

Lifting Monocular Events to 3D Human Poses

Gianluca Scarpellini^{1,2}, Pietro Morerio¹, Alessio Del Bue^{1,3}

¹Pattern Analysis & Computer Vision - Istituto Italiano di Tecnologia, Italy

²University of Genova, Genoa, Italy

³Visual Geometry and Modelling, Istituto Italiano di Tecnologia, Italy

{gianluca.scarpellini, pietro.morerio, alessio.delbue}@iit.it

Abstract

This paper presents a novel 3D human pose estimation approach using a single stream of asynchronous events as input. Most of the state-of-the-art approaches solve this task with RGB cameras, however struggling when subjects are moving fast. On the other hand, event-based 3D pose estimation benefits from the advantages of event-cameras, especially their efficiency and robustness to appearance changes. Yet, finding human poses in asynchronous events is in general more challenging than standard RGB pose estimation, since little or no events are triggered in static scenes. Here we propose the first learning-based method for 3D human pose from a single stream of events. Our method consists of two steps. First, we process the event-camera stream to predict three orthogonal heatmaps per joint; each heatmap is the projection of the joint onto one orthogonal plane. Next, we fuse the sets of heatmaps to estimate 3D localisation of the body joints. As a further contribution, we make available a new, challenging dataset for event-based human pose estimation by simulating events from the RGB Human3.6m dataset. Experiments demonstrate that our method achieves solid accuracy, narrowing the performance gap between standard RGB and event-based vision. The code is freely available at https://iit-pavis.github.io/lifting_events_to_3d_hpe.

1. Introduction

Natural selection has empowered us with an efficient perception system, enabling our brain to process visual information and respond to threats promptly. Biological evidence suggests that humans and other animals process visual cues differently from traditional cameras [34, 61]. Instead of handling frames at fixed time intervals, mammals collect visual cues asynchronously and elaborate information on demand. This observation pushed the research community and engineers to develop new sensors,

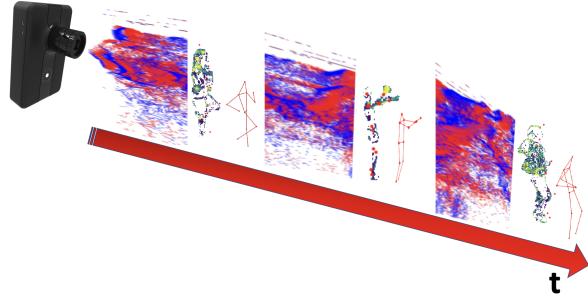


Figure 1: Our method computes the 3D pose of a subject from event-camera streams. We first aggregate events into meaningful representations that are then used to estimate the final 3D pose of the subject.

event-cameras, with a neuromorphic inspiration that provide crucial advantages in time-critical tasks and applications [31, 4].

Indeed, one of the most important activities we are daily involved in is interacting with other human beings. For this reason, we developed the ability to forecast human motion and adapt our behavior accordingly [50, 25, 15]. However, in order to encode asynchronous quick reactions to human activities, a basic but fundamental task to solve is the estimation of human pose from event-based streams. Human pose estimation is already widely adopted in action recognition [29, 30], human tracking [28], sport assistance [56], and virtual reality [55]. Most of the adopted solutions involve using multiple cameras and require the subjects to wear special markers suites [55]. Despite their efficiency and broad adoption, these techniques rely on delicate synchronization and are difficult to deploy in real environments. For these reasons, monocular human pose estimation represents a fascinating research challenge with growing interests in the industry [10, 51, 48]. There are two different families of solutions to solve 3D human pose es-

timation: skeleton-based and model-based solutions. The former regresses skeletal 3D joints from a planar image [38, 43, 44], while the latter fits a tri-dimensional parametric model of the human body to the subjects in the scene [16, 3]. Recently, Xu *et al.* adapted a model-based approach to event-cameras [57]. Although they underline an interesting solution, their approach requires RGB images to guide the tracking and cannot be used for real-time applications, as it relies on a time-consuming optimization phase.

We propose instead an event-only approach to predict skeletal poses from a single stream of events (Figure 1). Our pipeline consists of two steps. First, a Convolutional Neural Network predicts the projection of each joint of the skeleton onto three orthogonal planes. Instead of predicting the positions directly, we constrain our approach onto estimating intermediate heatmaps of probabilities for each joint. Second, we triangulate the sets of 2D positions of each joint to predict the 3D joint pose. Similar workss adopt raw events to solve pose estimation tasks [54, 57]. On the other hand, we aggregate events into tensor-like representations. Although event-representations have been widely investigated and validated [52, 27, 62, 13], no previous work has explored these approaches for monocular human pose estimation. Moreover, differently from standard computer vision, where transfer learning across different tasks has been widely investigated [60], it is still unclear whether pre-training on related tasks can improve event-based human pose estimation. To fill these gaps, we compare different pre-training tasks and different event-representations.

Experiments on natural and synthetic events validate our approach. For validating performance on real event-camera recordings, we adopt the recent DHP19 dataset [5]. DHP19 provides recordings of 33 activities from four different points of view. Despite the excellent contribution to event-based vision, DHP19 provides few self-occlusions or hard situations, as most of the activities are conducted on the spot. To fill these gaps, we propose a new, challenging event-based dataset for Human Pose Estimation by simulating events from the standard Human3.6M dataset [21]. The event-camera community proposed numerous simulation tools to tackle the absence of data [46], and these solutions have been successfully adopted in recent work [47]. Human3.6m provides challenging scenarios, such as people walking and moving extensively in the scene, that are intrinsically harder. In fact, we test our proposal on both DHP19 and Event-Human3.6 and provide different ablations and experiments to support our claims. To summarize, our proposal consists of three main contributions:

- A pipeline to predict human poses from a single stream of events;
- A new, synthetic dataset for benchmarking event-based Human Pose Estimation;

- Extensive experiments to validate transfer learning and pre-training approaches for event-based human pose estimation.

2. Related work

In this Section, we discuss skeleton-based approaches for solving monocular Human Pose Estimation and underline their critical points. To solve the limitations, previous works have focused on high-speed cameras; these approaches suffer, however, from high computational and storage limitations. Event-cameras can be a solution to these problem. Indeed, recent works have adopted event-by-event approaches to track objects and subjects in real-time. On the other hand, our approach aggregates events into tensor-like representation, which can be fed to standard Deep Learning models. Moreover, we recognize a gap of challenging datasets for event-based human pose estimation and discuss events simulation and its benefits.

Monocular Human Pose Estimation. Industry and academia are looking at human pose estimation with increasing interest [10, 51, 48]. Commercial solutions usually require special markers suite to track subjects from multiple point of views [55]. Despite their satisfactory performances, these approaches are extremely costly and require careful setup choices to perform well at high speeds [39]. For these reasons, monocular approaches have been widely researched [58]. Along with background, light conditions, texture, and image imperfection, monocular solutions must also handle the intrinsic ambiguity of monocular vision and therefore pose unanswered challenges to the research community. Model-based solutions are an established line of work on this problem. These approaches estimate the full 3D body and shape of the subjects by fitting a model of human body [32]. Although recent model-based works achieve impressive performances [16, 3, 18], here we will focus on skeleton-based solutions. Skeleton-based approaches aim to regress 3D joints of the skeleton directly from images. As machine learning models deal with probability better than with scalar values, recent solutions predict dense probability maps (denominated heatmaps) of the location of the skeletal joints onto the image plane. In particular, Newell *et al.* made a break-trough in the field by proposing Stacked-hourglass model [42]. The authors stack multiple Convolutional Neural Networks to extract expressive heatmaps and apply a differentiable sort-argmax operator to retrieve the 2D pixel location of each joint. Although we can adapt stacked-hourglass models to predict 3D (volumetric) heatmaps [44], this path is widely open for improvements, especially since Volumetric Heatmaps are computational and memory demanding [33]. Indeed, Mehta *et al.* factorize volumetric heatmaps into three 2D heatmaps to lower computational costs [38]. The authors train a deep learning model (VNect) to predict x, y, and z axes as dense

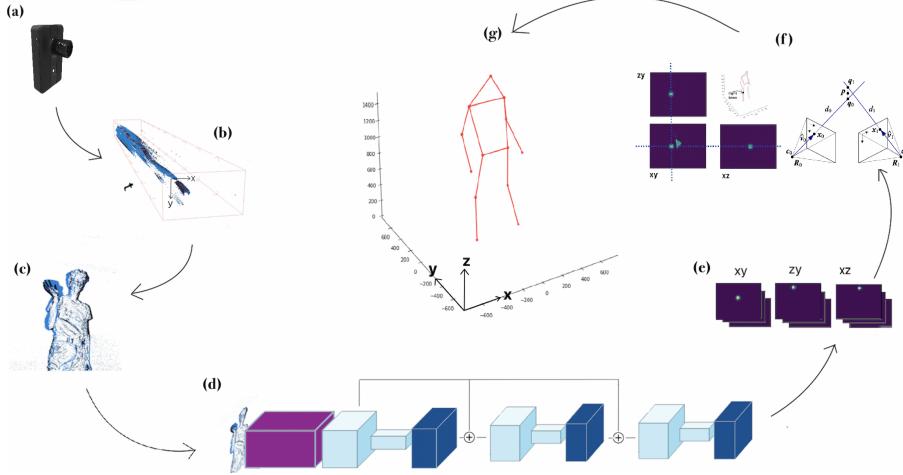


Figure 2: (a) A moving subject is recorded with an event-camera. (b) The recording is an asynchronous train of events; each event is characterized by an image plane coordinate (x , y), a timestamp (t), and a positive or negative polarity (respectively, blue and black in the figure). (c) batches of events are accumulated to build frames. Our model processes frames of events (d) in multiple stages. The model output are three set of independent planes; each subject’s joint (i.e., head, left and right wrists, and so on) is onto three independent planes(e). Next, it triangulates the planar predictions (f) and estimates the position of each joint. The output (g) is the subject’s skeleton in tridimensional coordinate.

2D heatmaps and combine the predictions through triangulation. The computational and resources savings of VNect come with a price in terms of accuracy, as this method reaches higher Mean Per-Joint Precision Error (MPJPE) on common benchmarks. Nibali *et al.* develop this approach further and propose a model (Margipose) to predict xy , zy , and xz heatmaps and regress the final 3D pose [43]. Others advancements in 3D human pose estimation include GANs [7] and temporal convolutions [9].

Event-based approaches. Real-time applications require a careful design to meet strong computational, speed, and energy requirements. This premise is especially true when fast-moving human subjects are involved, such as in sport assistance and virtual reality. Monocular solutions involving RGB-D sensors [59] and high-speed cameras [26] have been explored, although they cannot meet the computational requirements of real-time applications. On the other hand, event-cameras achieve high recording speed without saturating bandwidth and resources. For these reasons, human pose estimation performed with event-cameras is both interesting and challenging for the community. Approaches that extrapolate information from single events would be ideal, as these methods allow to exploit the interesting advantages of event-cameras. Initial proposals leveraged event-cameras to match events with known objects in the scene [54, 23]. Rebecq *et al.* exploit a similar caveat [45] to predict semi-dense 3D structure of a scene. More recently, Xu *et al.* [57] employ events to (1) track features across frames and (2) enhance the intensity outputs of a DAVIS camera. Next, they predict human-poses

with VNect [38] and Openpose [6] models and optimize a multi-step optimization scheme to refine the prediction. Despite its efficiency, their approach relies on a heavy pre-processing phase to extract 3D mesh of the subjects and involves multiple components, each with its own hyperparameters. Instead of processing events in small batches, numerous works accumulate events into tensors representations, conducting events in the realm of synchronous deep learning models [12]. To predict human poses through event-cameras, previous works aggregate events to predict 2D poses from multiple point of views and finally triangulate subjects’ 3D poses [5, 2]. On the other hand, our approach is the first attempt to estimate 3D human pose based on a single DVS camera. We prove that human pose estimation from event-only DVS camera is feasible. For an in-depth discussion of event-cameras and their applications, we refer to the excellent event-cameras summary [12].

Datasets for event-based Human Pose Estimation.

Few datasets have been recorded using event-cameras, especially if compared with the huge amount of RGB datasets. For human pose estimation, Calabrese *et al.* released DHP19, a dataset with recordings of 17 subjects and 33 movements. On the other hand, simulating events from RGB images is a promising path of research, especially since multiple works proved the soundness of training on simulated events. Mueggler et al. [40] generate synthetic events from RGB images and compare real and synthetic events for ego-pose estimation in various scenarios. More recently, simulated events have been employed for image reconstruction [47], depth estimation [14], and motion seg-

mentation [53], especially in high-speed scenarios where RGB ground-truth are hard to collect. In this work, we propose a pipeline to generate synthetic events from the Human3.6m dataset [20, 21] and compare our approach with standard RGB methods to establish a strong benchmark for further research.

3. Method

Our goal in this paper is to fill the gap between RGB-based and event-based monocular human pose estimation. In particular, we propose an end-to-end pipeline to predict the skeleton of a subject from the stream of a single event-camera. Figure 2 provides an overview of our methodology. An event-camera collects an asynchronous stream of events of a subject moving in the scene. Instead of tracking events as previous works [57], we aggregate them into tensor-like frames. Next, we predict three heatmap planes of the cuboid surrounding the subject and finally build his final 3D pose through triangulation.

Events. Event-cameras have peculiar pixel sensors that capture information asynchronously. In particular, event-cameras have no central clock; each pixel senses the light variations of the scene independently according to

$$\begin{aligned} \Delta L(x_k, t_k) &> p_k C, \text{ where} \\ \Delta L(x_k, t_k) &\doteq L(x_k, t_k) - L(x_k, t_k - \Delta t_k), \end{aligned} \quad (1)$$

where at each pixel x_k we compute the difference in light intensity $\Delta L(x_k, t_k)$ between the current and previous time instance every Δt_k seconds. If this difference exceeds a fixed threshold C , the pixel emits an event. An event-camera stream is a sequence of events, each characterized by the image coordinate pair (x, y) , a polarity (related to a positive or a negative change of intensity), and a timestamp.

Events aggregation. Instead of relying on raw asynchronous events, recent literature has shifted toward aggregating events together to build synchronous events representation. Common approaches range from simply integrating batch of events (constant-count) to representations involving stochastic modelling of events [52] and temporal sparsity [62]. As temporal information is critical in 3D human pose estimation [8], our first question is to understand if 3D Human Pose Estimation benefits from specific spatio-temporal representations. To provide an answer, we compare constant-count representation with spatio-temporal voxel grids [62]. While constant-count simply aggregates a constant number of events into an image, spatio-temporal voxel-grid preserves the timestamp contribution of events by building B temporal bins and have been already adopted in image reconstruction [47, 49] and depth estimation [14]. Given a set of N events $\{(x_k, t_k, p_k)\}_{k=0\dots N}$, we compute t_k^* as the normalized timestamp of event k into range $[0, B - 1]$. Each event (x_k, t_k, p_k) contribute to each bin B

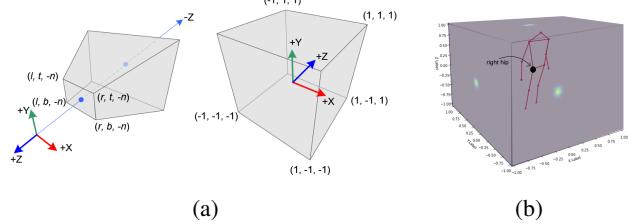


Figure 3: (a) We define the canonical 3D skeleton pose into a normalized cube [36, 1] and reproject the cube into the camera image plane using camera calibration parameters. (b) For each joint, our method extracts the three orthogonal faces of the cube to generate three *marginal heatmaps*.

of voxel V proportionally to its normalized timestamp t_k^* , as:

$$V(x, t) = \sum_{k=0}^N p_k \max(0, 1 - |t - t_k^*|), \quad (2)$$

where $t_k^* \doteq \frac{B-1}{t_N - t_0}$.

We set $N = 7500$ for both representations and $B = 4$ for spatio-temporal voxel-grid.

Skeleton normalization and projection. Instead of regressing 3D joints directly, our method relies, as a proxy, on their 2D projections onto specific planes [43]. We generate ground-truth as follows. First, we project the coordinates p_{xyz} of a joint on a plane parallel to the image plane and placed at depth z_{ref} (we adopt the z value of the *head joint* as z_{ref}). After that, we map the space to a normalized cube p_{xyz}^{NDC} (Normalized Device Coordinate - NDC [36, 1]): the three coordinates assume values in the range [-1, 1], as in Figure 3a. Last, we project p_{xyz}^{NDC} onto the three orthogonal faces of the cube and blur the projection on each face with a Gaussian Filter to generate ground-truth *marginal heatmaps* H_{xy} , H_{zy} and H_{xz} (Figure 3b).

Predicting marginal heatmaps. We design our approach upon marginal heatmaps [37, 43] and first predict three 2D heatmaps from our monocular input. Figure 4 summarizes our model. We first process the event-frame input with a backbone to extract intermediate representations. In particular, we adopt a ResNet-34 [17] which is cut after the second residual block. The feature extractor initialization is a critical design choice of our approach and we experimentally ablate possible alternatives in Section 4.3, where we compare different initialization and pre-training strategies and provide evidence of the benefits of RGB-to-events transfer learning. The main model consists in three branches, one for each marginal projection (xy , zy , and xz). Each branch is further made of three stages (Figure 4(a)), each consisting in a hourglass-like CNN, as de-

tailed in Figure 4(b). For each stage we compute an intermediate loss. The result of each stage is also aggregated (summation) with the previous output to feed the next stage, in a residual-like fashion. According to [42], intermediate losses help alleviating the problem of vanishing gradients.

Aggregating marginal heatmaps. Our model is trained jointly to predict the intermediate heatmaps as well as the normalized skeletal coordinates. We apply the soft-argmax operator [41] to extract the normalized coordinates of each joints onto the xy , xz , and yz planes. We choose the predictions from the xy -plane for the xy coordinate of the final prediction \hat{p}_{xyz} , as they match naturally with the input image. For z , we average the zy and xz predictions. Eq. 3 summarizes these steps as:

$$\begin{aligned} \hat{H}_{xy}^i, H_{xz}^i, \hat{H}_{yz}^i &= \text{Model}(\mathbf{x}) \\ [x_{xy}^i, y_{xy}^i] &= \text{soft-argmax}(\hat{H}_{xy}^i) \\ [x_{xz}^i, z_{xz}^i] &= \text{soft-argmax}(\hat{H}_{xz}^i) \\ [y_{zy}^i, z_{zy}^i] &= \text{soft-argmax}(\hat{H}_{yz}^i) \\ \hat{p}_{xyz}^i &= \left[x_{xy}^i, y_{xy}^i, \frac{z_{xz}^i + z_{zy}^i}{2} \right]. \end{aligned} \quad (3)$$

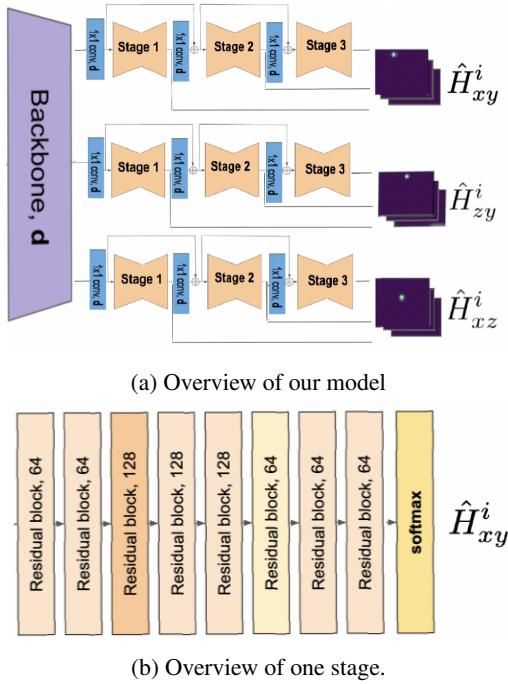


Figure 4: (a) We process event-frames with a **backbone** that outputs features of depth \mathbf{d} . Next, we adopt three sequential stages to output $3 \times J$ intermediate heatmaps. We apply an intermediate loss to each stage [41] and accumulate the losses to solve the vanish gradient problem. (b) Each stage process its input with three deep Convolutional Neural Network through an auto-encoder architecture.

Losses. As the full pipeline is differentiable, we can back-propagate the geometrical error between joints predictions and ground-truths and train our model end-to-end. Moreover, we can interpret marginal heatmaps as probability distributions of joints locations. In this framework, we apply the Jensen–Shannon divergence (Equation 4) between predicted heatmaps \hat{H}^i for stage i and ground-truth heatmaps H . JSD is based on the Kullbeck-Leibler divergence (KL), it is symmetric and has only finite values given by:

$$\text{JSD}(H, \hat{H}) = \frac{1}{2} \text{KL}(H\| \hat{H}) + \frac{1}{2} \text{KL}(\hat{H}\| H). \quad (4)$$

The Jensen-Shannon divergence and the geometrical loss for each stage i are aggregated into the final loss L as:

$$L = \sum_i L_{\text{geometrical}}(\hat{p}_{xyz}^i, p_{xyz}) + \text{JSD}(H_{xy}, \hat{H}_{xy}^i) + \text{JSD}(H_{xz}, \hat{H}_{xz}^i) + \text{JSD}(H_{zy}, \hat{H}_{zy}^i), \quad (5)$$

where $L_{\text{geometrical}}(\hat{p}_{xyz}^i, p_{xyz}) = \|\hat{p}_{xyz}^i - p_{xyz}\|_2$.

4. Experiments

We test our approach on our novel Event-Human3.6m dataset and provide extensive comparison to support our claims. Moreover, we experiment on real events from the event-based DHP19 dataset. For both the dataset, we address the scale-depth ambiguity using a ground-truth depth point and calculate the Mean Per-Joint Precision Error (MPJPE) between the de-normalized predictions and the ground-truths [43, 33].

4.1. Datasets

DHP19 dataset. DHP19 [5] contains 33 recordings of 17 subjects of different sex, age, and size. Each subject is recorded with four DVS cameras from different angles. Nevertheless, the range of movements in the recordings is narrow. Most of the activities, such as *legs kicking* and *arms abductions*, are conducted on the spot, with the exception of *slow jogging* and *walking*. Moreover, few recordings spot real life activities. These gaps in the data limit its applications in real scenarios.

Event-Human3.6m dataset. In the previous section we highlight some limitations of the DHP19 dataset [5], especially related to the narrowness of movements and activities that it provides. To solve these gaps, we contribute with a new simulated datasets based on the Human3.6m dataset [20, 21]. Human3.6 recordings include 11 subjects and different activities from real scenarios, such as *walking with a dog*, *talking at the phone*, and *giving directions*. Consequently, extensive research has adopted the standard Human3.6m dataset to evaluate monocular Human Pose Estimation methods [43, 38, 44]. We believe event-based re-

search will benefit from our Event-Human3.6m, as it extends DHP19 with more challenging scenarios and provides a new benchmark for monocular human pose estimation algorithms. We adopt the ESIM-Py simulator [46] to convert the RGB recordings of Human3.6m into events and synchronize raw joints ground-truth with events frames through interpolation (Figure 5). As a result, Event-Human3.6m and DHP19 have comparable ground-truths and event frames. In the following sections we report extensive experiments on both DHP19 and Event-Human3.6m to test the benefits of our proposal.

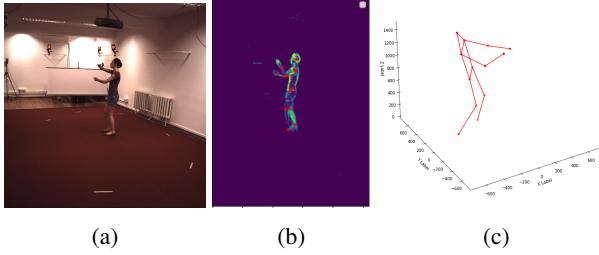


Figure 5: We simulate raw events from Human3.6m recordings (a) with the open-source simulator ESIM-Py [46]. We set the simulator’s parameters $cp = cn = 0.2$, $\log\text{-}\epsilon = 1e-3$, and $\text{refractory-period} = 1e-4$, as this setting produces synthetic events similar to DHP19 event-camera recordings. Next, we accumulate events into event-frames (b) and interpolate ground-truths to match timestamps (c).

Training details. We explore different hyper-parameters settings empirically. In the following experiments, we train our method on 4 Tesla V100 16Gb GPUs and adopt a batch size (per GPU) of 32. For updating the gradients, we opt for Adam optimizer [24] with learning rate of 0.0003. We interrupt the train at local convergence through an early-stopping strategy. We evaluate our approach with 1 stage (7M of parameters, 91 MB of storage) and 3 stages (21M parameters, 300MB of storage).

4.2. Results

Here we discuss the performance of our approach and validate it on the two datasets.

Evaluation on DHP19. We test 1-stage and 3-stage models with spatio-temporal voxel-grid and constant-count representations. Table 1 reports the Mean Per Joint Precision Error (MPJPE, in mm) and summarizes the results. As reference, we compare to the stereo approach of Calabrese *et al.* [5]. As a first observation, our methodology performs only slightly worse than the stereo approach (13mm difference). In this setting, constant-count representation performs better than voxel-grid. In the ablations, we elaborate on the differences between the two representations when different backbones are adopted as feature extractors.

Moreover, we provide results for our single stage and 3-stages model and compare them. Table 1 shows that multiple stacked stages and intermediate losses provide sensible performance benefits, at the cost of an increase in computational costs and model size.

Table 1: Results refer to DHP19 dataset [5]. We compare our approach with 1 and 3 stack of stages across constant-count and voxel-grid representation.

Method	input	MPJPE(mm)
Calabrese <i>et al.</i> [5]	stereo	79.63
Constant-count – stage 3	monocular	92.09
Voxel-grid – stage 3	monocular	95.51
Constant-count – stage 1	monocular	96.69
Voxel-grid – stage 1	monocular	105.24

Evaluation on Event-Human3.6m. For each subject, we keep 13 out of the 32 provided joints to build skeletons that are compatible with DHP19 ground-truths and evaluate our approach on a cross-subject protocol. We train our models on subjects 1, 3, 5, 7, 8 and test on subjects 9 and 11. Similar works [43, 33, 44] evaluate monocular approaches on every 64th frame of the recordings. We adapt this evaluation protocol to our asynchronous Event-Human3.6m by taking event-frames corresponding to the same testing frames. Table 2 reports the results of our approach with constant-count and voxel-grid representations. Moreover, we compare our methodology to state-of-the-art RGB approaches [22, 43, 44, 33]. Despite the gap with standard computer-vision techniques, our approach performs fairly against existing RGB approaches.

Table 2: Comparison between RGB approaches on Human3.6m and our approach on its synthetic counterpart. We adopt a standard cross-subject protocol to validate on the same testing strategy as RGB approaches.

Method	input	MPJPE(mm)
Metha <i>et al.</i> [38] (ResNet-50)	RGB	80.50
Kanazawa <i>et al.</i> [22]	RGB	88.00
Nibali <i>et al.</i> [43]	RGB	57.00
Pavlakos <i>et al.</i> [44]	RGB	71.90
Luvizon <i>et al.</i> [33]	RGB	53.20
Cheng <i>et al.</i> [9]	RGB	40.10
Spatio-temporal voxel-grid (Ours)	Events	119.18
Constant-count (Ours)	Events	116.40

4.3. Ablation study

In this Section, we deepen different aspects of our approach in more detail. In particular, we are interested to explore *what movements cause our approach to fail and how*

backbone initialization impacts performance. In the following, we discuss these questions in more details.

Transfer learning and pre-training tasks. Event representations and RGB images share some commonalities, especially edges and corners. However, if we compare them closely, we find subtle differences, since event-cameras recordings are highly correlated to the dynamic of the scene. If the RGB/event-frames analogy held, event-based vision could benefit widely from advancements in standard computer vision. As an example, recent computer vision research provides strong evidence in support of transfer learning from large dataset, e.g., the ImageNet dataset [11, 19]. Further works explore and validate the correlation between 3D Human Pose Estimation and reconstruction tasks [60]. These insights are supported by common intuition, as both tasks involve an understanding of the structure of the scene. Despite the differences between event and standard cameras, recent works validate the transfer learning hypothesis from RGB to constant-count representation [35] and learnable representations [13]. Moreover, Rebecq *et al.* provide evidence for direct transfer learning by predicting natural images from spatio-temporal event-frames [47].

Our work contributes further to this line of research with two evaluations. First, we compare ImageNet and random initialized models for solving monocular human pose estimation with both constant-count and voxel-grid representations. Second, we attempt to validate if different pre-training tasks help with event-based Human Pose Estimation. For this purpose, we train an auto-encoder consisting of a ResNet-34 as encoder and a small DeconvCNN as decoder. For comparison, we train a ResNet-34 and a ResNet-50 CNN on action recognition task, which has lower correlation with human pose estimation. Next, we test our approach with 4 backbones (random-initialized, action recognition task, reconstruction task, and ImageNet initialized) and compare the results on DHP19 dataset. Table 3 reports the MPJPE for both constant-count and voxel-grid representations. Constant-count frames benefit more from standard computer vision, especially from ImageNet transfer-learning. In fact, our model with ImageNet-pretrained ResNet34 outperforms all others approaches when we adopt constant-count representation.

Spatio-temporal frames have few similarities with standard RGB images; in fact, it is unclear if this approach can benefit from ImageNet transfer learning. Our experiments reflects these differences, as ImageNet pretrained ResNet-34 and ResNet-50 backbones have lower performance than the random-initialised counterpart.

We discuss Table 3 to explore further if recent research in pre-training tasks [60] is valid in event-based vision. Despite the correlations evidences in RGB settings, we find that auto-encoders backbones are performing worse than the classification counterpart; this conclusion is valid from both

representations. Indeed, action-recognition pre-training emerges favorably, especially for spatio-temporal voxel-grid. Our interpretation is that pre-training assumptions fail because of the spatial sparsity of event-representations. Further research is mandatory to unlock better pre-training strategies for event-based vision.

Table 3: We report the Mean Per Joint Precision Error (MPJPE, in mm) of our 3-stages approach equipped with different initialization strategies. ResNet-34 with ImageNet initialization emerges favorably for constant-count representation. Moreover, we find no benefits in adopting a reconstruction task as pre-training task, although standard computer vision research suggests the opposite[60].

Repr.	Model	Initialization	MPJPE (mm)
constant-count	ResNet-34	Random initialized	92.22
		Action recognition	95.19
		Reconstruction	98.89
		ImageNet	92.09
voxel-grid	ResNet-50	Random initialized	92.22
		Action recognition	92.26
		ImageNet	92.51
		Random initialized	93.06
	ResNet-34	Action recognition	95.26
		Reconstruction	105.44
		ImageNet	95.51
		Random initialized	93.88
	ResNet-50	Action recognition	93.54
		ImageNet	93.98

Per-movements comparison. Events are highly coupled with the dynamic of the scene. If parts of the body are static, fewer events are recorded. As a consequence, spatial sparsity increases and makes prediction tasks more challenging. To evaluate the impact of static body parts on our approach, we propose a per-movements study for our ImageNet-pretrained method. Table 4 compares our constant-count and spatio-temporal voxel-grid approaches with DHP19 [5] event-based stereo approach. Differently from [5], our approach is based upon the more recent state of the art solutions [42, 43] and reaches a higher per-movement accuracy and lower per-movement standard deviation. As expected, performance decreases when subjects perform movements with only parts of the body (e.g., *Punch up forwards left* implies static legs). This drop in performance matches the results of the stereo-vision approach (e.g., *Punch forwards left/right*). On the other hand, we notice above average performances for movements that involve the whole body, such as *knee lift* and *hand movements* (during these movements, subjects move on the spot and the whole body generates events).

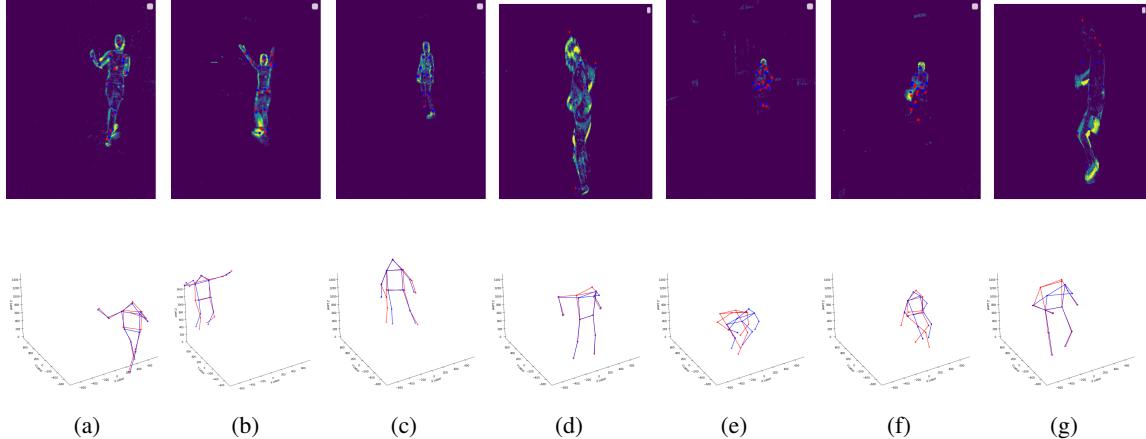


Figure 6: Our approach achieves good performance when subjects are actively moving, as in (a)–(d), but fails to predict the skeletons satisfactorily when some parts of the body remain static during the movements, as in (e)–(g).

Table 4: We compare the per-movement MPJPE between ours and DHP19 [5] stereo approach. Both fail when parts of the body are static and shine when the scene is more dynamic. In bold we highlight **worst results per column** while with an underline we show best results per column.

	Stereo [5]	Voxel-grid	Constant-count
Left arm abduction	115.04	82.32	80.41
Right arm abduction	99.65	81.92	79.68
Left leg abduction	84.65	110.07	105.39
Right leg abduction	78.35	99.87	93.81
Left arm bicep curl	103.29	90.49	86.40
Right arm bicep curl	121.06	80.75	95.73
Left leg knee lift	74.97	<u>71.60</u>	<u>72.14</u>
Right leg knee lift	71.95	78.47	<u>72.49</u>
Walking 3.5 km/h	58.75	86.88	84.74
Single jump up-down	82.23	80.11	76.73
Single jump forwards	80.53	89.92	85.10
Multiple jumps	<u>53.57</u>	99.47	93.83
Hop right foot	<u>55.56</u>	89.51	84.16
Hop left foot	<u>54.21</u>	97.86	91.60
Punch forward left	148.57	114.97	117.87
Punch forward right	135.92	98.35	93.69
Punch up forwards left	111.35	124.89	124.81
Punch up forwards right	131.46	103.01	106.56
Punch down forwards left	106.92	105.98	105.04
Punch down forwards right	98.28	90.02	89.90
Slow jogging	<u>55.16</u>	98.05	89.11
Star jumps	76.23	108.89	106.77
Kick forwards left	111.66	117.92	93.07
Kick forwards right	112.49	117.91	109.85
Side kick forwards left	118.00	128.38	120.39
Side kick forwards right	104.67	115.76	111.86
Hello left hand	96.22	89.08	87.22
Hello right hand	101.32	<u>71.82</u>	<u>69.83</u>
Circle left hand	110.59	99.17	95.89
Circle right hand	112.44	84.00	76.55
Figure-8 left hand	110.69	90.95	88.10
Figure-8 right hand	123.59	<u>72.42</u>	<u>72.49</u>
Clap	122.93	<u>81.03</u>	77.77
Mean (standard deviation)	98.06 (± 16.60)	95.51 (± 15.30)	92.09 (± 14.49)

5. Discussion and Conclusions

We have proposed a deep learning approach for event-based human pose estimation from a single event-camera. Our method aggregates events into synchronous tensor representations to feed a multi-stage Convolutional Neural Network. Our architecture predicts three orthogonal heatmaps which are triangulated to obtain the final 3D pose. We validated our approach on the event-based DHP19 dataset, where it showed satisfactory per-movement performance against DHP19 stereo approach [5]. Moreover, we proposed Event-Human3.6m, a new dataset of simulated events from the standard Human3.6m [21]. Event-Human3.6m extends DHP19 with more challenging movements and actions. We conducted experiments on the synthetic dataset and adopted a cross-subject protocol which is comparable to the standard RGB testing. Although we recognize the differences between synthetic and RGB datasets, our proposal achieved an accuracy comparable to RGB approaches. These experiments demonstrated the effectiveness of our method.

Figure 6 reports challenging examples where our method underperforms. Static parts of the body generated fewer events and are difficult to predict accurately. We leave this issue for further investigations. Next, we conducted extensive ablations studies to understand how event-based vision can benefit from RGB transfer learning and pre-training. Experiments showed that ImageNet pre-training boosts our approach more than pre-training tasks. Moreover, action recognition pre-training task archived higher performances than reconstruction pre-training, although extensive computer vision research suggests the opposite. Future research should consider closely the relationships between events and RGB cameras in transfer-learning and multi-task learning settings. Further works to answer these open questions can benefit from our synthetic Event-Human3.6m.

References

- [1] Song Ho Ahn. OpenGL projection onto frustum space. 4
- [2] R. Wes Baldwin, Ruixu Liu, Mohammed Almatrafi, Vijayan Asari, and Keigo Hirakawa. Time-ordered recent event (tore) volumes for event cameras, 2021. 3
- [3] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In *European conference on computer vision*, pages 561–578. Springer, 2016. 2
- [4] C. Brandli, R. Berner, M. Yang, S.-C. Liu, and T. Delbruck. A 240x180 130db 3 μ s latency global shutter spatiotemporal vision sensor. *IEEE Journal of Solid-State Circuits*, 49(10):2333–2341, 2014. 1
- [5] Enrico Calabrese, Gemma Taverni, Christopher Awai Easthope, Sophie Skribine, Federico Corradi, Luca Longinotti, Kynan Eng, and Tobi Delbruck. Dhp19: Dynamic vision sensor 3d human pose dataset. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019. 2, 3, 5, 6, 7, 8
- [6] Z. Cao, G. Hidalgo, T. Simon, S. E. Wei, and Y. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(1):172–186, 2021. 3
- [7] Ching-Hang Chen, Ambrish Tyagi, Amit Agrawal, Dylan Drover, Rohith MV, Stefan Stojanov, and James M. Rehg. Unsupervised 3d pose estimation with geometric self-supervision. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jun 2019. 3
- [8] Yucheng Chen, Yingli Tian, and Mingyi He. Monocular human pose estimation: a survey of deep learning-based methods. *Computer Vision and Image Understanding*, 192(nil):102897, 2020. 4
- [9] Yu Cheng, Bo Yang, Bo Wang, and Robby T. Tan. 3d human pose estimation using spatio-temporal networks with explicit occlusion training. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07):10631–10638, apr 2020. 3, 6
- [10] Andrew I Comport, Éric Marchand, and François Chaumette. A real-time tracker for markerless augmented reality. In *The Second IEEE and ACM International Symposium on Mixed and Augmented Reality, 2003. Proceedings.*, pages 36–45. IEEE, 2003. 1, 2
- [11] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009. 7
- [12] Guillermo Gallego, Tobi Delbruck, Garrick Michael Orchard, Chiara Bartolozzi, Brian Taba, Andrea Censi, Stefan Leutenegger, Andrew Davison, Jorg Conradt, Kostas Daniilidis, and et al. Event-based vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, page 1–1, 2020. 3
- [13] Daniel Gehrig, Antonio Loquercio, Konstantinos G Derpanis, and Davide Scaramuzza. End-to-end learning of representations for asynchronous event-based data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5633–5643, 2019. 2, 7
- [14] Daniel Gehrig, Michelle Rüegg, Mathias Gehrig, Javier Hidalgo Carrio, and Davide Scaramuzza. Combining events and frames using recurrent asynchronous multimodal networks for monocular depth prediction, 2021. 3, 4
- [15] I. Hasan, F. Setti, T. Tsesmelis, V. Belagiannis, S. Amin, A. Del Bue, M. Cristani, and F. Galasso. Forecasting people trajectories and head poses by jointly reasoning on tracklets and vislets. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(4):1267–1278, 2021. 1
- [16] Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J Black. Resolving 3d human pose ambiguities with 3d scene constraints. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2282–2292, 2019. 2
- [17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015. 4
- [18] Yinghao Huang, Federica Bogo, Christoph Lassner, Angjoo Kanazawa, Peter V Gehler, Javier Romero, Ijaz Akhter, and Michael J Black. Towards accurate marker-less human shape and pose estimation over time. In *2017 international conference on 3D vision (3DV)*, pages 421–430. IEEE, 2017. 2
- [19] Mi-Young Huh, Pulkit Agrawal, and Alexei A. Efros. What makes imagenet good for transfer learning? *CoRR*, abs/1608.08614, 2016. 7
- [20] Catalin Ionescu, Fuxin Li, and Cristian Sminchisescu. Latent structured models for human pose estimation. In *2011 International Conference on Computer Vision*, page nil, 11 2011. 4, 5
- [21] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, 2014. 2, 4, 5, 8
- [22] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, jun 2018. 6
- [23] Hanme Kim, Stefan Leutenegger, and Andrew J. Davison. Real-time 3d reconstruction and 6-dof tracking with an event camera. *Lecture Notes in Computer Science*, page 349–364, 2016. 3
- [24] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017. 6
- [25] Mark L Knapp, Judith A Hall, and Terrence G Horgan. *Nonverbal communication in human interaction*. Cengage Learning, 2013. 1
- [26] Adarsh Kowdle, Christoph Rhemann, Sean Fanello, Andrea Tagliasacchi, Jonathan Taylor, Philip Davidson, Mingsong Dou, Kaiwen Guo, Cem Keskin, Sameh Khamis, David Kim, Danhang Tang, Vladimir Tankovich, Julien Valentin, and Shahram Izadi. The need 4 speed in real-time dense visual tracking. *ACM Transactions on Graphics*, 37(6):1–14, jan 2019. 3
- [27] Xavier Lagorce, Garrick Orchard, Francesco Galluppi, Bertram E. Shi, and Ryad B. Benosman. HOTS: A Hierarchy of Event-Based Time-Surfaces for Pattern Recognition.

- IEEE Transactions on Pattern Analysis and Machine Intelligence*, 126(7):1346–1359, 7 2017. 2
- [28] Bo Li, Huahui Chen, Yucheng Chen, Yuchao Dai, and Mingyi He. Skeleton boxes: Solving skeleton based action detection with a single deep convolutional neural network. In *2017 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, page nil, 7 2017. 1
- [29] Bo Li, Yuchao Dai, Xuelian Cheng, Huahui Chen, Yi Lin, and Mingyi He. Skeleton based action recognition using translation-scale invariant image mapping and multi-scale deep cnn. In *2017 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, page nil, 7 2017. 1
- [30] Bo Li, Mingyi He, Yuchao Dai, Xuelian Cheng, and Yucheng Chen. 3d skeleton based action recognition by video-domain translation-scale invariant mapping and multi-scale dilated cnn. *Multimedia Tools and Applications*, 77(17):22901–22921, 2018. 1
- [31] Patrick Lichtsteiner, Christoph Posch, and Tobi Delbrück. A $1288 \times 1288 \times 120$ dB 15 μ s latency asynchronous temporal contrast vision sensor. *IEEE Journal of Solid-State Circuits*, 43(2):566–576, 2008. 1
- [32] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. Smpl. *ACM Transactions on Graphics*, 34(6):1–16, 2015. 2
- [33] Diogo C. Luvizou, David Picard, and Hedi Tabia. 2d/3d pose estimation and action recognition using multitask deep learning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 2, 5, 6
- [34] Misha A. Mahowald and Carver Mead. The silicon retina. *Scientific American*, 264(5):76–82, may 1991. 1
- [35] Ana I. Maqueda, Antonio Loquercio, Guillermo Gallego, Narciso Garcia, and Davide Scaramuzza. Event-based vision meets deep learning on steering prediction for self-driving cars. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun 2018. 7
- [36] TOM McREYNOLDS and DAVID BLYTHE. Chapter 2 - 3d transformations. In TOM McREYNOLDS and DAVID BLYTHE, editors, *Advanced Graphics Programming Using OpenGL*, The Morgan Kaufmann Series in Computer Graphics, pages 19–34. Morgan Kaufmann, San Francisco, 2005. 4
- [37] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3d human pose estimation in the wild using improved cnn supervision. In *2017 International Conference on 3D Vision (3DV)*, page nil, 10 2017. 4
- [38] Dushyant Mehta, Srinath Sridhar, Oleksandr Sotnychenko, Helge Rhodin, Mo-Hammad Shafiei, Hans-Peter Seidel, Weipeng Xu, Dan Casas, Christian Theobalt, and Rey Juan Carlos. VNect: Real-time 3D Human Pose Estimation with a Single RGB Camera. Technical report, 2017. 2, 3, 5, 6
- [39] Pierre Merriaux, Yohan Dupuis, Rémi Boutteau, Pascal Vasseur, and Xavier Savatier. A study of vicon system positioning performance. *Sensors*, 17(7):1591, 2017. 2
- [40] Elias Mueggler, Henri Rebecq, Guillermo Gallego, Tobi Delbrück, and Davide Scaramuzza. The event-camera dataset and simulator: Event-based data for pose estimation, visual odometry, and slam. *The International Journal of Robotics Research*, 36(2):142–149, 2017. 3
- [41] Daniel Neil, Michael Pfeiffer, and Shih-Chii Liu. Phased LSTM: Accelerating Recurrent Network Training for Long or Event-based Sequences. 10 2016. 5
- [42] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked Hourglass Networks for Human Pose Estimation. 3 2016. 2, 5, 7
- [43] Aiden Nibali, Zhen He, Stuart Morgan, and Luke Pendlagast. 3D Human Pose Estimation with 2D Marginal Heatmaps. 6 2018. 2, 3, 4, 5, 6, 7
- [44] Georgios Pavlakos, Xiaowei Zhou, Konstantinos G. Derpanis, and Kostas Daniilidis. Coarse-to-fine volumetric prediction for single-image 3d human pose. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, page nil, 7 2017. 2, 5, 6
- [45] Henri Rebecq, Guillermo Gallego, Elias Mueggler, and Davide Scaramuzza. EMVS: Event-Based Multi-View Stereo—3D Reconstruction with an Event Camera in Real-Time. *International Journal of Computer Vision*, 126(12):1394–1414, 12 2018. 3
- [46] Henri Rebecq, Daniel Gehrig, and Davide Scaramuzza. ESIM: an Open Event Camera Simulator. Technical report, 2018. 2, 6
- [47] Henri Rebecq, Rene Ranftl, Vladlen Koltun, and Davide Scaramuzza. High speed and high dynamic range video with an event camera. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, nil(nil):1–1, 2020. 2, 3, 4, 7
- [48] Bodo Rosenhahn, Christian Schmalz, Thomas Brox, Joachim Weickert, Daniel Cremers, and Hans-Peter Seidel. Markerless motion capture of man-machine interaction. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2008. 1, 2
- [49] Cedric Scheerlinck, Henri Rebecq, Daniel Gehrig, Nick Barnes, Robert Mahony, and Davide Scaramuzza. Fast image reconstruction with an event camera. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, March 2020. 4
- [50] Albert E Schefflen. Body language and the social order; communication as behavioral control. 1972. 1
- [51] Ashish Shingade and Archana Ghotkar. Animation of 3d human model using markerless motion capture applied to sports. *arXiv preprint arXiv:1402.2363*, 2014. 1, 2
- [52] Amos Sironi, Manuele Brambilla, Nicolas Bourdis, Xavier Lagorce, and Ryad Benosman. HATS: Histograms of Averaged Time Surfaces for Robust Event-based Object Classification. 3 2018. 2, 4
- [53] Timo Stoffregen, Guillermo Gallego, Tom Drummond, Lindsay Kleeman, and Davide Scaramuzza. Event-Based Motion Segmentation by Motion Compensation. Technical report. 4
- [54] David Reverter Valeiras, Garrick Orchard, Sio-Hoi Ieng, and Ryad B. Benosman. Neuromorphic event-based 3d pose estimation. *Frontiers in Neuroscience*, 9(nil):nil, 2016. 2, 3
- [55] Vicon. Motion capture systems, 2020. 1, 2
- [56] Jianbo Wang, Kai Qiu, Houwen Peng, Jianlong Fu, and Jianke Zhu. Ai coach: Deep human pose estimation and

- analysis for personalized athletic training assistance. In *Proceedings of the 27th ACM International Conference on Multimedia*, MM '19, page 374–382, New York, NY, USA, 2019. Association for Computing Machinery. 1
- [57] Lan Xu, Weipeng Xu, Vladislav Golyanik, Marc Habermann, Lu Fang, and Christian Theobalt. Eventcap: Monocular 3d capture of high-speed human motions using an event camera. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page nil, 6 2020. 2, 3, 4
 - [58] Weipeng Xu, Avishek Chatterjee, Michael Zollhöfer, Helge Rhodin, Dushyant Mehta, Hans-Peter Seidel, and Christian Theobalt. Monoperfcap. *ACM Transactions on Graphics*, 37(2):1–15, 2018. 2
 - [59] Ming-Ze Yuan, Lin Gao, Hongbo Fu, and Shihong Xia. Temporal upsampling of depth maps using a hybrid camera. *IEEE Transactions on Visualization and Computer Graphics*, 25(3):1591–1602, 2019. 3
 - [60] Amir R. Zamir, Alexander Sax, William B. Shen, Leonidas J. Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling task transfer learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2018. 2, 7
 - [61] Semir Zeki. A massively asynchronous, parallel brain. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 370(1668):20140174, may 2015. 1
 - [62] Alex Zihao Zhu, Liangzhe Yuan, Kenneth Chaney, and Kostas Daniilidis. Unsupervised event-based optical flow using motion compensation. In Laura Leal-Taixé and Stefan Roth, editors, *Computer Vision – ECCV 2018 Workshops*, pages 711–714, Cham, 2019. Springer International Publishing. 2, 4