uns20/