SDA 시니어 논문세션

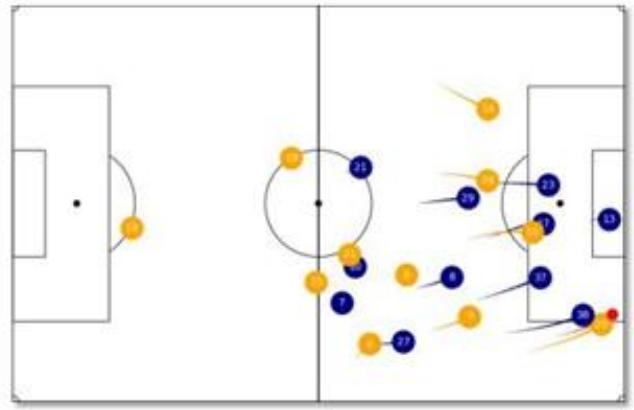
Approaching In-Venue Quality Tracking from Broadcast Video using Generative AI

2026년 3월 20일(금)

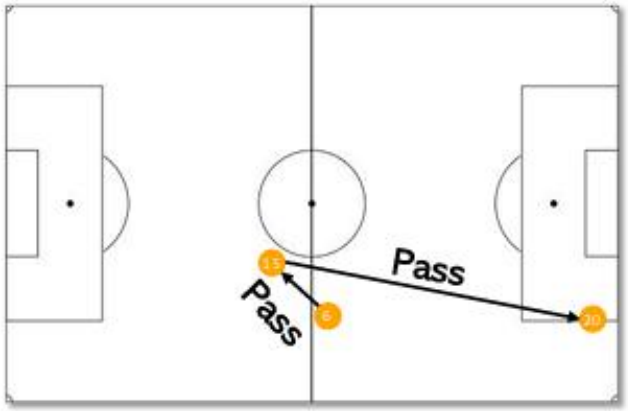
Introduction

- cv 기술의 발전으로 Soccer Tracking data를 이용한 실시간 데이터 분석 가능해짐
- 축구는 득점 기회가 sparse한 스포츠이기에 정확한 맥락 정보가 반드시 필요함
- 즉 고품질의 Soccer Tracking data가 반드시 필요하다

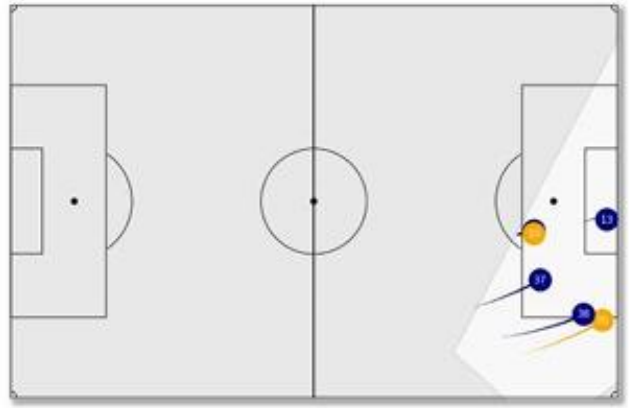
(a) Broadcast footage



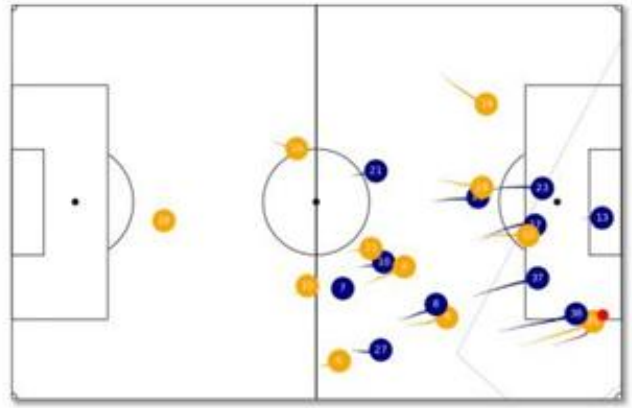
(b) In-Venue Tracking



(c) Event Data



(d) Raw Broadcast Tracking



(e) Imputed Tracking

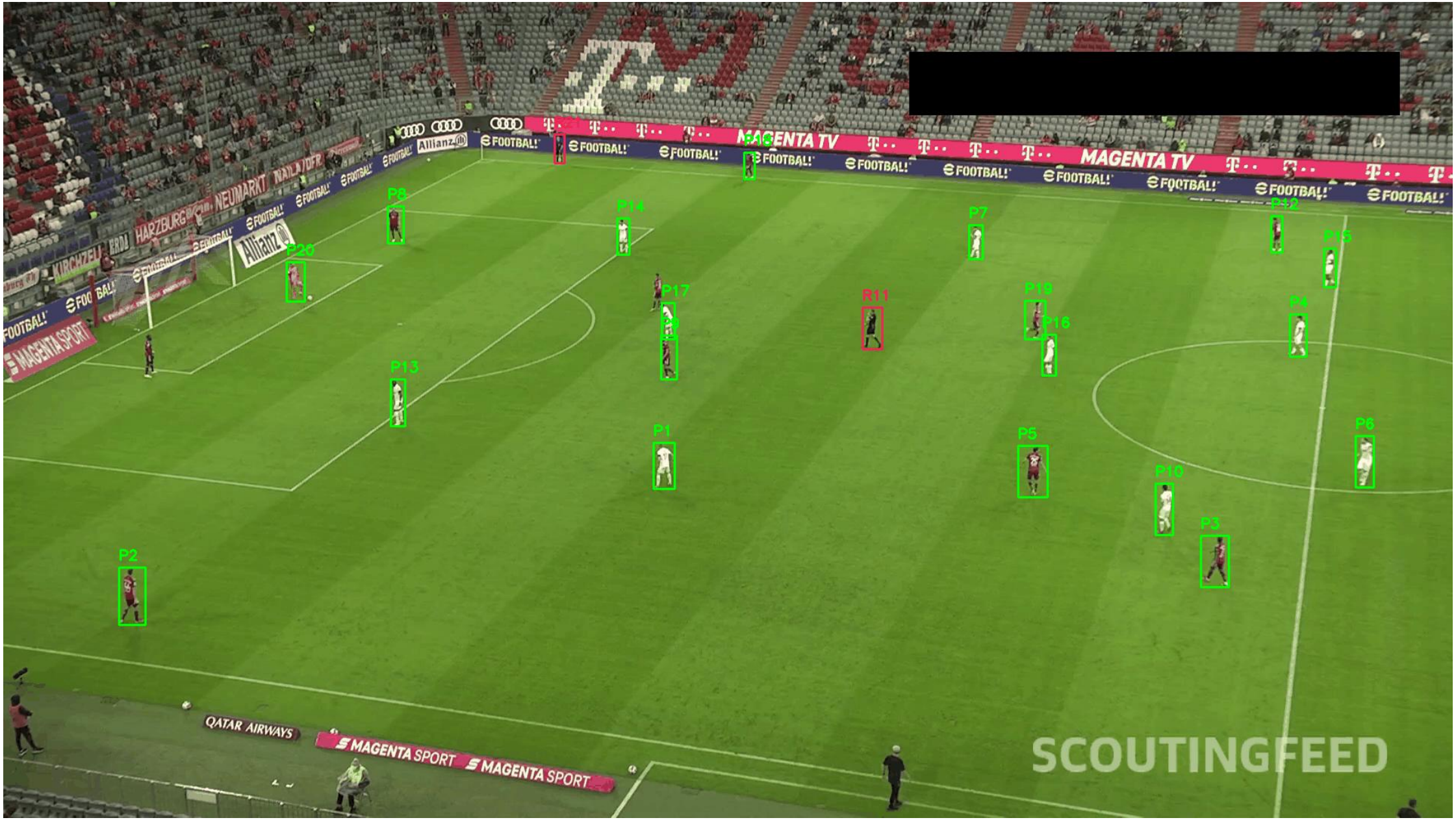
Introduction

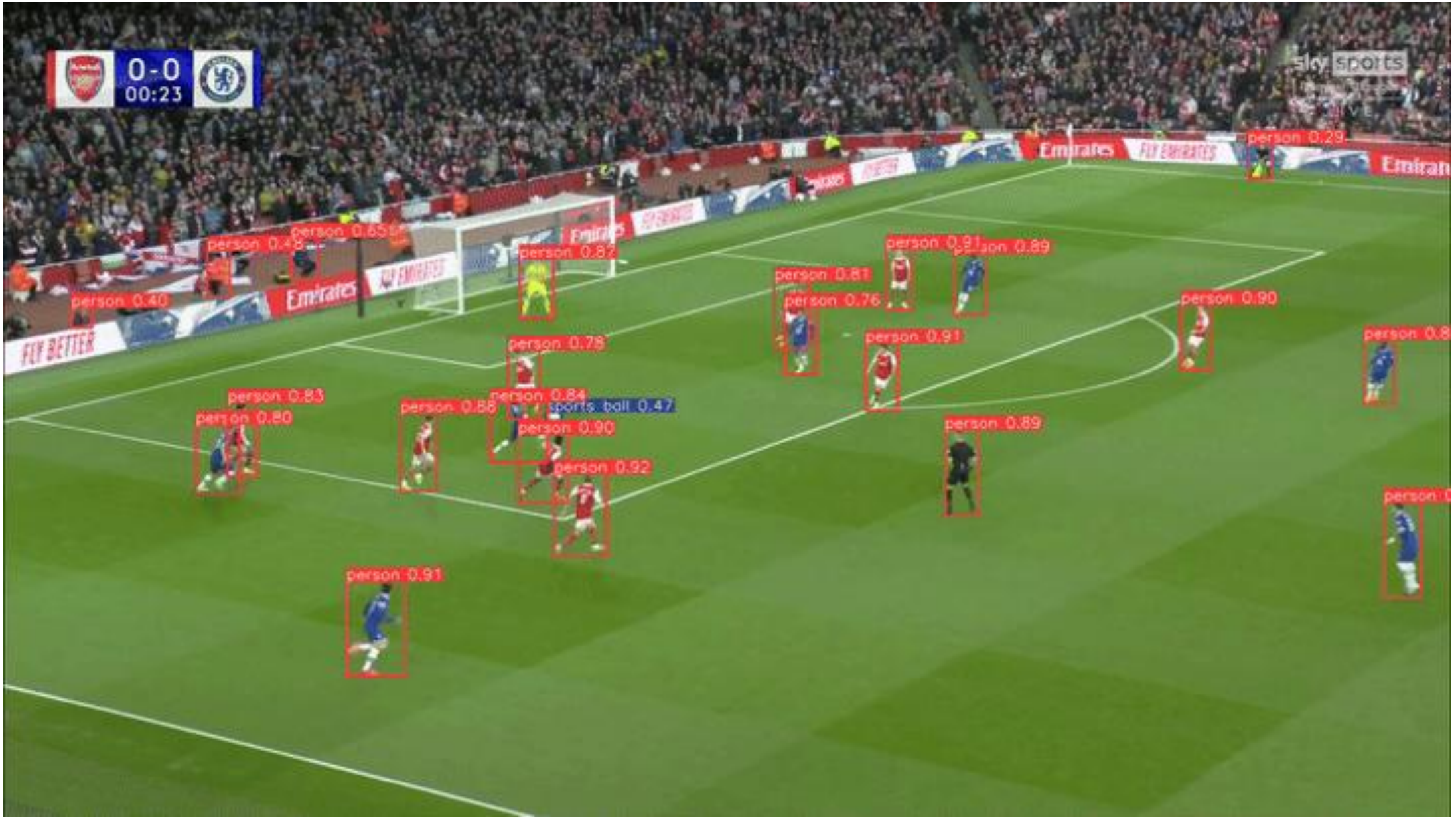
- In-venue tracking data는 고가의 카메라 장비를 경기장에 설치해야함

- Event data는 오프 더 볼 움직임을 감지하지 못한다

- Broadcast tracking data는 위의 문제를 어느 정도 해결하지만 근본적으로 불완전한 데이터이다(선수 가려짐, 클로즈업 뷰, 리플레이, 방송 품질)

- 본 논문은 디퓨전을 활용하여 broadcast tracking data의 선수, 볼 움직임의 불완전한 부분을 그럴듯하게 생성하여 분석에 활용할 수 있는 완전한 데이터로 만든다


P8
P14
P7
P12
P15
P20
P17
R11
P19
P16
P4
P13
P1
P5
P6
P10
P3
P2
MAGENTA TV
FOOTBALL!
Allianz
QATAR AIRWAYS
MAGENTA SPORT
SCOUTINGFEED



YOLOv8 object detection on a soccer game

Introduction

• Approaching In-Venue Quality Tracking from Broadcast Video using Generative AI (2024)

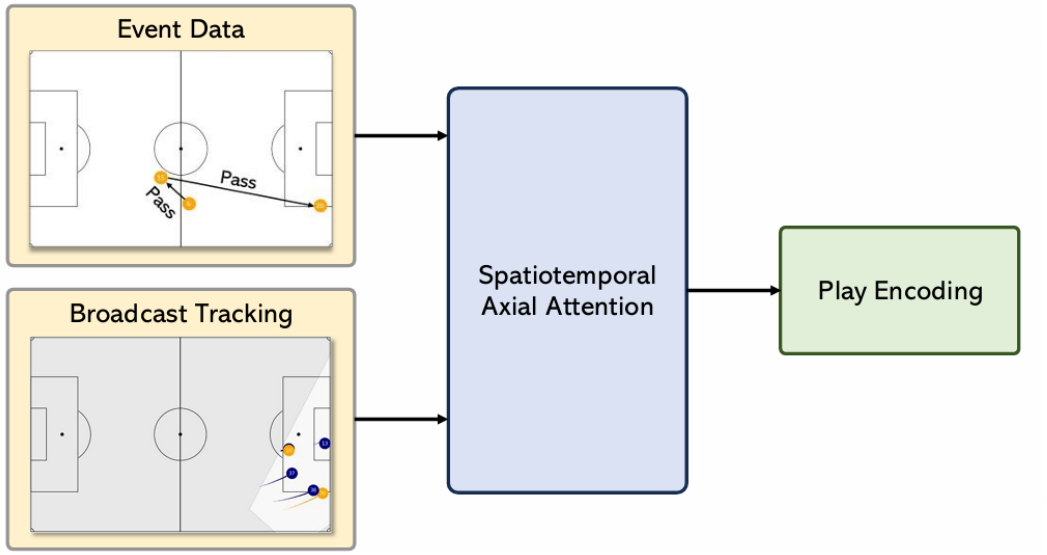
- MIT Sloan Sports Analytics Conference라는 스포츠 분석 컨퍼런스

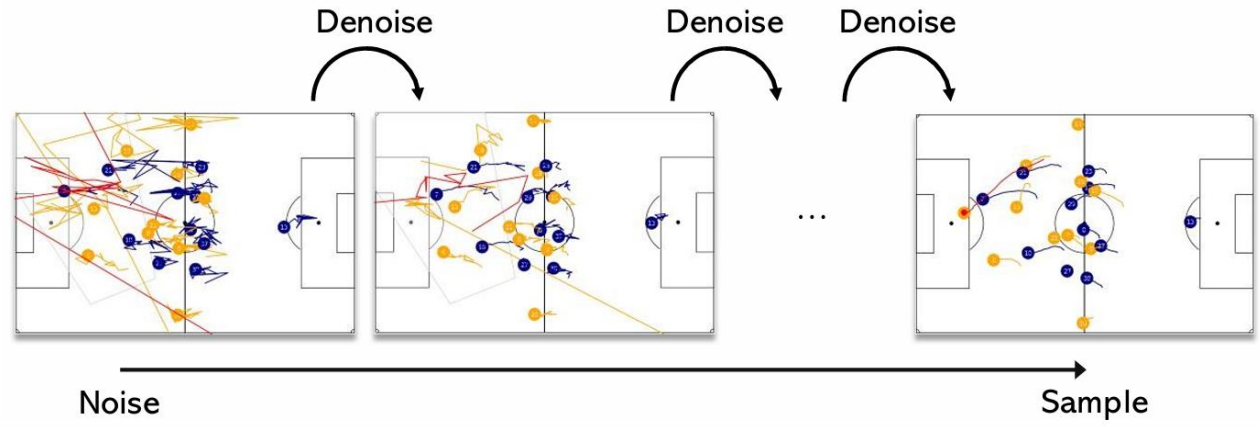
• Multimodal transformer–diffusion framework for large-scale reconstruction of soccer tracking data (2026)

- 컴퓨터 비전 전문 학술 저널(Computer Vision and Image Understanding)
- AI 모델의 내부 구조를 정확히 어떻게 만들었고 어떻게 학습시켰는지

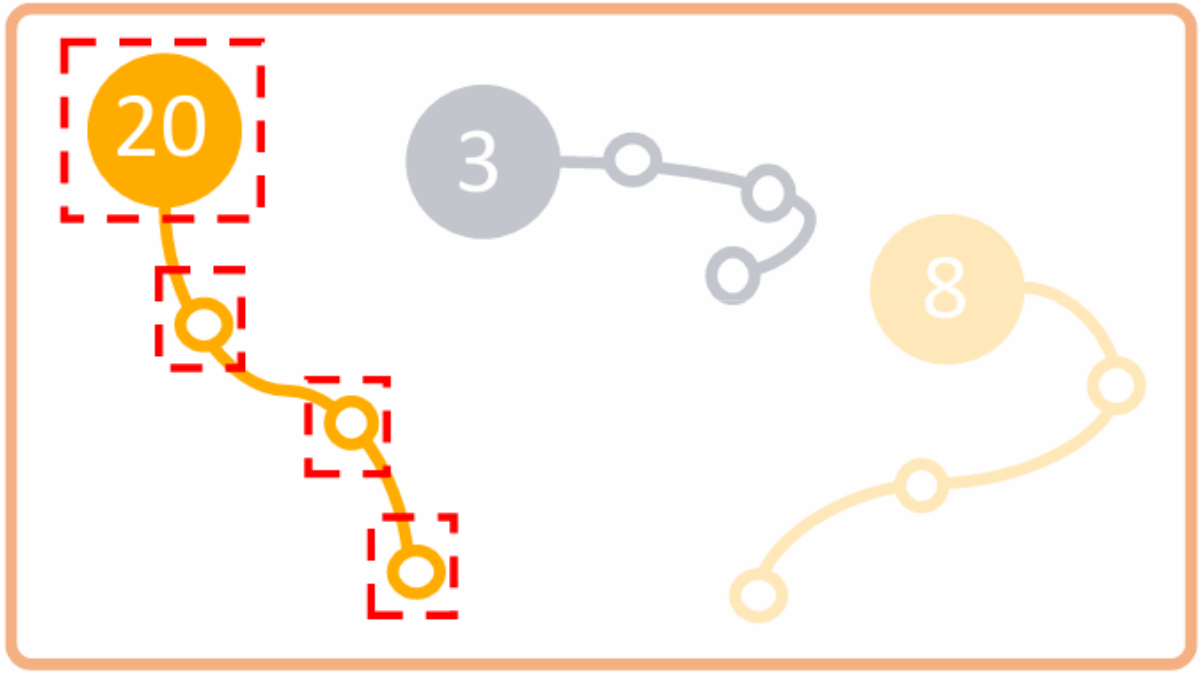
Generating Complete Tracking Data via Diffusion

- 주요 아키텍쳐
- 1. Encoding block(**Spatiotemporal Axial Attention)**
- 2. Diffusion

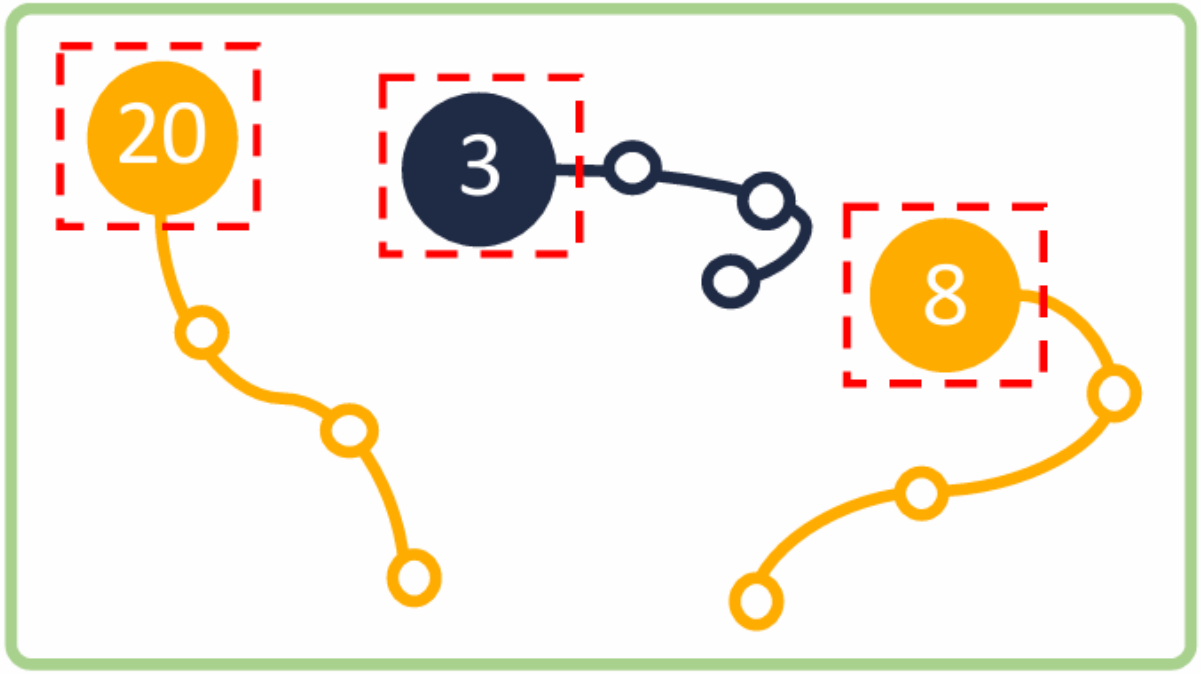




SAA(Spatiotemporal Axial Attention)

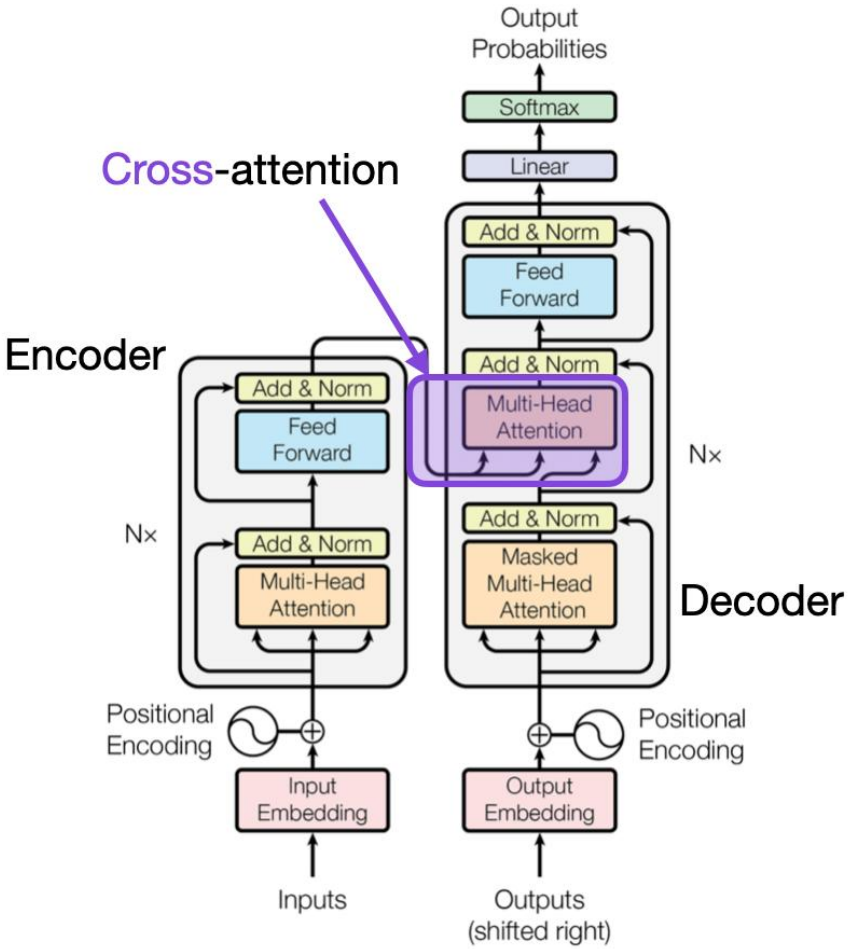


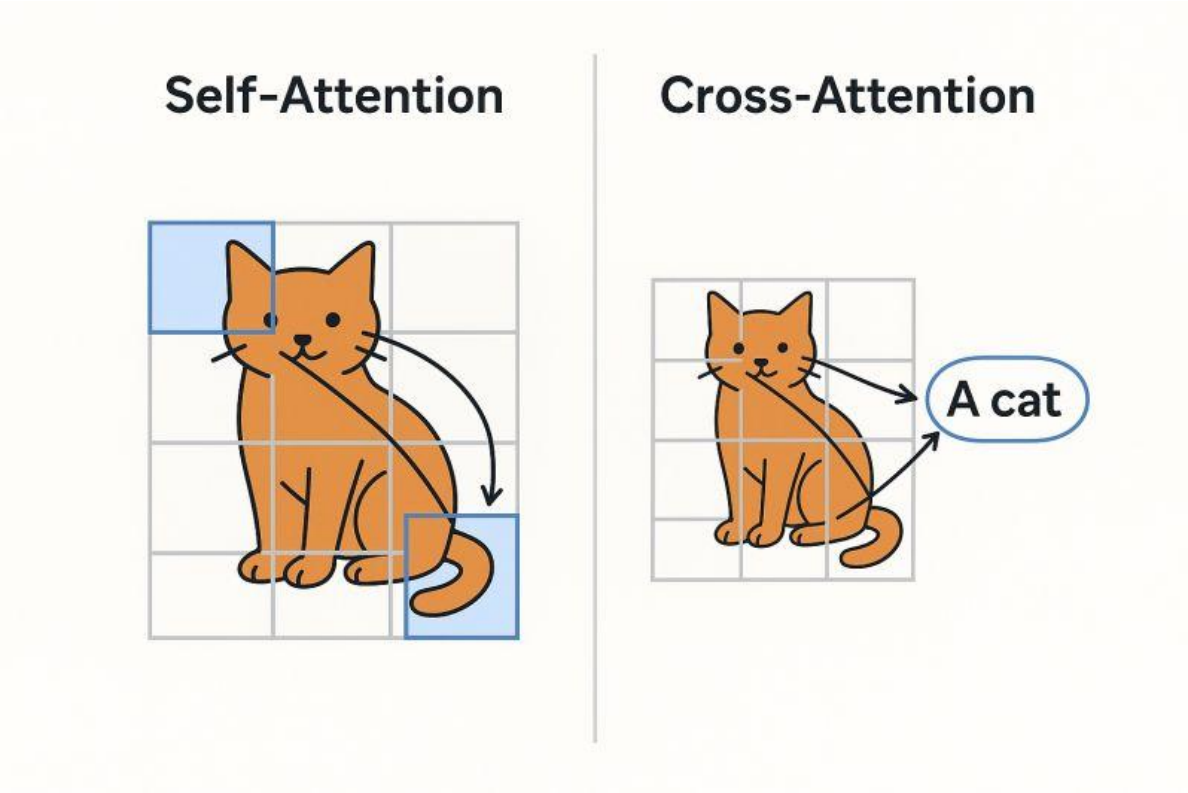
(a) Temporal Attention



(b) Spatial Attention

SAA(Spatiotemporal Axial Attention)

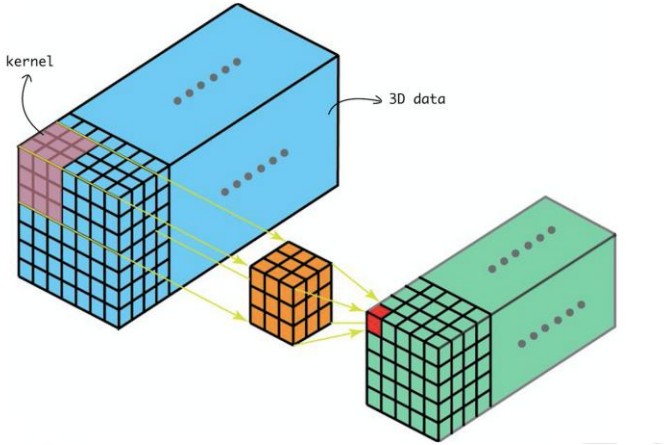




Source: “Attention Is All You Need” (https://arxiv.org/abs/1706.03762)

SAA(Spatiotemporal Axial Attention)

- 3d video understanding
    - cnn 없이 self-attention으로 충분한가?



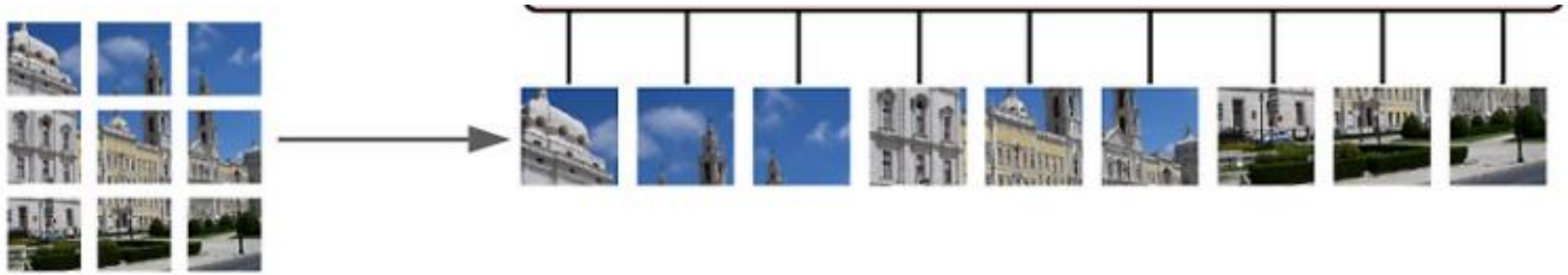
- 3D CNN의 문제점
    - 강한 Inductive bias 존재(local connectivity, translation equivariance)
    - Long-range dependency 모델링의 한계(cnn은 local receptive field 기반)
    - Deep 3d cnn의 높은 training, inference cost

- TimeSformer: pure transformer만으로 video 이해에 적용 시도
- SAA: temporal self attention 적용 후 spatial self attention 적용

SAA(Spatiotemporal Axial Attention)

2D

- ViT: pxp patch



3D

- TimeSformer: pxp patch
- ViViT , VideoMAE: pxpxt patch (t = 2, t = 4, ...)

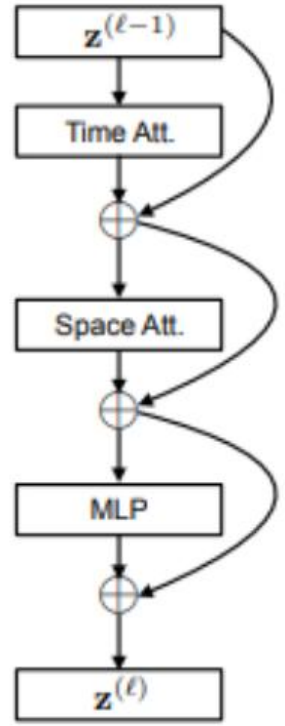
SAA(Spatiotemporal Axial Attention)

- TimeSformer

$$x_{(p,t)} \in \mathbb{R}^{3P^2}$$

1. Patch embedding
2. Spatiotemporal positional embedding
3. Divided Space-time attention(transformer encoder)

- temporal과 spatial에 대해 별도 parameteriation을 둠으로써 learning capacity가 커지고 가장 성능이 좋았음



Divided Space-Time Attention (T+S)

출처: https://allaboutdeeplearning.tistory.com/entry/REVIEW-Is-Space-Time-Attention-All-You-Need-for-Video-Understanding

SAA(Spatiotemporal Axial Attention)

- TimeSformer는 time과 space를 한 번에 섞기보다, 분리해서 차례대로 처리하는 것이 더 효과적임을 보여준다

- 추가 ablation으로, divided attention에서 temporal → spatial 순서가 spatial → temporal 순서보다 0.5% 정도 더 성능이 좋았음

- 먼저 같은 위치에서 시간 변화를 보고, 그다음 한 frame 안에서 공간 관계를 본다
- spatially-biased benchmark에서는 매우 강하고, temporal reasoning이 복잡한 benchmark에서는 충분한 데이터와 긴 input이 중요하다
- long-term video modeling에서 특히 강하다

SAA(Spatiotemporal Axial Attention)

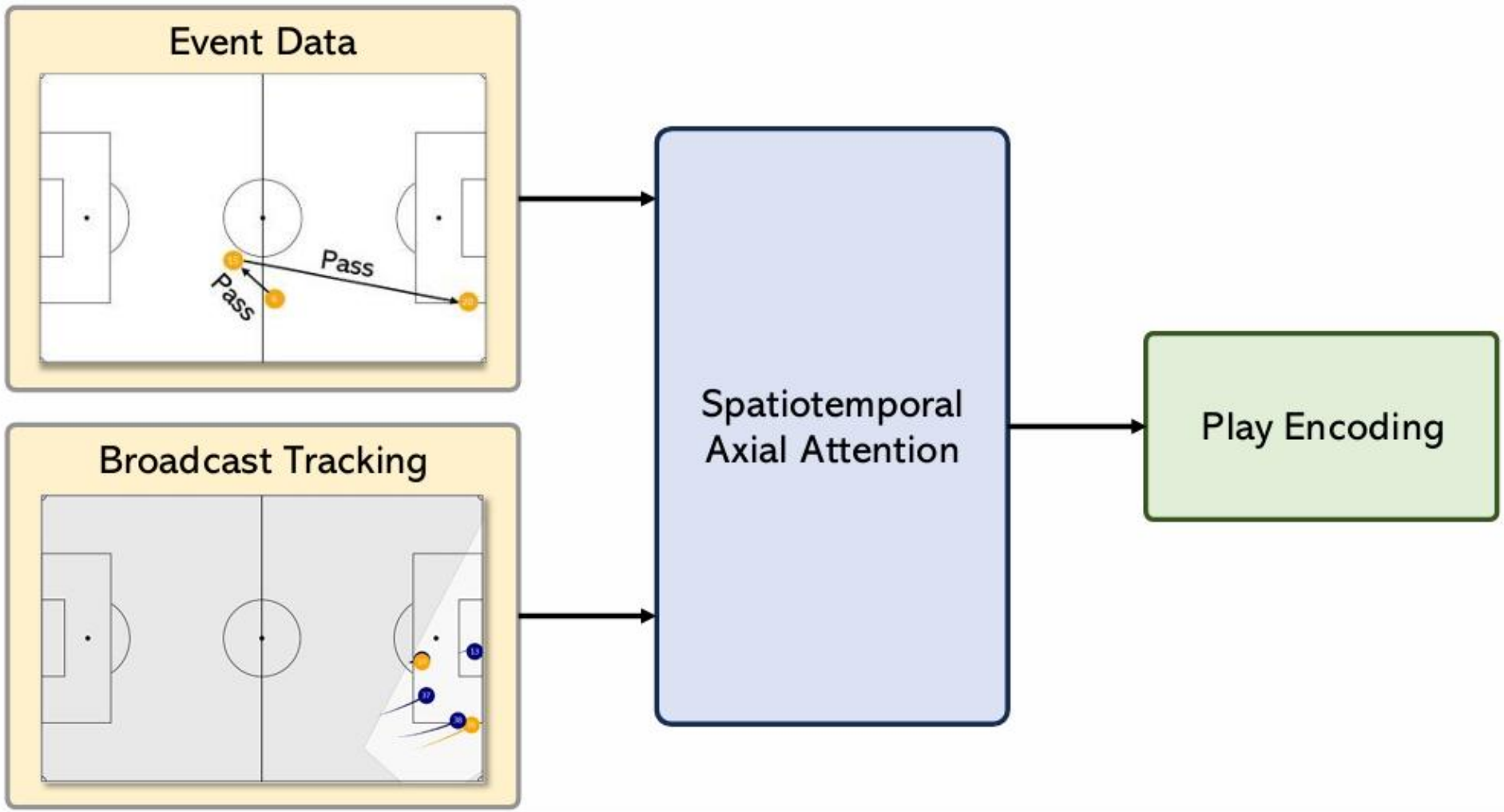
- SAA(Spatiotemporal Axial Attention): 시공간 축방향 어텐션

목표: encode broadcast tracking data

기존: tracking data -> 2D이미지로 시각화 -> cnn

논문: tracking data -> SAA

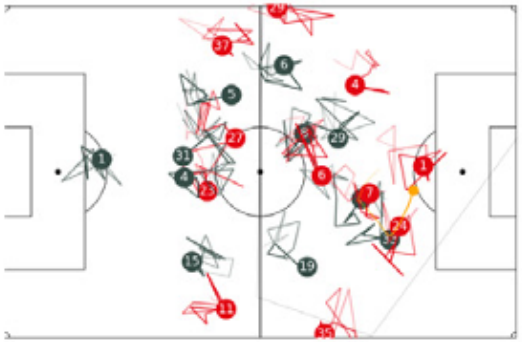
Multi-modal


Event Data
Pass
Pass
Broadcast Tracking
Spatiotemporal
Axial Attention
Play Encoding

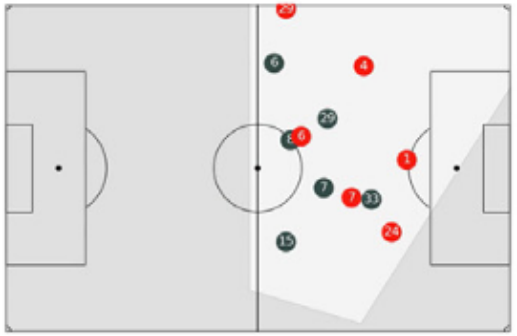
Multi-modal

- event data + broadcast tracking data
- 기존: 스포츠를 unimodal domain으로 간주
- 논문: 스포츠를 multimodal domain으로 간주

Perturbed Tracking



Broadcast Tracking



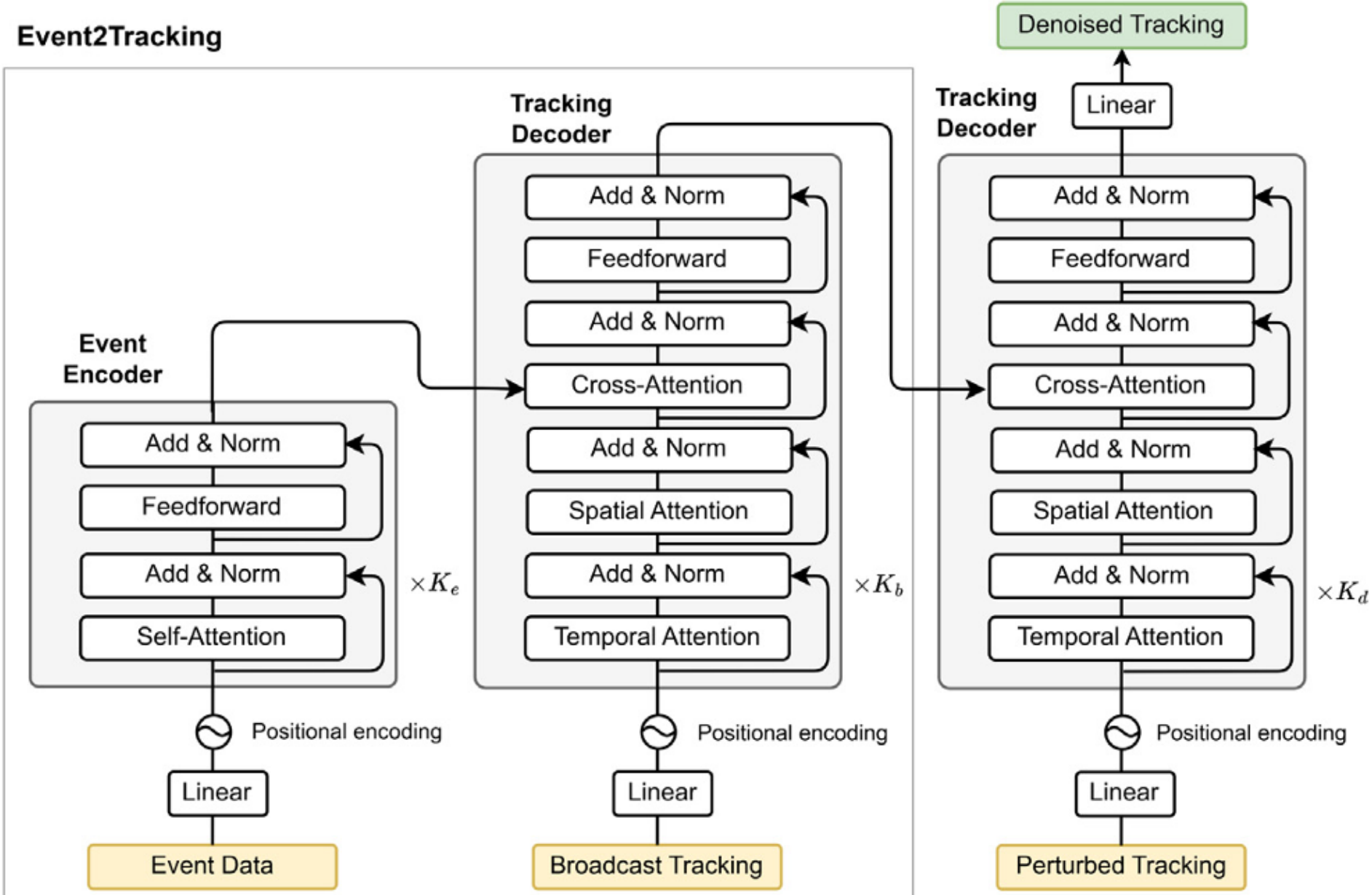
Event Data

category	team	player	x	y
Pass	1	3	32	21
Pass	1	4	11	13
Intercept	0	2	30	42

⋮

(a) Model Inputs

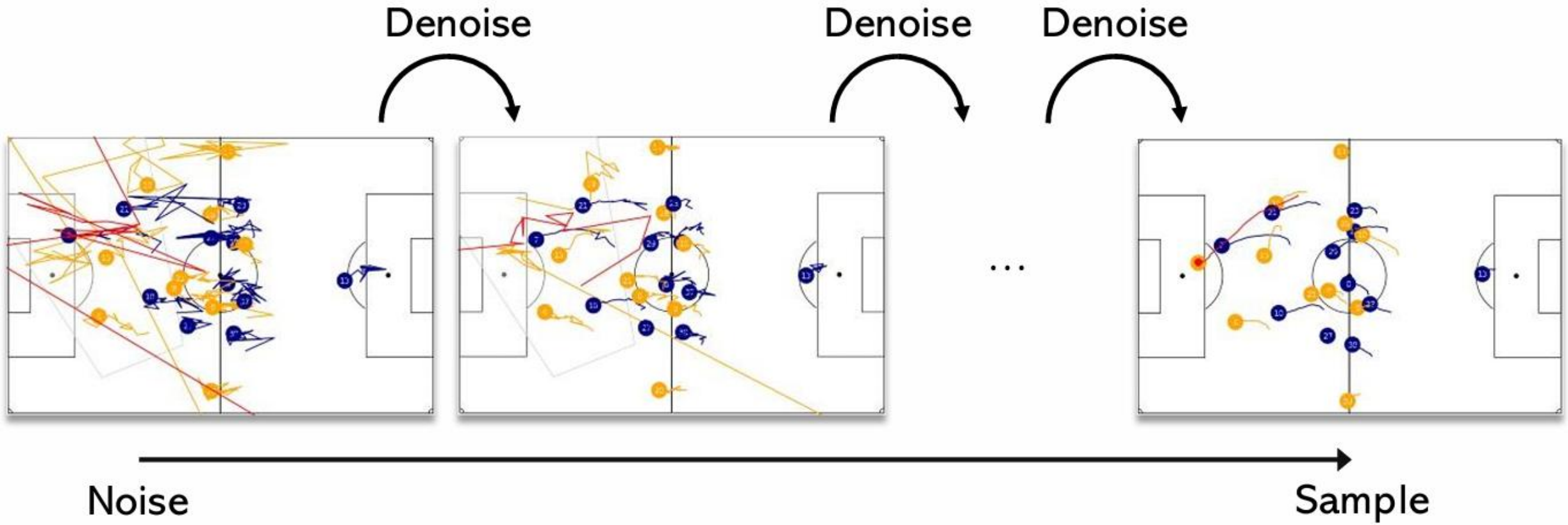
Event2Tracking



(b) Network Architecture

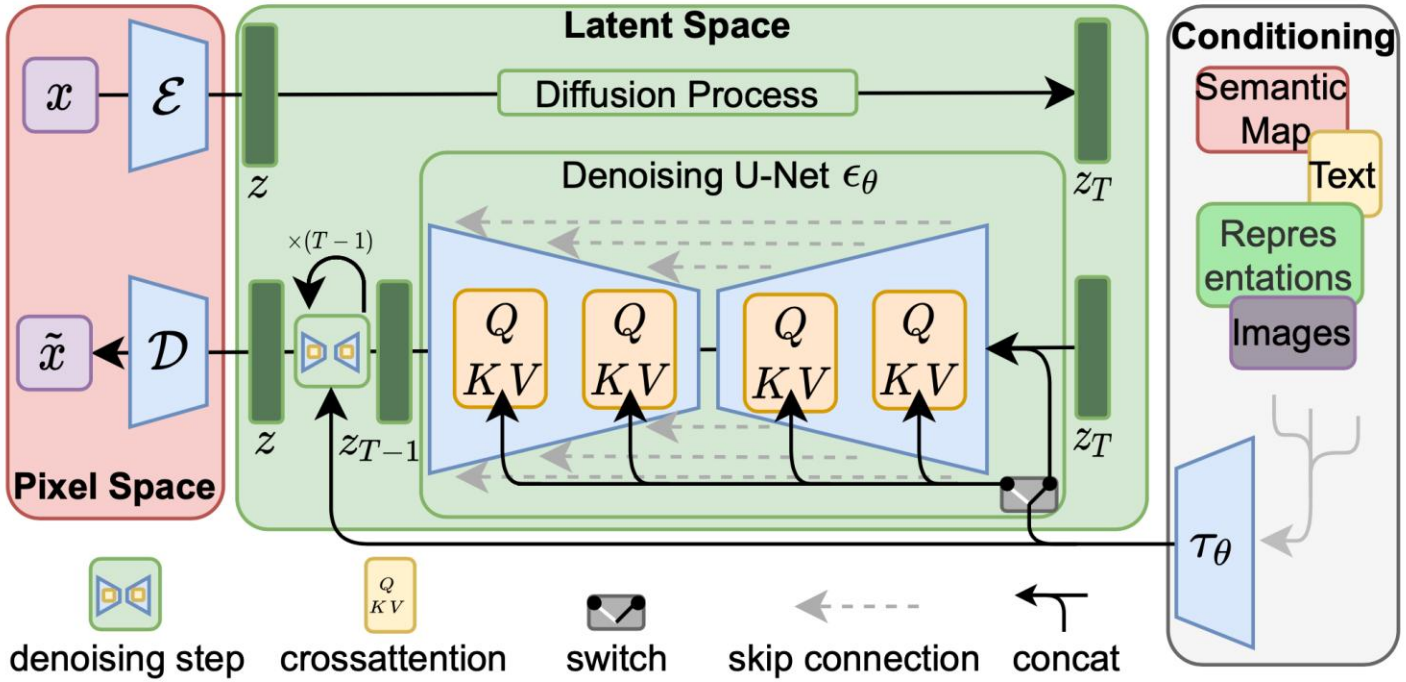
GT +가우시안 노이즈

Diffusion



Diffusion

- DDPM(Denoising Diffusion Probabilistic Models)
- Generative AI
- SAA를 통해 얻은 문맥 정보를 diffusion의 조건부로 줌
- VAE -> GAN -> Diffusion
- Iterative denoising



Performance

• 특정 패스에 대해 다른 선수가 패스를 받을 확률을 예측하는 태스크로, 생성된 tracking data의 성능을 평가함 (xReceiver metric)

• xReceiver Dataset: 130경기의 tracking data

• xReceiver model: SAA 구조

- Tracking data -> SAA -> linear projection -> softmax

- 한 모델은 in-venue tracking 데이터로 훈련, 다른 모델은 broadcast tracking

• 테스트 시 생성된 데이터는 in-venue tracking 데이터로 훈련된 모델에 적용

Performance(Top-k, IoU)

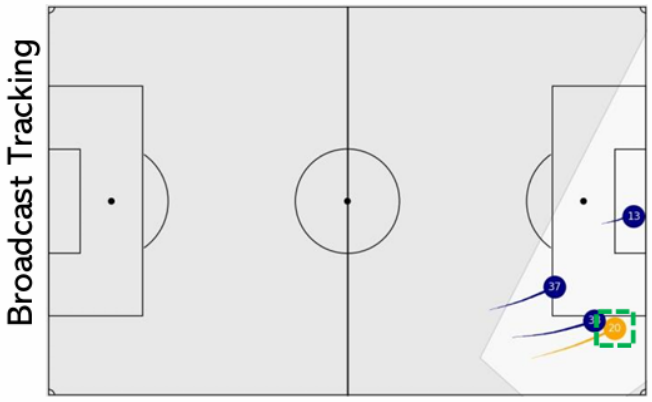
- Top – k : 상위 k개 안에 정답이 속하는지 여부

- IoU: 교집합 / 합집합 비율

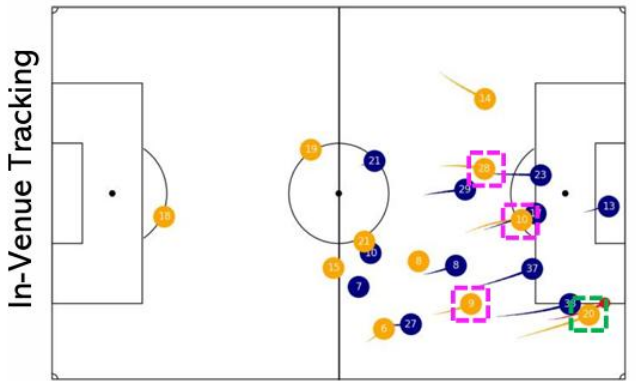
- in-venue and raw broadcast tracking 비교
- in-venue and imputed tracking 비교

Metric	Raw Broadcast	Imputed	In-Venue
Top-1	42.03	55.43	59.06
Top-2	59.42	76.81	79.35
Top-3	71.01	88.04	90.94
Top-4	80.43	93.84	93.48
Intersection over Union	0.5105	0.7121	-

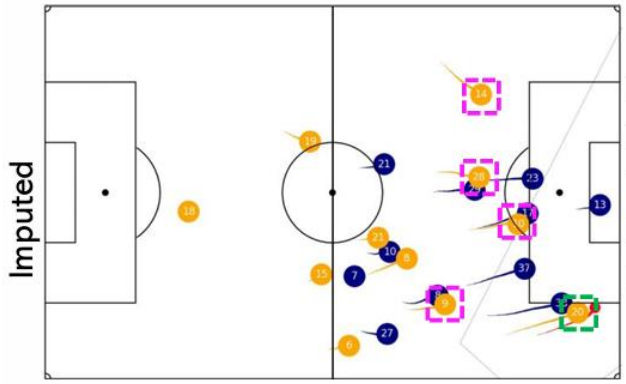
Performance



Passer	20
Receiver	28
Top xReceivers	8,9,10
Receiver in Top xReceivers?	No
IoU	0.5



Passer	20
Receiver	28
Top xReceivers	9,10,28
Receiver in Top xReceivers?	Yes
IoU	-



Passer	20
Receiver	28
Top xReceivers	9,10,14,28
Receiver in Top xReceivers?	No
IoU	0.75

- Broadcast tracking 데이터는 화면에서 가장 최근까지 보였던 선수들에게 확률을 높게 부여하는 편향
- 중계 카메라 사각지대에 있는 선수 전혀 감지 X

- AI가 복원한 데이터는 오랫동안 화면에 잡히지 않았던 선수의 움직임까지도 매우 현실적으로 추론

Critic

(+)

- Soccer Tracking data의 민주화
- 최신 모델 적용 시도
- spatiotemporal axial attention를 스포츠 트래킹 데이터 분석에 있어서 de facto(표준) approach로 제안
- semantic 데이터(event)와 fine-grained 데이터(tracking)를 결합해서 스포츠 도메인에서의 멀티모달 접근법 제안

(-)

- 논문에서 사용한 모델에 대한 자세한 아키텍처 설명 x
- 데이터 민주화를 선언했지만 논문의 내용이 다소 추상적이어서 직접 재현하기 힘듦

Critic

- 영국의 스포츠 기업 STATS PERFORM
- 논문의 저자이자 STATS PERFORM 수석 과학자 패트릭 루시
- https://www.statsperform.com/ko/resource/applications-of-generative-ai-in-sport-q2-update/

Future work

- tracking data 생성(imputed) 모델 성능 개선

- xReceiver 외에 다른 스포츠 분석 태스크로의 확장(전술, 선수들의 피지컬 지표 분석)