# SDA 시니어 논문세션

Precise Event Spotting in Sports Videos: Solving Long-Range Dependency and Class Imbalance

2026년 3월 29일(일)

# Contents.


01 Introduction

02 Method

03 Experiment

04 Conclusion

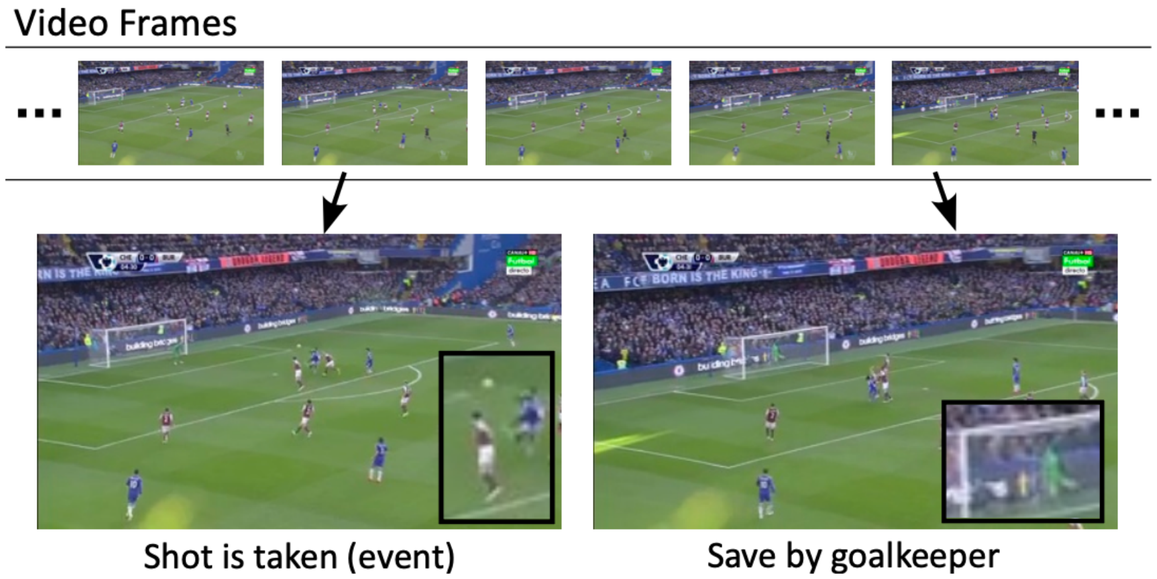
01

# Introduction

# 1.1 Precise Event Spotting (PES)

## 정의

> 긴 스포츠 영상에서 이벤트의 클래스와 정확한 발생 시점을 찾는 것
> 하이라이트 생성, 자동 이벤트 태깅, 경기 영상 검색, 스포츠 영상 분석 자동화



# 1.1 Precise Event Spotting (PES)

## Action Recognition

| 영상 구간이 어떤 행동인지 분류

## Temporal Localization

| 행동이 발생한 구간을 찾음

## Precise Event Spotting

| 이벤트가 발생한 정확한 frame을 찾음

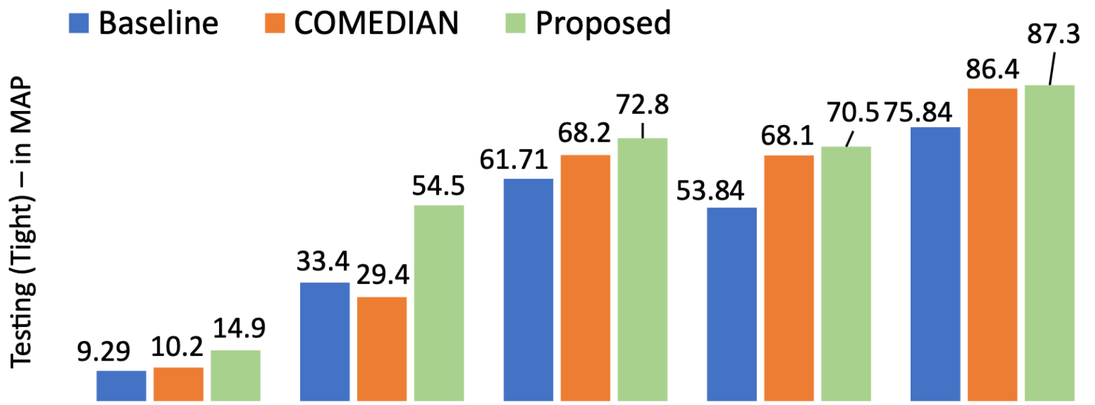
# 1.2 기존 방법의 한계

## 1. Long-range temporal dependancy

현재 프레임만 봐서는 이벤트 판별이 어려움, 과거/미래 프레임 정보 필요

## 2. Class imbalance

자주 발생하는 이벤트와 드문 이벤트의 수 차이가 큼




| No of samples | Yellow->Red card | Red card | Yellow card | Shots on target | Ball out of play |
|---|---|---|---|---|---|
| Training | 24 | 34 | 1238 | 3463 | 19097 |
| Testing | 14 | 8 | 431 | 1175 | 6460 |

# 1.2 기존 방법의 한계

## 기존 방법들: E2E-Spot, T-Deed, ASTRA, COMEDIAN, SpotFormer

> 큰 pre-trained feature extractor 사용
> transformer 기반 temporal modeling 사용

| | | |
|---|---|---|
| E2E-Spot | end-to-end spotting | 어려운 데이터셋에서 제한적 |
| ASTRA | transformer + augmentation | rare class 대응 부족 |
| COMEDIAN | distillation + transformer | 고비용, 대형 모델 |
| 기존 transformer 계열 | global context | 긴 시퀀스에서 비효율 가능 |

# 1.3 해결하려는 문제

## 목표

1. Long-range dependency 해결
2. class imbalance 해결
3. 효율적이면서도 정확한 PES 모델 제안

## 아이디어

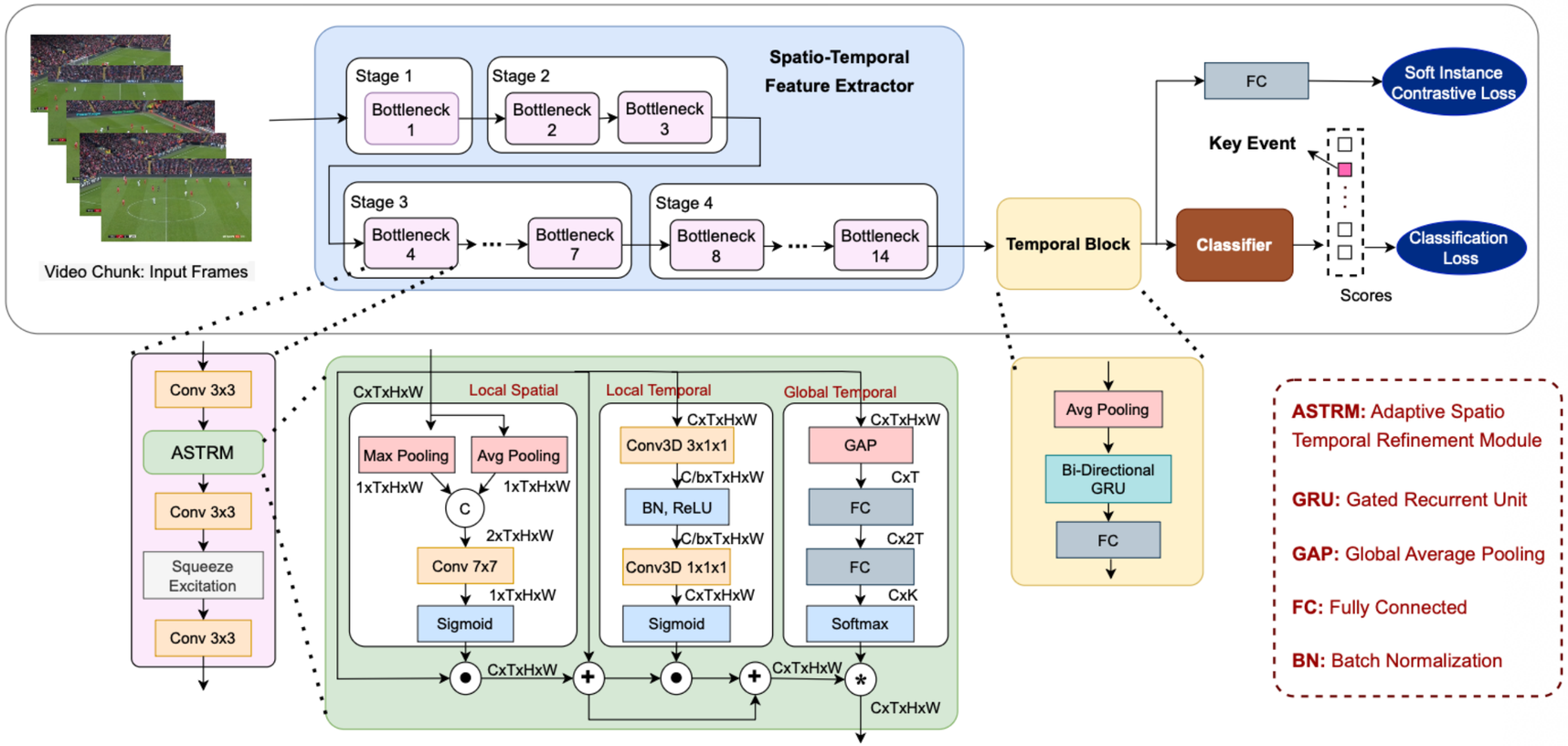
1. ASTRM: spatial + temporal feature refinement
2. Bi-GRU temporal block: 장기 시간 문맥 모델링
3. Soft-IC loss: rare class를 위한 feature compactness / seprability 강화

02

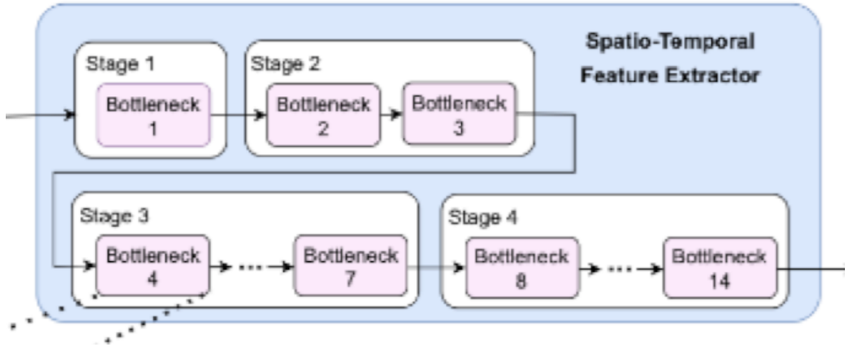
# Method

# 2.1 전체 모델 구조

입력: Video Clip
Backbone: RegNetY
각 Bootleneck에 ASTRM 삽입
temporal block: Bi-GRU
classifier: frame-wise classification
loss: BCE+ Soft-IC



# 2.2 Backbone: RegNetY

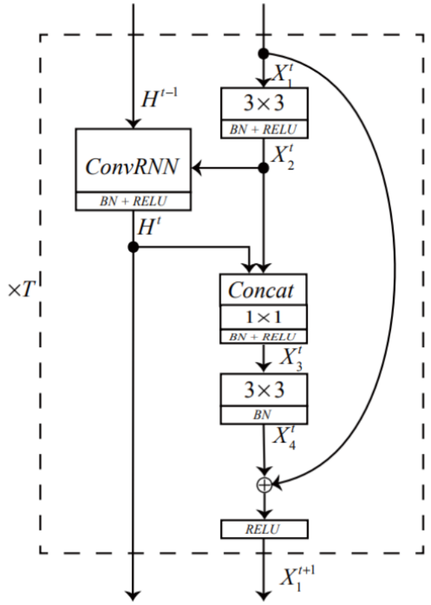


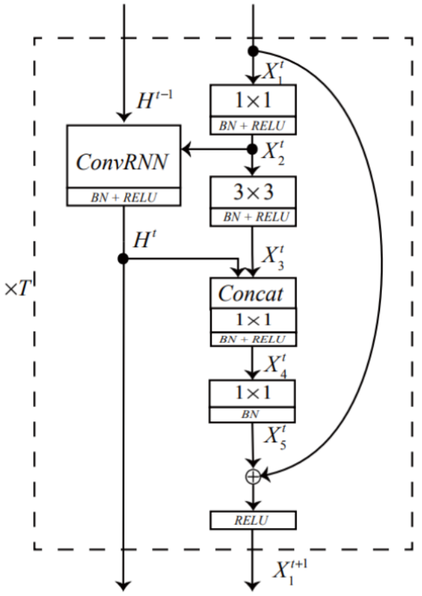

## Backbone

입력 영상에서 기본적인 시각 feature을 뽑아내는 기본 신경망

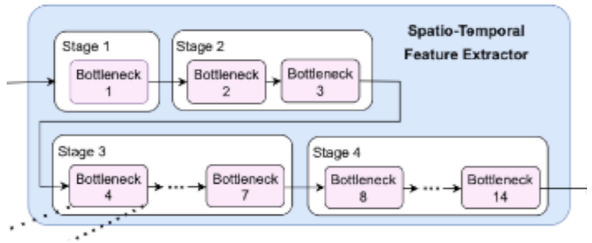
## RegNet

CNN 계열 backbone 중 하나
성능 대비 연산량이 우수
구조가 안정적이라서 영상 feature extractor로 쓰기 좋음



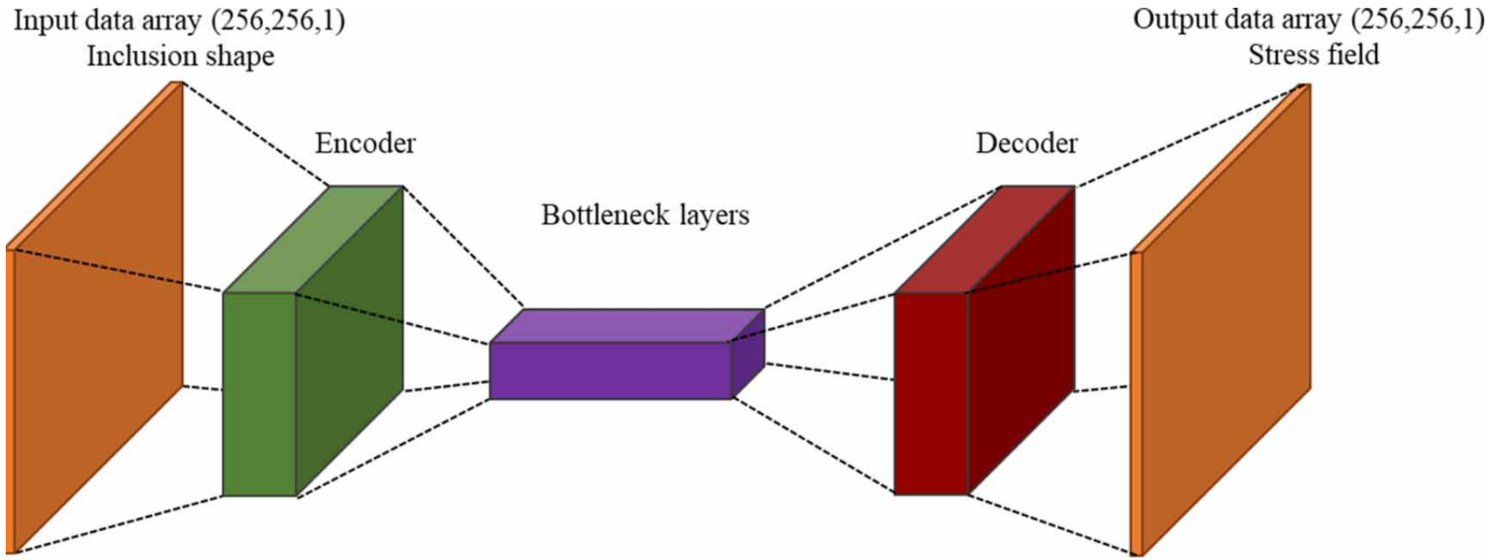



# 2.3 Bottleneck에 ASTRM 삽입

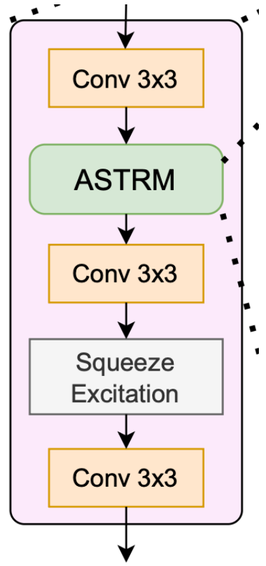



## Bottleneck

CNN 안에서 feature을 효율적으로 처리하는 작은 기본 블록

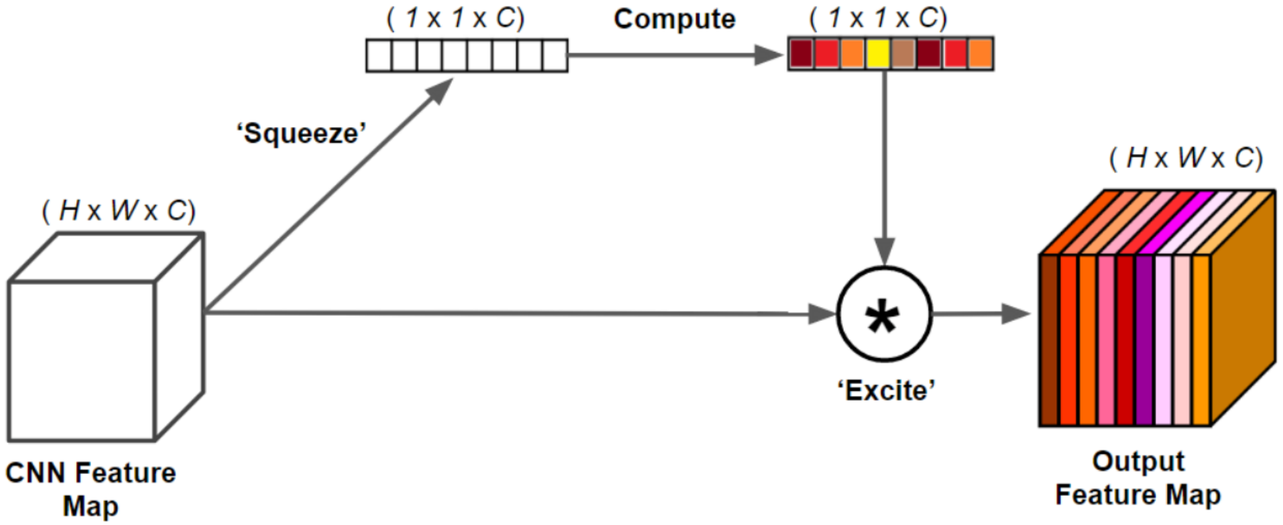


# 2.3 Bottleneck에 ASTRM 삽입




## Squeeze Excitation

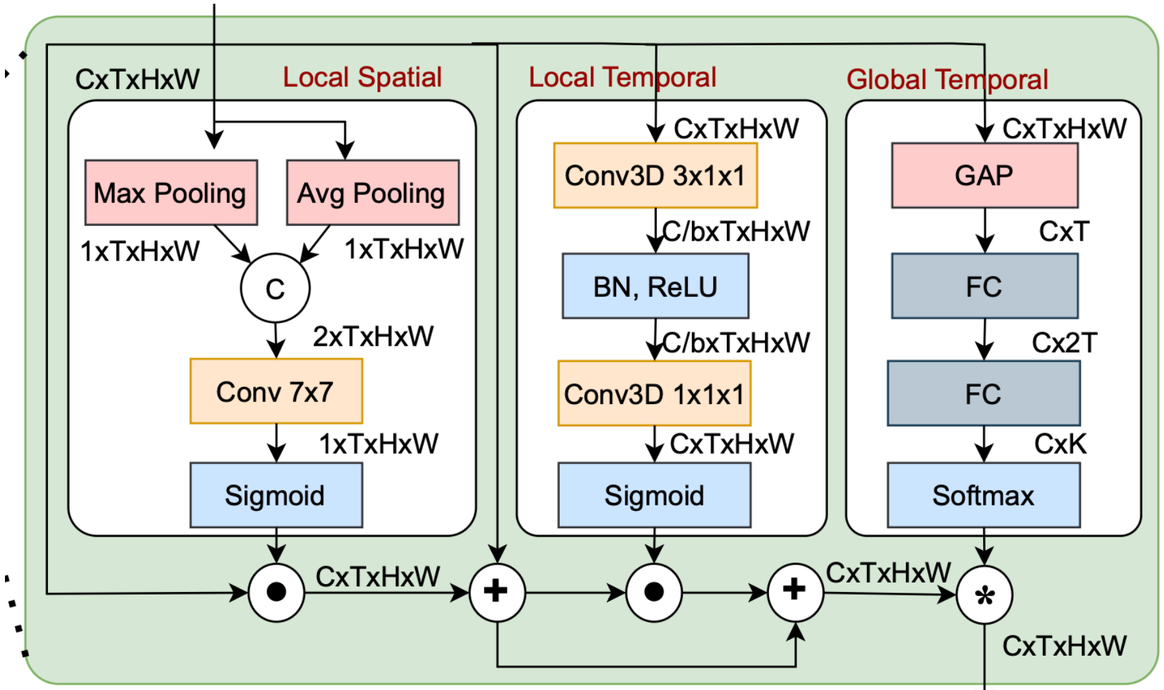
| 채널 별로 어떤 feature가 중요한지 다시 가중치를 주는 모듈



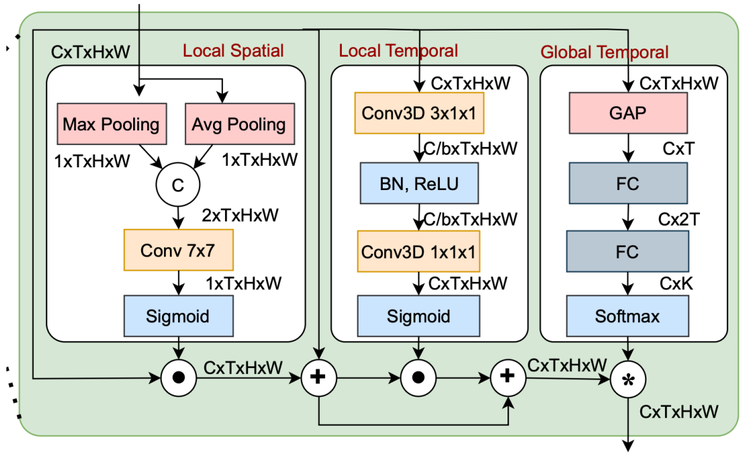
# 2.3.1 ASTRM

## Adaptive Spatio-Temporal Refinement Module

backbone이 뽑은 feature을 공간 정보와 시간 정보를 함께 반영해 더 정교하게 다듬는 모듈

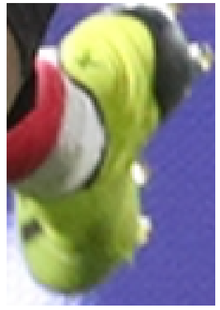


# 2.3.1 ASTRM

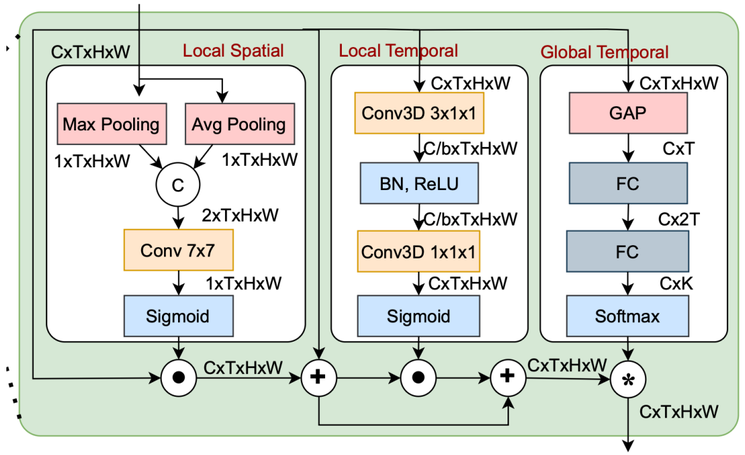



## 1. Local Spatial block

현재 프레임 안에서 어떤 공간 영역이 중요한지 강조하는 부분

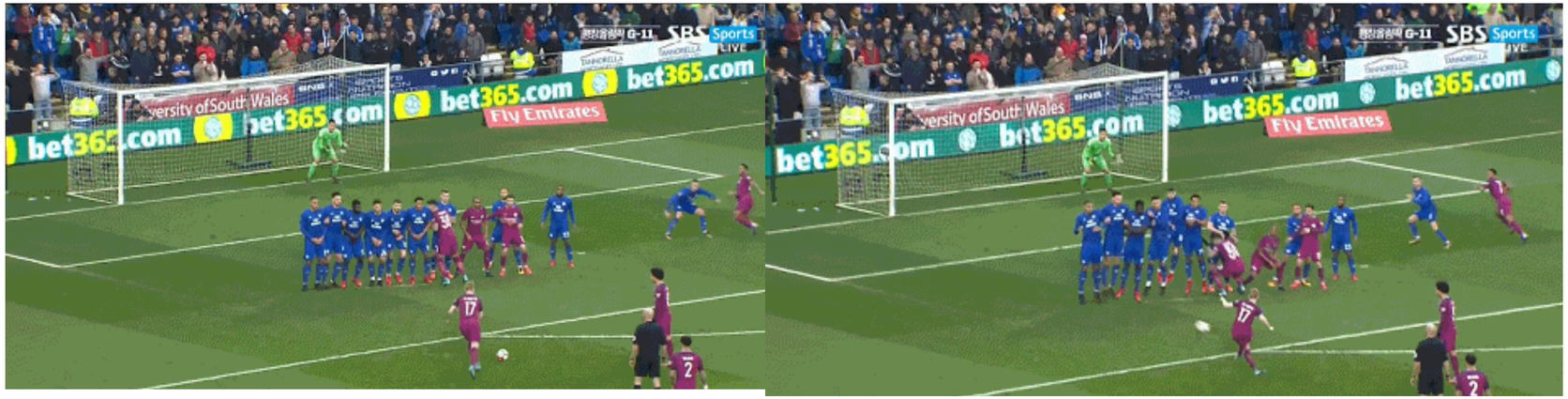


# 2.3.1 ASTRM

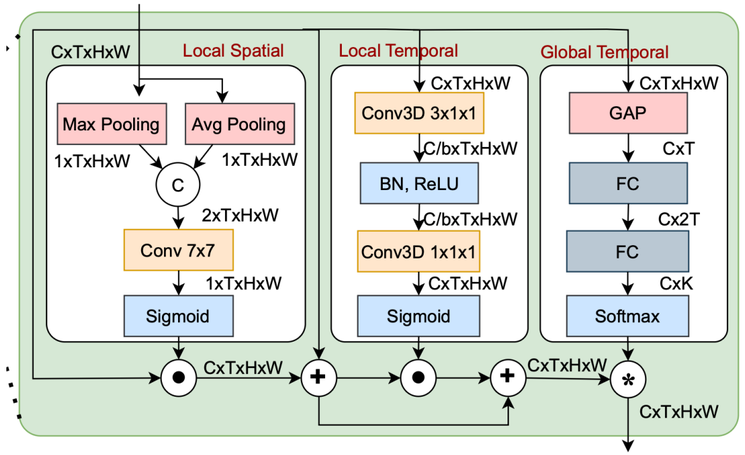



## 2. Local Temporal block

가까운 프레임들 사이의 시간 변화를 반영하는 부분



# 2.3.1 ASTRM



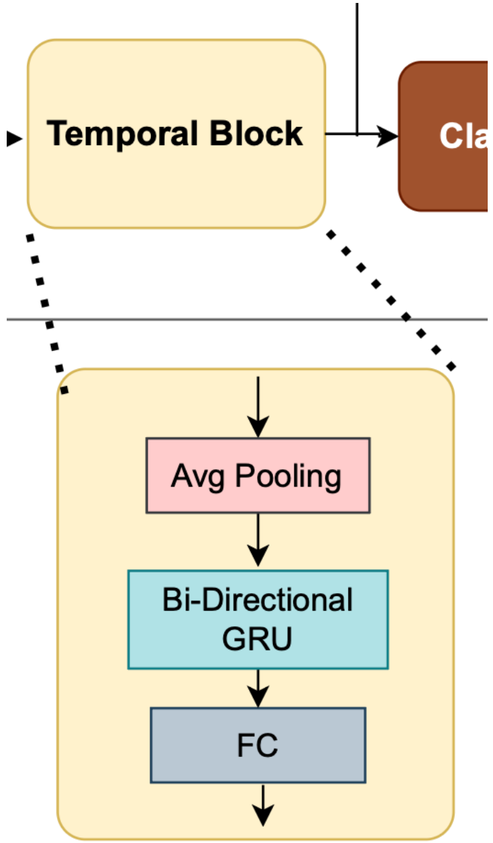

## 3. Global Temporal block

더 멀리 떨어진 프레임까지 고려해 현재 feature을 보정하는 부분

# 2.4 Temporal Block: Bi-GRU

## Temporal bolock

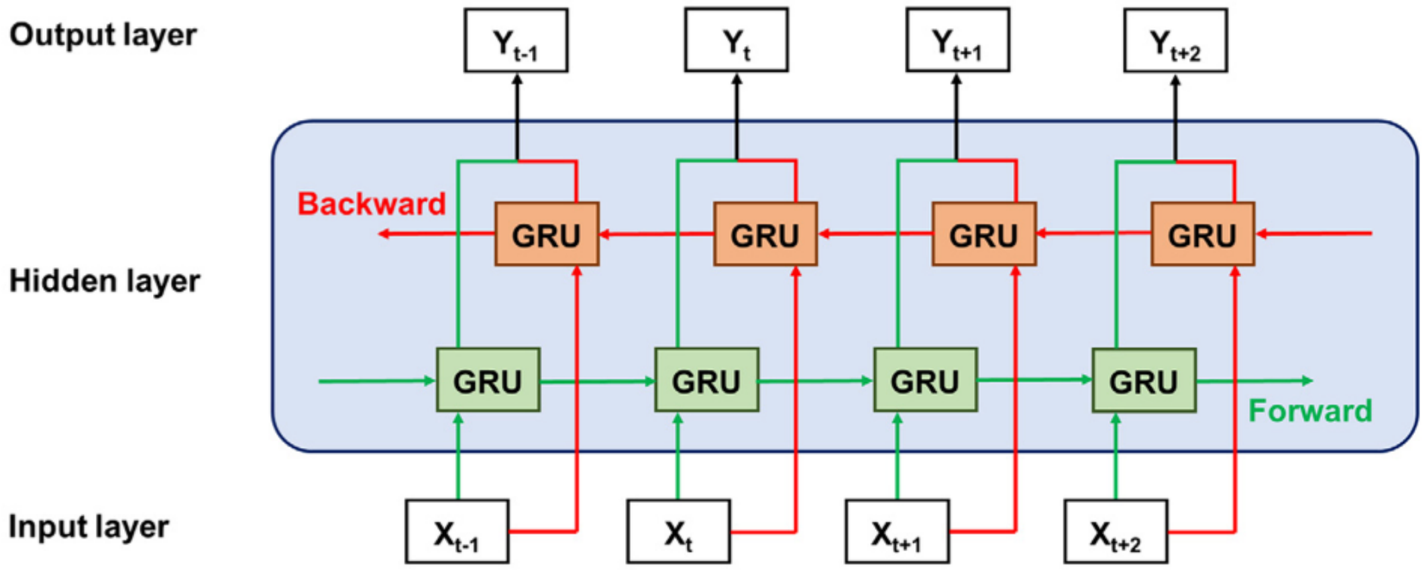
> ASTRM이 feature을 더 잘 만들게 해주더라도, 긴 시퀀스 전체의 흐름을 이해하려면 별도의 temporal block 필요

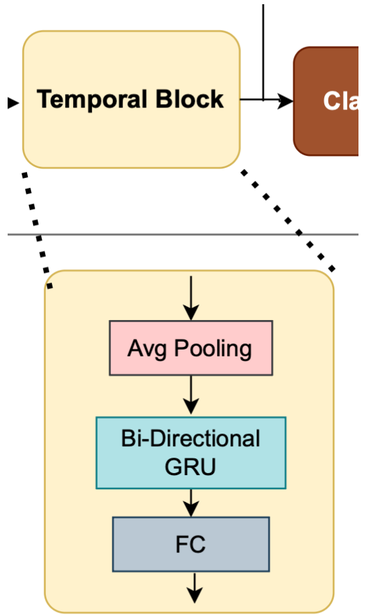


# 2.4 Temporal Block: Bi-GRU

## Bi-GRU (Bidirectional Gated Recurrent Unit)

앞뒤 시간 정보를 모두 보면서 시퀀스를 이해하는 순환신경망



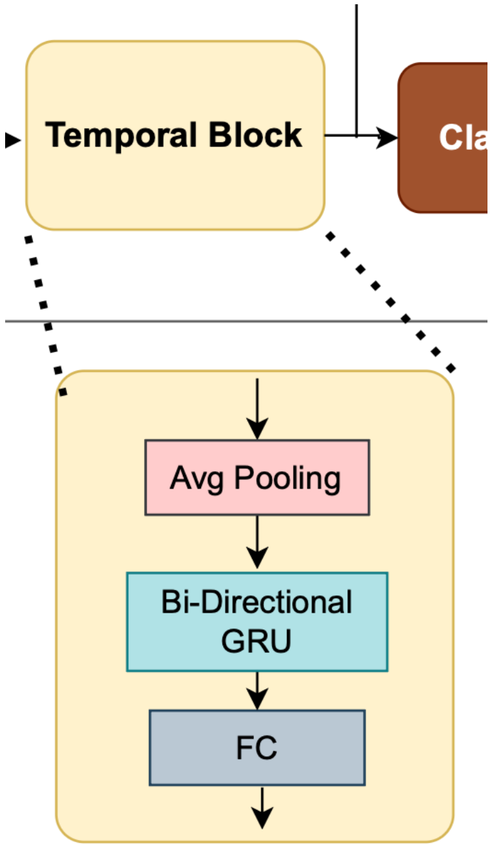
# 2.4 Temporal Block: Bi-GRU

## Why not Transformer?

1. 영상 길이
   self-attention이 긴 시퀀스에서 계산량, 학습 난이도 커짐

2. 데이터 규모가 무한히 크지 않음
   데이터가 충분하지 않으면 불안정하거나 오버피팅

3. PES는 frame-level precision이 중요
   너무 복잡한 global modeling보다 안정적인 sequence modeling이 더 유리



# 2.5 BCE Loss

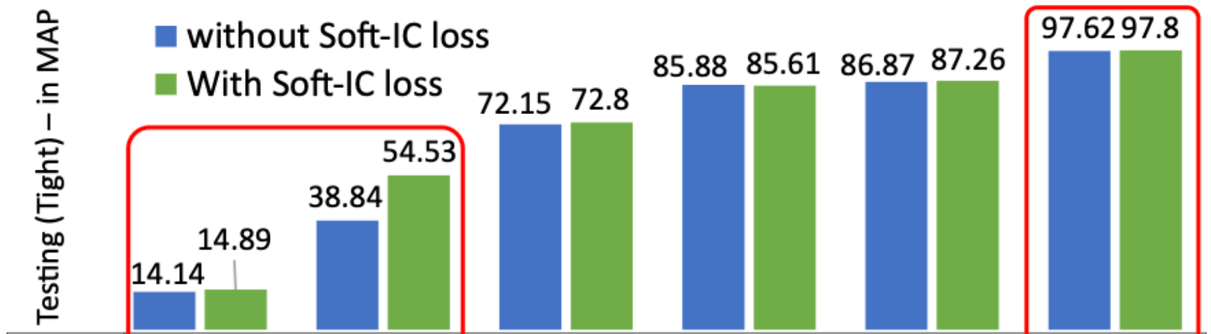
Binary Cross Entropy loss 예측 확률과 정답의 차이를 계산하는 손실함수

$$\mathrm{BCE} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log(p_i) + (1-y_i)\log(1-p_i)\right]$$

class imbalance 문제로 BCE는 기본 분류 오차만 보고,
rare class를 위해 다른 Loss 추가

# 2.6 Soft-IC Loss

> Soft Instance Contrastive loss는 soft label이 생기는 상황에서도 같은 클래스 feature는 더 compact하게 모으고 다른 클래스와는 더 잘 분리 되도록 만드는 손실함수




| No of samples | Yellow->Red card | Red card | Yellow card | Throw-in | Ball out of play | Penalty |
|---|---|---|---|---|---|---|
| Training | 24 | 34 | 1238 | 11391 | 19097 | 96 |
| Testing | 14 | 8 | 431 | 3809 | 6460 | 41 |

03

# Experiment

# 3.1 사용 데이터셋

SoccerNet V2

Tennis

Figure Skating (FS-Comp / FS-Perf)

FineGym

# 3.2 입력 및 실험 세팅

Backbone: RegNetY200

Clip length: 128

Optimizer: AdamW + ASAM

Framework: PyTorch

Hardware: 4×A100 GPU

# 3.3 평가 지표

SoccerNet V2: tight-mAP / loose-mAP

Tennis, FS, FineGym: mAP at $\delta$=0,1,2

04

# Conclusion

# 4.1 SoccerNet V2 결과

| Model | Params (M) | GFLOPs | Tight | Loose |
|---|---|---|---|---|
| Baidu (TF) [35] | — | — | 47.05* | 73.77* |
| SpotFormer [2] | — | — | 60.90 | 76.10 |
| E2E-Spot (RegNet-Y 200MF) [14] | 4.46 | 39.61 | 61.19 | 73.25 |
| E2E-Spot (RegNet-Y 800MF) [14] | 12.64 | 151.40 | 61.82 | 74.05 |
| Spivak [23] | 17.46 | 461.89 | 65.10 | *78.50* |
| ASTRA [29] | 44.33 | 8.83** | 66.63 | 78.14 |
| T-DEED (RegNet-Y 200MF) [30] | 16.36 | 21.96 | 39.43‡ | 52.93‡ |
| T-DEED (RegNet-Y 800MF) [30] | 46.22 | 60.25 | 21.57‡ | 30.49‡ |
| UGLF [27] | 4.46 | 39.61 | 63.51 | 73.98‡ |
| COMEDIAN (ViSwin-T) [6] | 70.12 | 222.76 | *71.33*† | 77.07† |
| Proposed | 6.46 | 60.25 | **73.74** | **79.11** |

# 4.2 다른 스포츠 데이터셋 결과

| Model | Tennis | | | FS-Comp | | | FS-Perf | | | FineGym | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $\delta = 0$ | $\delta = 1$ | $\delta = 2$ | $\delta = 0$ | $\delta = 1$ | $\delta = 2$ | $\delta = 0$ | $\delta = 1$ | $\delta = 2$ | $\delta = 0$ | $\delta = 1$ | $\delta = 2$ |
| TSP [1] + MS-TCN [8] | — | 90.90 | 95.10 | — | 72.40 | 87.80 | — | 76.80 | 89.90 | — | 40.50 | 58.50 |
| TSP [1] + ASFormer [32] | — | 89.80 | 95.50 | — | 77.70 | 94.10 | — | 80.20 | 94.50 | — | 38.80 | 57.60 |
| 2D-VDP [12] + MS-TCN [8] | — | — | — | — | 83.50 | **96.20** | — | 85.20 | *96.40* | — | — | — |
| VC-Spot (RegNet-Y 200MF) [14] | — | 92.40 | 96.00 | — | 61.80 | 75.50 | — | 56.20 | 75.30 | — | 18.70 | 28.60 |
| E2E-Spot (RegNet-Y 200MF) [14] | *45.34** | *96.10* | 97.70 | 22.08* | 81.00 | 93.50 | 22.33* | 85.10 | 95.70 | 10.01* | 47.90 | 65.20 |
| E2E-Spot (RegNet-Y 800MF) [14] | 44.64* | 95.16* | *97.72** | 23.95* | 83.40 | 94.90 | 22.17* | 83.30 | 96.00 | 10.15* | 50.11* | 67.23* |
| T-DEED (RegNet-Y 200MF) [30] | 43.49 | 96.17 | 97.95 | *28.92** | *85.15* | 91.70 | *31.87** | 86.79 | 96.05 | *13.54** | *53.73** | 66.43* |
| T-DEED (RegNet-Y 800MF) [30] | 44.08 | 96.42 | 97.05 | 27.59* | 84.77 | 92.86 | 31.25* | *88.17* | 95.87 | 12.42* | **54.00*** | *66.70** |
| UGLF [27] | —- | — | —- | — | — | — | — | — | — | — | 50.20 | **67.80** |
| Proposed | **61.01** | **96.21** | **97.75** | **37.54** | **86.67** | *95.53* | **40.36** | **89.57** | **97.38** | **15.24** | 52.31 | 66.57 |

# 4.3 Ablation Study

| Method | Tight | Loose |
|---|---|---|
| **Analysis on proposed modules** | | |
| Baseline: E2E-Spot (RegNet-Y 200MF + GSM) | 61.19 | 73.25 |
| RegNet-Y 200MF + ASTRM | 68.63 | 75.57 |
| RegNet-Y 200MF + ASTRM + SAM | 71.89 | 77.50 |
| RegNet-Y 200MF + ASTRM + ASAM | 72.65 | 78.58 |
| RegNet-Y 200MF + ASTRM + ASAM + Focal Loss | 70.36 | 76.86 |
| RegNet-Y 200MF + ASTRM + ASAM + IC Loss | 72.05 | 78.30 |
| RegNet-Y 200MF + ASTRM + ASAM + Soft-IC loss | **73.74** | **79.11** |
| **Analysis on proposed ASTRM module** | | |
| With local spatial only | 69.91 | 76.39 |
| With local temporal only | 70.41 | 76.79 |
| With global temporal only | 71.65 | 77.06 |
| With local spatial + local temporal | 71.06 | 77.10 |
| With local spatial + global temporal | 71.82 | 77.85 |
| With local temporal + global temporal | 71.55 | 76.93 |
| Proposed ASTRM module | **73.74** | **79.11** |
| **Analysis on Different Temporal networks** | | |
| Baseline with Bi-GRU | 61.19 | 73.25 |
| Proposed with Bi-GRU | **73.74** | **79.11** |
| with Deformable Attention | 67.43 | 72.15 |
| with Bi-GRU 2 Layers | 67.85 | 73.03 |
| with Transformer (L1H8) | 65.89 | 70.76 |
| with Transformer (L2H8) | 67.60 | 72.35 |
| with Bi-LSTM | 68.15 | 73.35 |
| with MSTCN | 68.75 | 73.50 |
| **Analysis on Clip Length** | | |
| Clip length = 64 | 43.82 | 52.09 |
| Clip length = 100 (Used in baseline E2E-Spot) | 71.20 | 76.73 |
| Clip length = 128 (proposed) | **73.74** | **79.11** |
| Clip length = 192 | 71.74 | 77.02 |

# Thank you