# Video Object Detection Using Object's Motion Context and Spatio-Temporal Feature Aggregation

*Jaekyum Kim[1]*, Junho Koh[1]*, Byeongwon Lee[2], Seungji Yang[2], and Jun Won Choi[1]*

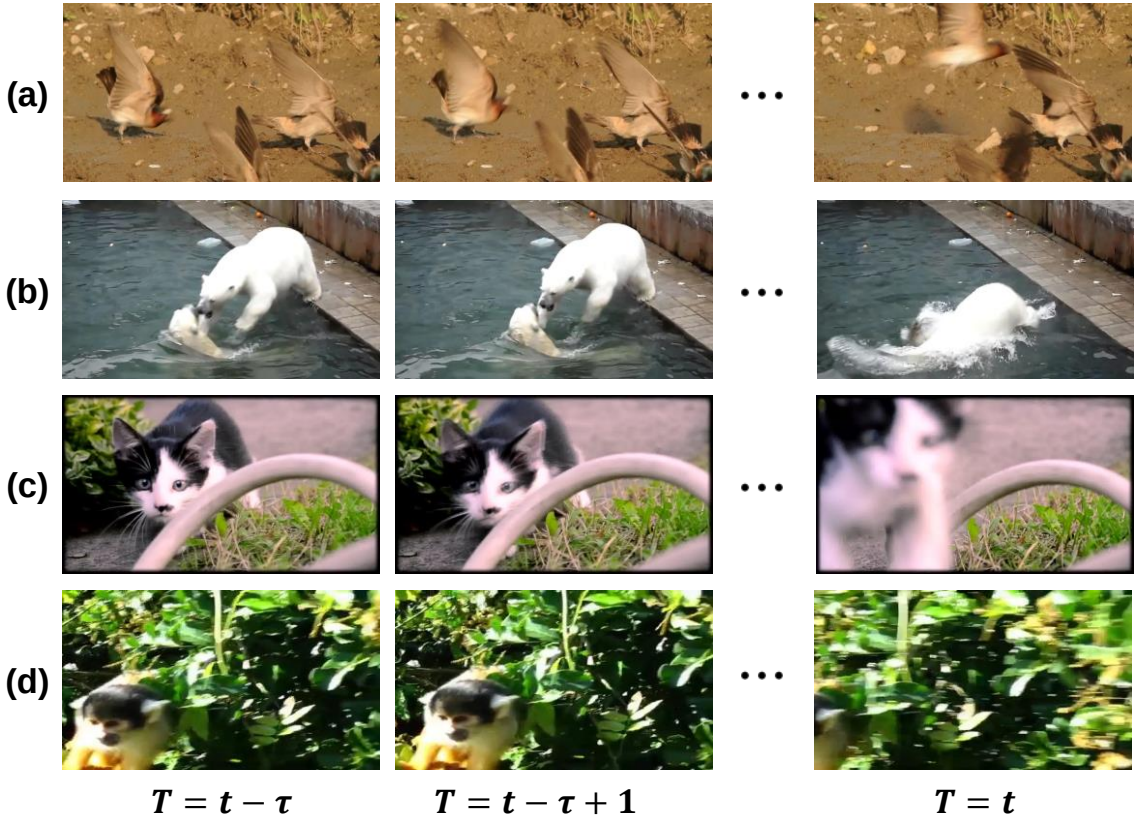
*[1] Hanyang University, [2] Seoul and SK Telecom*

** Indicates the equal contribution*

# Contents

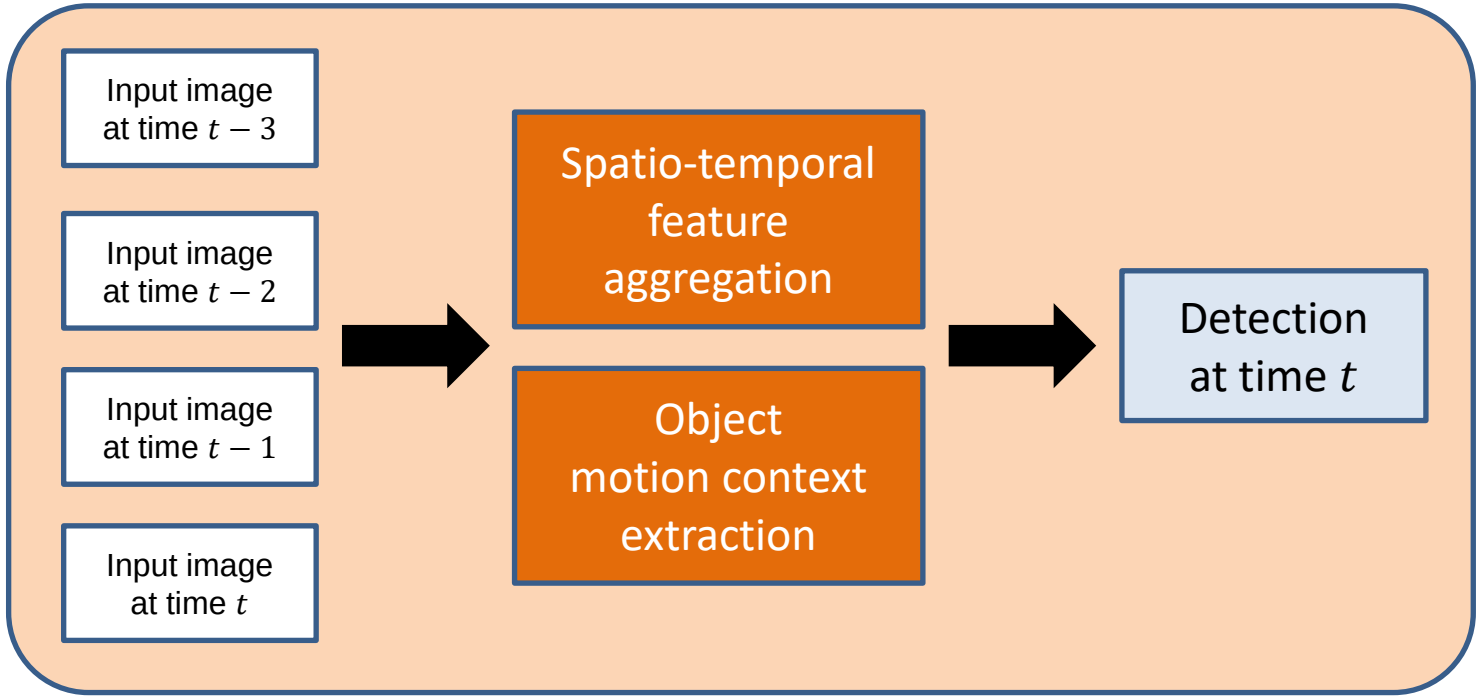
# Video object detection

- Static object detection
  - Static object detector uses **only single shot image**.
- Challenges of video object detection
  - **Degraded image quality** due to object motion and camera moving in video data
    - Motion blur ((a)) and camera defocusing ((c))
    - Anomalous poses ((a), (b), (c)) and object occlusion ((d))
  - Exploit **abundant temporal information** to solve the problems



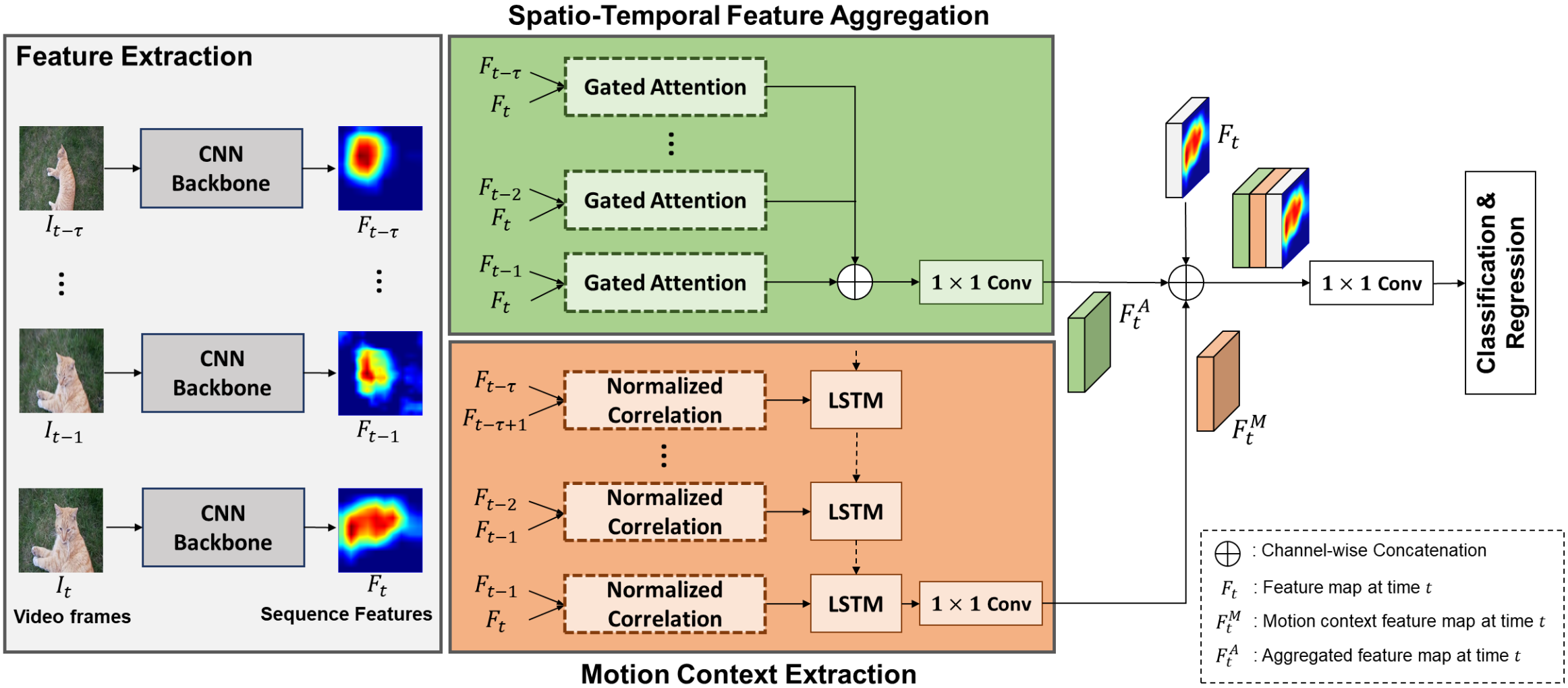
# Proposed method (VOD-MT)

- Key contribution
  - Two approach to exploit the useful temporal information
    - **Temporal redundancy** from spatio-temporal feature aggregation to recover the visual defect such as motion blur and camera defocusing
    - Contextual information captured in the **object motion** to recover the anomalous poses and object occlusion
  - Applicable to every one-stage detectors.

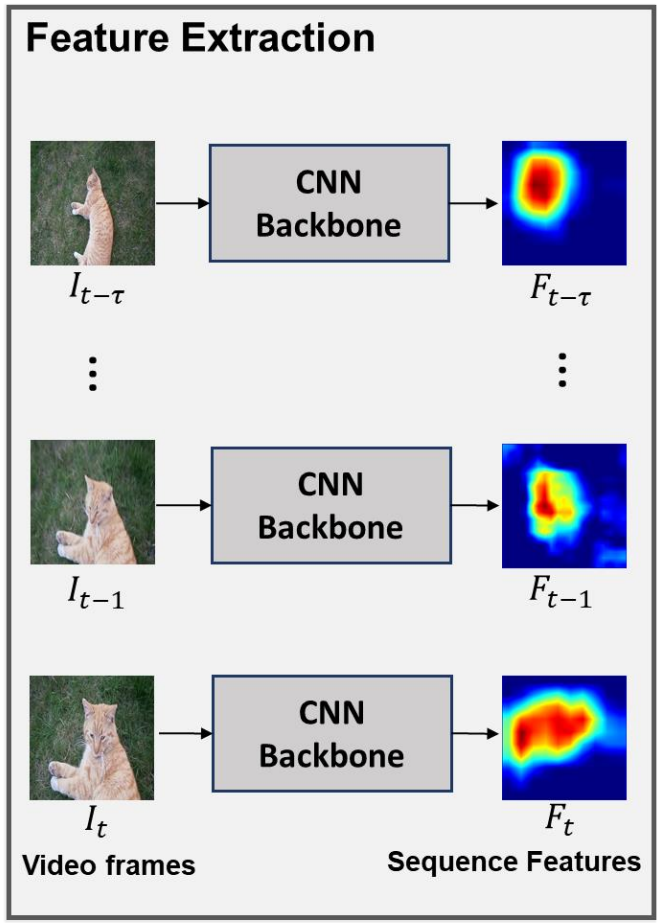


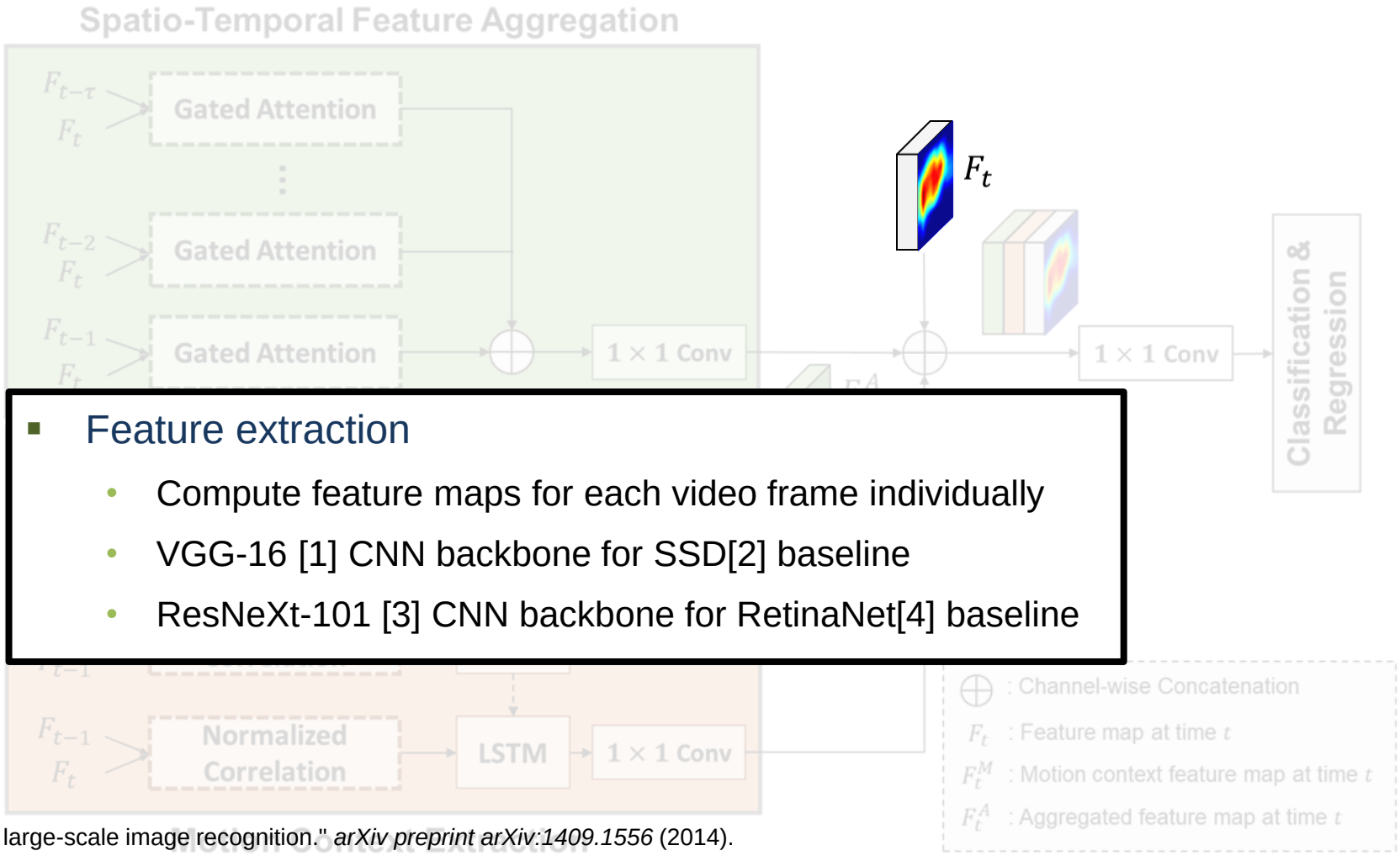
# Proposed method (VOD-MT)

- Overall architecture



# Feature extraction

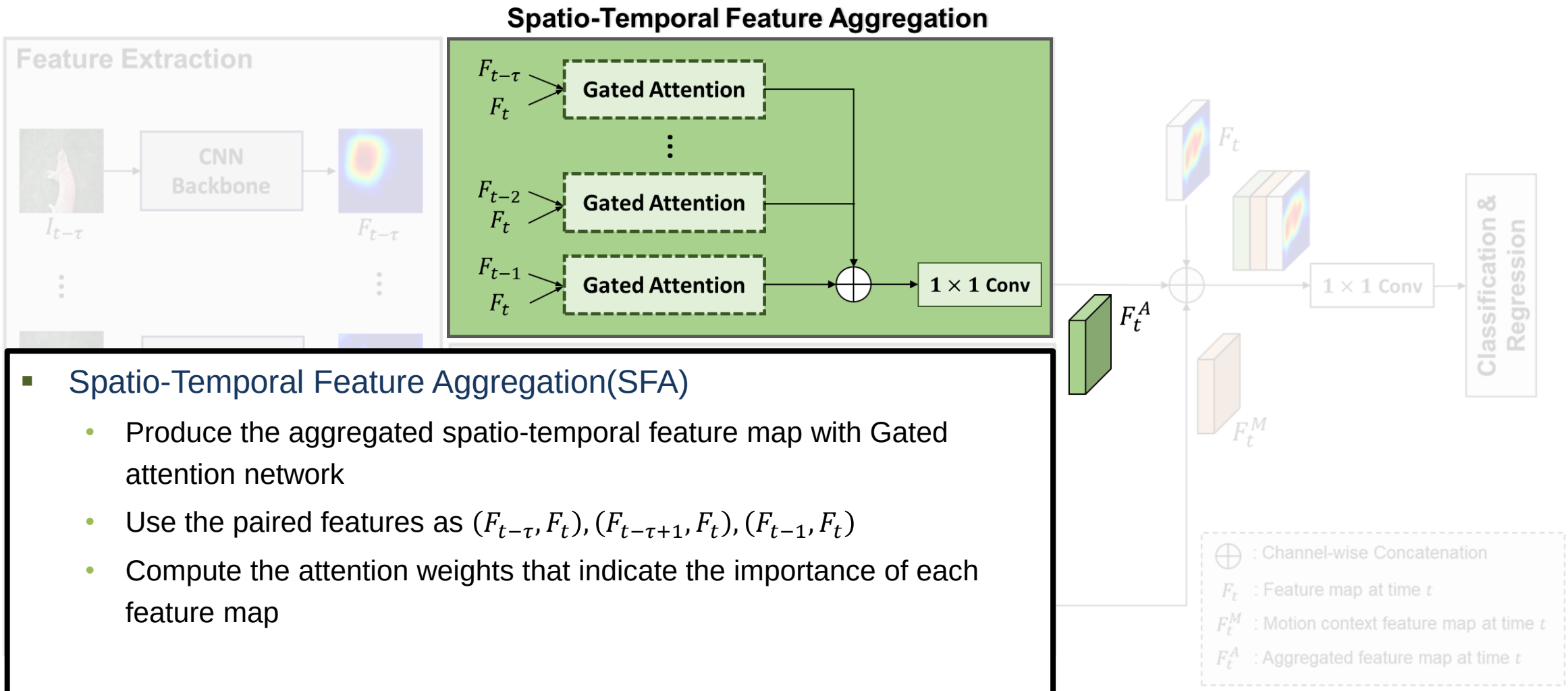



- Feature extraction
  - Compute feature maps for each video frame individually
  - VGG-16 [1] CNN backbone for SSD[2] baseline
  - ResNeXt-101 [3] CNN backbone for RetinaNet[4] baseline

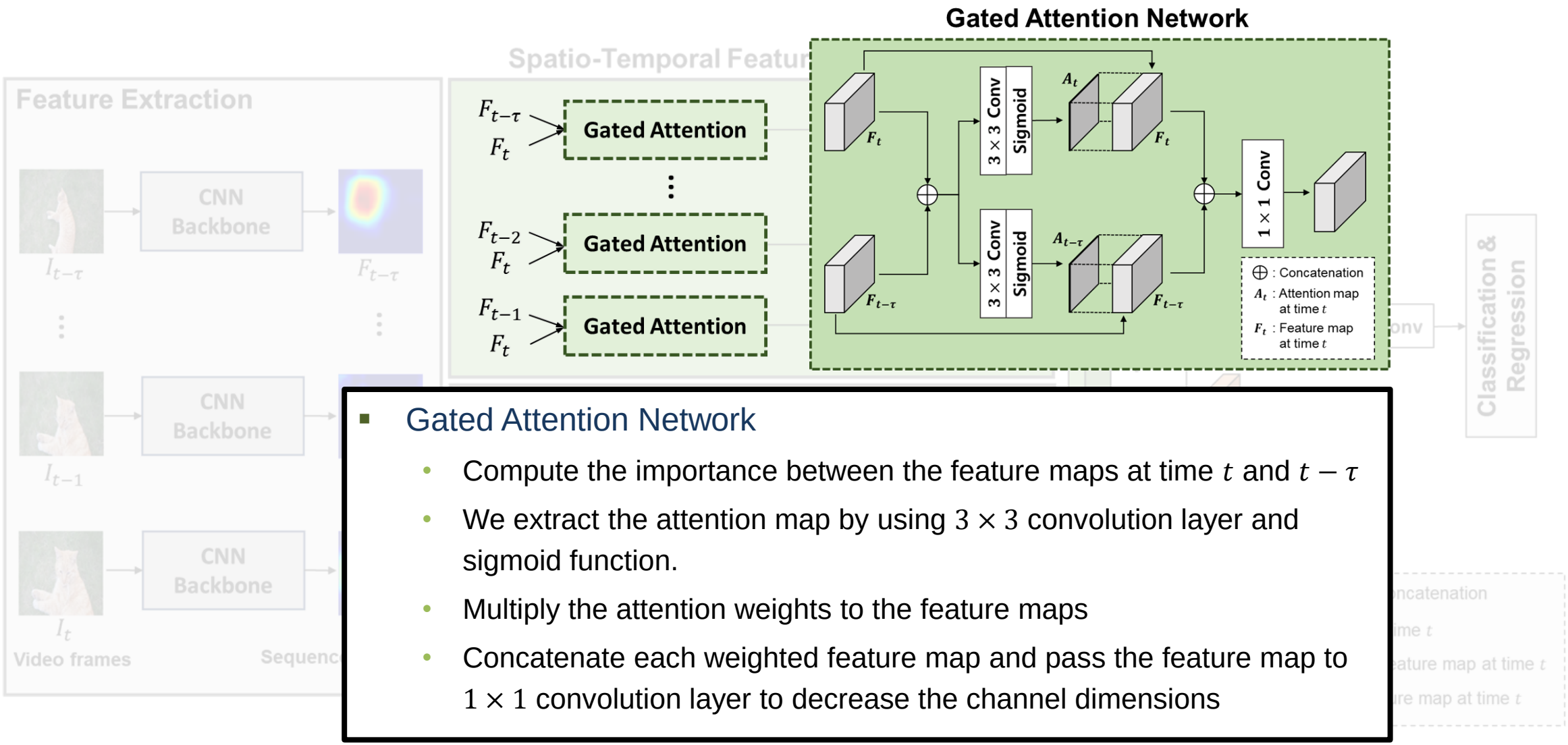

[1] Simonyan, et al. "Very deep convolutional networks for large-scale image recognition." *arXiv preprint arXiv:1409.1556* (2014).
[2] Liu, Wei, et al. "Ssd: Single shot multibox detector." *European conference on computer vision*. Springer, Cham, 2016.
[3] He, Kaiming, et al. "Deep residual learning for image recognition." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.
[4] Lin, Tsung-Yi, et al. "Focal loss for dense object detection." *Proceedings of the IEEE international conference on computer vision*. 2017.

# Spatio-Temporal Feature Aggregation(SFA)




- Spatio-Temporal Feature Aggregation(SFA)
  - Produce the aggregated spatio-temporal feature map with Gated attention network
  - Use the paired features as $(F_{t-\tau}, F_t), (F_{t-\tau+1}, F_t), (F_{t-1}, F_t)$
  - Compute the attention weights that indicate the importance of each feature map

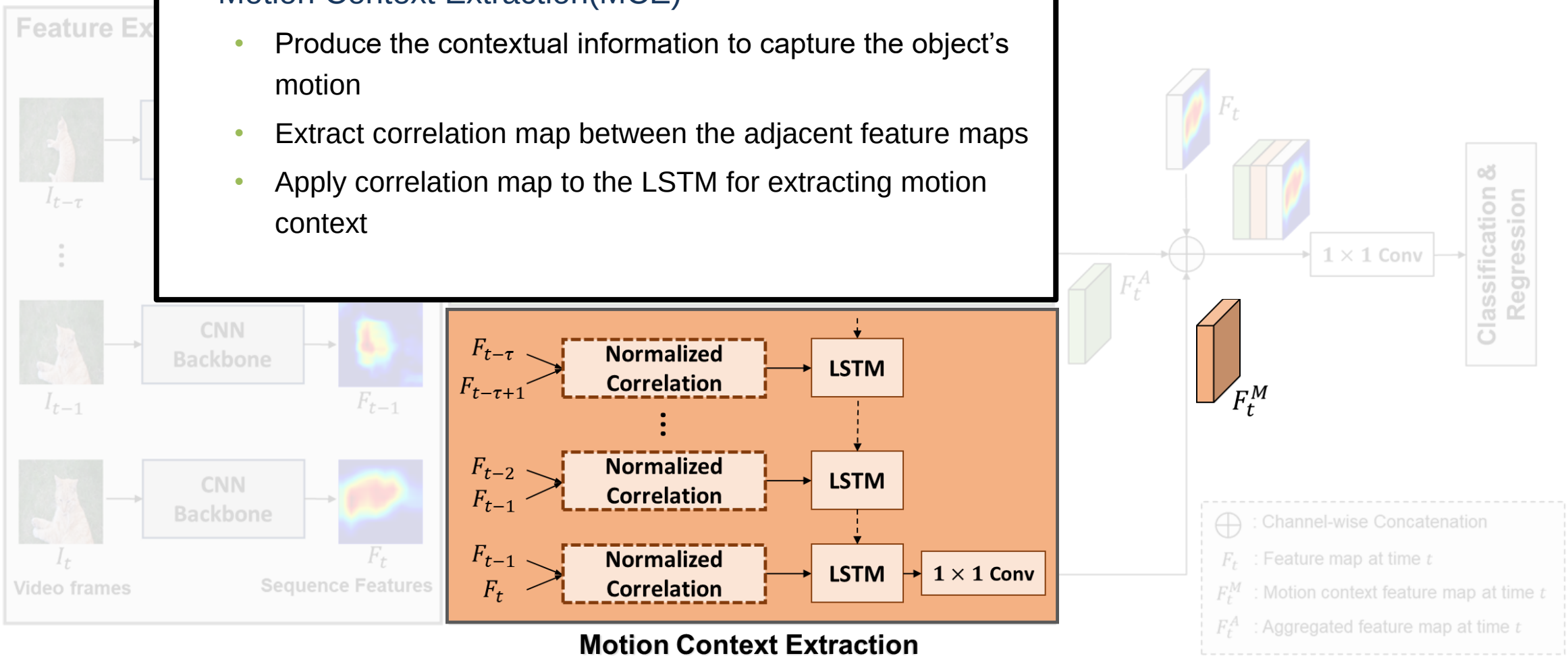
# Spatio-Temporal Feature Aggregation(SFA)




- Gated Attention Network
  - Compute the importance between the feature maps at time $t$ and $t-\tau$
  - We extract the attention map by using $3 \times 3$ convolution layer and sigmoid function.
  - Multiply the attention weights to the feature maps
  - Concatenate each weighted feature map and pass the feature map to $1 \times 1$ convolution layer to decrease the channel dimensions

# Proposed method (VOD-MT)

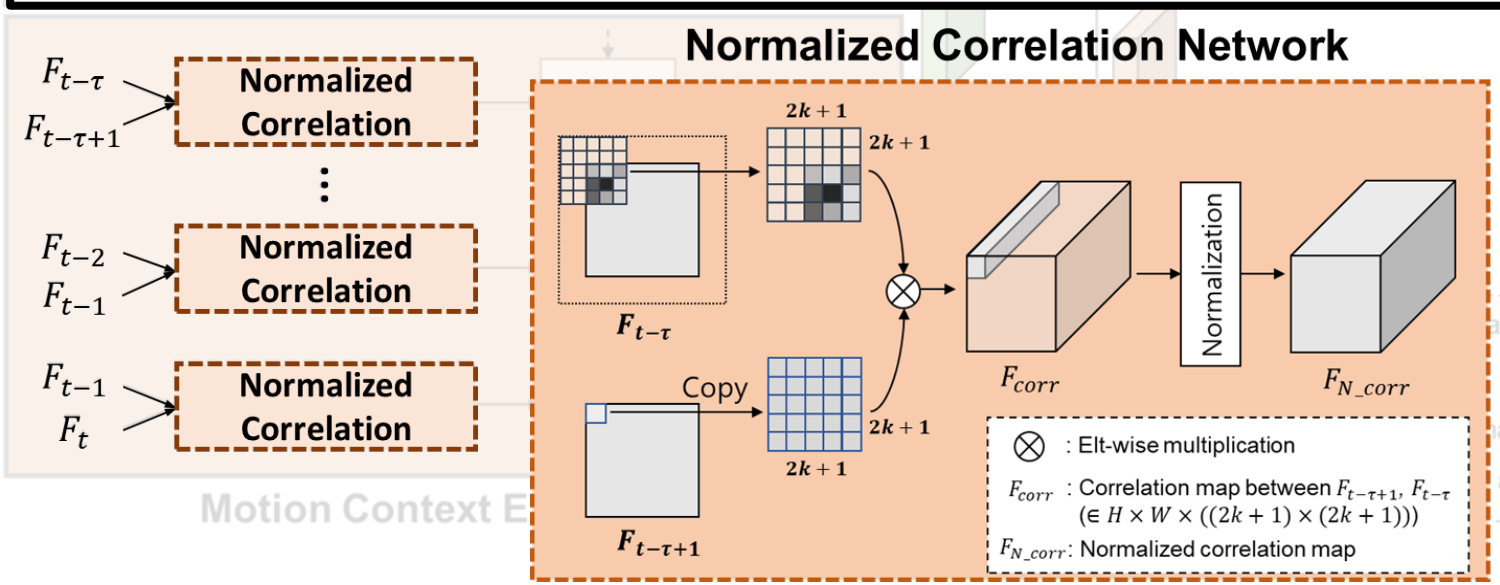
- Motion Context Extraction(MCE)
  - Produce the contextual information to capture the object's motion
  - Extract correlation map between the adjacent feature maps
  - Apply correlation map to the LSTM for extracting motion context



# Proposed method (VOD-MT)

- Normalized correlation Network
  - Compute correlation operation to extract correlation map between $F_{t-\tau}$ and $F_{t-\tau+1}$
  - $F_{corr}(p,q,i,j) = F_{t-\tau+1}(p,q) \otimes F_{t-\tau}(p+i,q+j)$
  - $-k \le i,j \le k$
  - $F_{t-\tau+1}(p,q)$: Feature vector from $F_{t-\tau+1}$ at $(p,q)$ lacation
  - $F_{corr}(p,q)$: Correlation vector between $F_{t-\tau}$ and $F_{t-\tau+1}$



# Experiment results on the ImageNet VID validation set [5]

- Effectiveness of each sub-network

| | Proposed VOD-MT with SSD300 | | | | |
|---|---|---|---|---|---|
| Methods | (a) | (b) | (c) | (d) | (e) |
| Spatio-Temporal Feature Aggregation | | ✓ | | ✓ | ✓ |
| Motion Context Extraction | | | ✓ | ✓ | ✓ |
| Post-processing (Seq-NMS) | | | | | ✓ |
| mAP (%) | 66.72 | 69.04 (↑ 2.32) | 69.00 (↑ 2.28) | 71.03 (↑ 4.33) | **73.20 (↑ 6.50)** |
| mAP (%) (Slow) | 77.45 | 78.91 (↑ 1.46) | 80.27 (↑ 2.82) | 82.08 (↑ 4.63) | **83.47 (↑ 6.02)** |
| mAP (%) (Mudium) | 64.38 | 67.54 (↑ 3.16) | 66.36 (↑ 1.98) | 67.99 (↑ 3.61) | **70.82 (↑ 6.44)** |
| mAP (%) (Fast) | 41.52 | 44.34 (↑ 2.82) | 43.18 (↑ 1.66) | 46.13 (↑ 4.61) | **49.83 (↑ 8.31)** |
| Run-time (ms) | 23 | 32 | 38 | 55 | - |

* score = IoU($boxes_{previous}, boxes_{current}$ )

- ✓ Slow : score > 0.9
- ✓ Medium : 0.7< score < 0.9
- ✓ Fast : score < 0.7

[5] Russakovsky, Olga, et al. "Imagenet large scale visual recognition challenge." *International journal of computer vision* 115.3 (2015): 211-252.

# Experiment results on the ImageNet VID validation set

- Effectiveness of each sub-network

| | Proposed VOD-MT with ResNeXt-101 | | | | |
|---|---|---|---|---|---|
| Methods | (a) | (b) | (c) | (d) | (e) |
| Spatio-Temporal Feature Aggregation | | ✓ | | ✓ | ✓ |
| Motion Context Extraction | | | ✓ | ✓ | ✓ |
| Post-processing (Seq-NMS) | | | | | ✓ |
| mAP (%) | 77.89 | 78.60 (↑ 0.71) | 78.33 (↑ 0.44) | 79.23 (↑ 1.34) | **80.17 (↑ 2.28)** |
| mAP (%) (Slow) | 87.29 | 87.57 (↑ 0.28) | 87.67 (↑ 0.38) | 88.22 (↑ 0.93) | **89.01 (↑ 1.72)** |
| mAP (%) (Mudium) | 74.47 | 75.00 (↑ 0.53) | 74.76 (↑ 0.29) | 75.98 (↑ 0.98) | **76.84 (↑ 2.37)** |
| mAP (%) (Fast) | 55.70 | 57.22 (↑ 1.52) | 55.91 (↑ 0.21) | 57.51 (↑ 1.81) | **58.88 (↑ 3.18)** |
| Run-time (ms) | 110 | 32 | 38 | 55 | - |

* score = IoU($boxes_{previous}, boxes_{current}$ )

- ✓ Slow : score > 0.9
- ✓ Medium : 0.7< score < 0.9
- ✓ Fast : score < 0.7

# Experiment results on the ImageNet VID validation set

- Ablation study with different hyper parameter setting of MCE

| Methods | Upsampling rate ($r$) | Maximum displacement ($k$) | mAP (%) |
|---|---|---|---|
| Baseline (SSD) | - | - | 66.7 |
| Method (c) | $r = 0$ | $k = 2$ | 67.6 |
| | $r = 2$ | | 69.0 |
| | $r = 3$ | | 67.8 |
| | $r = 4$ | | 66.8 |
| | $r = 2$ | $k = 1$ | 67.7 |
| | | $k = 2$ | 69.0 |
| | | $k = 4$ | 68.8 |
| | | $k = 6$ | 67.2 |

# Experiment results on the ImageNet VID validation set

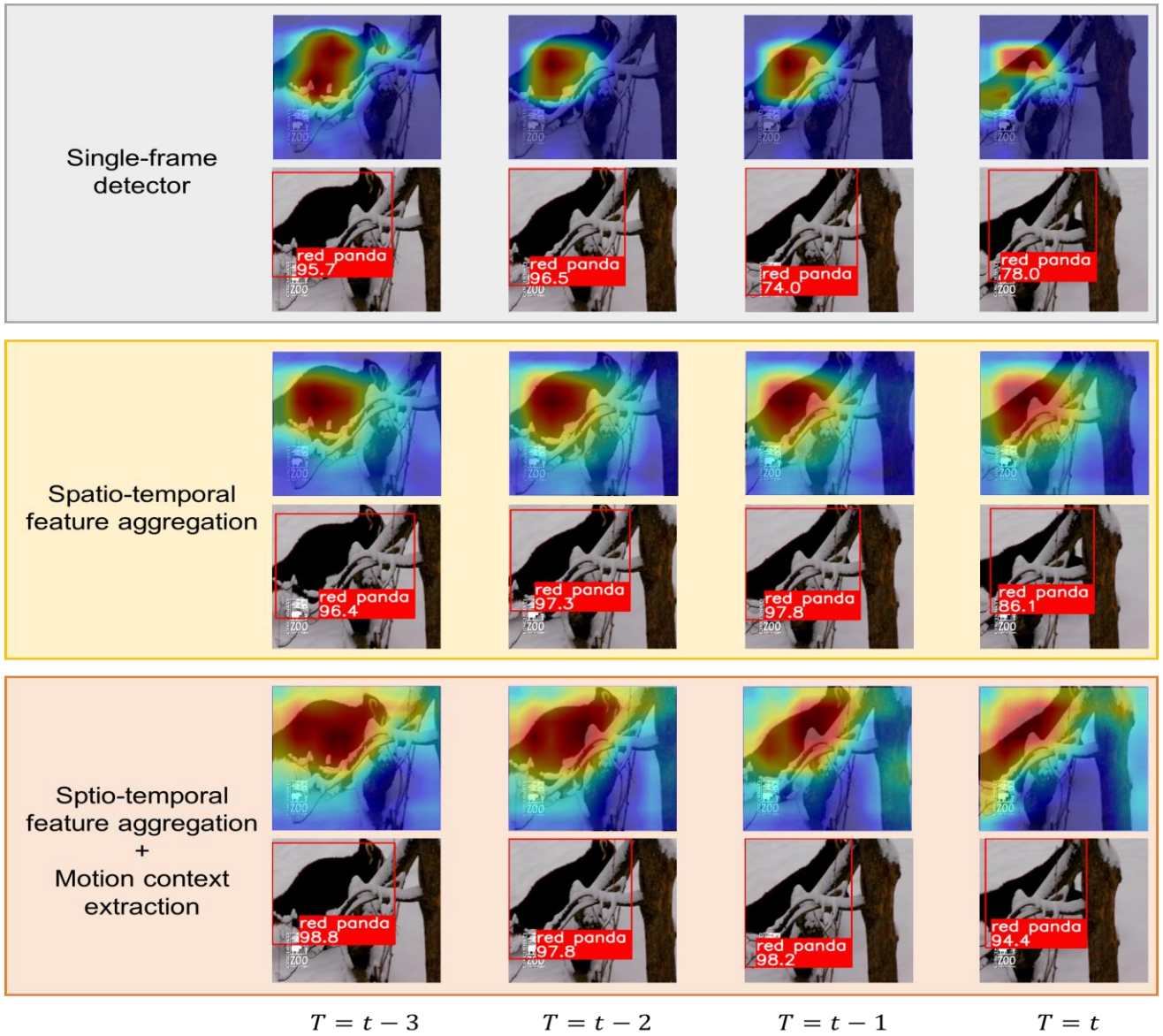
- Performance comparison

| Network | Backbone | Base Detector | mAP (%) |
|---|---|---|---|
| LSTM-SSD [6] | MobileNet | SSD | 54.4 |
| TSSD(-OTA) [7] | VGG-16 | SSD | 65.4 |
| Method of [8] | VGG-16 | SSD | 69.5 |
| Ours with SSD300 | VGG-16 | SSD | 71.0 |
| Ours with SSD300 * | VGG-16 | SSD | 73.2 |

[6] Liu, Mason, et al. "Mobile video object detection with temporally-aware feature maps." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018.
[7] Chen, Xingyu, et al. "Temporally identity-aware SSD with attentional LSTM." *IEEE transactions on cybernetics* 50.6 (2019): 2674-2686.
[8] Zhao, Baojun, et al. "Deep spatial-temporal joint feature representation for video object detection." *Sensors* 18.3 (2018): 774.

# Experiment results on the ImageNet VID validation set

- Visualization



# Conclusions

- New one-stage video object detector, called VOD-MT, exploits **temporal redundancy** and **motion context** as the temporal information.

- **Spatio-Temporal Feature Aggregation (STA)** block is employed for producing temporal redundancy information to solve visual defects such as motion blur camera defocusing.

- **Motion Context Extractor (MCE)** block is employed for producing contextual information which can capture the object motion.

- Our network is **applicable** to every one-stage object detector and **outperforms** over the

Jaekyum Kim: jkkim@spa.hanyang.ac.kr

Junho Koh (Presenter): jhkoh@spa.hanyang.ac.kr

# Thank You!!