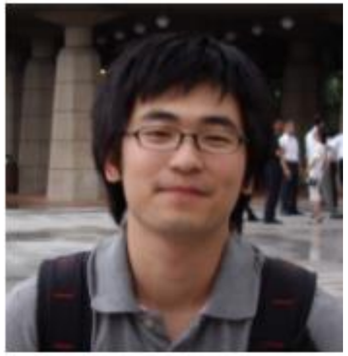
Pixel-Level Matching for Video Object Segmentation using Convolutional Neural Networks

Jae Shin Yoon

Francois Rameau

Seokju Lee

Junsik Kim



Seunghak Shin

In So Kweon

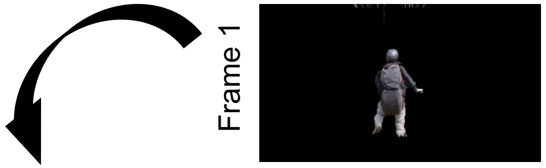
Korea Advanced Instittue of Science and Technology

Robotics and Computer Vision laboratory

ICCV 17

International Conference on Computer Vision

[Initial supervision]



Video object segmentation

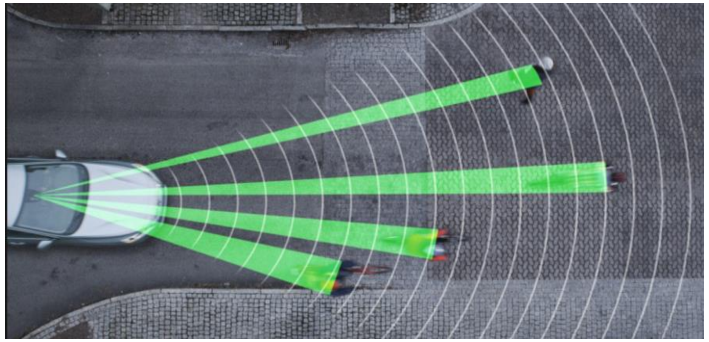
Frame 1 Frame 2 Frame 3 Frame 4 Frame 5 Frame 6 Frame 7

Frame 74 Frame 75 Frame 76 Frame 77 Frame 78 Frame 79 Frame 80

- Video object segmentation is to propagate the initial object(s) mask from the first to the last frame of a video sequence.
- Users can separate foreground from the background in the video.

Industrial Machinary Camera

Foreground vs. Background

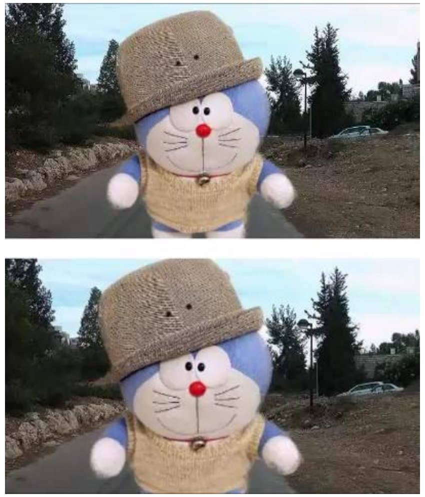


Personal Mobile Device Camera

:: Visually pleasing video application

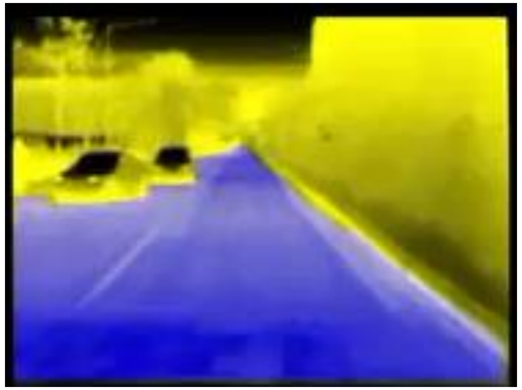
:: Practical robot application

__Video stabilization__



__Visual Tracking__

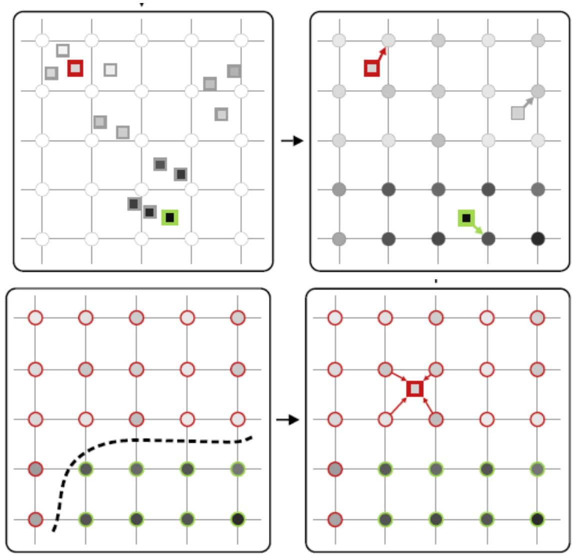
__Drivable region detection__



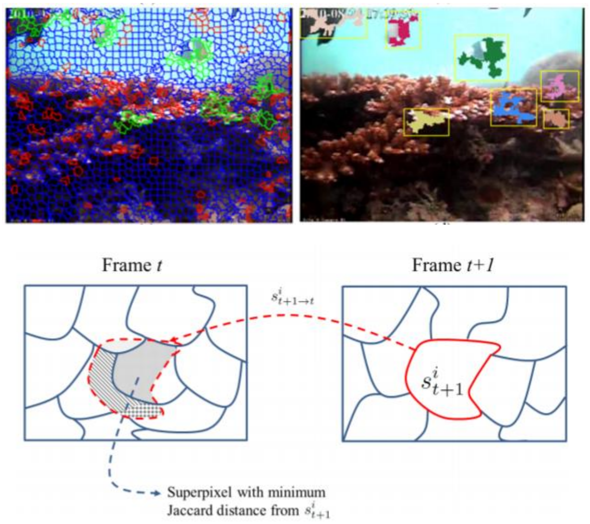
(~ 2016)

Traditional Video Object Segmentation Method

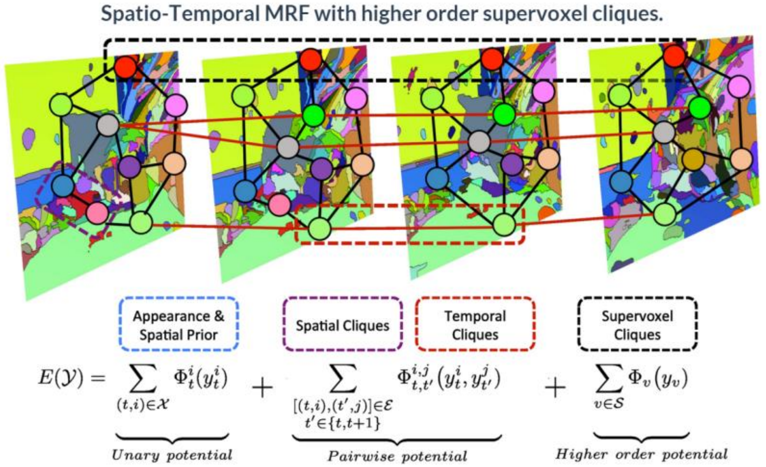
:: Design and optimize energy equation considering (*pixel* / *superpixel* / *supervoxel*) connectivity in the spatio-temporal domain.



Pixel optimization [1]



Super-pixel optimization [2]



Super-voxel optimization [3]

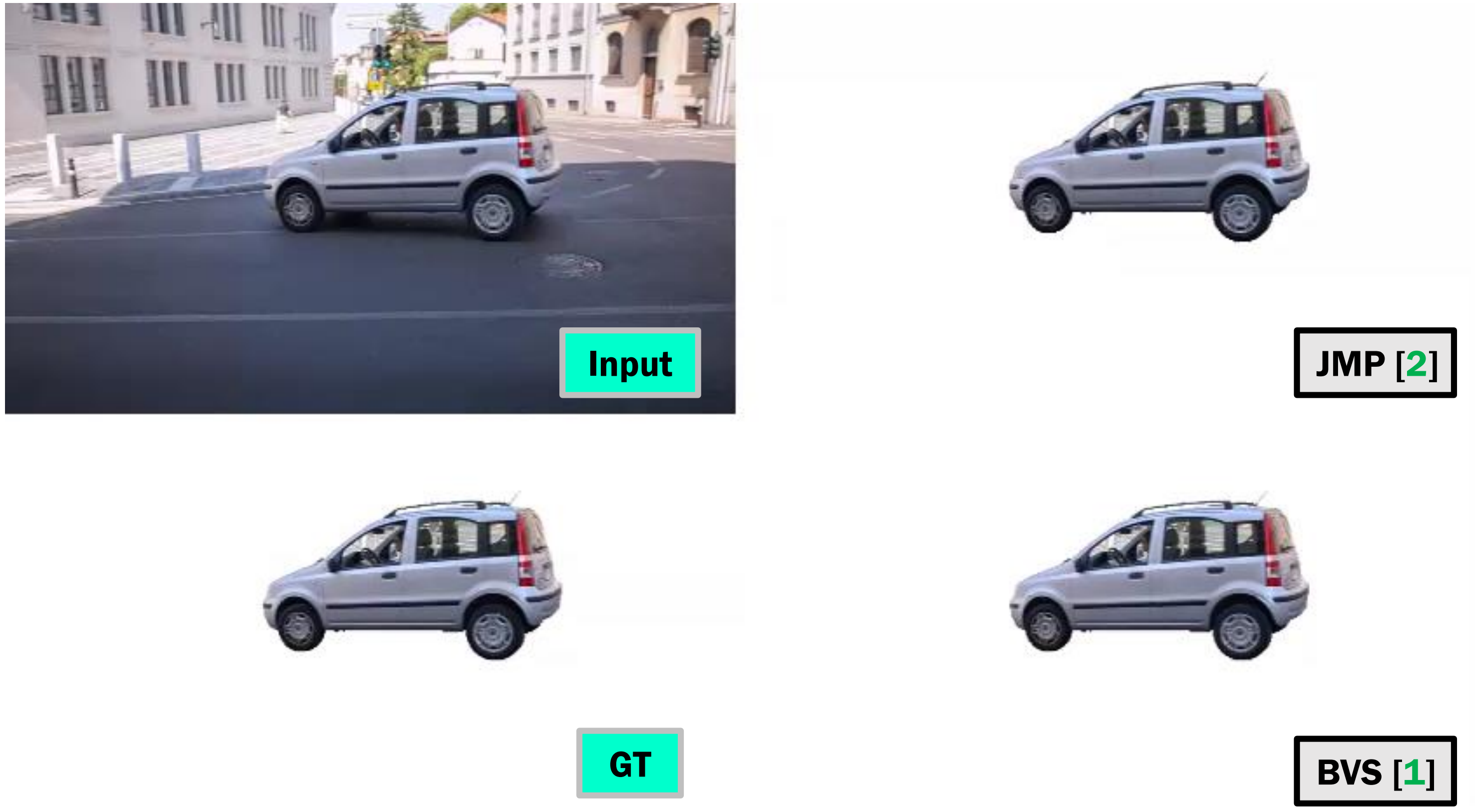
**suvpervoxel* : the space-time analog of spatial superpixels

[1] Märki, Nicolas, et al. "Bilateral space video segmentation." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016.

[2] Giordano, Daniela, et al. "Superpixel-based video object segmentation using perceptual organization and location prior." *IEEE Conference on Computer Vision and Pattern Recognition*. 2015.

[3] Jain, Suyog Dutt, and Kristen Grauman. "Supervoxel-consistent foreground propagation in video." *European Conference on Computer Vision*. Springer International Publishing, 2014.

However,...

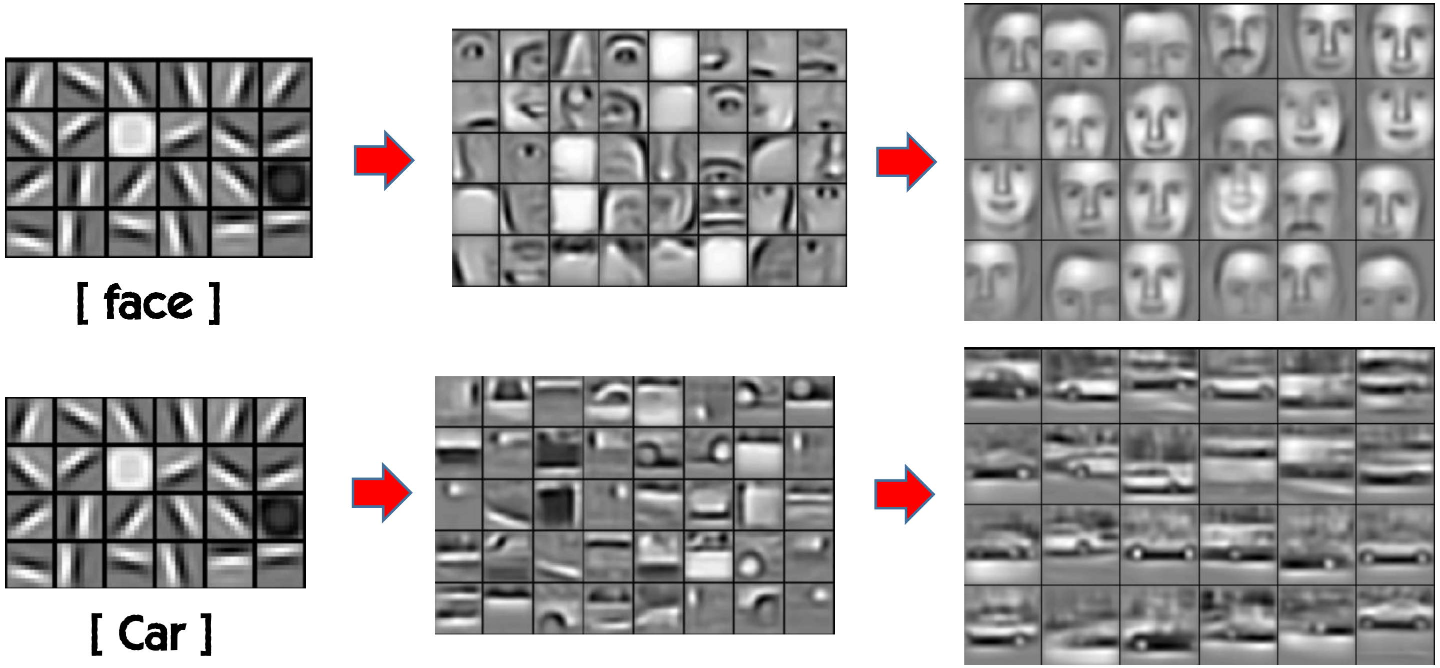


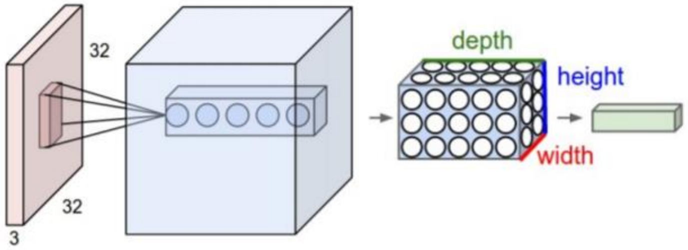
... previous hand-crafted feature based approaches are too sensitive to prevent background drift effect.

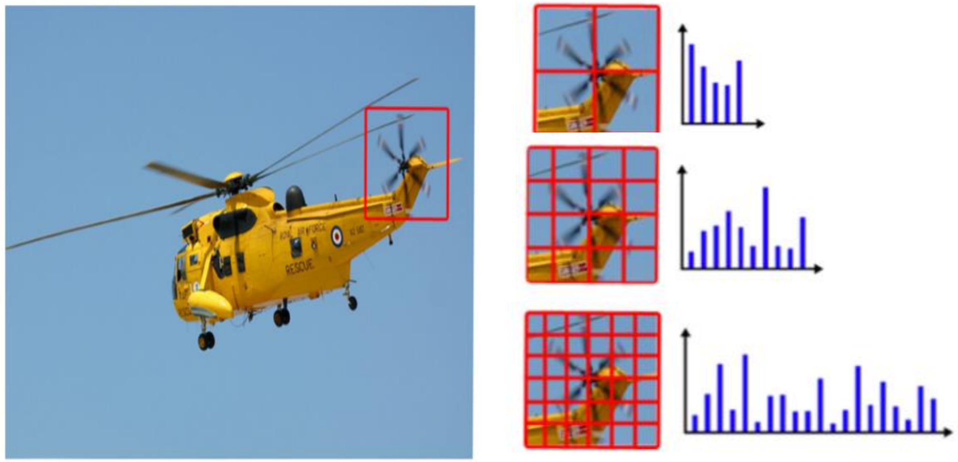
[1] Märki, Nicolas, et al. "Bilateral space video segmentation." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016.

[2] Fan, Qingnan, et al. "JumpCut: non-successive mask transfer and interpolation for video cutout." *ACM Transactions on Graphics (TOG)* 34.6 (2015): 195.

Why Convolutional Neural Networks (CNN) ?





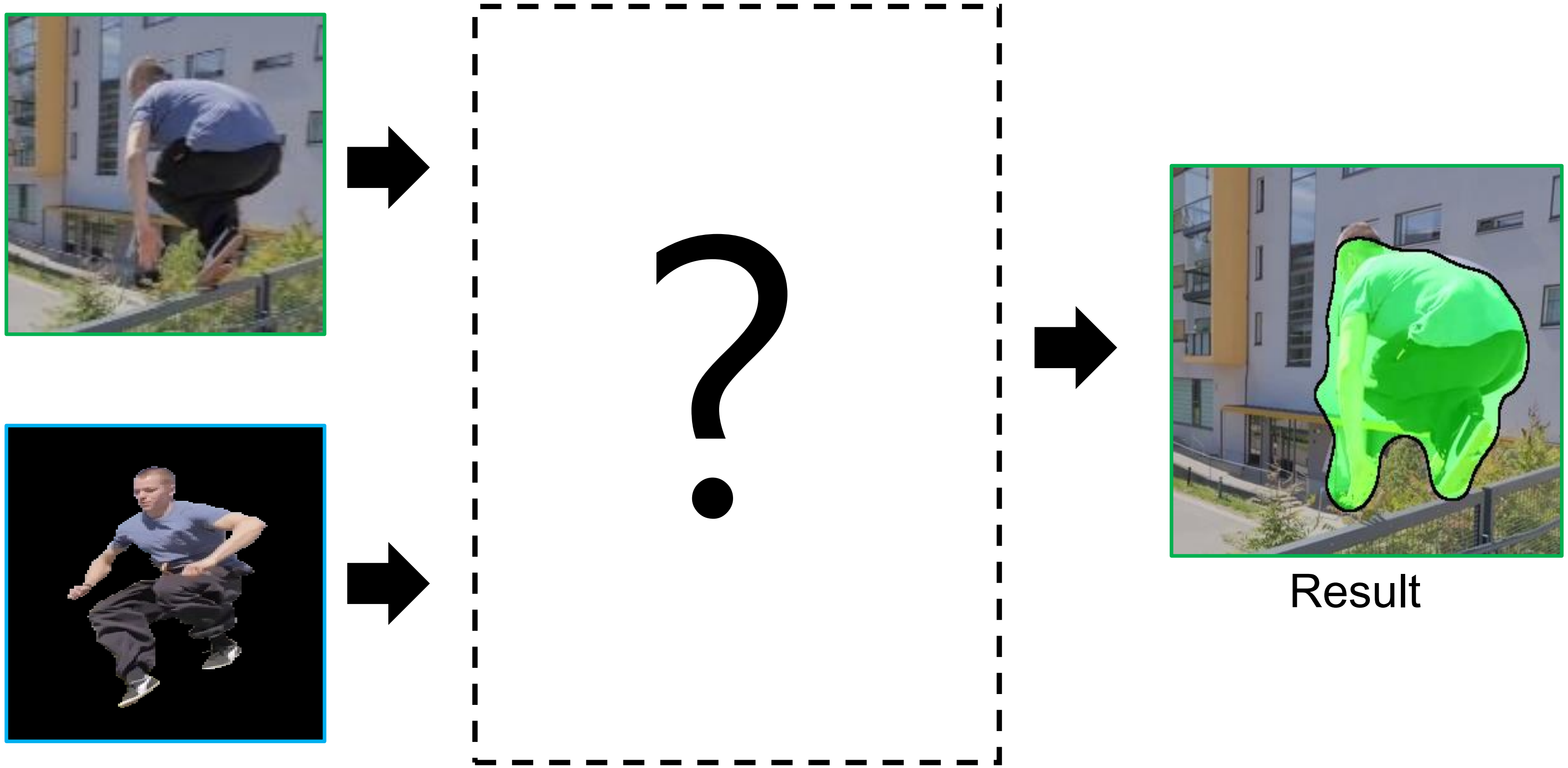


Strong feature representation power

: Can describe a target object with category-level semantic information

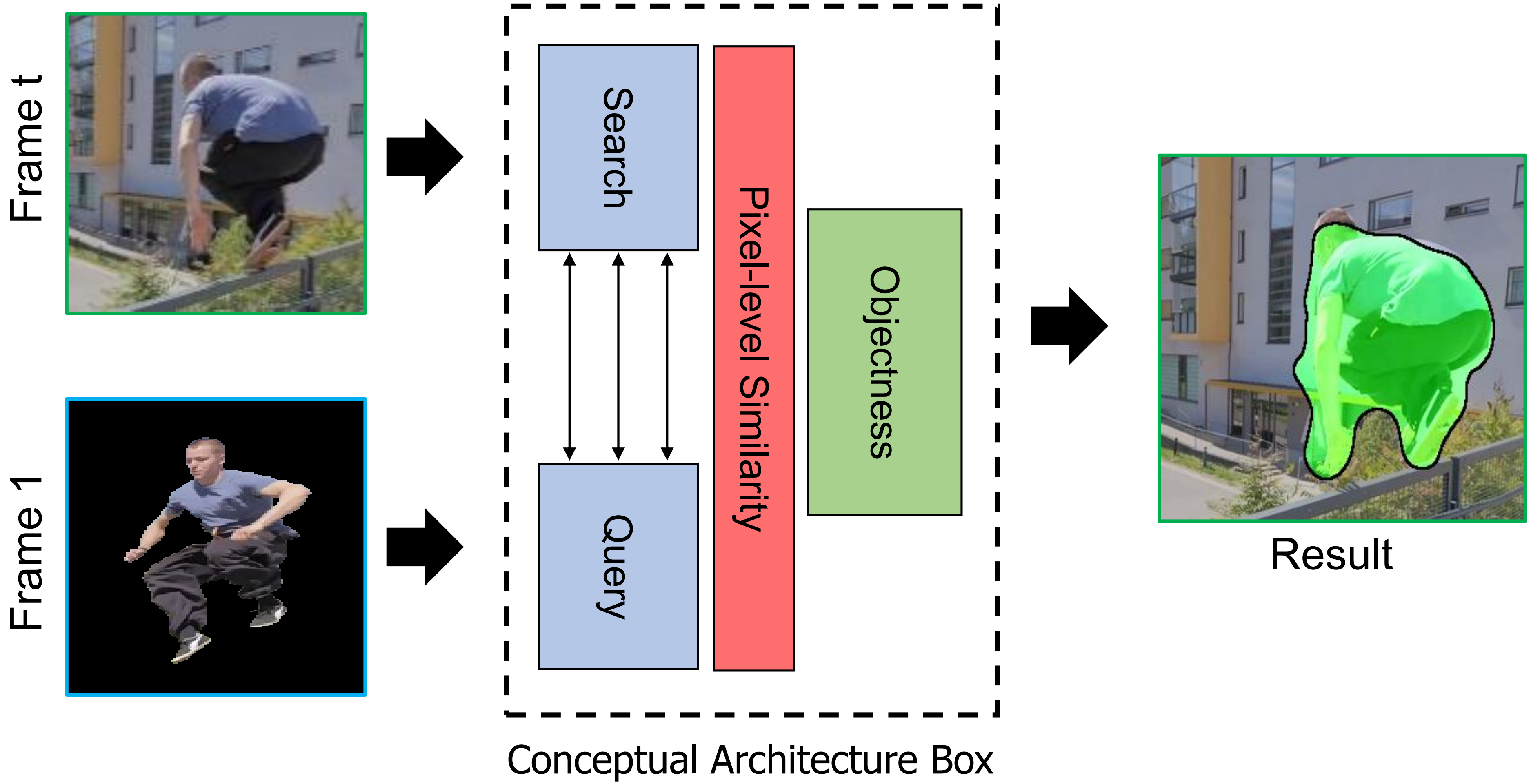
: Foreground can be more discriminative from background

GOAL

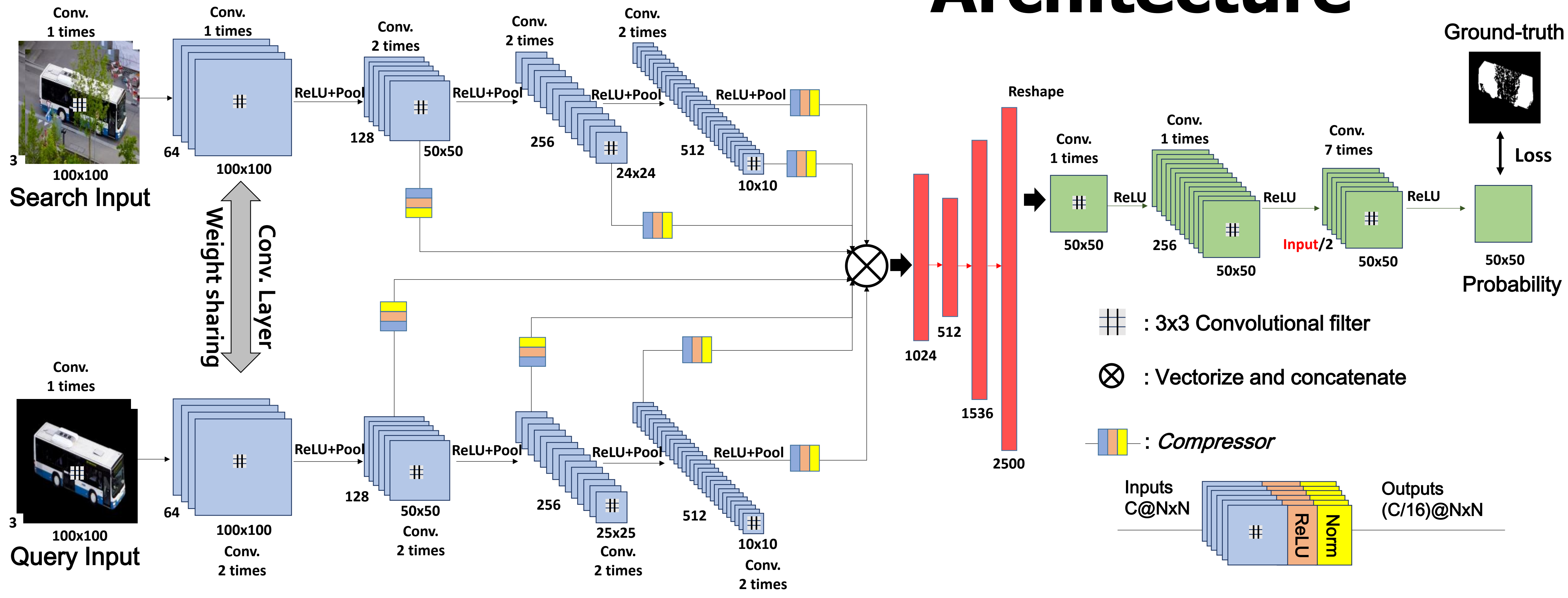


GOAL

Pixel-Level Matching Network

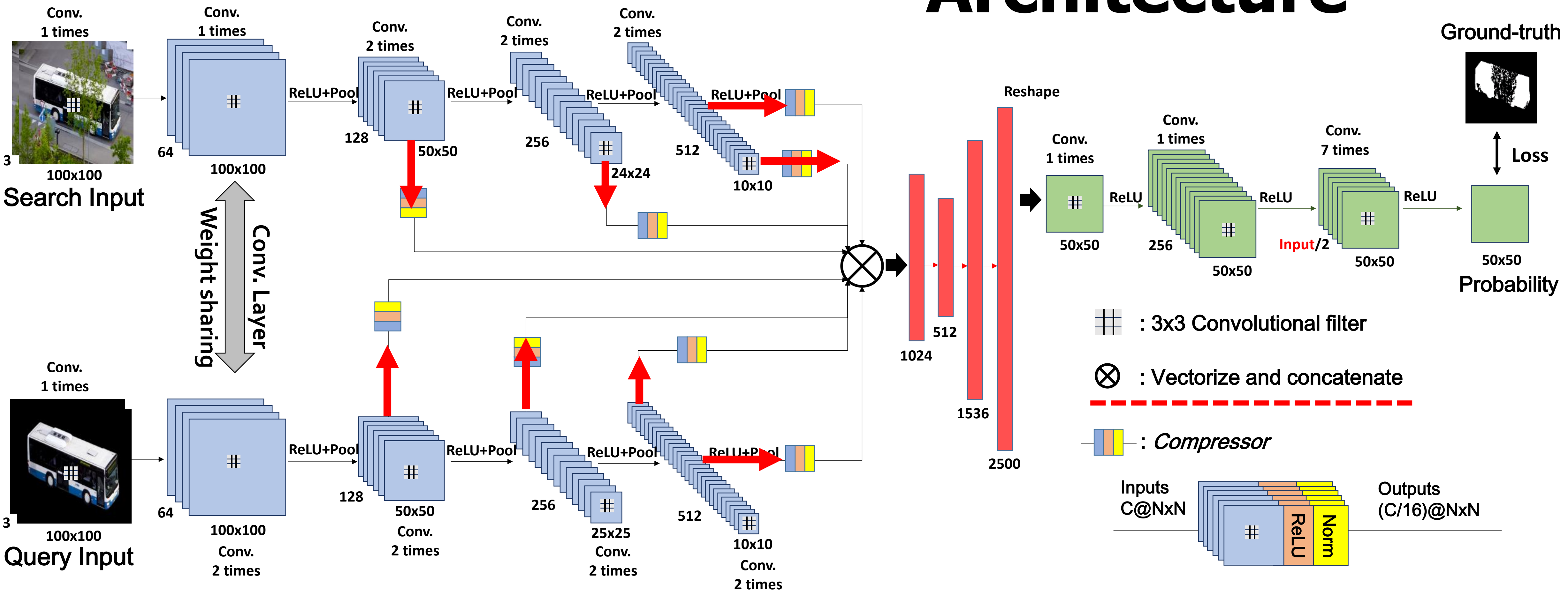


Architecture



Feature extraction → Similarity encoding → Objectness decoding

Architecture

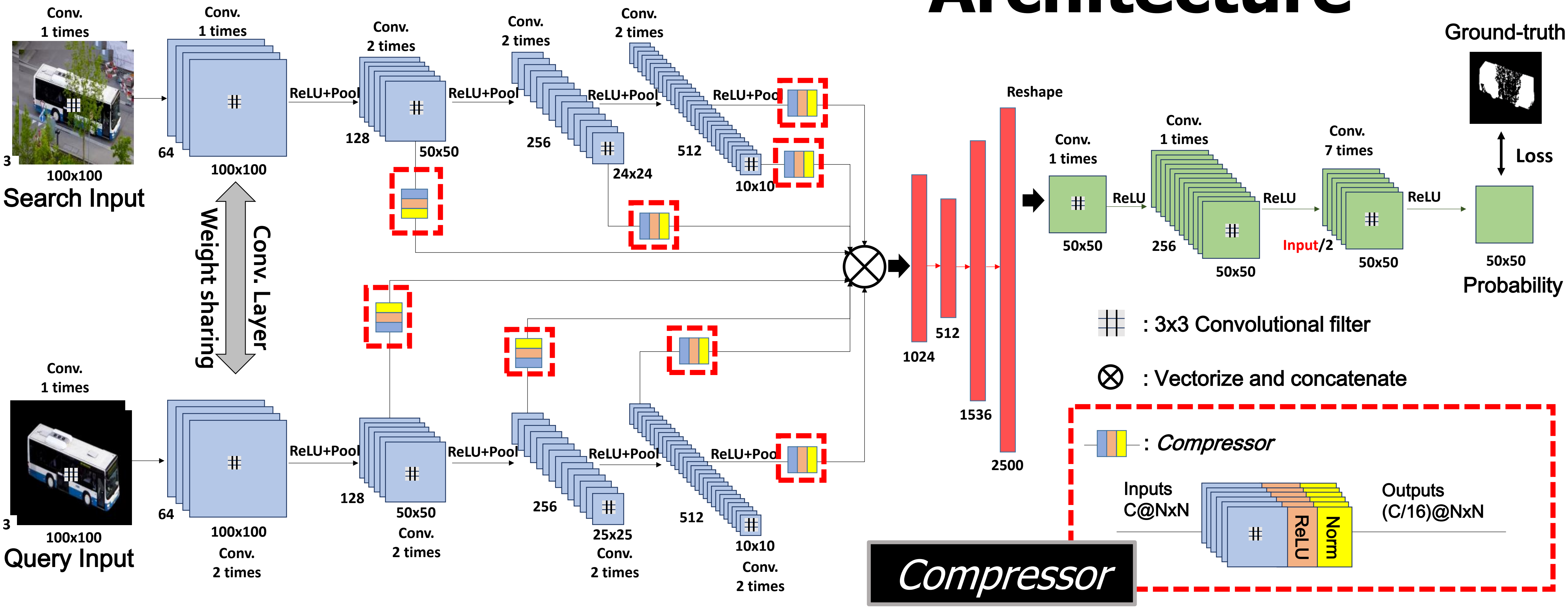


- Exploit features from multiple layers with different depths

  :: to handle both category-level semantic information and spatial details at the same time

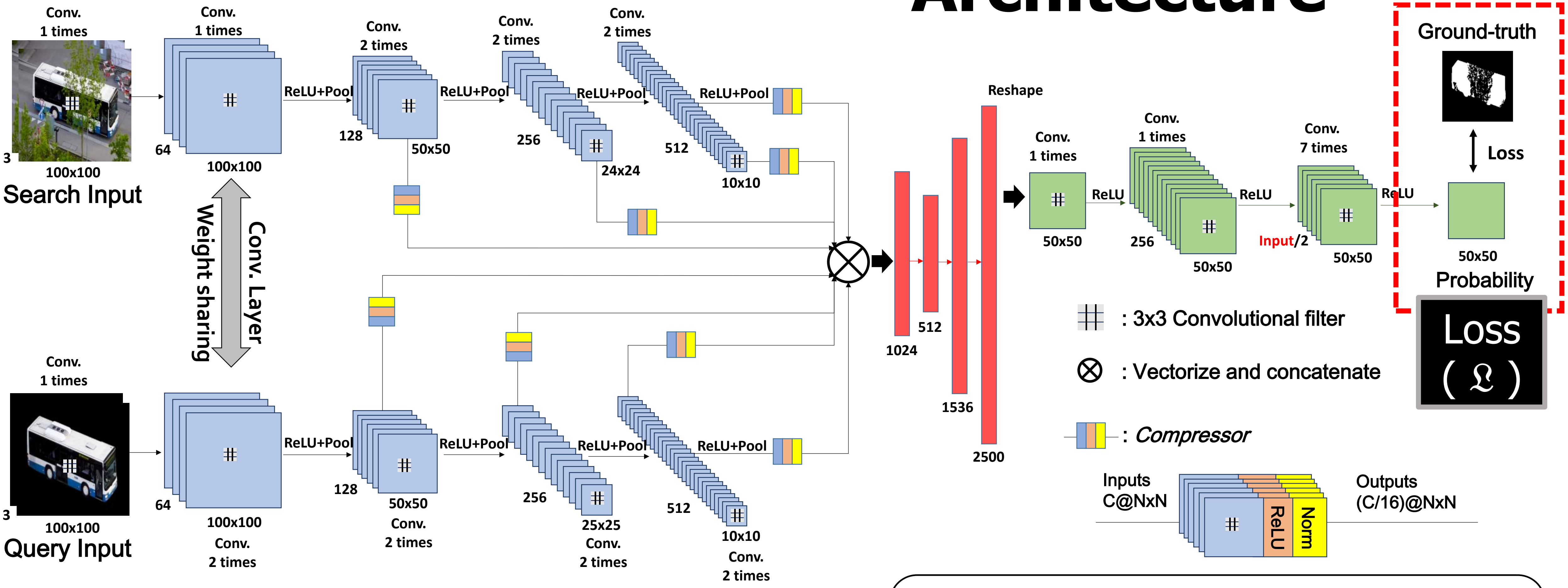
- Make 1d concatenated feature vector

Architecture



- *Compress* the feature volume into 16 times smaller than the original one.
- Consist of 1-Convolution, 1-ReLU, 1-Normalization.
- Normalization prevents the unbounded ReLU scale variation.

Architecture



- Pixel-wise Euclidian distance loss
- Probability range is [0 1]

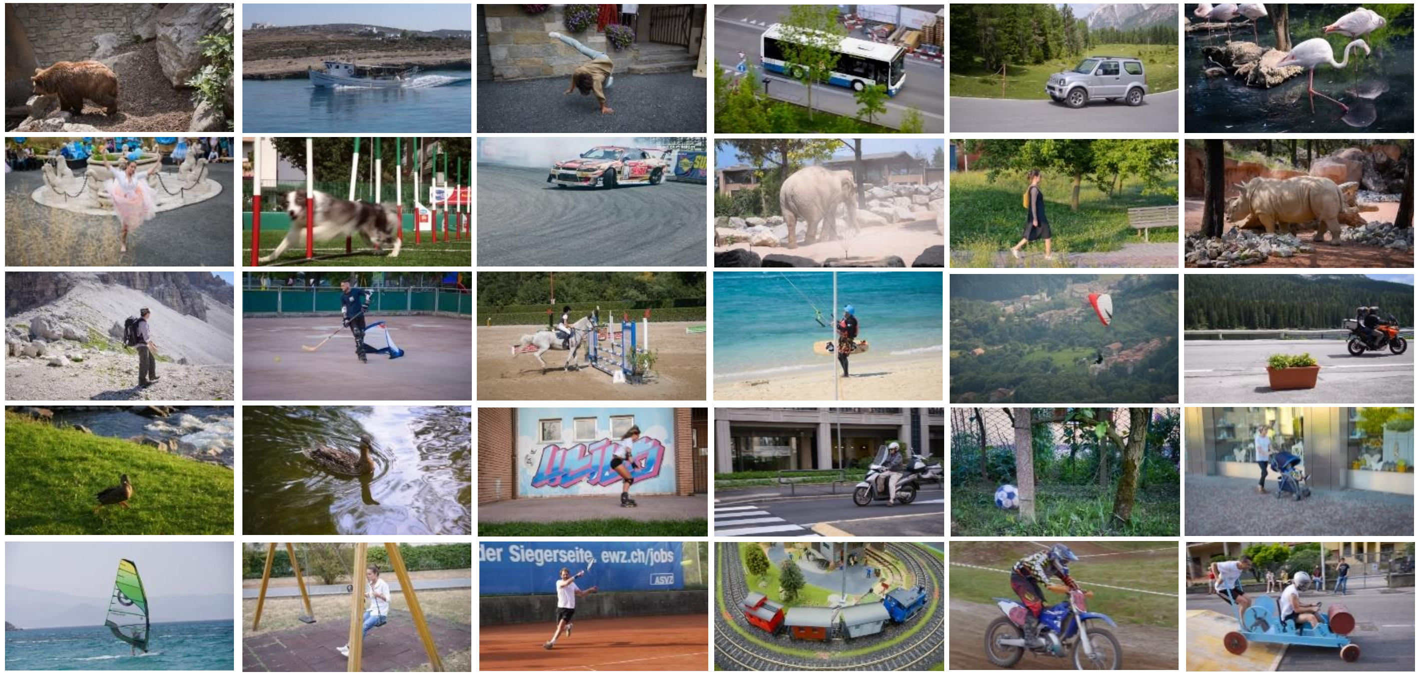
Background Foreground

$$\mathfrak{L} = \frac{1}{N^2}\sum_{x=1}^{N}\sum_{y=1}^{N} \| P(x,y) - L(x,y) \|_2,$$

L : ground-truth label P : probability

(x, y) : pixel-location N : map size (50)

Two-stage Training with



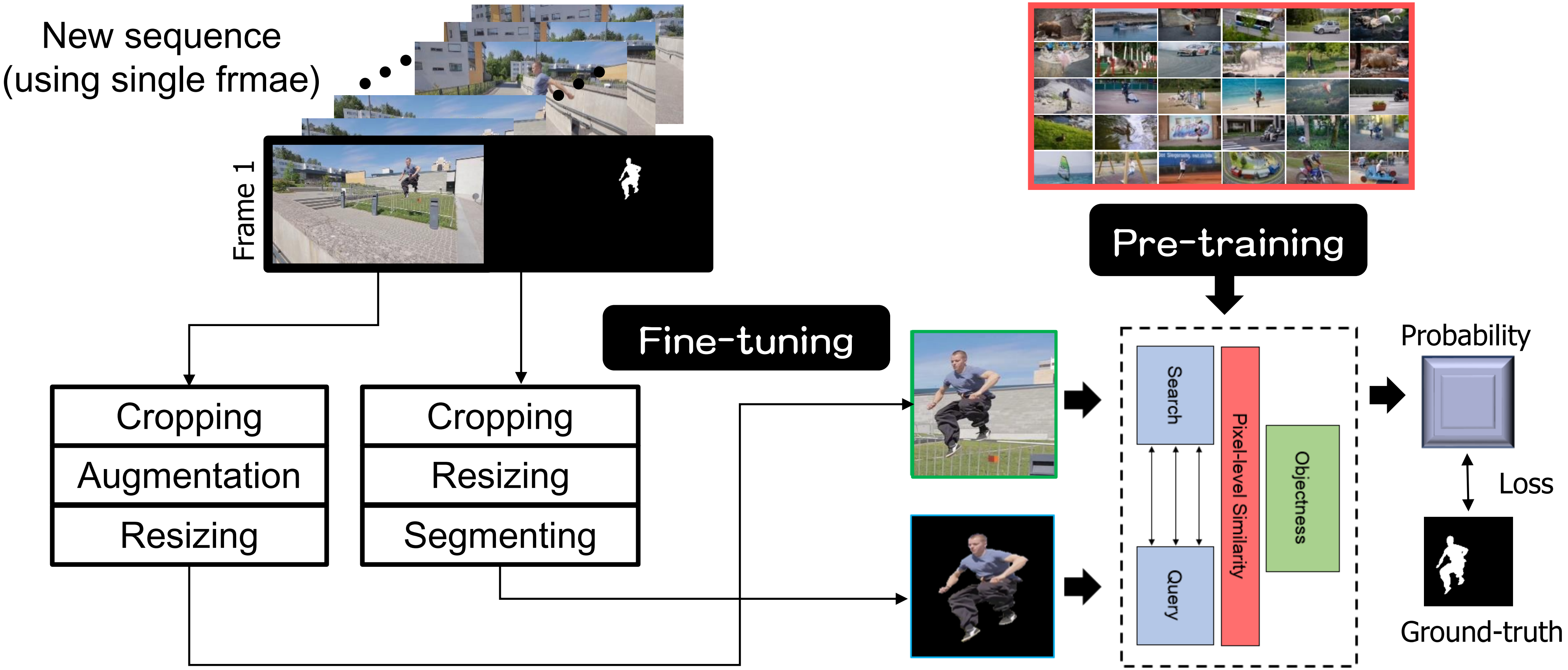
DAVIS [1] benchmark [CVPR 2016]

- 30 sequences for training, 20 sequences for testing
  (Augmentation : translation, rotating, scale variation, fliping)

- Scenarios : Background clutter, Deformation, Motion Blur, Fast-motion, Low resolution, Occlusion, Out-of-view, Scale-variation, Appearnce change, Edge ambiguity, Shape complexity.

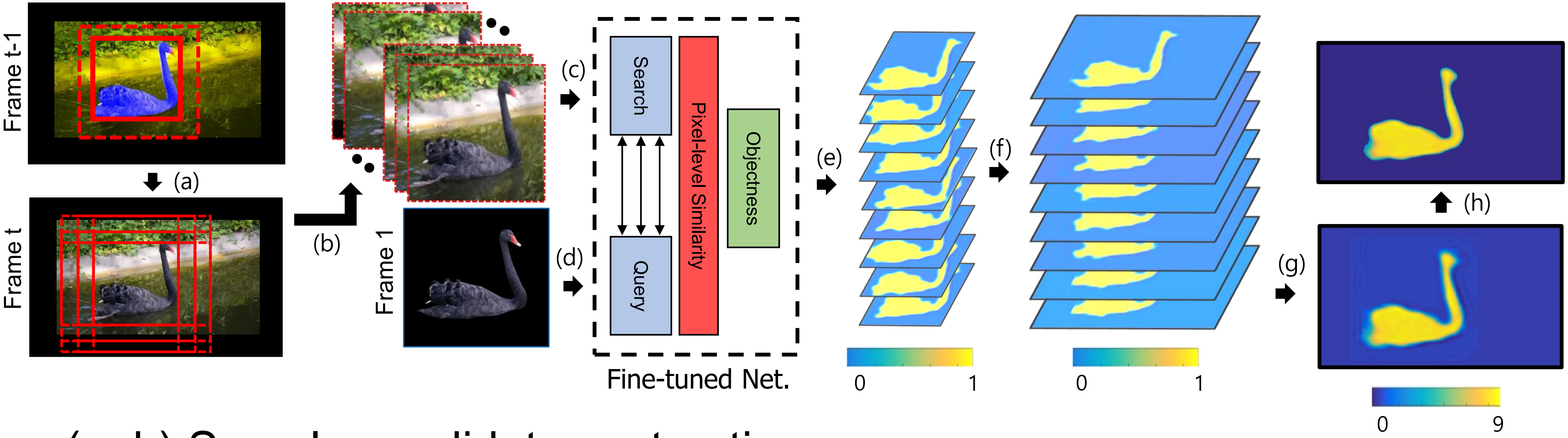
[1] Perazzi, F., et al. "A Benchmark Dataset and Evaluation Methodology for Video Object Segmentation.“. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016.

Two-stage Training



- In **pre-training**, the networks learn semantic matching inference dealing with appearance variation.
- In **fine-tuning**, the networks learn specific target object appearance.
- Without fine-tuning, the network will be distracted by a background similar to the trained objects.
- Our network can handle any kind of objects with only an initial supervision.

Overall Step : Video Object Segmetataion

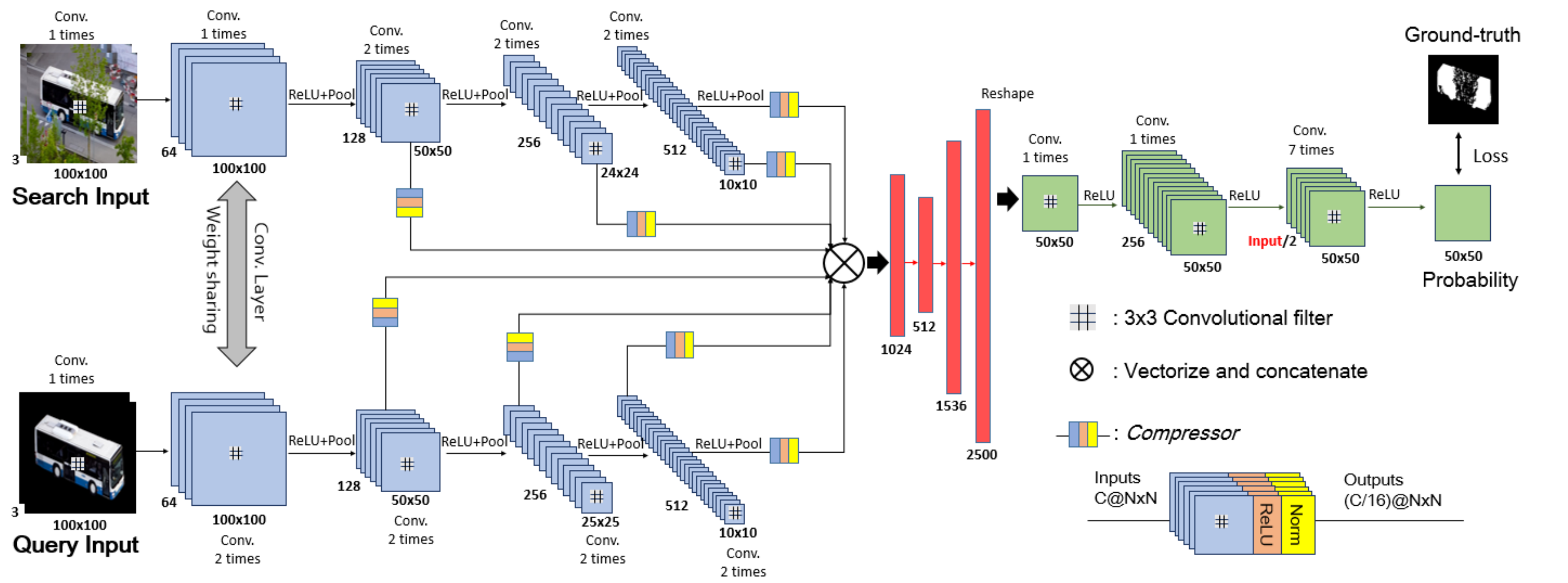


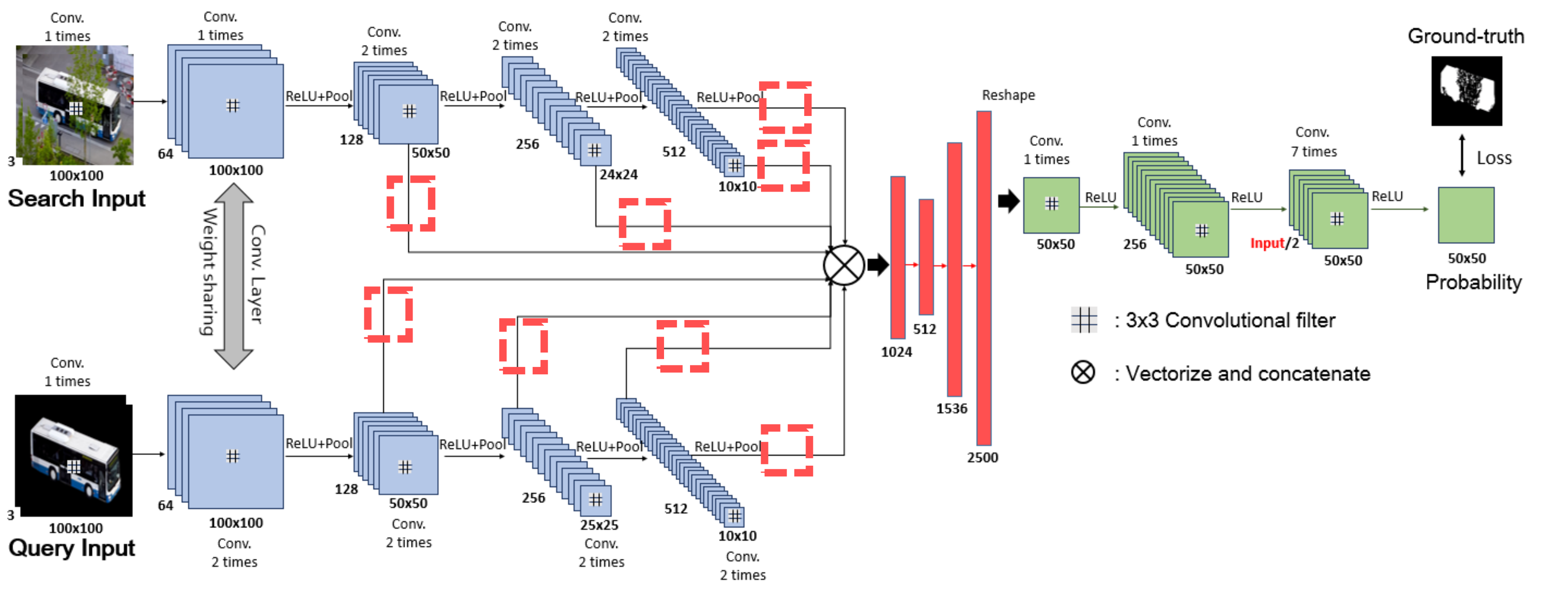
(a, b) Sample candidates extraction.

(c, d, e) Response extraction by feeding search and query inputs.

(f) Size restoreing. (g) Stacking (h) Thresholding

Experiments :: *DAVIS* benchmark :: Self-structure evaluation



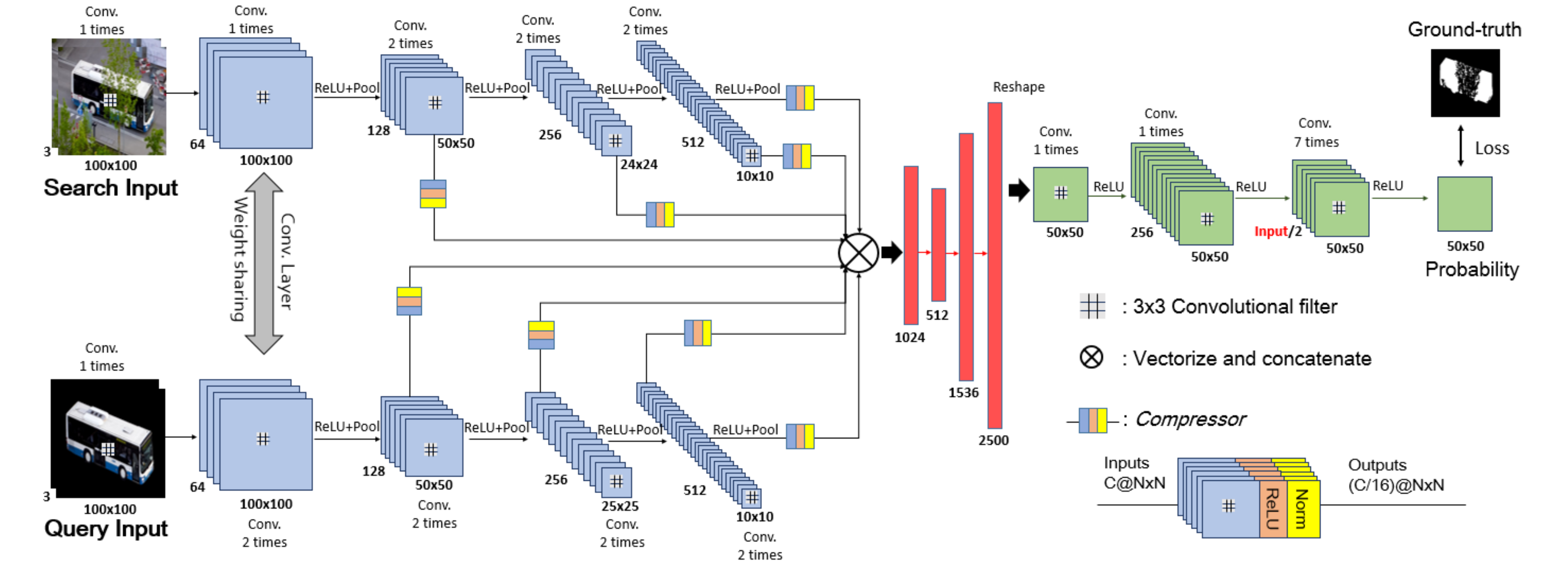


With *Compressor*

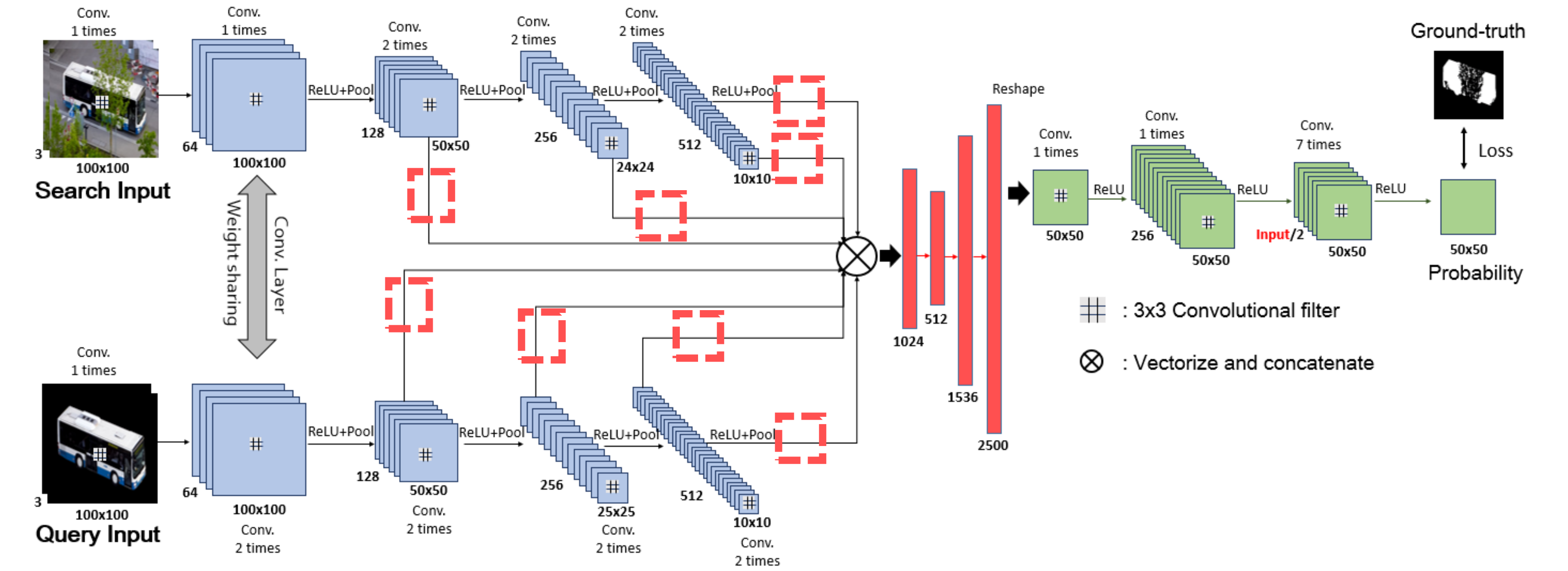
vs.

Without *Compressor*

Experiments :: *DAVIS* benchmark :: Self−structure evaluation



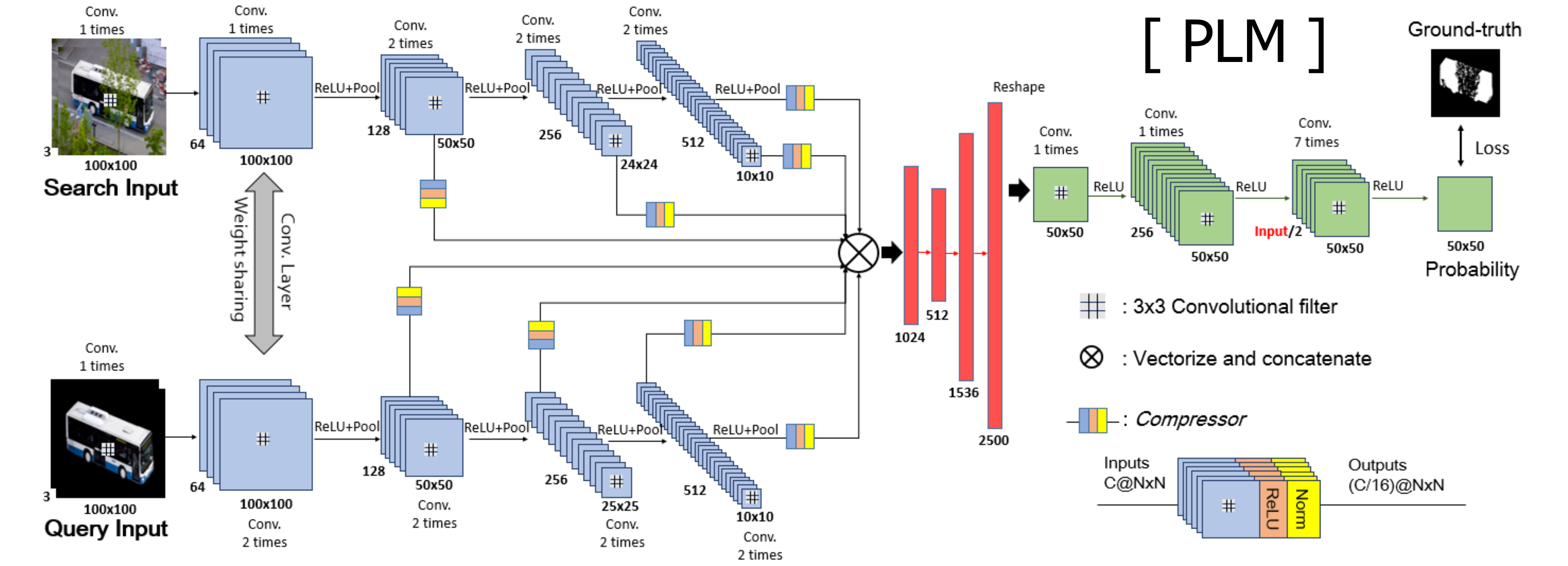
About 3.7GB (batch size: 32)

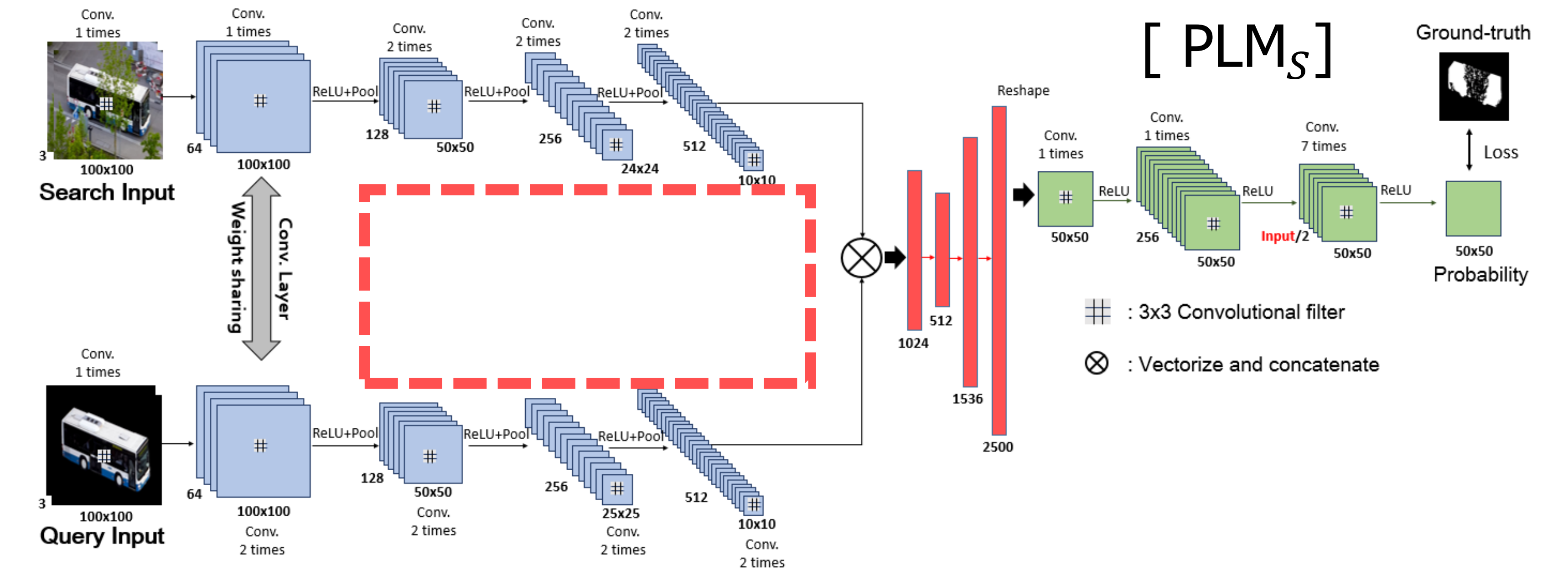


More than 12GB (batch size: 1)

TITAN X cannot handle this…

Experiments :: *DAVIS* benchmark :: Self−structure evaluation



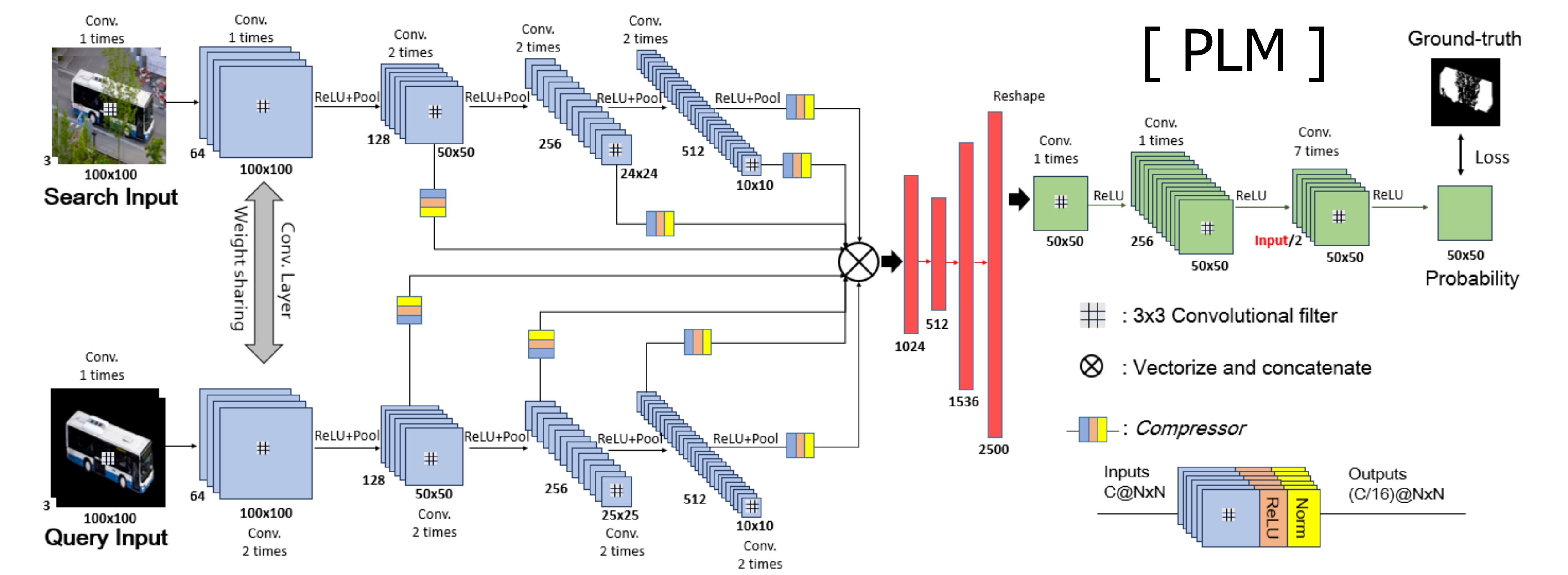


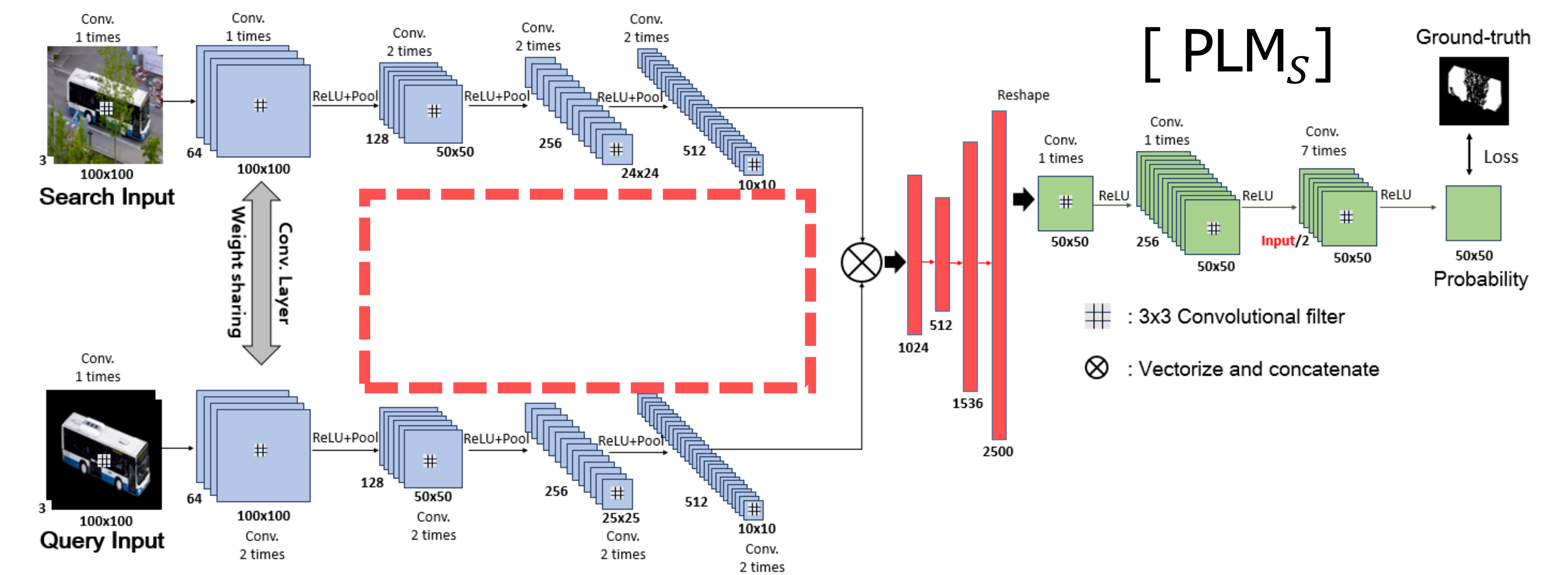
Feature from multiple layers

vs.

Feature from single deep layer

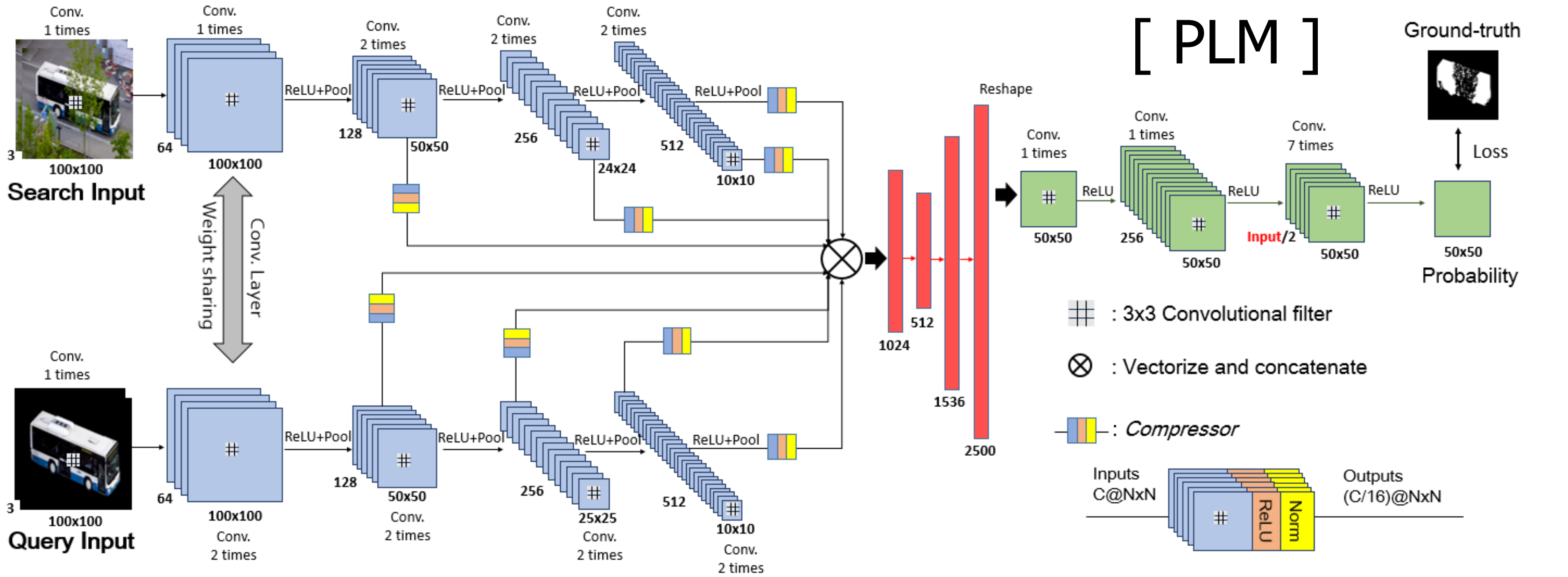
Experiments :: *DAVIS* benchmark :: Self−structure evaluation

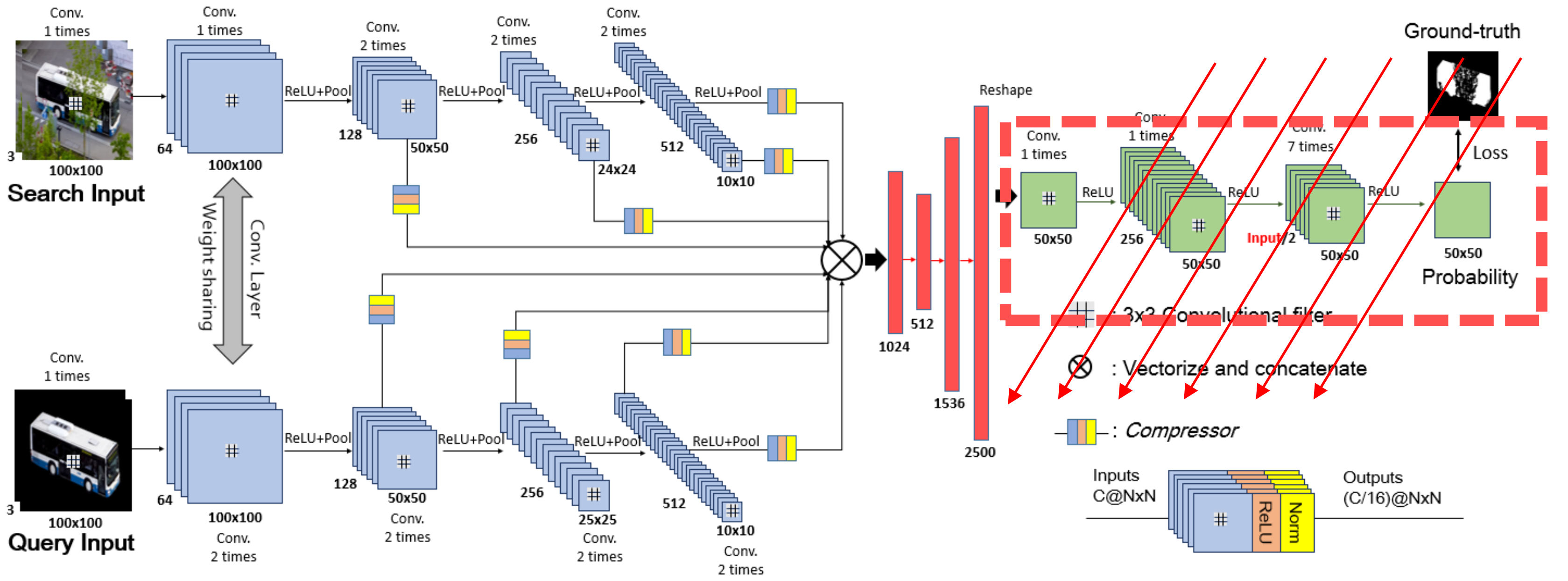




	PLM$_S$	PLM$_U$	PLM
Blackswan	0.27	0.86	0.86
	0.20	0.87	0.87
Bmx	0.27	0.40	0.47
Tree	0.43	0.55	0.66
Break	0.06	0.50	0.47
Dance	0.08	0.36	0.35
Came	0.40	0.62	0.65
	0.20	0.49	0.54
Car	0.27	0.87	0.86
Roundabout	0.16	0.68	0.64
Car	0.43	0.77	0.77
Shadow	0.32	0.63	0.62
Cows	0.15	0.79	0.81
	0.10	0.65	0.71
Dance	0.44	0.61	0.59
Twirl	0.35	0.49	0.50
Dog	0.29	0.79	0.73
	0.15	0.63	0.61
Drift	0.03	0.61	0.73
Chicane	0.03	0.60	0.79
Drift	012	0.63	0.73
Straight	0.05	0.50	0.52
Goat	0.05	0.74	0.76
	0.05	0.60	0.63
Horse	0.38	0.71	0.72
Jump	0.24	0.70	0.73
Kite	0.20	0.25	0.59
Surf	0.15	0.17	0.45
Libby	0.23	0.43	0.52
	0.21	0.45	0.57
Motorcross	0.27	0.44	0.49
Jump	0.06	0.27	0.32
Paragliding	0.41	0.54	0.55
Launch	0.11	0.16	0.15
Parkour	0.05	0.57	0.77
	0.08	0.56	0.75
Scooter	0.09	0.69	0.73
Black	0.10	0.53	0.57
Soapbox	0.16	0.70	0.72
	0.18	0.49	0.54
Average	0.23	0.63	0.68
	0.16	0.52	0.58
Speed	0.2s	0.15s+α	0.15s

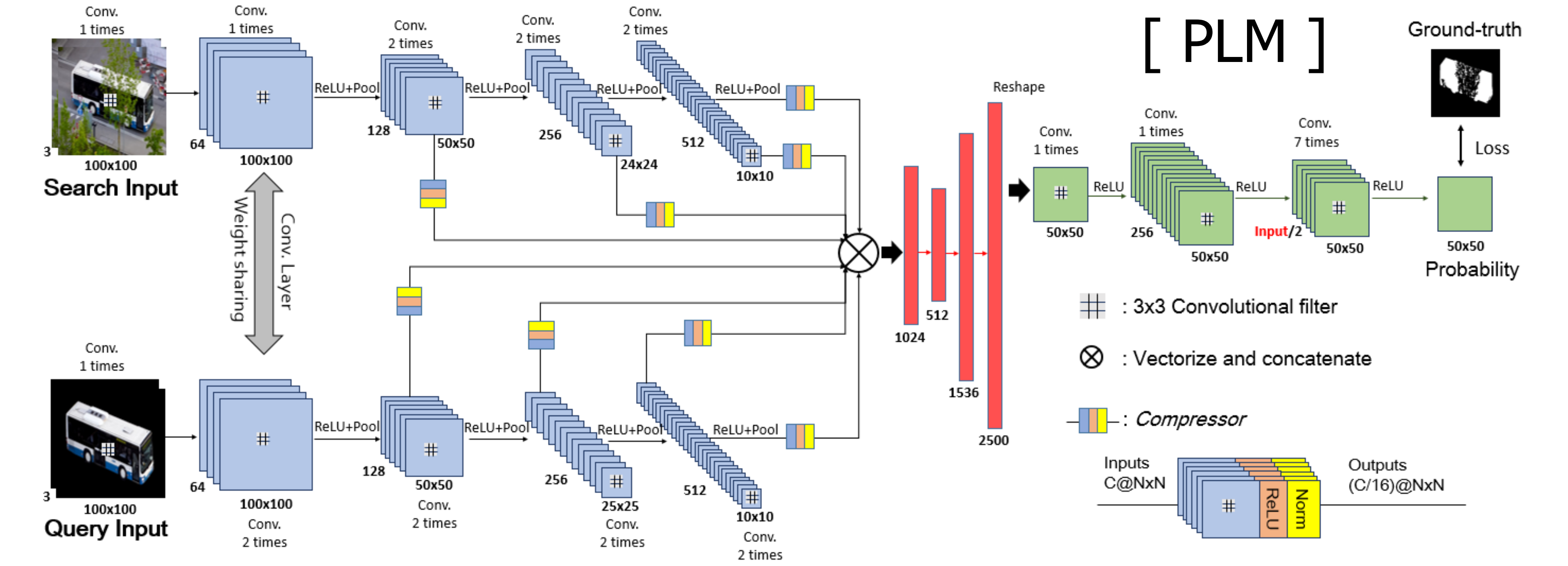
Experiments :: *DAVIS* benchmark :: Self−structure evaluation

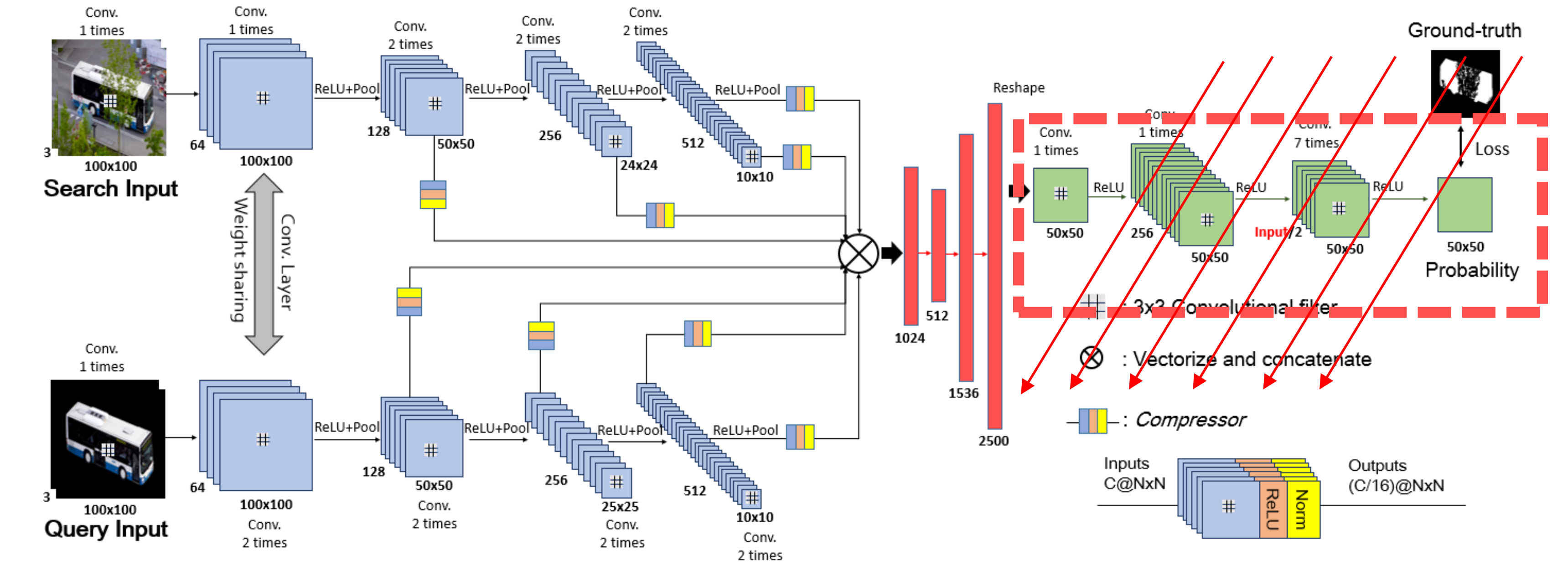


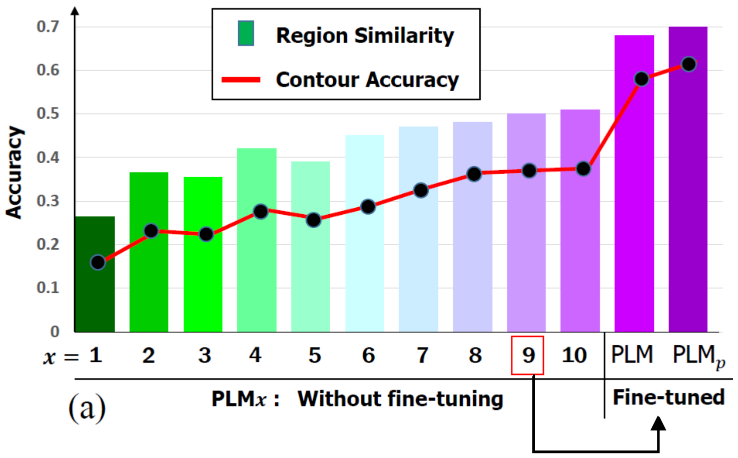


Ablation Test
for
Ojbect Decoder

Experiments :: *DAVIS* benchmark :: Self−structure evaluation







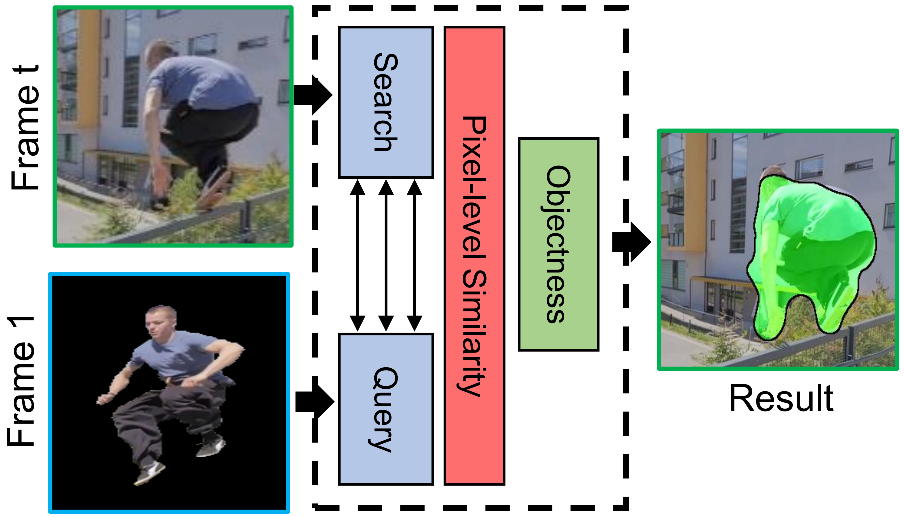
Saturated enough!

+ Two-stage training maxize the performance!

*x= the number of layers

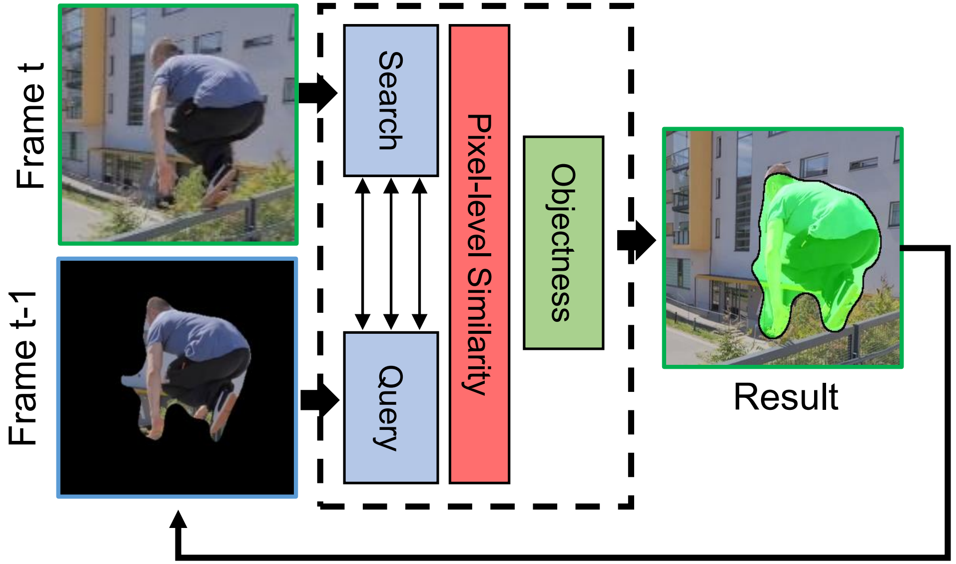
*PLMp = post-processed results using weighted median filter

Experiments :: *DAVIS* benchmark :: Self−structure evaluation



[PLM]

- **Matching frame t with frame 1**
- **No model update after fine-tuning**



[PLM_U]

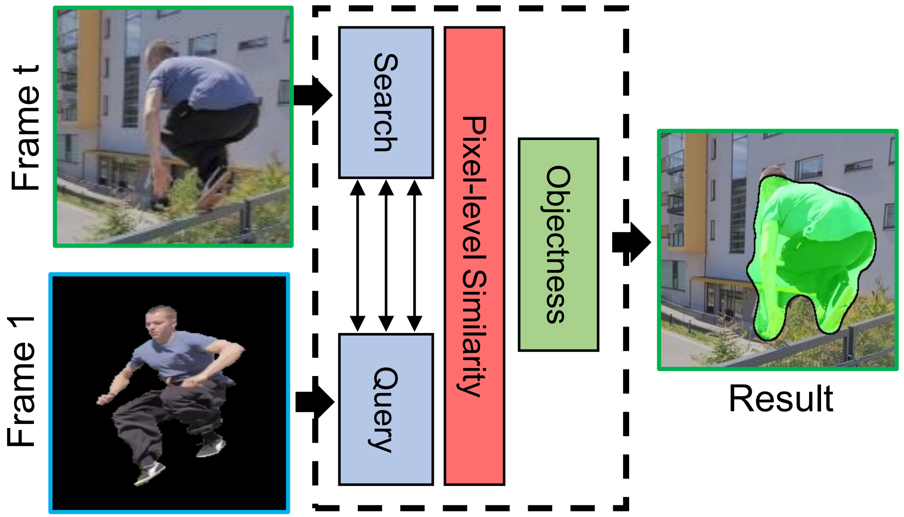
- **Matching frame t with frame t-1**
- **Model update after ten-frame processing**

Initial frame matching without model update

vs.

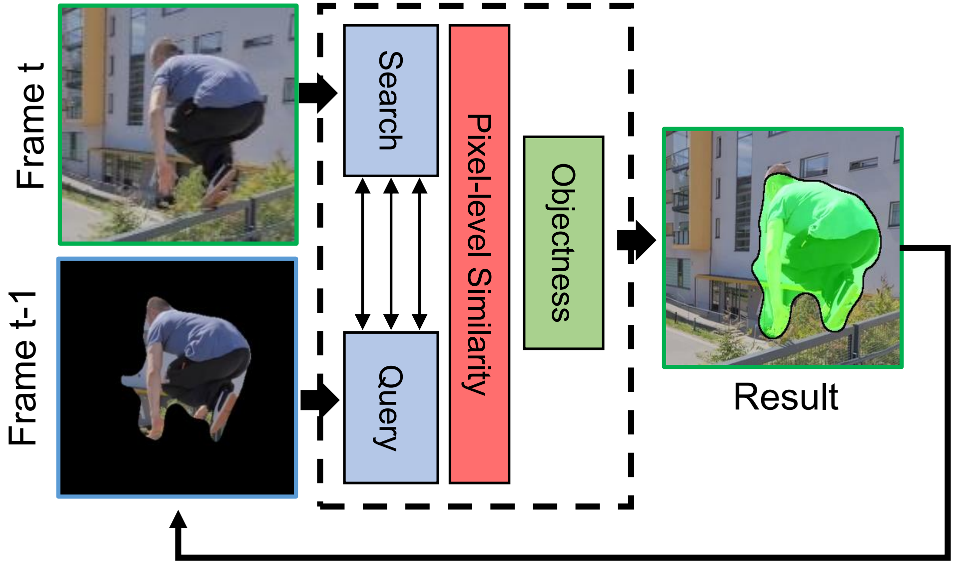
Successive frames matching with model update

Experiments :: *DAVIS* benchmark :: Self−structure evaluation



[PLM]

- **Matching frame t with frame 1**
- **No model update after fine-tuning**



[PLM_U]

- **Matching frame t with frame t-1**
- **Model update after ten-frame processing**

	PLM_S	PLM_U	PLM
Blackswan	0.27	0.86	0.86
	0.20	0.87	0.87
Bmx Tree	0.27	0.40	0.47
	0.43	0.55	0.66
Break Dance	0.06	0.50	0.47
	0.08	0.36	0.35
Came	0.40	0.62	0.65
	0.20	0.49	0.54
Car Roundabout	0.27	0.87	0.86
	0.16	0.68	0.64
Car Shadow	0.43	0.77	0.77
	0.32	0.63	0.62
Cows	0.15	0.79	0.81
	0.10	0.65	0.71
Dance Twirl	0.44	0.61	0.59
	0.35	0.49	0.50
Dog	0.29	0.79	0.73
	0.15	0.63	0.61
Drift Chicane	0.03	0.61	0.73
	0.03	0.60	0.79
Drift Straight	012	0.63	0.73
	0.05	0.50	0.52
Goat	0.05	0.74	0.76
	0.05	0.60	0.63
Horse Jump	0.38	0.71	0.72
	0.24	0.70	0.73
Kite Surf	0.20	0.25	0.59
	0.15	0.17	0.45
Libby	0.23	0.43	0.52
	0.21	0.45	0.57
Motorcross Jump	0.27	0.44	0.49
	0.06	0.27	0.32
Paragliding Launch	0.41	0.54	0.55
	0.11	0.16	0.15
Parkour	0.05	0.57	0.77
	0.08	0.56	0.75
Scooter Black	0.09	0.69	0.73
	0.10	0.53	0.57
Soapbox	0.16	0.70	0.72
	0.18	0.49	0.54
Average	0.23	0.63	0.68
	0.16	0.52	0.58
Speed	0.2s	0.15s+α	0.15s

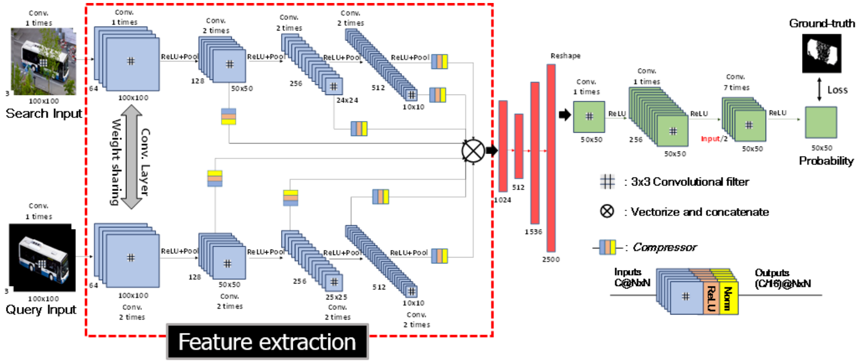
Experiments :: *DAVIS* benchmark :: Comparative evaluation

	HVS	NLC	JMP	SEA	BVS*	PLM	PLM_P
Blackswan	0.91	0.87	0.93	0.93	0.94	0.86	0.89
	0.91	0.82	0.94	0.95	0.96	0.87	0.89
Bmx Tree	0.18	0.21	0.23	0.11	0.38	0.47	0.48
	0.28	0.33	0.31	0.13	0.65	0.66	0.68
Break Dance	0.55	0.67	0.48	0.33	0.50	0.47	0.48
	0.47	0.66	0.51	0.39	0.49	0.35	0.41
Camel	0.87	0.76	0.64	0.65	0.67	0.65	0.68
	0.87	0.72	0.71	0.61	0.70	0.54	0.61
Car Roundabout	0.77	0.50	0.72	0.70	0.85	0.86	0.87
	0.55	0.25	0.61	0.71	0.62	0.64	0.71
Car Shadow	0.70	0.64	0.64	0.77	0.57	0.77	0.79
	0.59	0.54	0.62	0.75	0.47	0.62	0.73
Cows	0.78	0.88	0.75	0.71	0.89	0.81	0.83
	0.63	0.80	0.7	0.67	0.85	0.71	0.74
Dance Twirl	0.31	0.34	0.44	0.11	0.49	0.59	0.62
	0.51	0.36	0.52	0.21	0.48	0.50	0.54
Dog	0.72	0.81	0.67	0.58	0.72	0.73	0.74
	0.63	0.70	0.59	0.54	0.59	0.61	0.63
Drift Chicane	0.33	0.32	0.24	0.12	0.03	0.73	0.71
	0.54	0.31	0.33	0.16	0.07	0.79	0.79
Drift Straight	0.29	0.47	0.61	0.51	0.40	0.73	0.74
	0.26	0.38	0.47	0.50	0.41	0.52	0.56
Goat	0.58	0.01	0.73	0.53	0.66	0.76	0.77
	0.54	0.13	0.61	0.47	0.58	0.63	0.69
Horse Jump	0.76	0.83	0.58	0.63	0.80	0.72	0.77
	0.80	0.88	0.65	0.65	0.80	0.73	0.78
Kite Surf	0.40	0.45	0.50	0.48	0.42	0.59	0.64
	0.37	0.45	0.31	0.28	0.64	0.45	0.45
Libby	0.55	0.63	0.29	0.22	0.77	0.52	0.54
	0.64	0.74	0.36	0.21	0.84	0.57	0.58
Motorcross Jump	0.09	0.25	0.58	0.38	0.34	0.49	0.51
	0.13	0.30	0.54	0.40	0.37	0.32	0.37
Paragliding Launch	0.53	0.62	0.59	0.57	0.64	0.55	0.57
	0.20	0.24	0.17	0.18	0.32	0.15	0.18
Parkour	0.24	0.90	0.34	0.12	0.75	0.77	0.82
	0.32	0.91	0.41	0.27	0.67	0.75	0.81
Scooter Black	0.62	0.16	0.62	0.79	0.33	0.73	0.75
	0.57	0.22	0.53	0.72	0.40	0.57	0.64
Soapbox	0.68	0.63	0.75	0.78	0.79	0.72	0.76
	0.69	0.65	0.67	0.75	0.76	0.54	0.62
Average	0.54	0.55	0.56	0.50	0.59	0.68	0.70
	0.52	0.52	0.53	0.47	0.58	0.58	0.62
Speed	5s	20s	12s	6s	0.37s	0.15s	0.3s

- PLM_P : post-processed results using weighted median filter [1]

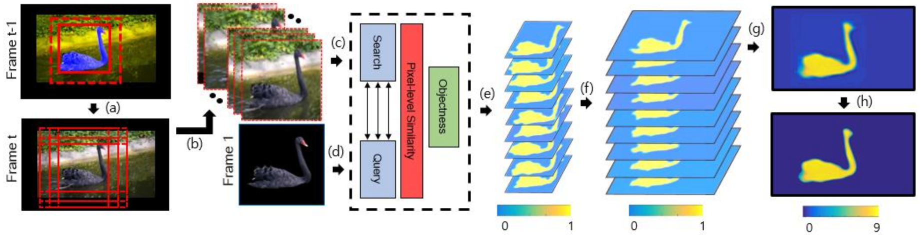
- Our method outperforms all the other approaches in terms of both accuracy and speed
- Our pixel-level matching network can encompass the semantic-level information as well as spatial details

  :: robust to many challenging scenarios



- Little temporal correlation between successive frames

  :: prevent drift effect



[1] Zhang, Qi, Li Xu, and Jiaya Jia. "100+ times faster weighted median filter (WMF)." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2014.

Experiments :: *DAVIS* benchmark :: Comparative evaluation

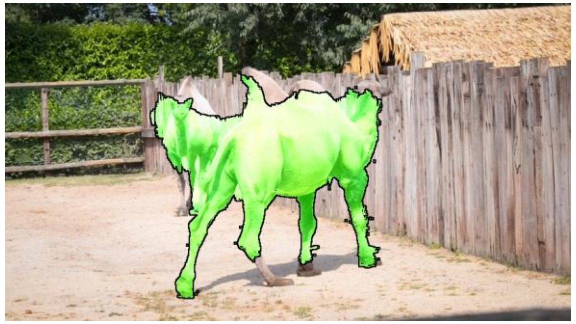
	HVS	NLC	JMP	SEA	BVS*	PLM	PLM_P
Blackswan	0.91	0.87	0.93	0.93	0.94	0.86	0.89
	0.91	0.82	0.94	0.95	0.96	0.87	0.89
Bmx Tree	0.18	0.21	0.23	0.11	0.38	0.47	0.48
	0.28	0.33	0.31	0.13	0.65	0.66	0.68
Break Dance	0.55	0.67	0.48	0.33	0.50	0.47	0.48
	0.47	0.66	0.51	0.39	0.49	0.35	0.41
Camel	0.87	0.76	0.64	0.65	0.67	0.65	0.68
	0.87	0.72	0.71	0.61	0.70	0.54	0.61
Car Roundabout	0.77	0.50	0.72	0.70	0.85	0.86	0.87
	0.55	0.25	0.61	0.71	0.62	0.64	0.71
Car Shadow	0.70	0.64	0.64	0.77	0.57	0.77	0.79
	0.59	0.54	0.62	0.75	0.47	0.62	0.73
Cows	0.78	0.88	0.75	0.71	0.89	0.81	0.83
	0.63	0.80	0.7	0.67	0.85	0.71	0.74
Dance Twirl	0.31	0.34	0.44	0.11	0.49	0.59	0.62
	0.51	0.36	0.52	0.21	0.48	0.50	0.54
Dog	0.72	0.81	0.67	0.58	0.72	0.73	0.74
	0.63	0.70	0.59	0.54	0.59	0.61	0.63
Drift Chicane	0.33	0.32	0.24	0.12	0.03	0.73	0.71
	0.54	0.31	0.33	0.16	0.07	0.79	0.79
Drift Straight	0.29	0.47	0.61	0.51	0.40	0.73	0.74
	0.26	0.38	0.47	0.50	0.41	0.52	0.56
Goat	0.58	0.01	0.73	0.53	0.66	0.76	0.77
	0.54	0.13	0.61	0.47	0.58	0.63	0.69
Horse Jump	0.76	0.83	0.58	0.63	0.80	0.72	0.77
	0.80	0.88	0.65	0.65	0.80	0.73	0.78
Kite Surf	0.40	0.45	0.50	0.48	0.42	0.59	0.64
	0.37	0.45	0.31	0.28	0.64	0.45	0.45
Libby	0.55	0.63	0.29	0.22	0.77	0.52	0.54
	0.64	0.74	0.36	0.21	0.84	0.57	0.58
Motorcross Jump	0.09	0.25	0.58	0.38	0.34	0.49	0.51
	0.13	0.30	0.54	0.40	0.37	0.32	0.37
Paragliding Launch	0.53	0.62	0.59	0.57	0.64	0.55	0.57
	0.20	0.24	0.17	0.18	0.32	0.15	0.18
Parkour	0.24	0.90	0.34	0.12	0.75	0.77	0.82
	0.32	0.91	0.41	0.27	0.67	0.75	0.81
Scooter Black	0.62	0.16	0.62	0.79	0.33	0.73	0.75
	0.57	0.22	0.53	0.72	0.40	0.57	0.64
Soapbox	0.68	0.63	0.75	0.78	0.79	0.72	0.76
	0.69	0.65	0.67	0.75	0.76	0.54	0.62
Average	**0.54**	**0.55**	**0.56**	**0.50**	**0.59**	**0.68**	**0.70**
	0.52	**0.52**	**0.53**	**0.47**	**0.58**	**0.58**	**0.62**
Speed	**5s**	**20s**	**12s**	**6s**	**0.37s**	**0.15s**	**0.3s**

Limitation

- The object having same class and same appearance with the target object leads to network confusion

[Input]

[Result]

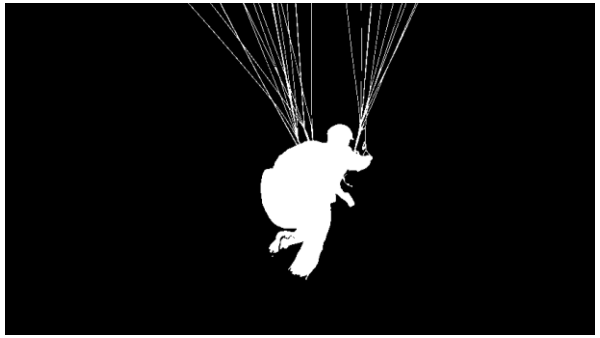


(sequence name: Camel)

- Difficult to handle thin object

[Result]

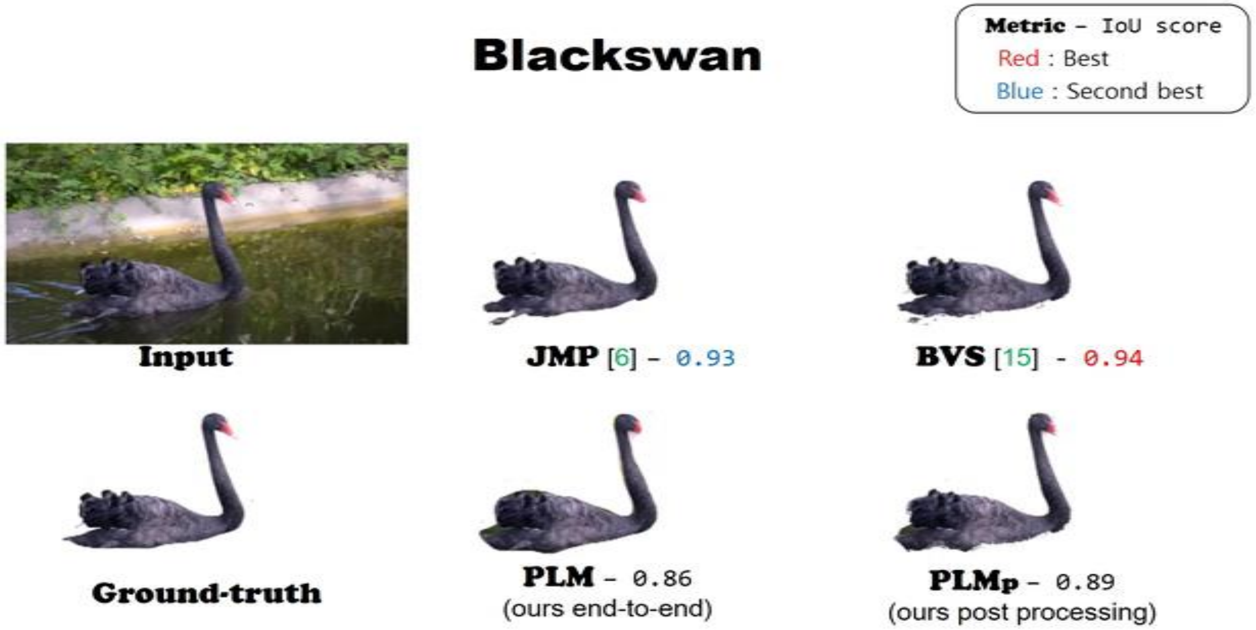
[Ground truth]

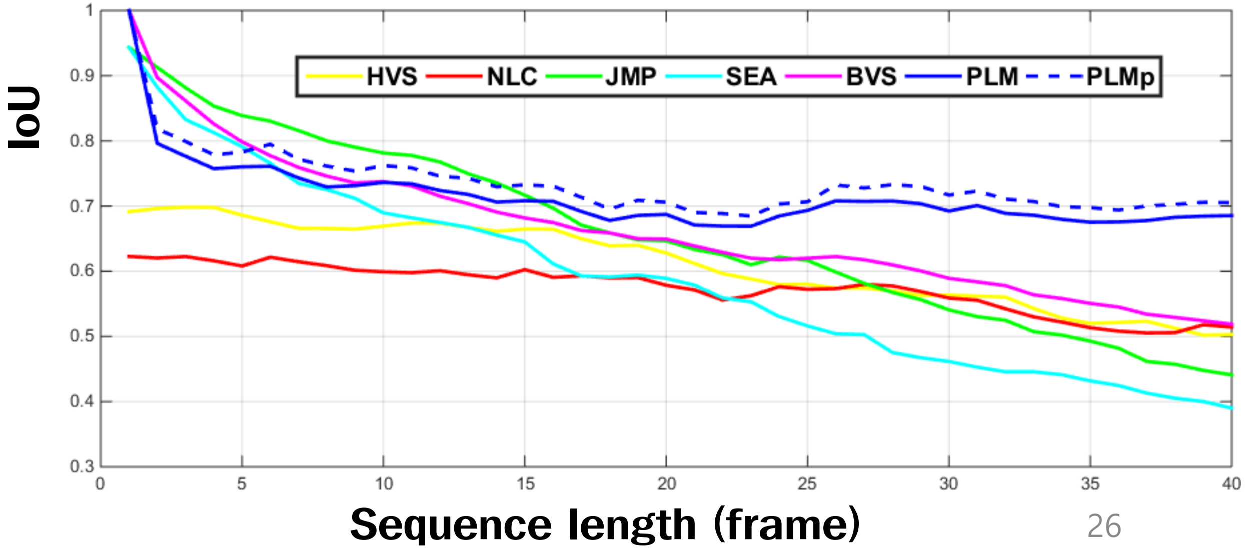


(sequence name: Paragliding Launch)

Experiments :: *DAVIS* benchmark :: Comparative evaluation

	HVS	NLC	JMP	SEA	BVS*	PLM	PLM_P
Blackswan	0.91	0.87	0.93	0.93	0.94	0.86	0.89
	0.91	0.82	0.94	0.95	0.96	0.87	0.89
Bmx Tree	0.18	0.21	0.23	0.11	0.38	0.47	0.48
	0.28	0.33	0.31	0.13	0.65	0.66	0.68
Break Dance	0.55	0.67	0.48	0.33	0.50	0.47	0.48
	0.47	0.66	0.51	0.39	0.49	0.35	0.41
Camel	0.87	0.76	0.64	0.65	0.67	0.65	0.68
	0.87	0.72	0.71	0.61	0.70	0.54	0.61
Car Roundabout	0.77	0.50	0.72	0.70	0.85	0.86	0.87
	0.55	0.25	0.61	0.71	0.62	0.64	0.71
Car Shadow	0.70	0.64	0.64	0.77	0.57	0.77	0.79
	0.59	0.54	0.62	0.75	0.47	0.62	0.73
Cows	0.78	0.88	0.75	0.71	0.89	0.81	0.83
	0.63	0.80	0.7	0.67	0.85	0.71	0.74
Dance Twirl	0.31	0.34	0.44	0.11	0.49	0.59	0.62
	0.51	0.36	0.52	0.21	0.48	0.50	0.54
Dog	0.72	0.81	0.67	0.58	0.72	0.73	0.74
	0.63	0.70	0.59	0.54	0.59	0.61	0.63
Drift Chicane	0.33	0.32	0.24	0.12	0.03	0.73	0.71
	0.54	0.31	0.33	0.16	0.07	0.79	0.79
Drift Straight	0.29	0.47	0.61	0.51	0.40	0.73	0.74
	0.26	0.38	0.47	0.50	0.41	0.52	0.56
Goat	0.58	0.01	0.73	0.53	0.66	0.76	0.77
	0.54	0.13	0.61	0.47	0.58	0.63	0.69
Horse Jump	0.76	0.83	0.58	0.63	0.80	0.72	0.77
	0.80	0.88	0.65	0.65	0.80	0.73	0.78
Kite Surf	0.40	0.45	0.50	0.48	0.42	0.59	0.64
	0.37	0.45	0.31	0.28	0.64	0.45	0.45
Libby	0.55	0.63	0.29	0.22	0.77	0.52	0.54
	0.64	0.74	0.36	0.21	0.84	0.57	0.58
Motorcross Jump	0.09	0.25	0.58	0.38	0.34	0.49	0.51
	0.13	0.30	0.54	0.40	0.37	0.32	0.37
Paragliding Launch	0.53	0.62	0.59	0.57	0.64	0.55	0.57
	0.20	0.24	0.17	0.18	0.32	0.15	0.18
Parkour	0.24	0.90	0.34	0.12	0.75	0.77	0.82
	0.32	0.91	0.41	0.27	0.67	0.75	0.81
Scooter Black	0.62	0.16	0.62	0.79	0.33	0.73	0.75
	0.57	0.22	0.53	0.72	0.40	0.57	0.64
Soapbox	0.68	0.63	0.75	0.78	0.79	0.72	0.76
	0.69	0.65	0.67	0.75	0.76	0.54	0.62
Average	**0.54**	**0.55**	**0.56**	**0.50**	**0.59**	**0.68**	**0.70**
	0.52	**0.52**	**0.53**	**0.47**	**0.58**	**0.58**	**0.62**
Speed	**5s**	**20s**	**12s**	**6s**	**0.37s**	**0.15s**	**0.3s**

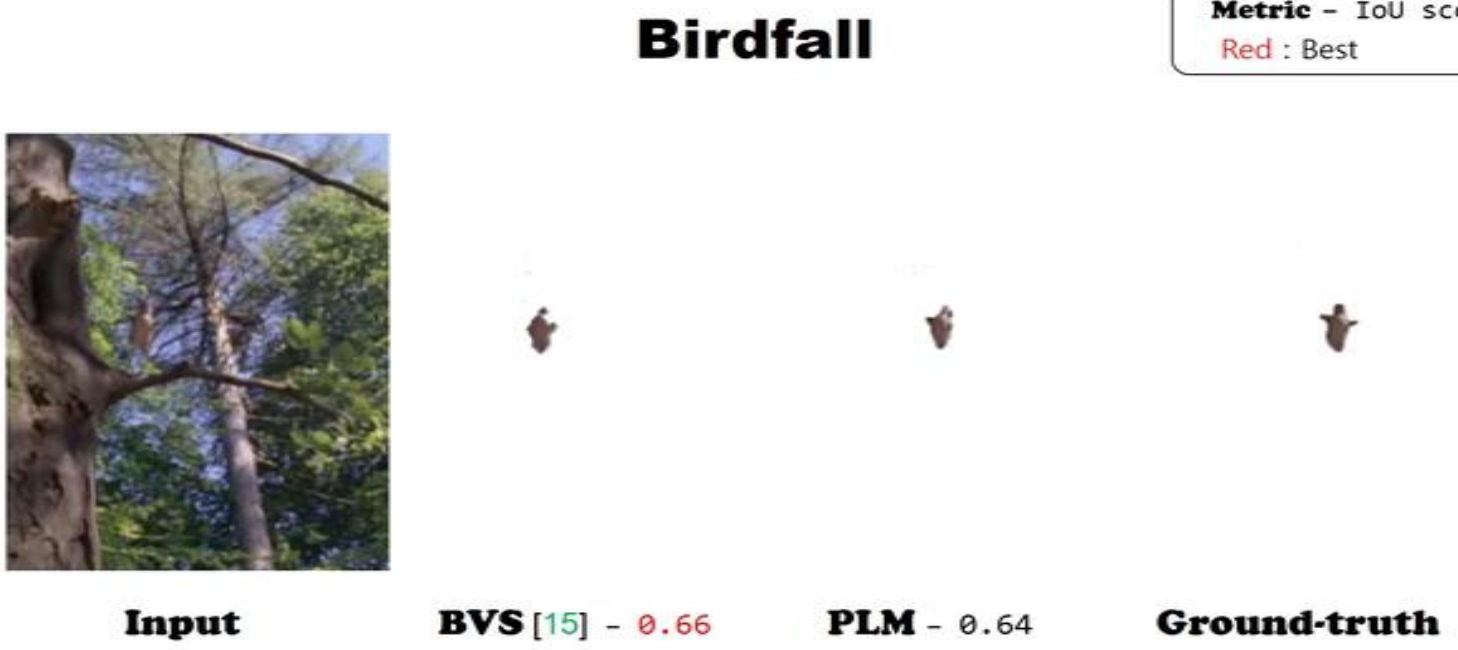




Experiments :: *SegTrack* benchmark

	HSV	FST	DAG	TMF	KEY	NLC	BVS	PLM	PLM_p
Birddfall	0.57	0.59	0.71	0.62	0.49	0.74	0.66	0.64	0.65
Cheetah	0.19	0.28	0.40	0.37	0.44	0.69	0.10	0.65	0.65
Girl	0.32	0.73	0.82	0.89	0.88	0.91	0.89	0.77	0.78
Monkeydog	0.68	0.79	0.75	0.71	0.74	0.78	0.41	0.61	0.72
Parachute	0.69	0.91	0.94	0.93	0.96	0.94	0.94	0.85	0.88
Average	0.49	0.66	0.72	0.70	0.70	0.81	0.70	0.70	0.73

Metric – IoU score
Red : Best

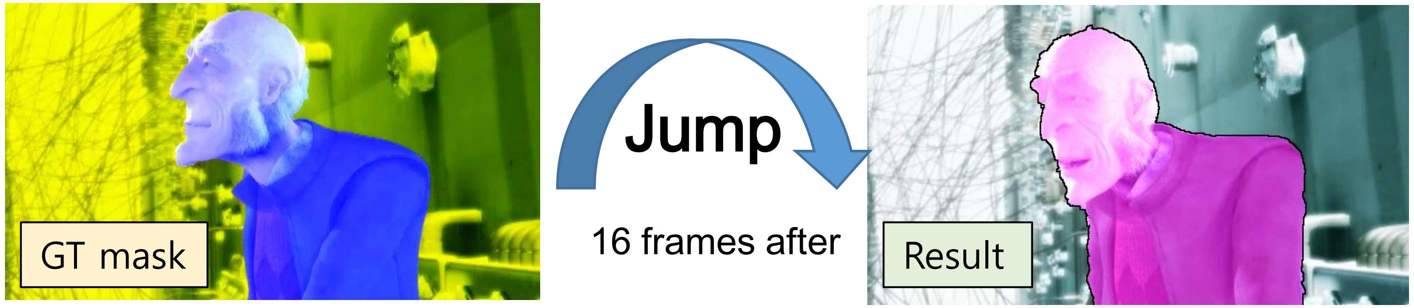


- Measure (higher is better)
  :: Intersection_over_Union
- Our results are comparable
- Low resolution video causes confusion of network

Experiments :: *JumpCut* benchmark

Why *JumpCut* [1] ?

- Mask propagation task validation on non-successive frames scenario such as…



- GT masks are given at each key frame in the middle of sequence, and propagate the mask to 16-after frame.

Measuare

$$\text{Err} = \frac{100}{N_i} \sum_i \frac{\#of\ \text{mislabled–pixels at } (i+d)^{th} frame}{\#of\ \text{foreground–pixels } at\ (i+d)^{th} frame}$$

*N_i = number of key frames

	RB	DA	SEA	JMP	PLM	PLM_P
Animation	11.9	6.38	7.78	4.55	9.38	5.86
Bball	18.4	8.47	8.89	3.90	12.8	8.04
Bear	4.58	4.48	4.21	4.00	8.60	3.45
Car	1.76	5.93	5.08	2.26	4.12	2.18
Cheetah	31.5	16.6	7.68	8.16	14.1	11.8
Couple	17.5	16.0	23.4	5.13	13.1	9.14
Cup	5.45	12.9	9.31	2.15	8.63	6.04
Dance	56.1	50.8	43.0	18.7	31.5	14.7
Fish	51.8	21.7	25.7	17.5	9.39	7.42
Giraffe	22.0	11.2	17.4	7.40	19.7	17.4
Goat	13.1	13.3	8.22	4.14	17.8	15.2
Hiphop	67.5	51.1	33.7	14.2	19.9	13.6
Horse	8.39	45.1	37.8	6.80	11.5	7.94
Kongfu	40.2	40.8	17.9	8.00	9.74	6.25
Park	11.8	6.54	6.91	5.39	16.5	10.2
Pig	9.22	9.85	10.3	3.43	9.09	5.15
Pot	2.43	5.03	2.98	2.95	5.46	2.66
Skater	38.7	40.8	29.6	22.8	16.8	12.6
Station	8.85	20.9	21.3	9.01	7.31	4.68
Supertramp	129	60.5	57.4	42.9	30.4	20.7
Toy	1.28	3.19	2.16	1.30	5.07	2.25
Tricking	79.4	70.9	35.8	21.3	21.9	15.7
Average	28.6	23.7	18.8	9.81	13.7	9.55

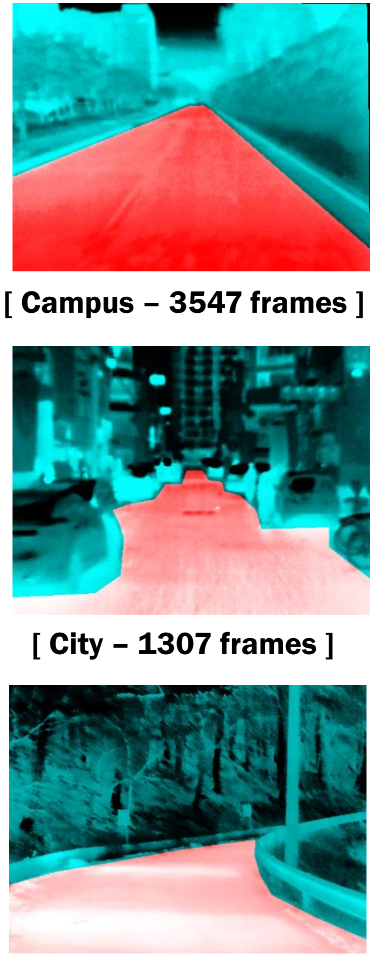
[1] Fan.Q., et al. Jumpcut: non-successive mask transfer and interpolation for video cutout. *ACM Transactions on Graphics (TOG), 34(6):195, 2015.*

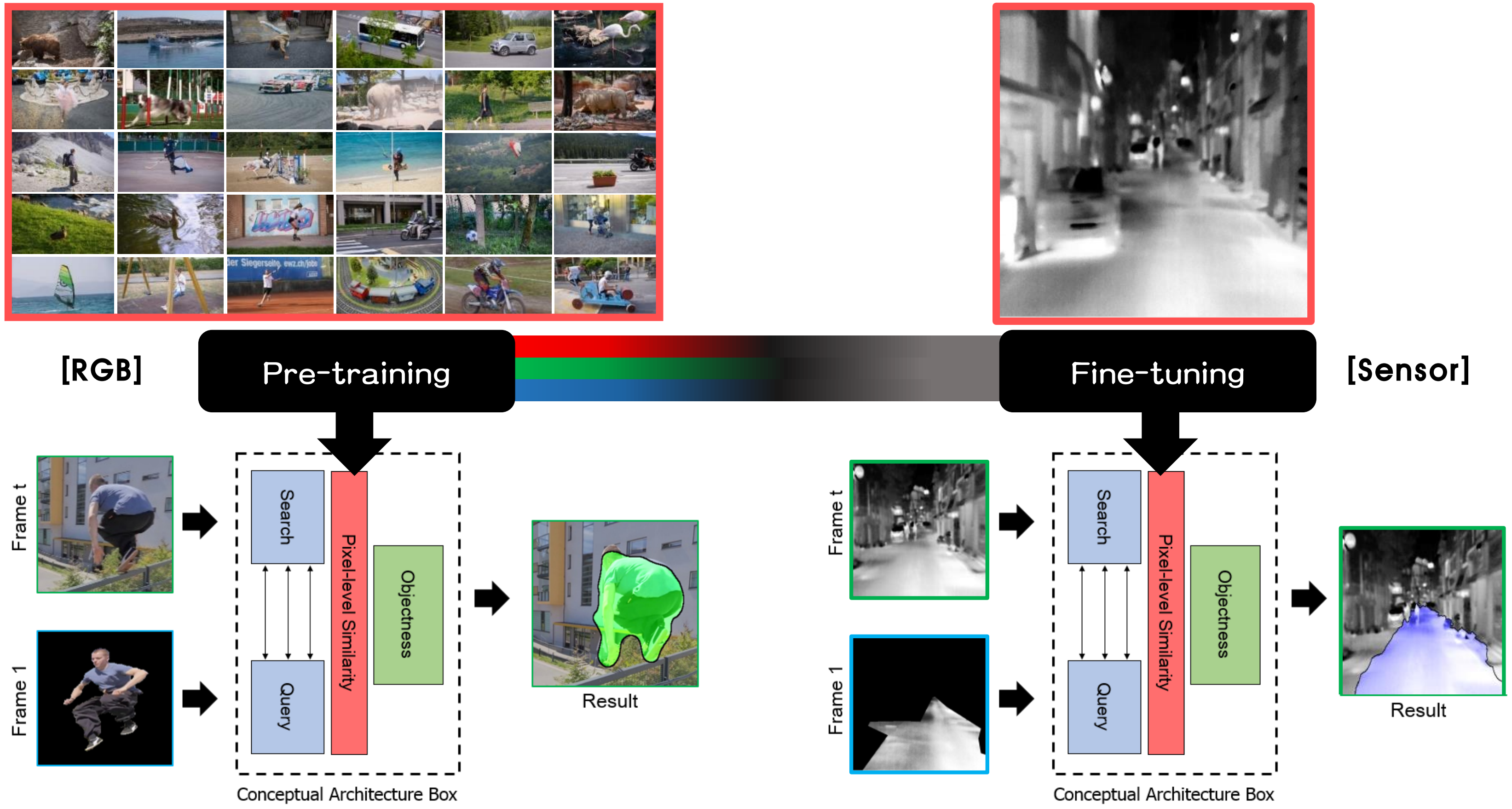
Experiments :: *JumpCut* benchmark

Experiments :: *JumpCut* benchmark

Experiments :: *Thermal−Road* benchmark

Drivable Region Tracking



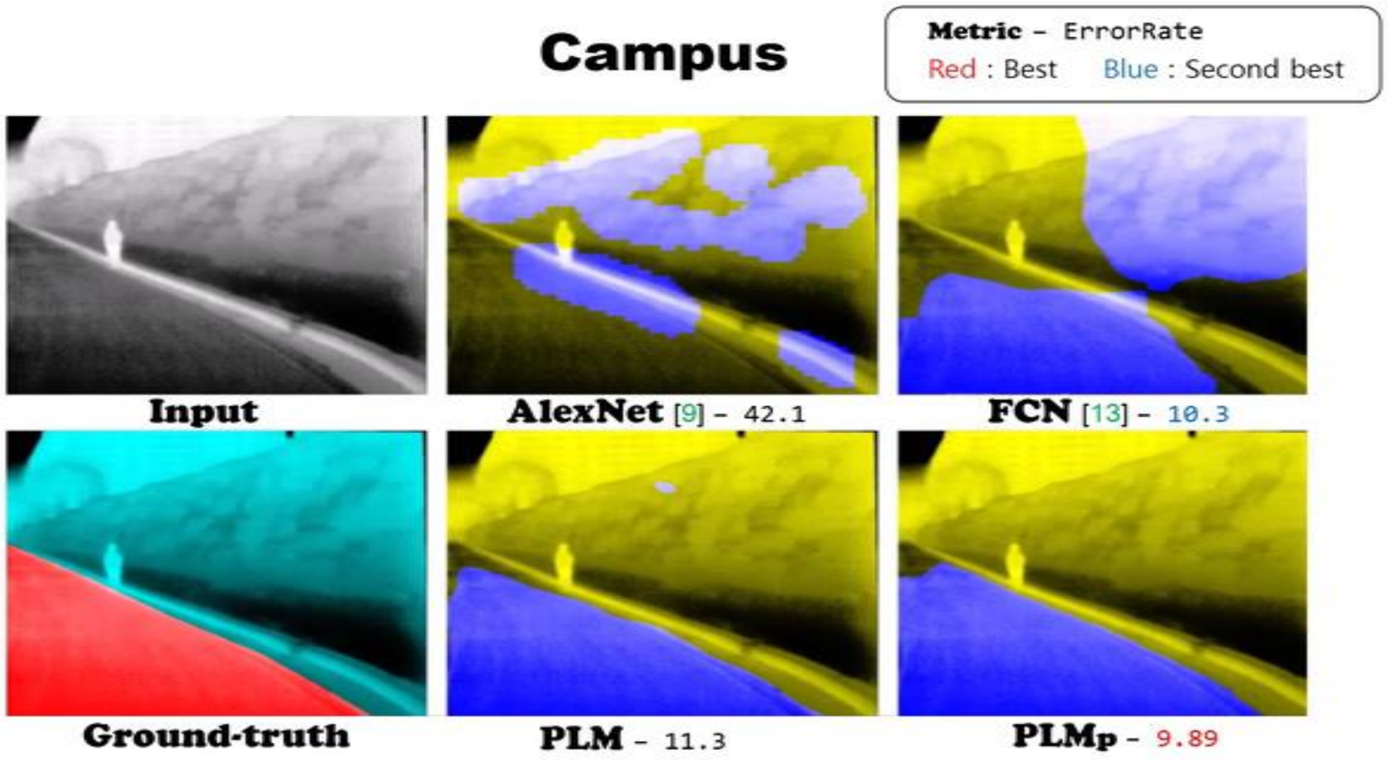


Network domain is transferred by fine-tuning with only a frame

[1] Yoon.J. S., et al. Thermal-infrared based drivable region detection. *In Intelligent Vehicles Symposium (IV), 2016 IEEE, pages 978–985. IEEE, 2016.*

Experiments :: *Thermal−Road* benchmark

	Classifier	GMM	VP	GrowCut	TD	AlexNet	CN24	FCN	PLM	PLM_P
Campus	17.6	22.9	15.1	19.8	10.7	42.1	36.4	10.3	11.3	9.89
Moutnain	29.8	24.1	43.9	26.1	42.1	41.1	21.1	6.34	11.4	10.4
City	19.7	21.9	16.6	21.3	12.6	29.1	28.0	9.77	11.9	11.5
Average	20.7	22.4	21.7	21.5	11.8	39.0	36.7	9.45	11.5	10.6
	:: Hand-crafted feature based approach					:: Deep learning based approach				



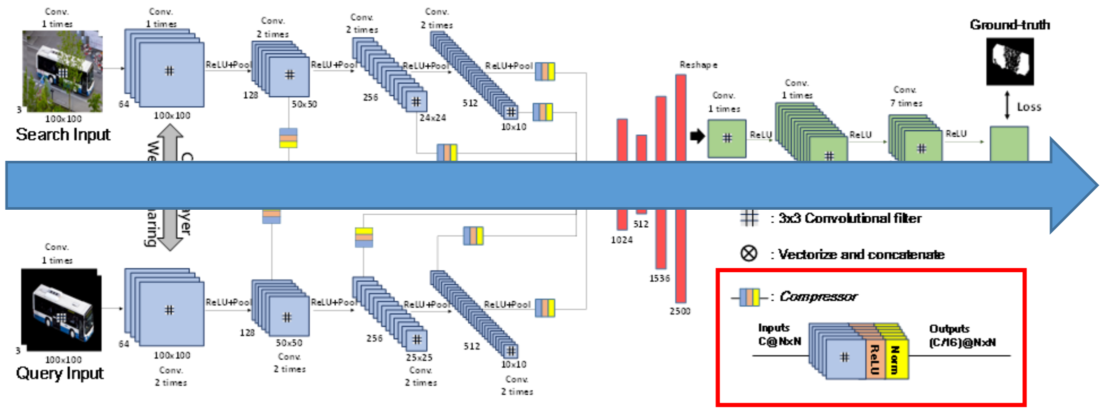
Our method,

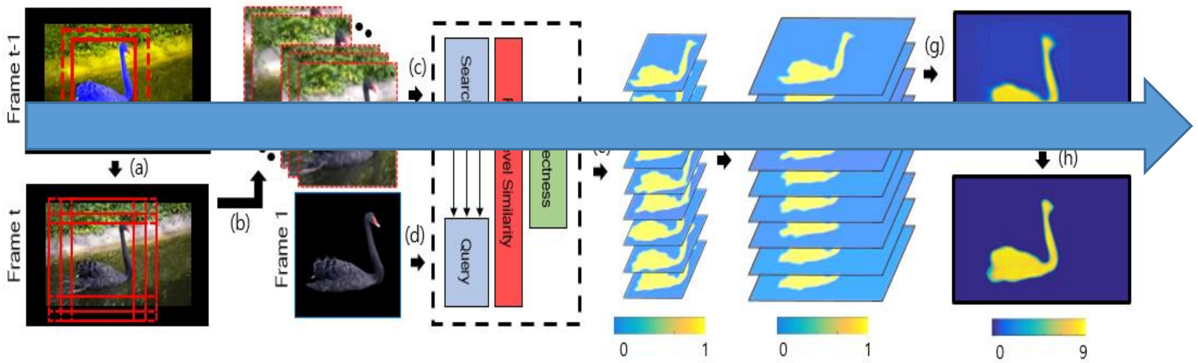
- outperforms state-of-the-art hand crafted feature based approach.
- takes the second best using only a initial frame fine-tuning.

AlexNet, CN24	FCN, Ours
Patch based approach	Encoding-decoding base approach
Local information	Global+local information

Contribution

Usability | **Transferability** | **Capability**





- **One feed forword : 0.008 s**

  [thanks to the *Compressor*]

- **Whole processing time : 0.15 s**

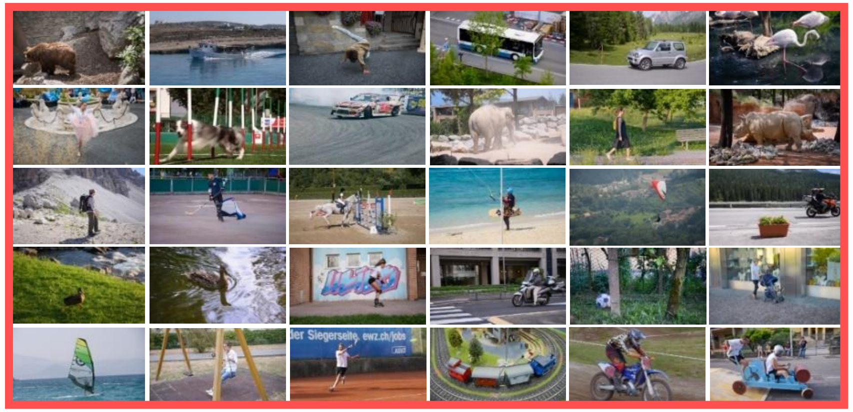
- **About 3.7GB (batch size: 32) for training**

  [training and testing are available with GTX 750 Ti (2GB)]

- **Usable with only a single frame fine tuning**

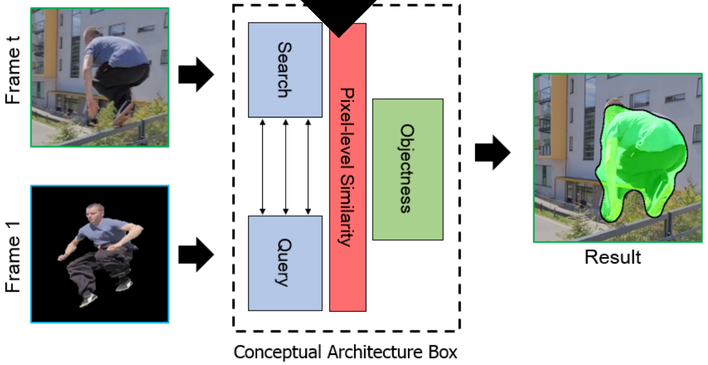
Contribution

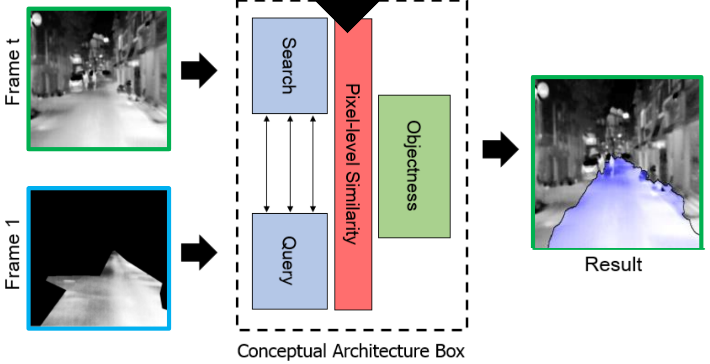
Usability | **Transferability** | **Capability**



[RGB data] **Pre-training**

Fine-tuning [Sensor data]

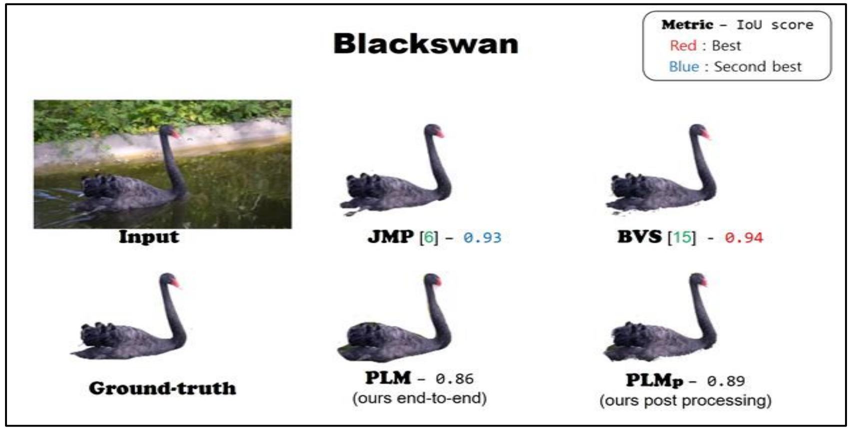


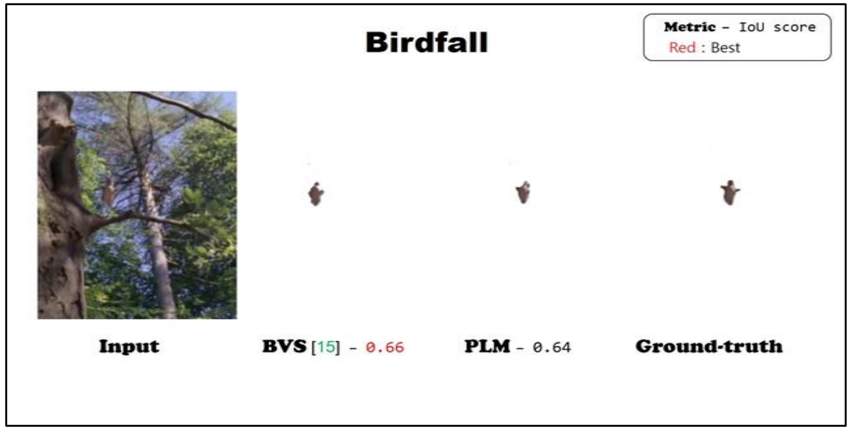


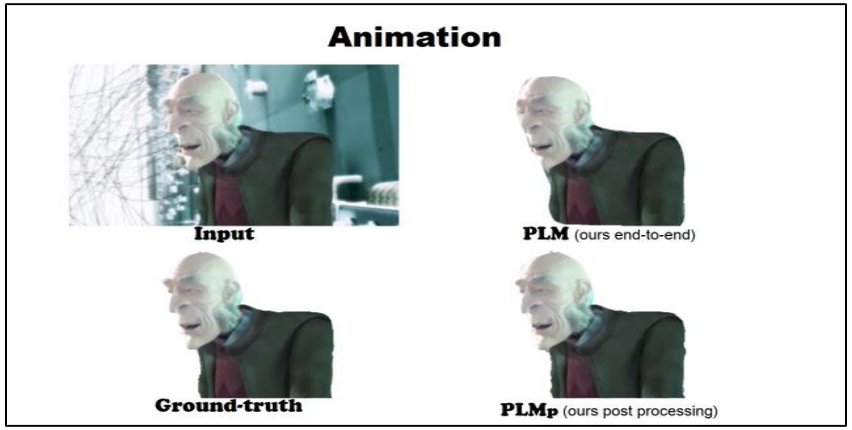
Network domain is transferred by fine-tuning with only a frame

Contribution

Usability **Transferability** **Capability**







DAVIS [1] — **Best**

SegTrack [2] — **Second**

JumpCut [3] — **Best**

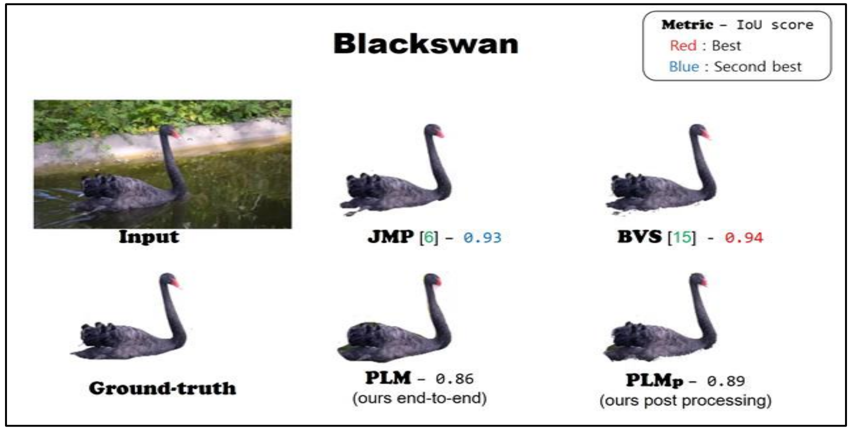
[1] Perazzi, F., et al. "A Benchmark Dataset and Evaluation Methodology for Video Object Segmentation.". *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016.

[2] F. Li, T. Kim, A. Humayun, D. Tsai, and J. M. Rehg. Video segmentation by tracking many figure-ground segments. In ICCV, 2013.

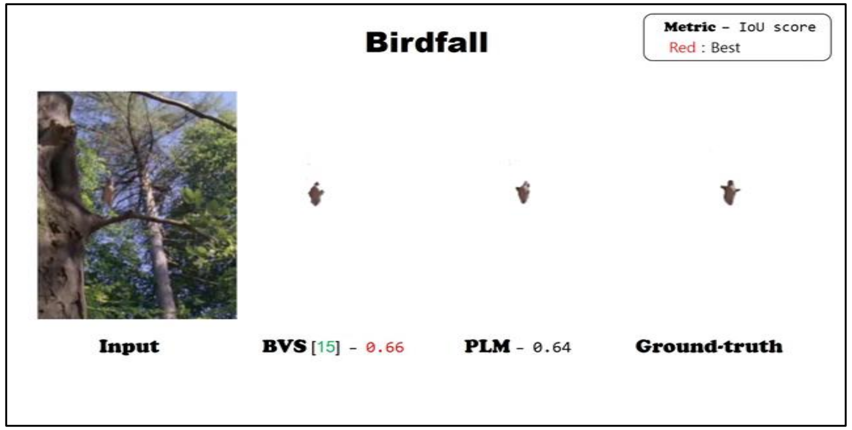
[3] Q. Fan, F. Zhong, D. Lischinski, D. Cohen-Or, and B. Chen. Jumpcut: non-successive mask transfer and interpolation for video cutout. ACM Transactions on Graphics (TOG), 34(6):195, 2015.

Contribution

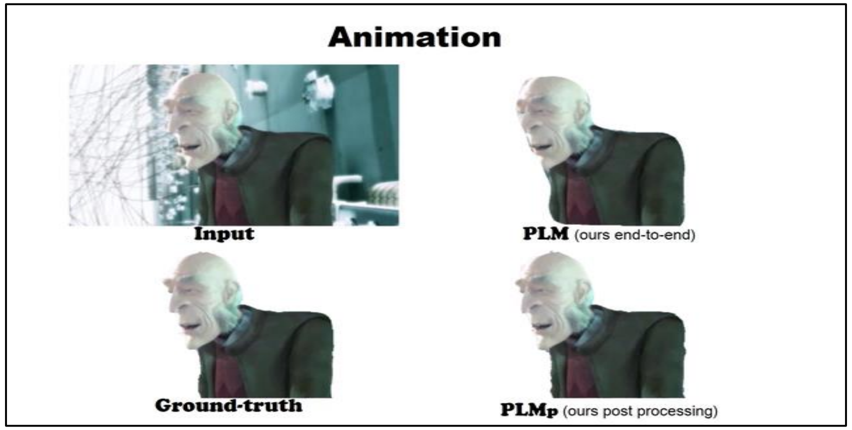
Usability | **Transferability** | **Capability**



DAVIS [1]



SegTrack [2]



JumpCut [3]

Thank you