A Benchmark Dataset for Event-Guided Human Pose Estimation and Tracking in Extreme Conditions

NeurIPS 2024

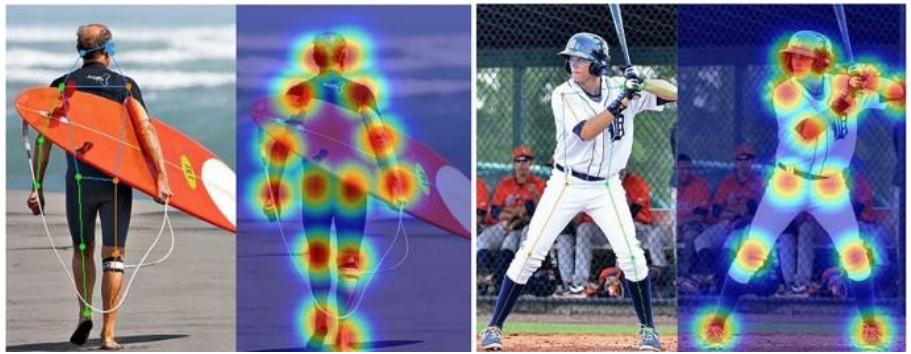
Hoonhee Cho*, Taewoo Kim*, Yuwhan Jeong, and Kuk-Jin Yoon

Visual Intelligence Laboratory, KAIST

KAIST

Motivations

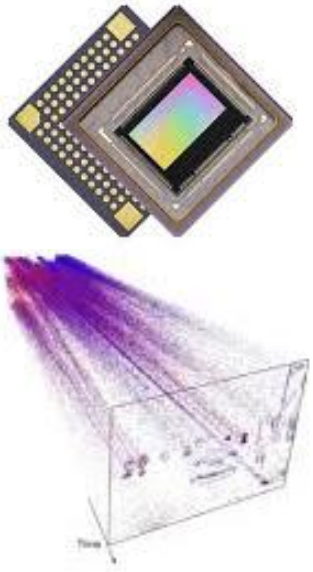
- Human pose estimation is a field in computer vision that has been studied extensively for a long time, achieving significant advancements.

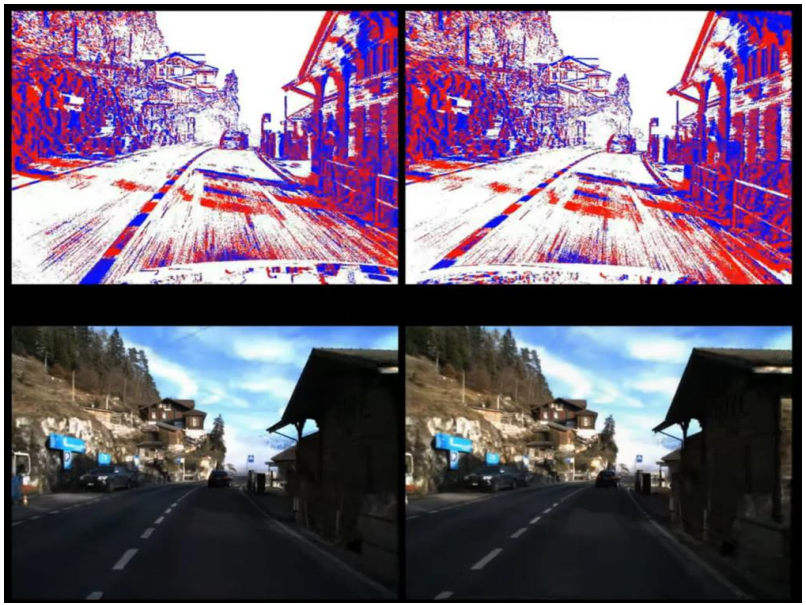


- However, many approaches have been developed using clean, refined data, and there's limited research focused on extreme conditions like **low light and motion blur**.

Motivations

- Event camera provides valuable information in extreme condition (night, fast moving), , making it easier to perform perception tasks in extreme conditions.
- We leverage the properties of event cameras to tackle human pose estimation in extreme conditions.



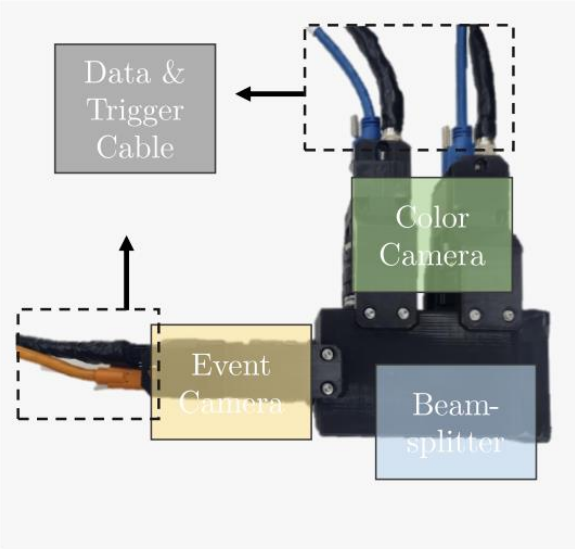


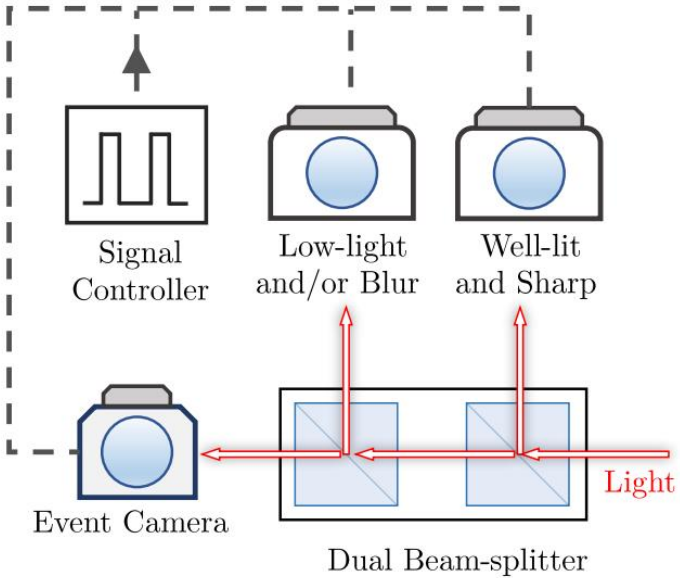
Comparison with Previous Datasets

- Our dataset, the Event-guided Human Pose Estimation and Tracking in eXtreme Conditions (**EHPT-XC**), is the **first multi-person pose dataset using an event camera**, capturing data in **low-light and motion blur** conditions.

Dataset	Total images	# Scenes	# Poses	# Boxes	Track ids	ppF poses	Indoor + Outdoor	Modality	Resolution	Extreme Conditions	
										Low Light	Motion Blur
MPII [3]	25K	491	40K	✗	✗	1-17	✓	RGB	1280×720	✗	✗
Penn Action [49]	160K	2326	160K	160K	✗	1	✓	RGB	640×480	✗	△
COCO [20]	200K	200K	250K	500K	✗	1-20	✓	RGB	640×480	✗	✗
MOT20 [9]	13K	8	✗	1.65M	✗	✗	✓	RGB	1920×1080	✗	✗
PoseTrack21 [10]	66K	514	177K	429K	✓	1-13	✓	RGB	1280×720	✗	△
ExLPose [17]	3K	251	15K	✗	✗	1-26	✓	RGB	1920×1200	✓	✗
mRI [2]	160K	300	160K	160K	✗	1	✗	RGB+depth+mmWave+IMU	512×424	✗	✗
GoPose [26]	676k	unk	676k	✗	✗	1	✗	RGB+WiFi	1920×1080	✗	✗
MM-Fi [43]	320K	1080	320K	✗	✗	1	✗	RGB+depth+LiDAR+mmWave+WiFi	1280×720	✗	✗
RELI11D [42]	239K	48	239K	✗	✗	1	✓	RGB+IMU+LiDAR+Event	1280×800	✗	✗
JRDB-Pose [33]	28K	54	636K	2.8M	✓	1-36	✓	RGB+LiDAR	752×480	✗	✗
NTU RGB+D [28]	57K	17	57K	✗	✗	1	✗	RGB+Depth	512×424	✗	✗
M^3FD [22]	4K	8	✗	34K	✗	✗	✗	RGB+IR	1024×768	✓	✗
WIHPD [41]	2K	*unk*	7.3K	✗	✗	1-12	✓	RGB+IR	1280×720	✓	✗
EHPT-XC (Ours)	16K	158	38K	38K	✓	1-13	✓	RGB+Event	1373×928	✓	✓

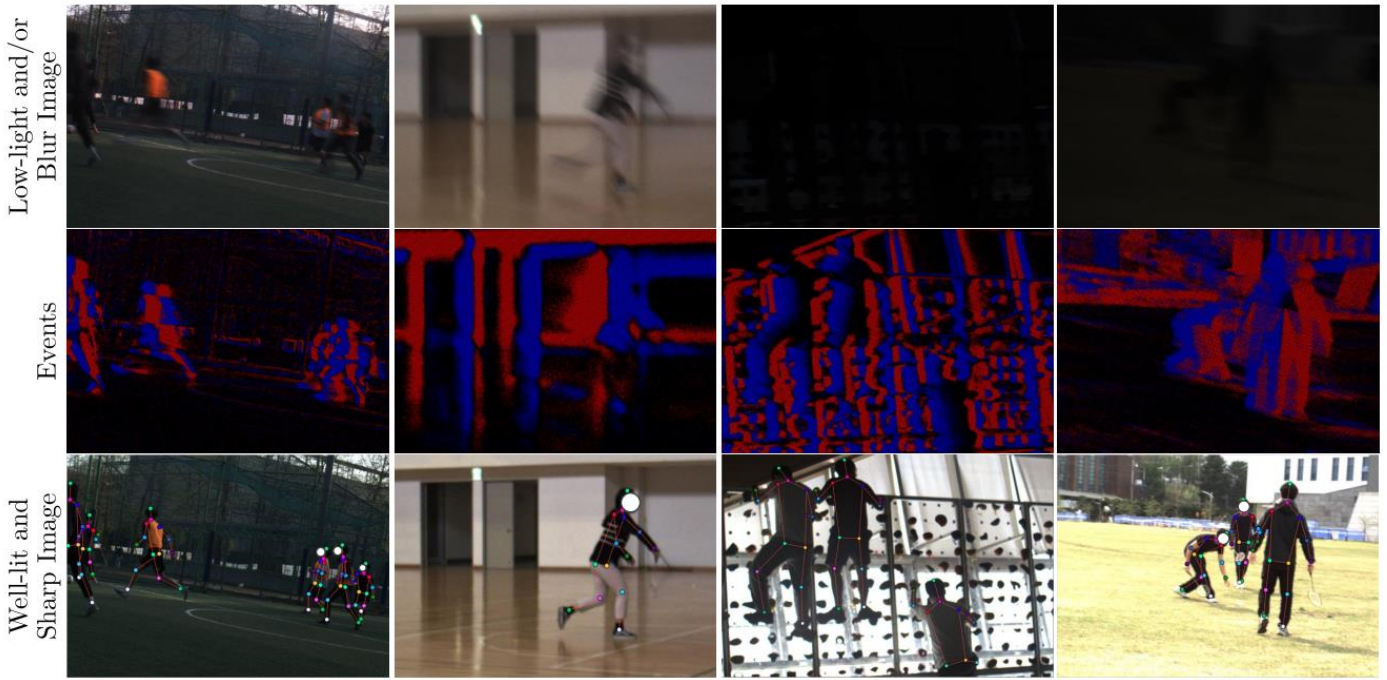
Camera System





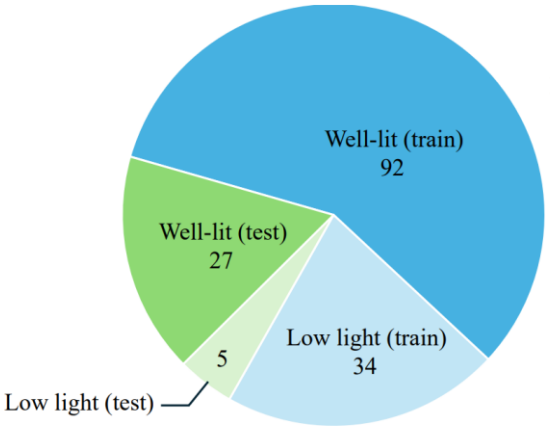
- Labeling degraded images is challenging, so we developed a triplet camera system for data acquisition.
- One RGB camera captures clean, well-lit images, while another captures degraded images.

Data Samples

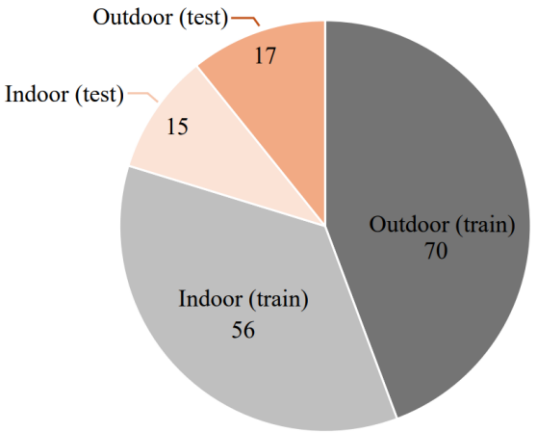


- The dataset provides degraded images, time-synced events, and well-lit images.

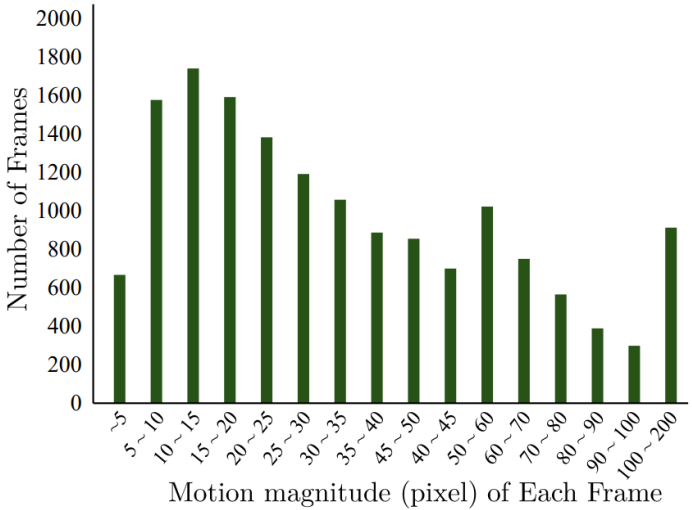
Dataset Distributions



Distribution of the light conditions

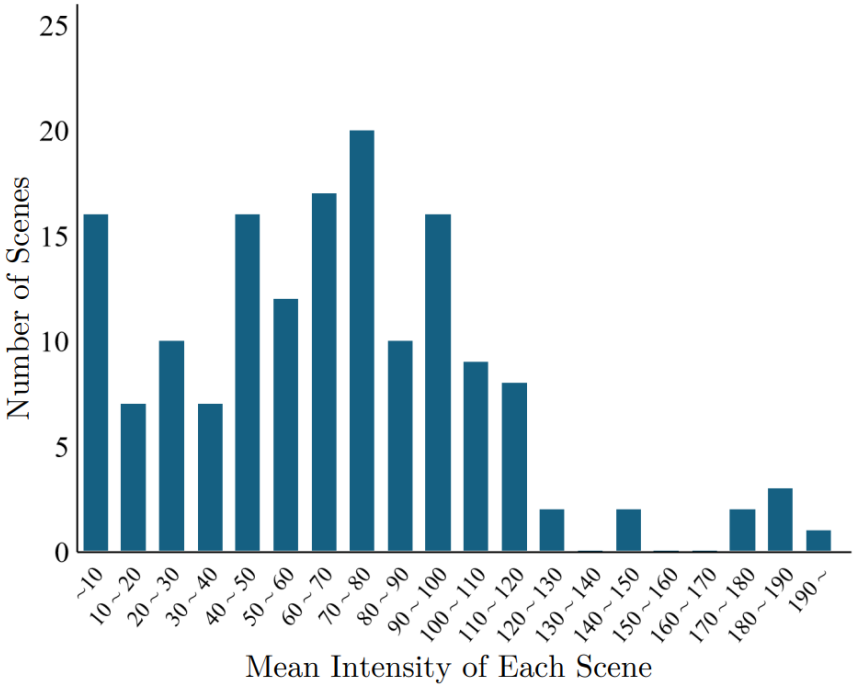


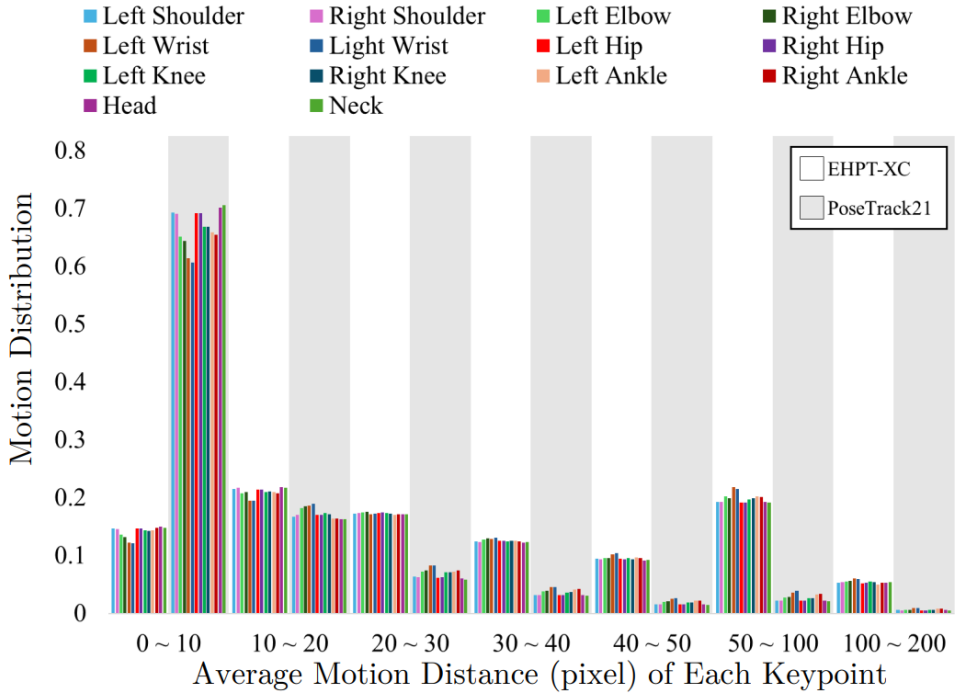
Distribution of the indoor/outdoor



- The EHPT-XC dataset includes both indoor and outdoor scenes, various lighting conditions, and a range of motions.

Dataset Distributions

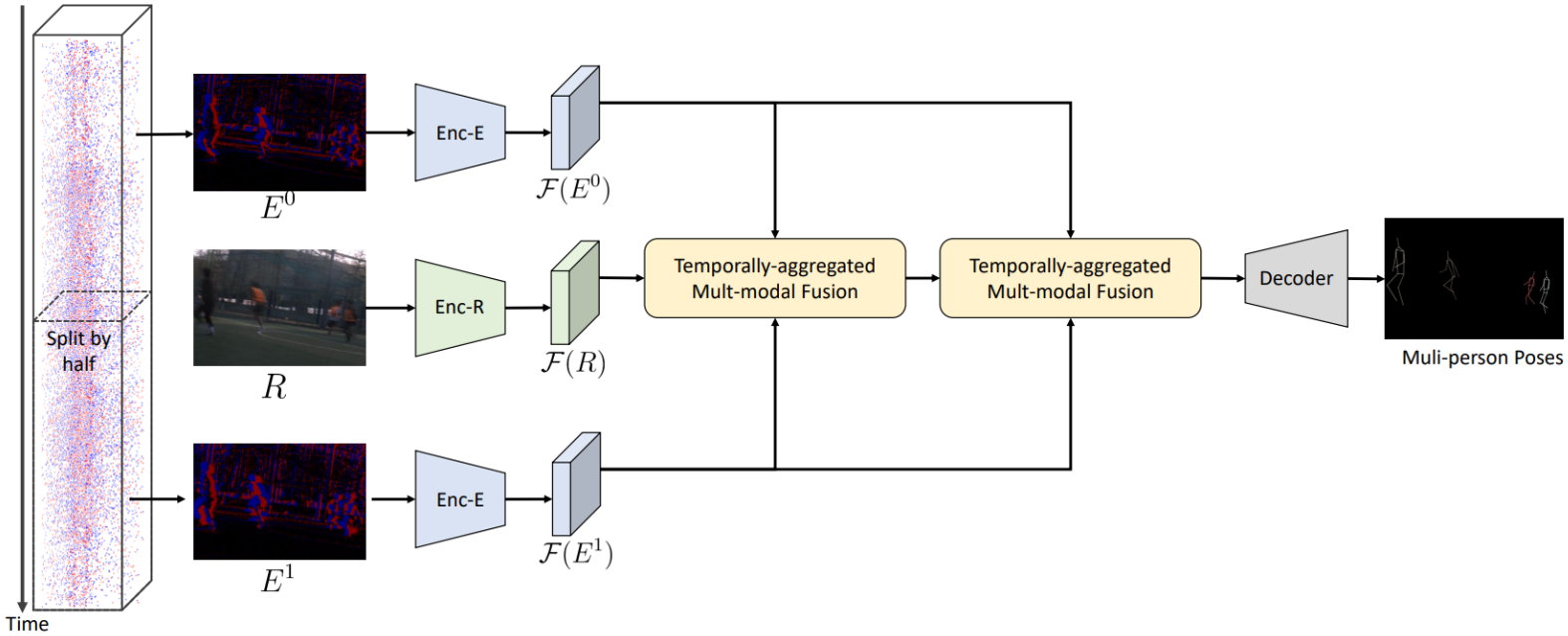




- The intensity distribution of our images is uniform and diverse.
- Compared to existing datasets (PoseTrack21), which are skewed towards smaller motions, ours has a balanced range of motion distances.

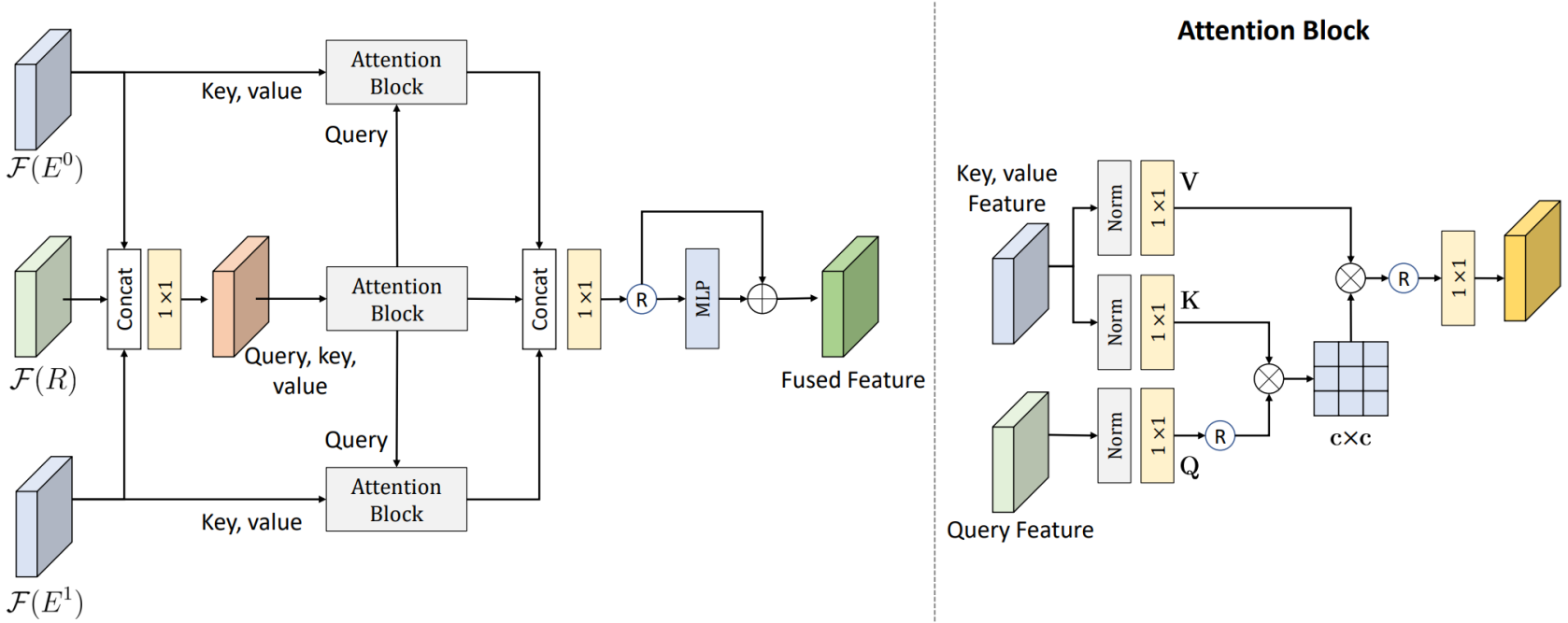
KAIST

Baseline Methods



- As a baseline fusion module for combining images and events, we propose a method that splits the events into two parts around the image timestamp for fusion.

Baseline Methods



- The two split event features and the image feature are fused into a single feature using a transformer.

Benchmark Results

Multi-person pose estimation results on the EHPT-XC dataset.

Modality	Method	mAP@0.5:0.95	mAP@0.5	mAP@0.75	mAR@0.5:0.95	mAR@0.5	mAR@0.75
RGB	HigherHRNet [7]	22.7	31.8	23.3	67.0	85.8	70.0
	DEKR [13]	25.1	34.8	26.5	63.8	85.5	66.9
	CID [34]	24.0	33.1	24.5	65.9	87.7	67.9
Event	HigherHRNet [7]	32.1	37.8	33.6	83.8	95.5	86.8
	DEKR [13]	33.1	39.4	34.6	84.0	96.6	87.6
	CID [34]	31.1	38.1	32.6	85.6	97.0	88.8
RGB + Event [29]	HigherHRNet [4]	33.8	39.0	34.3	84.4	94.6	85.9
	DEKR [9]	36.9	41.0	37.2	87.7	95.8	**88.9**
	CID [25]	33.7	39.3	34.5	**88.0**	**98.2**	**90.2**
RGB + Event (Ours)	HigherHRNet [4]	**34.3**	**39.7**	**34.8**	**86.0**	**97.0**	**87.6**
	DEKR [9]	**37.3**	**42.0**	**37.5**	**87.8**	**97.1**	**88.9**
	CID [25]	**34.5**	**40.9**	**36.0**	84.7	98.0	88.4

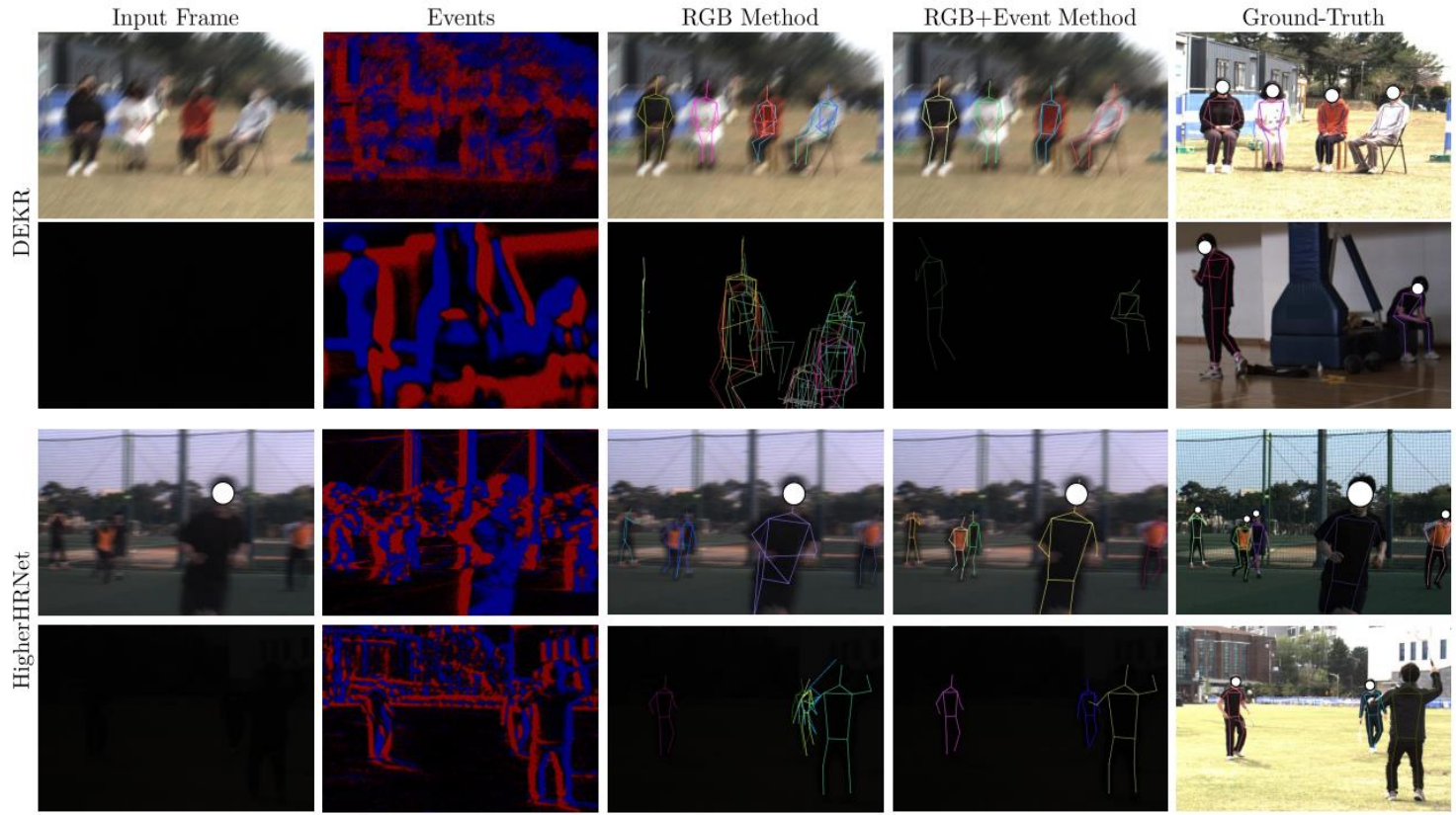
Benchmark Results

Multi-person pose tracking results on the EHPT-XC dataset.

Modality	Pose Estimation	Tracking	MOTA↑	IDF1↑	FP↓	IDSW↓	FN↓
RGB	DEKR [13]	ByteTrack [50]	33.19	18.73	328	316	4643
		UniTrack [36]	25.60	8.56	90	159	5611
		OC-SORT [6]	23.25	15.08	67	127	5880
RGB + Event	DEKR [13]	ByteTrack [50]	47.37 (+14.18)	20.46	461	405	3299
		UniTrack [36]	46.82 (+21.22)	7.93	205	374	3630
		OC-SORT [6]	42.34 (+19.09)	22.72	193	207	4163

Qualitative Results

Multi-person pose estimation results on the EHPT-XC dataset.



Conclusion

- We are the first to tackle human pose estimation in extreme conditions, including low light and motion blur, by utilizing event cameras.

- Our dataset, the Event-guided Human Pose Estimation and Tracking in eXtreme Conditions (EHPT-XC), has the following features:
  - **Multi-human Pose Dataset with Neuromorphic Cameras**: EHPT-XC is the first dataset of its kind using event cameras, featuring track IDs for multi-object tracking.
  - **Real-captured Data in Extreme Conditions:** The dataset is collected with a triplet camera setup, specifically designed to handle low-light and motion-blur conditions.
  - **Indoor/Outdoor Environments and Various Scenarios:** EHPT-XC encompasses diverse settings, including sports and different numbers of individuals, enhancing its versatility.

- Our work shows the great potential of event camera for human pose estimation!