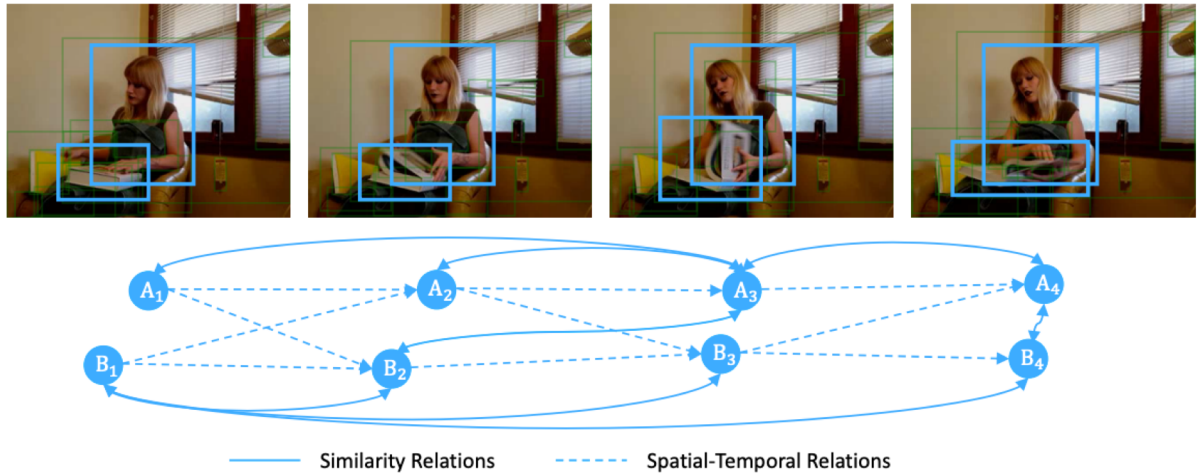
Videos as Space-Time Region Graphs

Jaedong Hwang

Computer Vision Lab.

Seoul National University

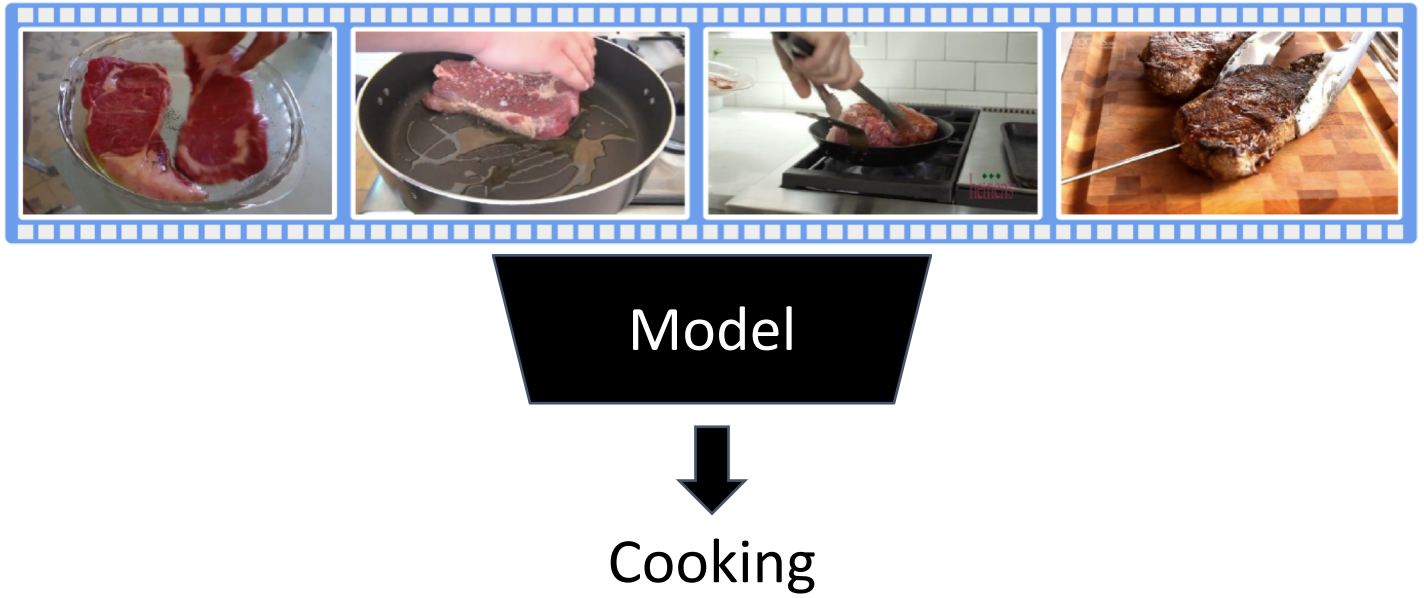


- TL;DR: Generate and utilize graphs of region proposals to capture relationship between objects.

Wang, Xiaolong, and Abhinav Gupta. "Videos as space-time region graphs." *ECCV*. 2018.

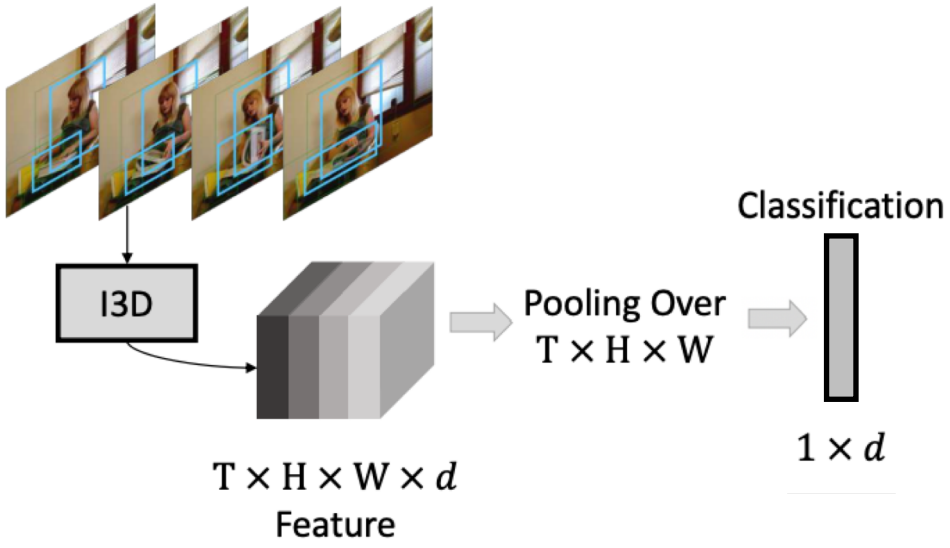
What is Action Recognition?

- Classify which actions in a videos.

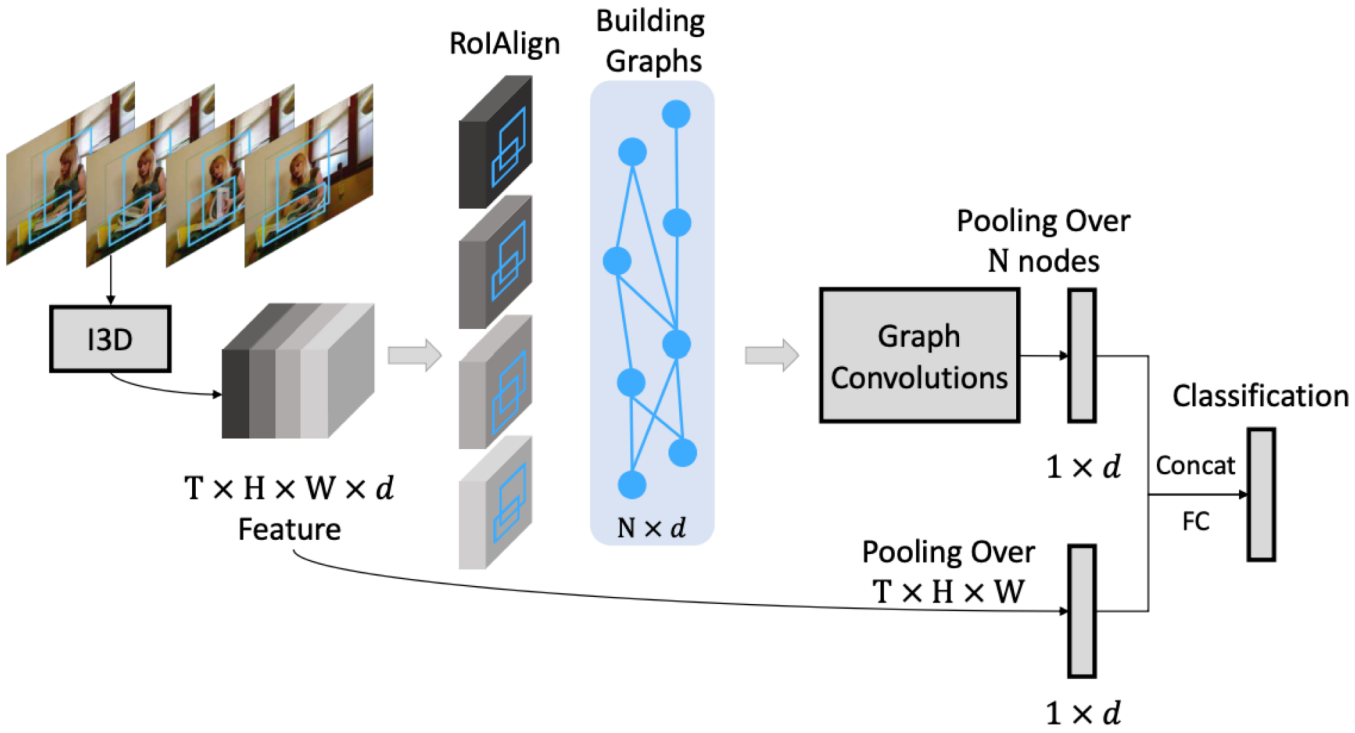


- Depending on temporal direction, label can be changed
  - Door open vs. Door close

Videos as Space-Time Region Graphs



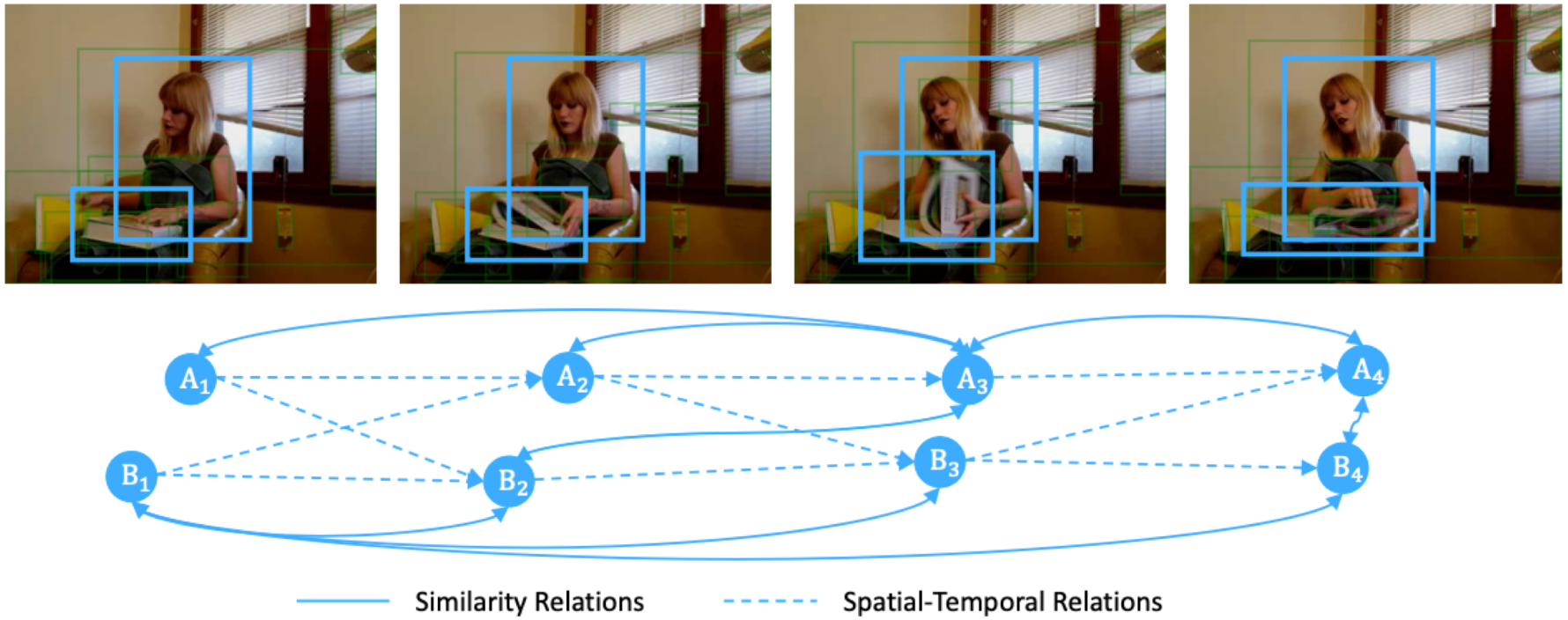
Previous architecture



Proposed architecture

Wang, Xiaolong, and Abhinav Gupta. "Videos as space-time region graphs." *ECCV*. 2018.

Two Graphs for region proposals



- Similarity graph – fully connected, similarity
- Spatiotemporal graph – partially connected, overlap

Wang, Xiaolong, and Abhinav Gupta. "Videos as space-time region graphs." *ECCV*. 2018.

Building Graphs – Similarity Graph

- Fully connected directed graph.
- Transition matrix, $\mathbf{G}_{ij}^{sim}$ is defined as correlation btw embedded features.

$$F(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i)^T \phi'(\mathbf{x}_j)$$

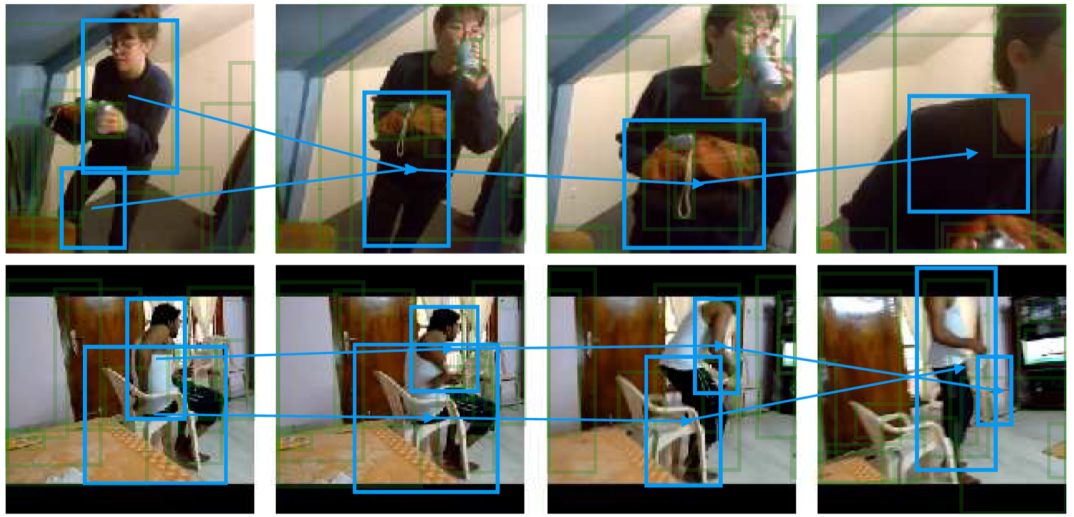
where $\phi(\mathbf{x}) = \mathbf{w}\mathbf{x}$, $\phi'(\mathbf{x}) = \mathbf{w}'\mathbf{x}$

$$\mathbf{G}_{ij}^{sim} = \frac{\exp F(\mathbf{x}_i, \mathbf{x}_j)}{\sum_{j=1}^{\mathrm{N}} \exp F(\mathbf{x}_i, \mathbf{x}_j)}$$

Wang, Xiaolong, and Abhinav Gupta. "Videos as space-time region graphs." *ECCV*. 2018.

Building Graphs – Spatiotemporal graph

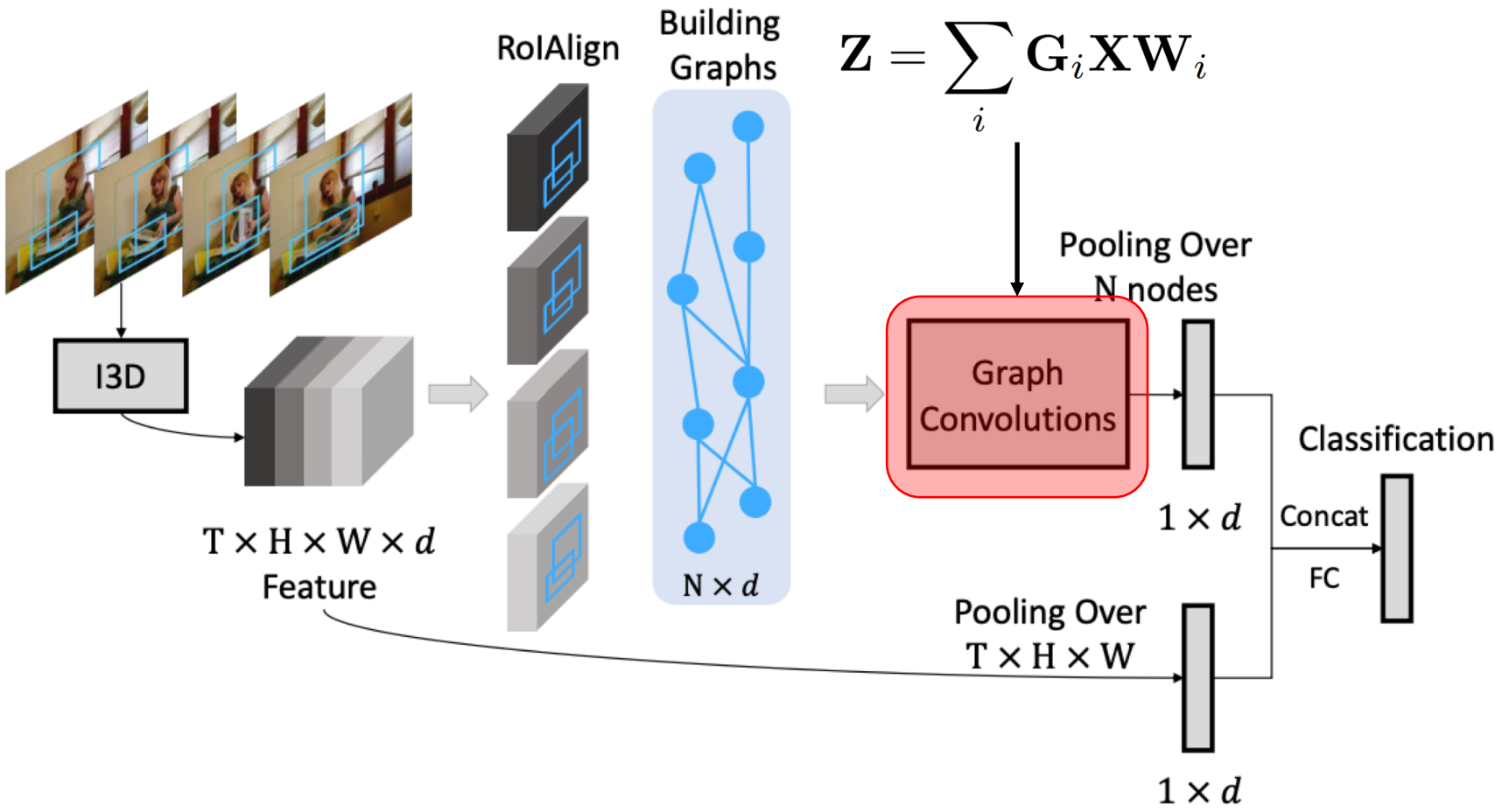
- Use spatio-temporal relation between region proposals.
- Two directed graph (forward graph, backward graph).
- If a proposal at t has an overlap with a proposal at $t+1$, connect them (vice versa in backward graph).



$$\mathbf{G}_{ij}^{front} = \frac{\sigma_{ij}}{\sum_{j=1}^{\mathrm{N}} \sigma_{ij}},$$

σ_{ij} is an IoU of $\mathbf{X}_i$ and $\mathbf{X}_j$

Wang, Xiaolong, and Abhinav Gupta. "Videos as space-time region graphs." *ECCV*. 2018.



Wang, Xiaolong, and Abhinav Gupta. "Videos as space-time region graphs." *ECCV*. 2018.

Result

- Dataset
  - Charades: 157 classes, multi-action, 8k train, 1.8k validation.
  - Something-Something: 174 classes, single-action 86k train, 12k validation, 11k test.

model, R50, I3D	mAP
baseline	31.8
Proposal+AvgPool	32.1
Spatial-Temporal GCN	34.2
Similarity GCN	35.0
Joint GCN	**36.2**

Wang, Xiaolong, and Abhinav Gupta. "Videos as space-time region graphs." *ECCV*. 2018.

Conclusion

- first uses a Graph Convolutional Network for reasoning with multiple relation edges on video action recognition.

- presents a novel graph representation with variant relationships between objects in a long range video.

Wang, Xiaolong, and Abhinav Gupta. "Videos as space-time region graphs." *ECCV*. 2018.