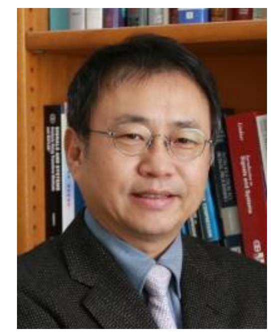
ICIP 2019
IEEE International Conference on Image Processing

Graph Regularization Network with Semantic Affinity for Weakly-supervised Temporal Action Localization







Jungin Park[1] Jiyoung Lee[1] Sangryul Jeon[1] Seungryong Kim[2] Kwanghoon Sohn[1]

1) YONSEI UNIVERSITY

2) EPFL

INTRODUCTION

…WEAKLY-SUPERVISED TEMPORAL ACTION LOCALIZATION

Main Task

Temporal action localization in untrimmed videos

- Detecting time interval which indicates *where is an action content*
- Detecting action class in that time which indicates *what action is contained*

Supervised setting

- Requiring the full annotation of the temporal boundary
- Annotating temporal boundaries for each action instance is very expensive and time-consuming

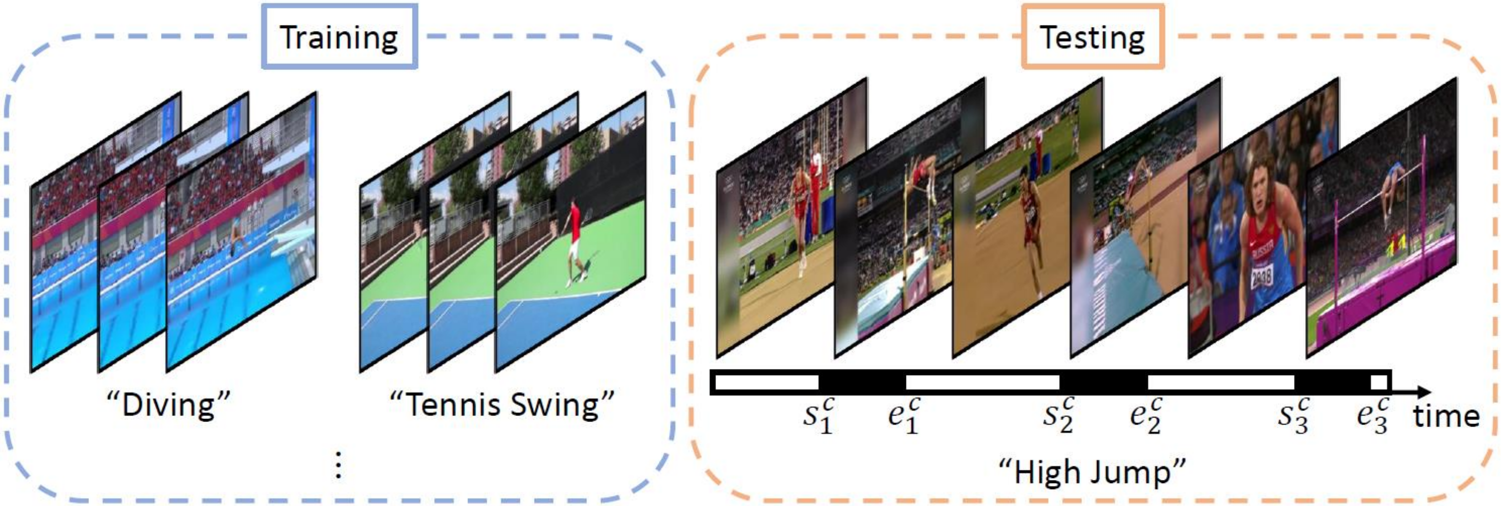
Introduction

…Weakly-supervised Temporal Action Localization

Main Task

Weakly-supervised temporal action localization

- Using only video-level action labels
- Much easier to collect compared to the temporal boundary annotations



INTRODUCTION

···WEAKLY-SUPERVISED TEMPORAL ACTION LOCALIZATION

Challenges

- How to deal with insufficient training data
- How to generate temporal proposals from the video-level classifier

Existing Work

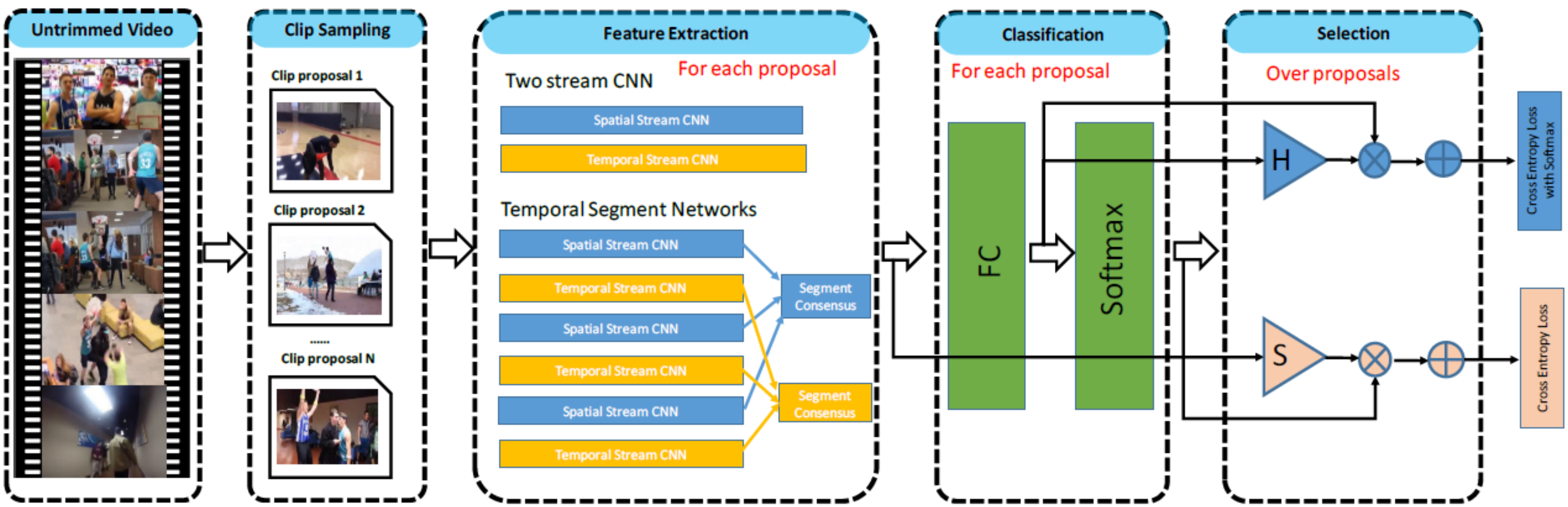
- UntrimmedNet [Wang, *et.al.*, CVPR’17]
- Sparse Temporal Pooling Network (STPN) [Nguyen, *et.al.,* CVPR’18]
- AutoLoc [Shou, *et.al.*, ECCV’18]
- W-TALC [Paul, *et.al.*, ECCV’18]

INTRODUCTION

…WEAKLY-SUPERVISED TEMPORAL ACTION LOCALIZATION

Existing Work

- ✓ UntrimmedNet [Wang, *et.al.*, CVPR'17]
- Sparse Temporal Pooling Network (STPN) [Nguyen, *et.al.*, CVPR'18]
- AutoLoc [Shou, *et.al.*, ECCV'18]
- W-TALC [Paul, *et.al.*, ECCV'18]

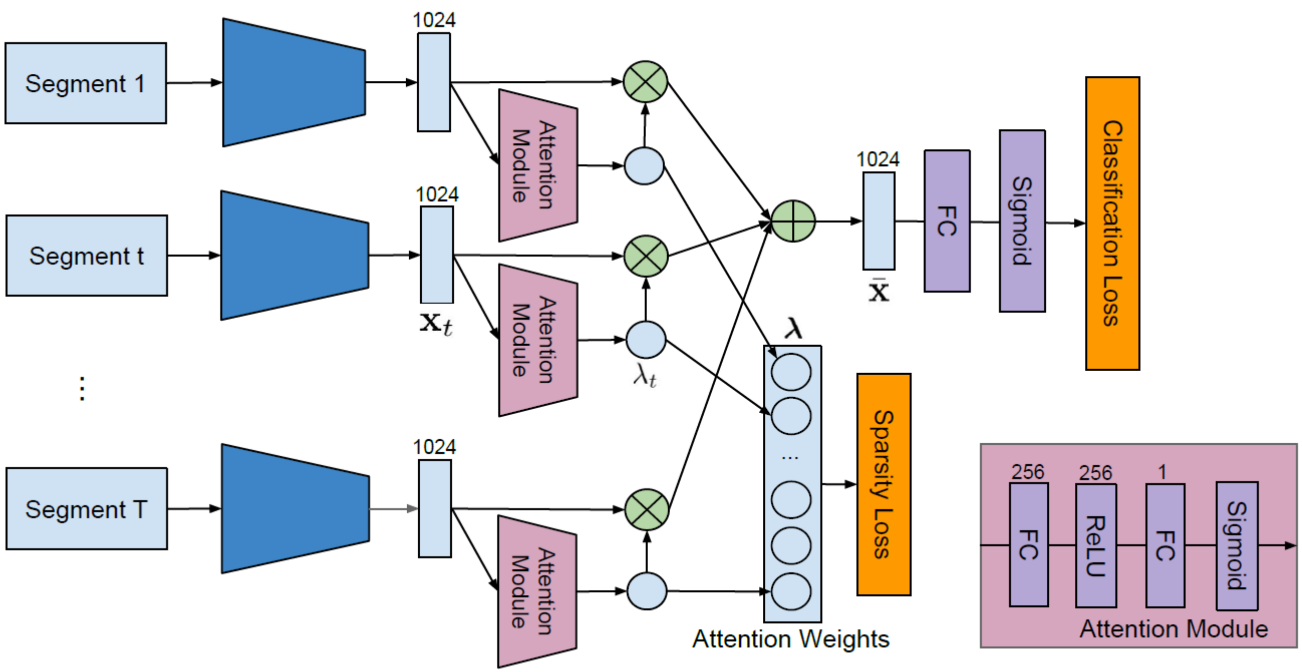


INTRODUCTION

…WEAKLY-SUPERVISED TEMPORAL ACTION LOCALIZATION

Existing Work

- UntrimmedNet [Wang, *et.al.*, CVPR’17]
- ✓ Sparse Temporal Pooling Network (STPN) [Nguyen, *et.al.*, CVPR’18]
- AutoLoc [Shou, *et.al.*, ECCV’18]
- W-TALC [Paul, *et.al.*, ECCV’18]

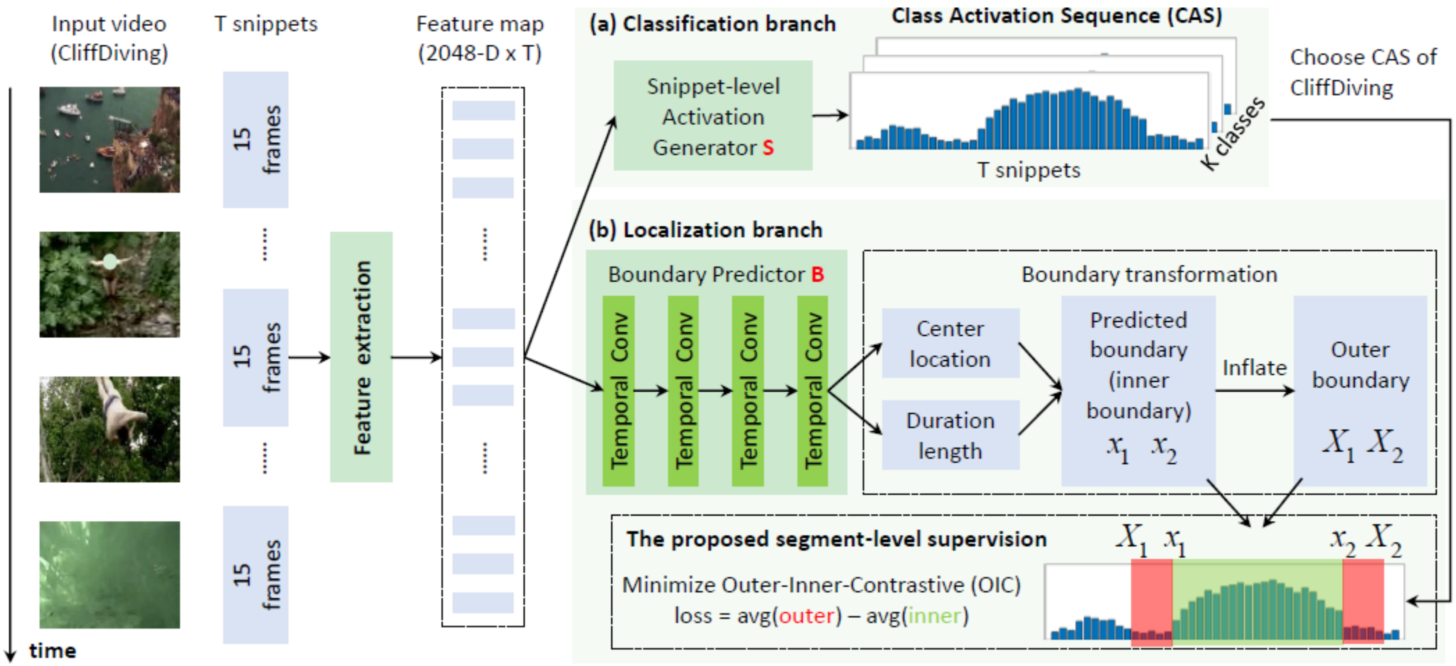


Introduction

…Weakly-supervised Temporal Action Localization

Existing Work

- UntrimmedNet [Wang, *et.al.*, CVPR'17]
- Sparse Temporal Pooling Network (STPN) [Nguyen, *et.al.*, CVPR'18]
- ✓ AutoLoc [Shou, *et.al.*, ECCV'18]
- W-TALC [Paul, *et.al.*, ECCV'18]

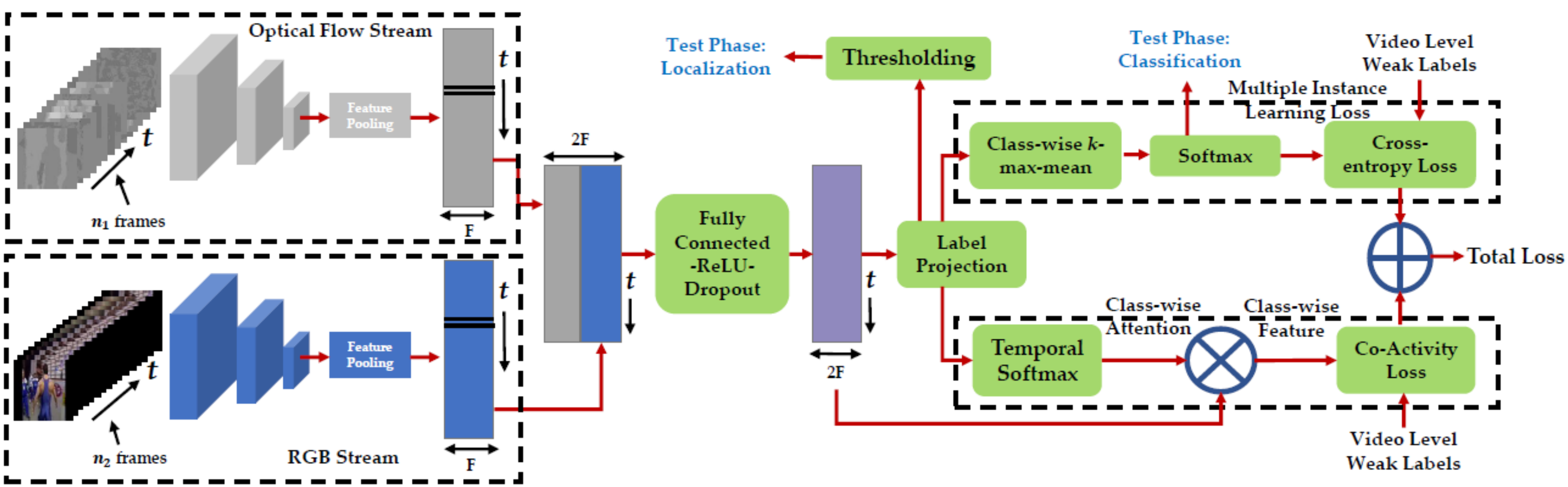


Introduction

…Weakly-supervised Temporal Action Localization

Existing Work

- UntrimmedNet [Wang, *et.al.*, CVPR'17]
- Sparse Temporal Pooling Network (STPN) [Nguyen, *et.al.*, CVPR'18]
- AutoLoc [Shou, *et.al.*, ECCV'18]
- ✓ W-TALC [Paul, *et.al.*, ECCV'18]



INTRODUCTION

…WEAKLY-SUPERVISED TEMPORAL ACTION LOCALIZATION

Existing Work

- UntrimmedNet [Wang, *et.al.*, CVPR'17]
- Sparse Temporal Pooling Network (STPN) [Nguyen, *et.al.*, CVPR'18]
- AutoLoc [Shou, *et.al.*, ECCV'18]
- W-TALC [Paul, *et.al.*, ECCV'18]

Limitations

1. Difficulties on remove noisy activities since only video-level supervisions are provided
2. Action score map is computed without considering the score consistency of clips of the same class

INTRODUCTION

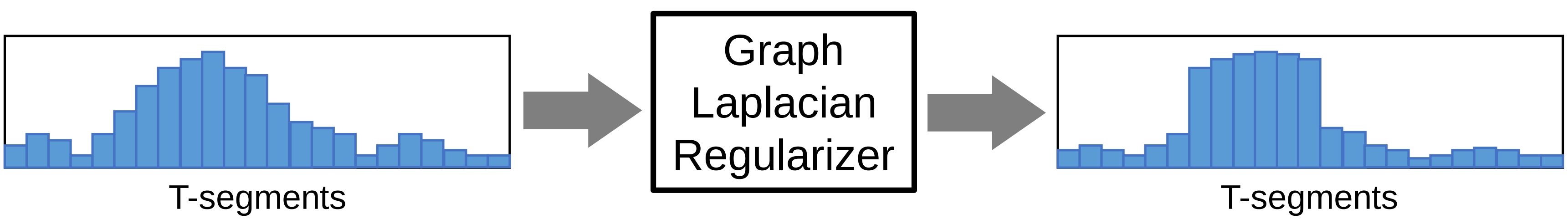
…GRAPH LAPLACIAN REGULARIZATION

Motivation

- Similar actions have similar class activation scores
- Graph which represents affinities between frames (or segments) can be used to refine the class activation map

Goal

- Learning the accurate graph (class agnostic)
- Refining the class activation map using graph Laplacian regularization



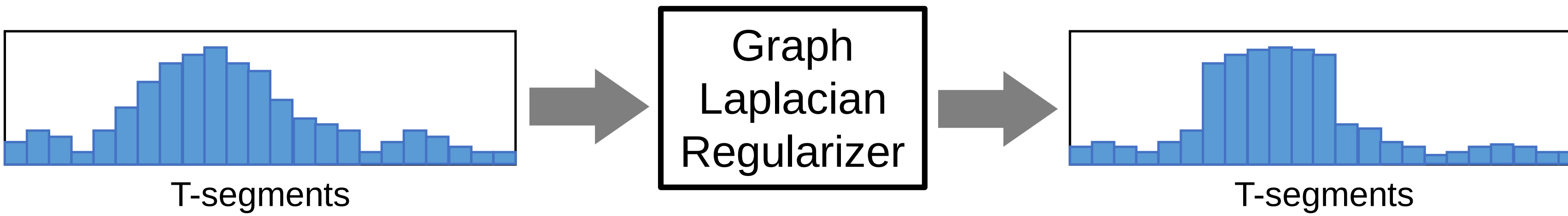
Class activation map for class k

Refined class activation map

Problem Formulation

···Graph Laplacian Regularization

Optimization



Class activation map for class k Refined class activation map

Optimization

$$\hat{s}^{c*} = \arg\min_{\hat{s}^c}\left(\left\|s^c - \hat{s}^c\right\|_2^2 + \mu \cdot (\hat{s}^c)^T \boldsymbol{L} \hat{s}^c\right)$$

Initial class activation map

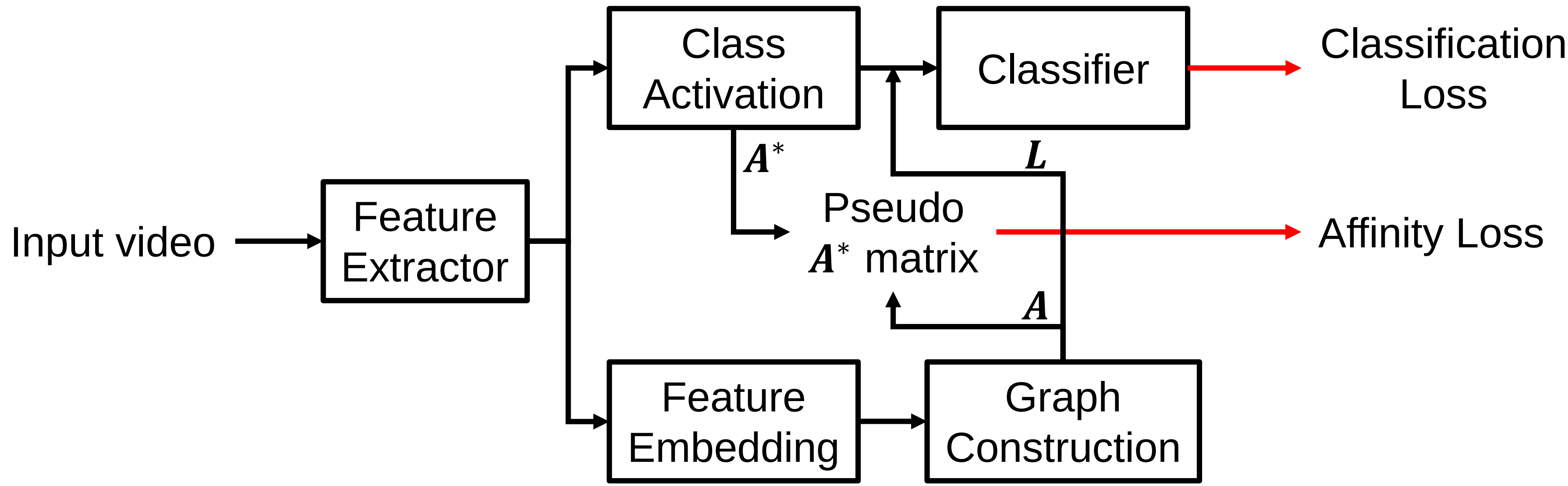
Laplacian matrix of the graph

refined class activation map

Problem Formulation

···Graph Laplacian Regularization

Overall Diagram



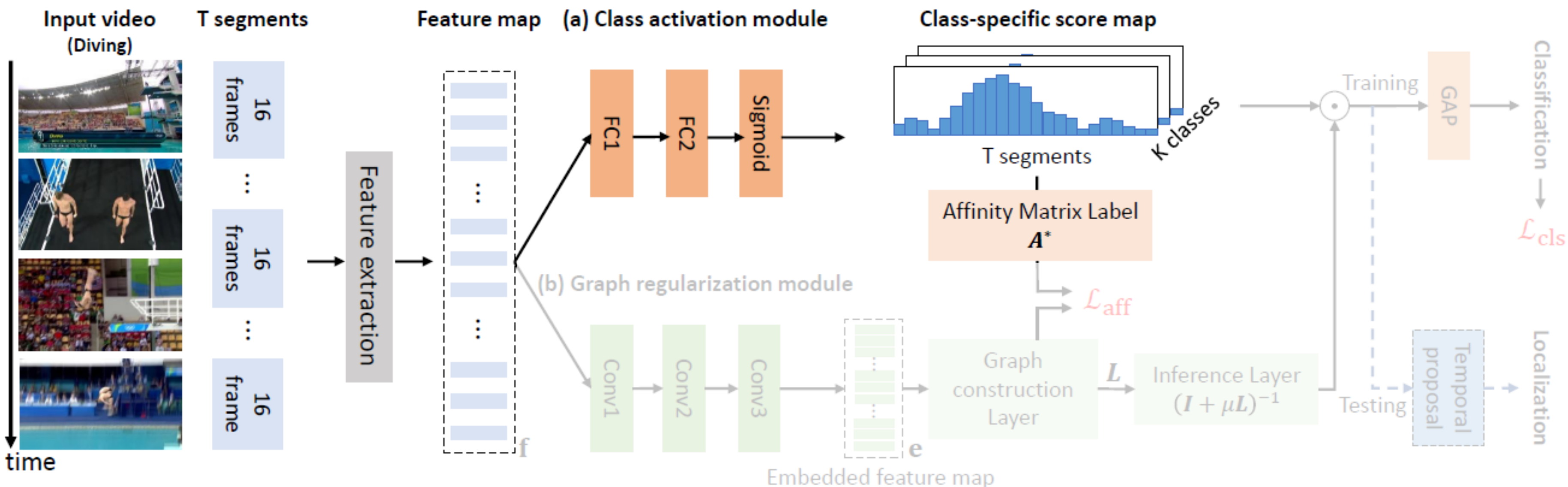
- How to learn the feature embedding network with only video-level labels?

➡ Generating pseudo ground truth of the affinity matrix from the class activation maps

MODEL

···GRAPH REGULARIZATION NETWORK

Overall Architecture



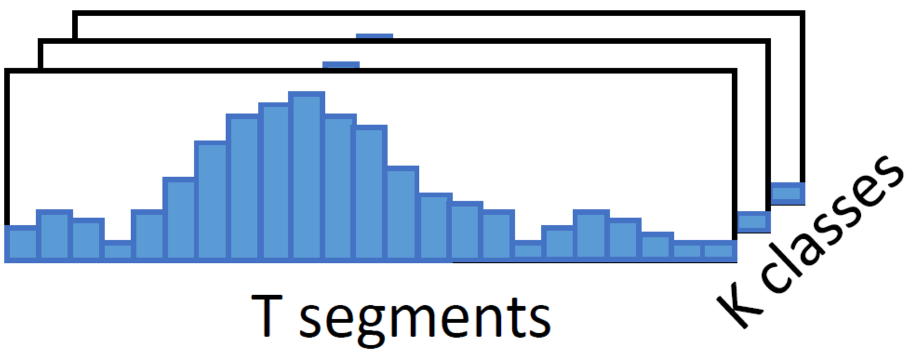
Ingredient 1 Class activation module

- Generating the class-specific activation map
- Generating the pseudo ground truth affinity matrix

Model

···Graph Regularization Network

Overall Architecture



- Collecting class-specific activation map $\boldsymbol{s}$, after sigmoid function
- Each clip has individual score for every action classes

Class-specific activation map

- Generating the class-specific activation map
- Generating the pseudo ground truth affinity matrix

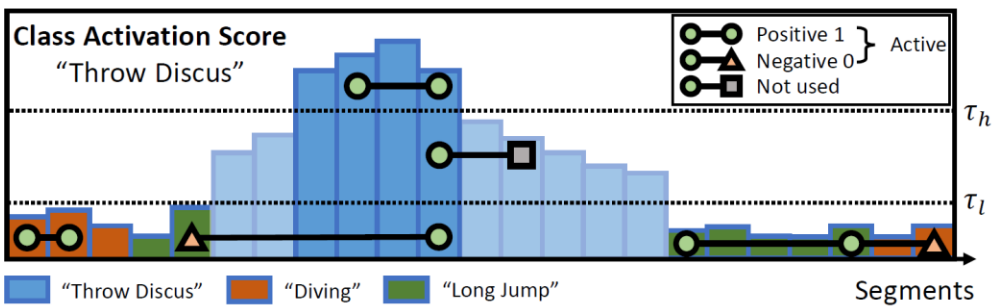
MODEL

…GRAPH REGULARIZATION NETWORK

Overall Architecture

Ingredient 1 Class activation module

- Generating the class-specific activation map
- Generating the pseudo ground truth affinity matrix



Clip-wise activation map

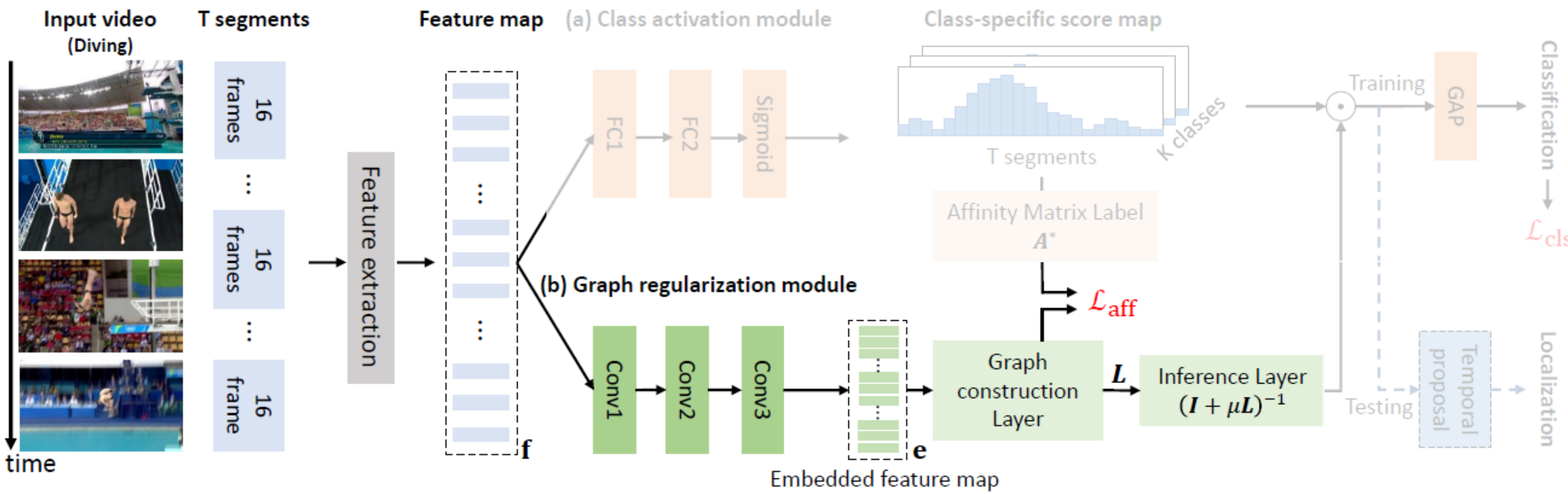
- Collecting highest score class for each clip $\boldsymbol{z} = \arg\max_{c} \boldsymbol{s}^c$
- Set *active samples* $\boldsymbol{v}$ which have scores over τ_h and under τ_l
- Make pseudo ground truth affinity matrix $\boldsymbol{A}^*$

$$\boldsymbol{A}^* = \begin{cases} \mathbf{1}, & \boldsymbol{if} \quad \boldsymbol{v_i} = \boldsymbol{v_j} \\ \mathbf{0}, & \boldsymbol{otherwise} \end{cases}$$

MODEL

…GRAPH REGULARIZATION NETWORK

Overall Architecture



Ingredient 2 Graph regularization module

- Embedding the features to the lower dimensional feature space
- Constructing the graph represents affinities between frames
- Solving the graph Laplacian regularization

Overall Architecture

Ingredient 2 Graph regularization module

Graph construction

- Affinity graph:

$$\mathcal{G}(\mathcal{E}, \mathcal{V})$$

- Embedded feature (as a node):

$$\mathbf{e} = \mathcal{F}(\mathbf{f}; \mathbf{w})$$

- Edge weight (elements of the adjacency matrix A):

$$w_{ij} = \exp(-\left|\left|\mathbf{e}_i - \mathbf{e}_j\right|\right|^2 / 2\epsilon^2)$$

- Degree matrix:

$$D = \mathrm{diag}(\sum w_{ij})$$

- Graph Laplacian matrix:

$$L = D - A$$

SOLUTION

…INVERSE SYSTEM PROBLEM

Reformulation

Graph Laplacian regularization

- With previous components, we can reformulate the optimization as an inverse system problem of the linear equation

Optimization

$$\hat{s}^{c*} = \arg\min_{\hat{\boldsymbol{s}}^{c}} \left(\left|\left| s^{c} - \hat{s}^{c} \right|\right|_2^2 + \mu \cdot (\hat{s}^{c})^{T} \boldsymbol{L} \hat{s}^{c} \right)$$

Inverse System Problem

$$\hat{s}^{c*} = (\boldsymbol{I} + \mu \boldsymbol{L})^{-1} s^{c}$$

Backpropagation details will be introduced extensions of this work!

Solution

···Learning Network Parameters

Loss Functions

- Affinity loss between the affinity matrix and the pseudo ground truth affinity matrix
- Classification loss with the video-level label

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{aff}} + \lambda \mathcal{L}_{\text{cls}}$$

$$\mathcal{L}_{\text{aff}} = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{|\mathcal{N}(i)|} \sum_{j \in \mathcal{N}(i)} \left|\left| \boldsymbol{A}_{ij} - \boldsymbol{A}^{*}_{ij} \right|\right|^{2}$$

Number of clips

Semantic neighbors of v_i

$$\mathcal{L}_{\text{cls}} = -\sum_{c=1}^{K} [y^c \log \hat{y}^c + (1 - y^c) \log(1 - \hat{y}^c)]$$

Video class GT

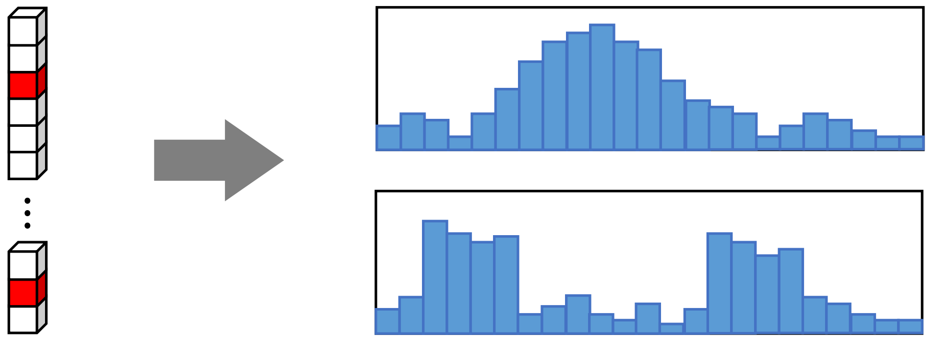
Predicted video class

SOLUTION

…TESTING PHASE

Localize Actions

- Employing the two-stream model (RGB and optical flow)
- Each stream is trained individually, and integrated in testing phase
- Temporal proposals are extracted by applying threshold to each stream



Output of the video classifier

Class activation map for class i and j

- Final class score can be represented as

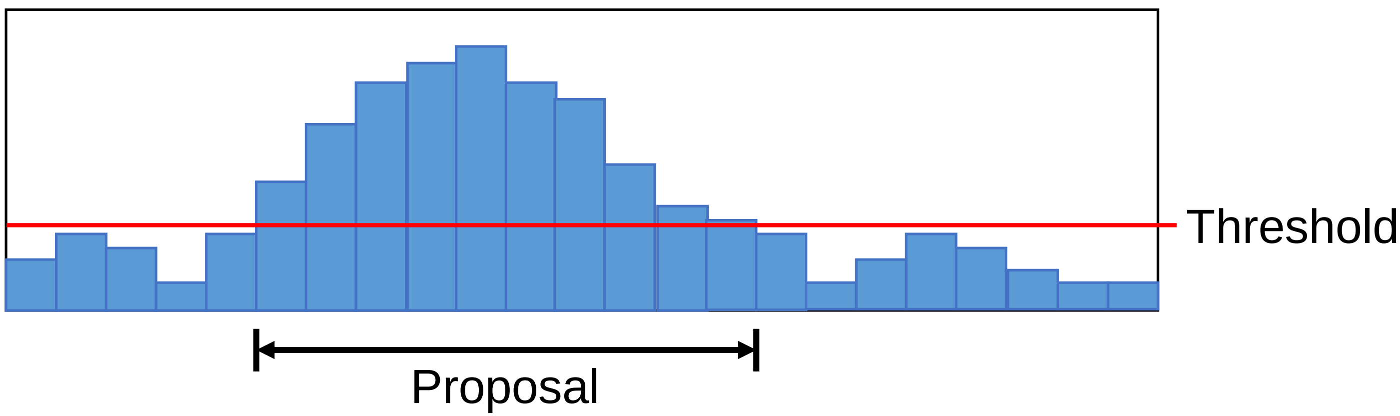
$$\boldsymbol{s}_{\text{final}} = \sum_{t=t_s}^{t_e} \frac{\alpha \hat{\boldsymbol{s}}_{t,\text{RGB}}^{c*} + (1-\alpha) \cdot \hat{\boldsymbol{s}}_{t,\text{FLOW}}^{c*}}{t_e - t_s + 1}$$

SOLUTION

…TESTING PHASE

Localize Actions

- Employing the two-stream model (RGB and optical flow)
- Each stream is trained individually, and integrated in testing phase
- Temporal proposals are extracted by applying threshold to each stream



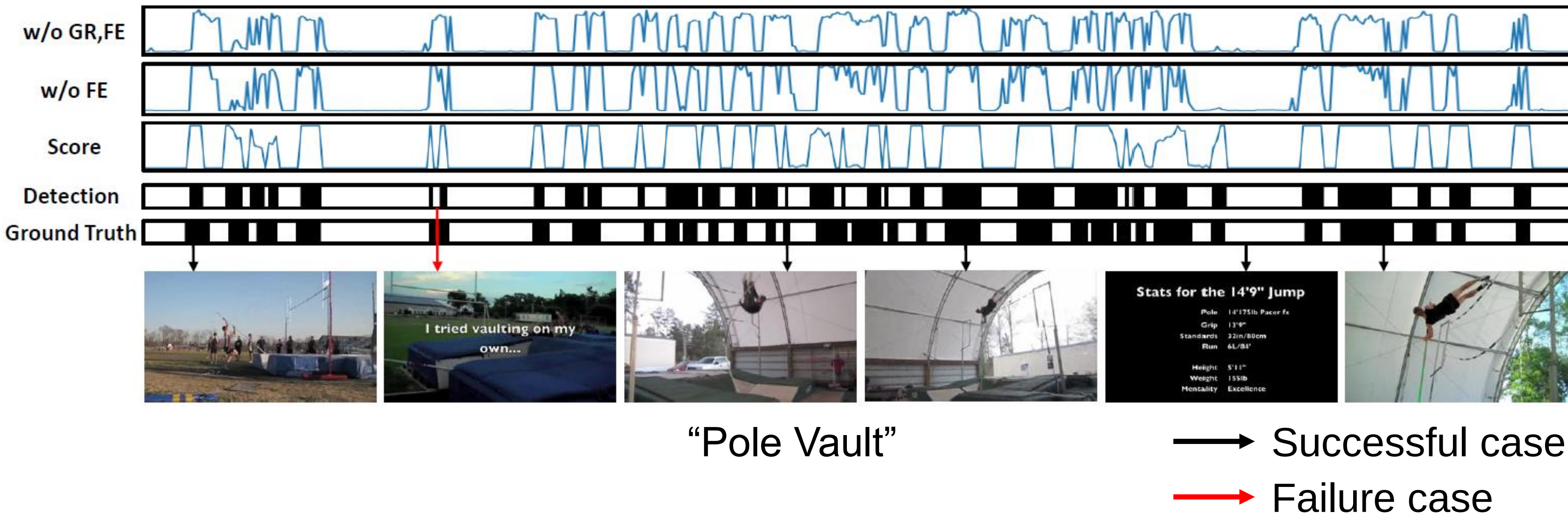
- Final class score can be represented as

$$s_{\text{final}} = \sum_{t=t_s}^{t_e} \frac{\alpha \hat{\boldsymbol{s}}_{t,\text{RGB}}^{c*} + (1-\alpha) \cdot \hat{\boldsymbol{s}}_{t,\text{FLOW}}^{c*}}{t_e - t_s + 1}$$

EXPERIMENTAL RESULTS

…QUALITATIVE RESULT

THUMOS14



- w/o GR, FE: without graph regularization and feature embedding
- w/o FE: without feature embedding (using features from feature extractor)

EXPERIMENTAL RESULTS

…QUANTITATIVE RESULT

THUMOS14

Supervision	Methods	AP@IoU 0.3	0.4	0.5	0.7
Full	Yuan *et al.* [7]	36.5	27.8	17.8	-
	Gao *et al.* [26]	50.1	**41.3**	**31.0**	9.9
	Zhao *et al.* [9]	**51.9**	41.0	29.8	**10.7**
Weak	Wang *et al.* [10]	28.3	21.1	13.7	-
	Nguyen *et al.* [11]	35.5	25.8	16.9	4.3
	Shou *et al.* [12]	35.8	29.0	21.2	5.8
	Paul *et al.* [13]	40.1	31.1	**22.8**	7.6
	Ours w/o GR, FE	25.2	17.8	9.6	2.7
	Ours w/o FE	35.4	26.1	16.7	4.2
	Ours	**40.2**	**32.2**	21.7	**9.2**

- w/o GR, FE: without graph regularization and feature embedding
- w/o FE: without feature embedding (using features from feature extractor)

- State-of-the-art performance on various threshold value for IoU
- Comparable performance to fully-supervised approach on higher threshold value

Further Work

…Extensions

Limitations

- High computational cost
- Weaknesses on occlusions

Further Work

- Developing the sparse graph regularization with higher performance
- Developing the general module for various applications such as video summarization and spatio-temporal action localization, etc.

Thank you!

Jungin Park, Ph.D. Candidate
Digital Image Media Lab.
Yonsei University, Seoul, Korea
Tel: +82-2-2123-2879
E-mail: newrun@yonsei.ac.kr