E3V-K5: An Authentic Benchmark for Redefining Video-Based Energy Expenditure Estimation

Shengxuming Zhang[1,4*], Lei Jin[2*], Yifan Wang[2], Xinyu Wang[2], Xu Wen[2], Zunlei Feng[3,4†], and Mingli Song[3,4]

[1] School of Software Technology, Zhejiang University, China
[2] Department of Sports Science, Zhejiang University, China
[3] College of Computer Science and Technology, Zhejiang University, China
[4] Key Laboratory of Visual Perception (Zhejiang University) , Ministry of Education and Microsoft

Abstract. Accurately estimating energy expenditure (EE) is crucial for optimizing athletic training, monitoring daily activity levels, and preventing sports-related injuries. Estimating energy expenditure based on video (E^3V) is an appealing research direction. This paper introduces E3V-K5, an authentic dataset of sports videos that significantly enhances the accuracy of EE estimation. The dataset comprises 16,526 video clips from various categories and intensity of sports with continuous calorie readings obtained from the COSMED K5 indirect calorimeter, recognized as the most reliable standard in sports research. Augmented with the heart rate and physical attributes of each subject, the volume, diversity, and authenticity of E3V-K5 surpass all previous video datasets in E^3V, making E3V-K5 a cornerstone in this field and facilitating future research. Furthermore, we propose E3SFormer, a novel approach specifically designed for the E3V-K5 dataset, focusing on EE estimation using human skeleton data. E3SFormer consists of two Transformer branches for simultaneous action recognition and EE regression. The attention of joints from the action recognition branch is utilized in assisting the EE regression branch. Extensive experimentation validates E3SFormer's effectiveness, demonstrating its superior performance to existing skeleton-based action recognition models. Our dataset and code are publicly available at https://github.com/zsxm1998/E3V.

1 Introduction

Regular physical activity (PA), especially habitual aerobic exercise with appropriate intensity and frequency, is beneficial to human health [18]. Physical inactivity may predispose to chronic diseases such as obesity, cardiovascular disease, or diabetes [3, 30, 37, 45], while prolonged exercise with high intensity often leads to sports injuries [49]. As one of the important physiological changes caused by exercise, energy expenditure (EE) can quantify the volume of exercise. Therefore,

* Equal contribution. † Corresponding author, email: zunleifeng@zju.edu.cn

accurately estimating EE is crucial for monitoring and controlling daily activity levels and scientific sports training [31], which has always been a concern in sports science [80].

Traditional methods for estimating actual EE include doubly labeled water (DLW) method [81], indirect calorimetry (IC) [82], and the use of wearable sensors such as heart rate monitors [35], accelerometers [11], etc. The reliability and validity of the first two methods are high, making the DLW and IC the "gold standard" for calorimetric measurements. However, the DLW method only measures total EE over time, and IC, requiring a stationary metabolic cart and mask, may interfere with natural behavior; both methods are expensive and limited in practicality [1,82]. Heart rate monitoring technology is relatively mature, but ensuring its accuracy can be challenging during high or low-intensity exercise [57]. Accelerometers are widely used in various physical activity studies due to their convenience and lower cost [11], but they usually need to be worn for a long time, which limits application scenarios, and their accuracy can be significantly influenced by the wearing positions and movement patterns [44]. With the rise of fitness tracking apps, many pioneering studies have focused on multiple signals obtained from wearable devices that combine physiological and biochemical indicators [10, 64, 70]. However, such sensors are not always readily available or comfortable to wear, which restricts their overall practicality.

In contrast, sports videos can be easily accessed and can accurately capture the movement of the entire body. From sports videos, we can obtain the kinematic parameters (such as velocity, acceleration, angle, etc.) to quantitatively describe any bodily movement and PA levels and then estimate EE. Thanks to the advancements in deep learning, there have been numerous remarkable visual works for predicting action types [6, 13, 21, 22, 24, 28, 41, 54, 60, 65–68, 74, 75], which has inspired us to estimate EE based on videos (E^3V). Currently, there is already some research delving into this. Some studies have demonstrated the effectiveness of video-based methods. In their study, Tao *et al.* [63] collected data in housework scenarios and employed an action-specific method to predict EE. Nakamura *et al.* [48] curated an egocentric video dataset along with heart rate and acceleration signals, and introduced a multi-modal method for predicting action categories and energy expenditure. Peng *et al.* [51] integrated multiple action recognition datasets and annotated videos with calorie estimates calculated by kinematic formulas.

However, contemporary studies in this field exhibit notable shortcomings. First, in the field of action recognition, existing datasets have constraints, such as deep learning models identifying actions from specific scenes in videos rather than focusing on human motion patterns. For precise estimation of energy expenditure, understanding the nuances of body movement patterns and their intensity is paramount. Second, existing datasets for estimating EE based on exercise videos either involve activities with limited intensity variation in restricted settings [63], or their EE labels are inaccurate and overlook individual differences due to their label generation methods [48, 51].

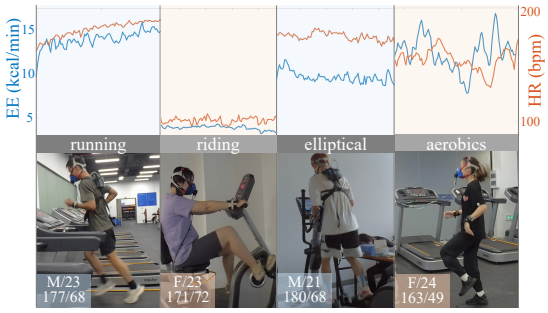

Fig. 1: Examples from E3V-K5 benchmark: sports videos with calorie readings from COSMED K5 (kcal/min, blue curve), heart rate (bpm, orange curve) and physical attributes of subjects in the box at the bottom left corner of each sample, including gender, age, height (cm), and weight (kg).

In this paper, we curate an authentic dataset called E3V-K5, which contains multiple sports videos with corresponding EE labels. For ground truth measurement, we employ the COSMED K5 indirect calorimeter, which relies on gas exchange. This method is widely recognized as the most reliable and universal standard in sports research [82]. This represents the inaugural sports video dataset with continuous calorie readings obtained from the K5, with the goal of constructing an authentic, diverse, and large-scale vision-EE benchmark. The dataset has been split into training, validation, and test sets based on the subjects. This cross-subject split allows us to evaluate whether the model has genuinely learned the relationship between motion and EE for various individuals. Some examples from our dataset are shown in the Figure 1.

In addition, we propose a method for estimating EE based on the human skeleton for the E3V-K5 dataset called the Energy Expenditure Estimation Skeleton Transformer (E3SFormer). First, we utilize an off-the-shelf pose estimation method [20] to extract the skeleton sequence of the exerciser from videos. Then, we input this skeleton sequence into a Spatio-Temporal Fusion Transformer backbone to extract features. These features are subsequently fed into two Transformer network branches, which are used for predicting the action category and energy expenditure, respectively. The features extracted by the backbone contain information in two dimensions: temporal dimension and spatial dimension (*i.e.* human joint dimension). Intuitively, we believe that the movement features of certain specific joints on the human body are key to action classification, and the motion intensity or characteristics of these joints over time have a greater correlation with EE. For instance, regardless of how the hands move, the running requires a rapid alternation of stepping forward and backward with both legs, so the EE of running is more related to legs than hands. Therefore, we transfer the attention of each joint from the action recognition branch to the EE regression branch to enhance its performance. Extensive experimentation was conducted to compare E3SFormer with many skeleton-based action recognition models, demonstrating the superiority of our method.

In conclusion, the main contributions of our work are summarized as follows:

- We curate the E3V-K5 dataset, featuring 16,526 video clips, significantly surpassing previous datasets in volume for EE estimation.
- E3V-K5 includes authentic EE measurements obtained from the COSMED K5, alongside comprehensive labels like heart rate and subjects' physical attributes, enabling more precise and comprehensive analyses of EE.
- We propose the E3SFormer that uses human skeleton data from videos for recognizing action categories and regressing EE. The importance of each joint for a specific action category, extracted from the action recognition branch, is utilized to boost the EE regression branch.
- Extensive experiments validate the challenging nature of E3V-K5 and the efficacy of E3SFormer, aiming to inspire further research based on this dataset.

2 Related Work

2.1 Deep Learning-Based Action Recognition

Video-Based Methods. In the domain of computer vision, action recognition from videos has become a pivotal area of research due to its applications in various fields such as intelligent surveillance and health care [53]. The most natural approach is to use video clips as input, training a neural network to predict the categories of actions involved in the video. Many works extend the architectural framework of 2D-CNNs to handle the comprehension of videos composed of time series of images. These extensions include transforming 2D-CNNs into 3D-CNNs in certain ways [6, 22, 33, 41, 66–68, 76], using optical flow as an additional input branch in dual-branch CNN networks and merging its features with RGB image features [6, 25, 26, 60, 72, 73, 78, 85], as well as employing some aggregation functions, such as averaging and RNNs, to fuse the sequences of features extracted from each frame using 2D-CNNs [13, 36, 85, 86]. In recent years, with the rise in prevalence of the Transformer [69] architecture, a substantial body of research [21, 24, 29, 40, 54, 55, 62, 65, 74, 75, 77, 79] has adopted self-supervised approaches to pre-train Vision Transformers [14] on large-scale datasets and fine-tune them on downstream tasks, achieving state-of-the-art results across multiple action recognition datasets.

Skeleton-Based Methods. Besides video, dynamic human skeleton also can be used for action recognition [84], which has demonstrated robustness and effectiveness in action recognition [71]. The skeleton data is typically acquired through the localization of 2D/3D body joints coordinates using depth sensors or readily available pose estimation algorithms [5]. Human skeleton can be perceived as a spatial-temporal graph of human body joints, thus, Yan *et al*. [84] proposed a method called ST-GCN that first applied GNN to skeleton-based action recognition. In the following years, numerous works were dedicated to enhancing the ST-GCN framework to enable more efficient capture of action representations within spatial-temporal graphs [8,9,15,16,38,39,43,58,59]. PoseConv3D [17] re-renders human pose sequences into videos, then utilizes these videos, which

exclusively contain the human skeleton or joints, to train a 3D-CNN for action recognition. MotionBERT [87] employs a unified pretraining framework to enhance skeleton-based action recognition and other human motion analysis tasks by learning from a broad range of human motion data.

2.2 Energy Expenditure Estimation

The origin of EE estimation technology can be traced back to the 17th century [47]. The early calorimetric methods dominated [34, 81, 82]. With the development of theory and technology, there are gradually derived a variety of non-calorimetric methods based on human physiological and biochemical signals [2, 11, 35, 83]. The most reliable method is considered to be direct calorimetry, but its use is limited by the high cost and inconvenient [34]. As a proxy, indirect calorimetry is an accurate, noninvasive and portable method [82]. It is the most widely used as a "gold standard" to assess the accuracy of other non-calorimetric methods [44, 64, 70].

As for estimating EE based on video, there is relatively limited research work in this area currently. Edgcomb and Vahid [19] used variations of bounding boxes of people in videos to estimate EE compared with a body-worn device. Tao *et al*. [63] curated an RGB-Depth video dataset called SPHERE-calorie in a home environment with EE labels obtained from gas exchange measurements, and proposed a method that first performs action recognition and then invokes a specific model based on the identified action category to estimate EE. Masullo *et al*. [46] proposed a dual-modal CNN to leverage human silhouette data and accelerometer data to predict EE on SPHERE-calorie [63] dataset. Perrett *et al*. [52] adopted a meta-learning method to achieve a more personalized estimation of EE on the above dataset. Nakamura *et al*. [48] collected an egocentric video dataset termed Stanford-ECM augmented with heart rate and acceleration signals, and proposed a multi-modal multi-task method to jointly predict action category and EE based on video and acceleration signals. Peng *et al*. [51] assembled four widely used video action recognition datasets to acquire Vid2Burn and assigned hourly EE labels through three predefined methods.

However, the datasets mentioned above all have certain shortcomings. The ground truth EE labels in both Stanford-ECM [48] and Vid2Burn [51] are calculated based on a predefined metabolic equivalent (MET) lookup table, which offers a standardized method for quantifying the absolute intensity of various physical activities [27]. However, the MET can only provide a rough estimation of EE, and the impact of individual physical attributes was not considered in the annotation construction process of these two datasets. Therefore, the label accuracy of these two datasets is clearly inadequate, which limits the development of E^3V. As for SPHERE-calorie [63], it used a calorimeter called COSMED K4b^2 that relies on the same technique (gas exchange) as ours to obtain the EE ground truth. But the calorimeter it used is not as advanced as ours. Moreover, it is limited to just 11 household activities with light to moderate intensity (MET$\leq$5.0). This dataset does not encompass the majority of daily exercise categories and cannot meet the data requirements in the field of E^3V.

Table 1: Comparison with existing vision-EE datasets. E3V-K5 is the only dataset that contains heart rate signals, physical attributes (abbr. as "Attr" in the table), and $\dot{V}O_2$-based ground truth. *Our dataset encompasses 6 categories of aerobic exercise, with 4 of them featuring 3 distinct speed levels, resulting in a total of 14 classes.

Dataset	Subject Num.	Clip Num.	Class Num.	Resolution	HR	Attr	$\dot{V}O_2$-based GT	Scenario
SPHERE-calorie	10	188	11	480p	-	✓	✓	Home
Stanford-ECM	10	113	24	720p	✓	-	-	Natural
Vid2Burn	4	9,789	72	variable	-	-	-	Natural
E3V-K5 (ours)	36	16,526	14*	2.7k	✓	✓	✓	Gym

3 E3V-K5 Dataset

To construct a comprehensive and authentic benchmark for E^3V, recruiting a large number of subjects and collecting video samples of various types of physical activities are indispensable. And the calorimeter based on oxygen consumption ($\dot{V}O_2$) is a more ideal manner than MET to measure EE labels. Additionally, the heart rate (HR) and physical attributes of the subject are also correlated with EE.

Therefore, we introduce an authentic dataset called E3V-K5 that contains videos of common exercises and corresponding authentic EE labels, with additional information such as HR and subjects' physical attributes. The EE labels of our dataset are obtained from the most advanced indirect calorimeter based on $\dot{V}O_2$. Table 1 shows a comparison of existing vision-EE datasets, illustrating that our E3V-K5 dataset has the largest number of subjects and video clips, the highest video resolution, and the most comprehensive annotations.

3.1 Dataset Collection

The E3V-K5 dataset is derived from over 112 original videos captured from different perspectives, with each original video having a duration of approximately thirty minutes. The recorded videos include 36 subjects with varying anthropometric measurements, containing 6 exercise categories. The categories are: running, skipping, riding, elliptical, aerobics, and HIIT, which are the most popular types of daily fitness activities. For more refinement and variety, the first 4 categories are further labeled with slow, medium, and fast speeds so that the dataset is subdivided into 14 classes.

As the E3V-K5 dataset aims to establish an authentic vision-EE benchmark, the EE was captured by the COSMED K5 portable metabolic system (K5), while HR was real-time recorded by the Polar H10 heart rate band. The K5 is capable of measuring respiratory gas exchange by the dynamic mixing chamber (DMC) or breath-by-breath (B×B) technique, and then calculates EE based on indirect calorimetry, which is the most effective and accurate way to estimate EE during rest and aerobic exercise [12]. The Polar H10 is a chest belt for HR monitoring synchronized with the K5. These ground truth sources are recognized as the "gold standard" and have been widely used in sports research. Simultaneously,

all RGB videos were captured by the EZVIZ S2 camera at 2.7k raw resolution and 30 fps.

3.2 Data Processing

Given the high resolution of the original videos, we downsample them to an 856×480 resolution for the convenience of processing. Our samples have two kinds of energy expenditure measurement techniques, namely DMC and B×B mentioned in Section 3.1. The former records EE and HR every ten seconds, while the latter records EE and HR with each breath. For the sake of uniformity and ease of processing, we intend to cut the video into clips every 10 seconds and label them with EE and HR. For the DMC video samples, the original record is sufficient to assign the labels for the clips, while for the B×B samples, we average the EE and HR records every ten seconds as the labels of the video clip. In this manner, we obtained 17,260 video clips with EE and HR labels, and matched them with the physical attributes of the subjects.

As described in [32], there is a delay between the time when the EE occurs in the muscles and the time it is recorded by the metabolic system. This delay varies for each individual. Therefore, we calculated the cross-correlation of EE and HR to obtain a mean delay time for each subject to revise the EE label, these manner is similar to [4]. We revised the EE label before cutting the videos.

In order to facilitate research on E3V-K5, we extract the human body skeleton sequence of subjects using the AlphaPose [20] framework. Concretely, we use the AlphaPose pretrained on the COCO dataset [42] as the pose estimator, and use the QDTrack [50] pretrained on the CrowdHuman dataset [56] as the tracker. We write a script to automatically assign the skeletons with tracking failures to the nearest sequence and extract the skeleton sequence of the subject in the video based on the amount of skeletal movement. A ten-second video clip contains 300 frames. We select video clips where the length of the detected posture sequences of the subject is greater than or equal to 290 as valid samples. Along with some manually removed video clips, we finally obtained 16,526 video clips as the release version of our dataset for training models.

Figure 2a illustrates the distribution of video clips across each class. *Running* is the most frequent category, comprising a total of 7,096 clips, with 1,733 in fast (_f), 2,676 in medium (_m), and 2,687 in slow (_s) speed variations. On the other hand, *Skipping* is the category with the fewest clips, totaling 1,942, which consist of 1,074 in fast, 692 in medium, and 176 in slow speed variations. The average number of video clips per class is 1,232.9. Figure 2b illustrates the distribution of energy expenditure measurements across exercise class, and Figure 2c depicts the distribution of heart rate measurements across exercise class. Generally speaking, the higher the exercise intensity, the higher the energy expenditure and heart rate. Among the categories, *Running_f* shows the highest average EE and HR, while *riding_s* has the lowest average EE and HR. It is evident that the heart rate signals between classes are generally close, indicating a limited discrimination ability for different exercises. Moreover, there is a

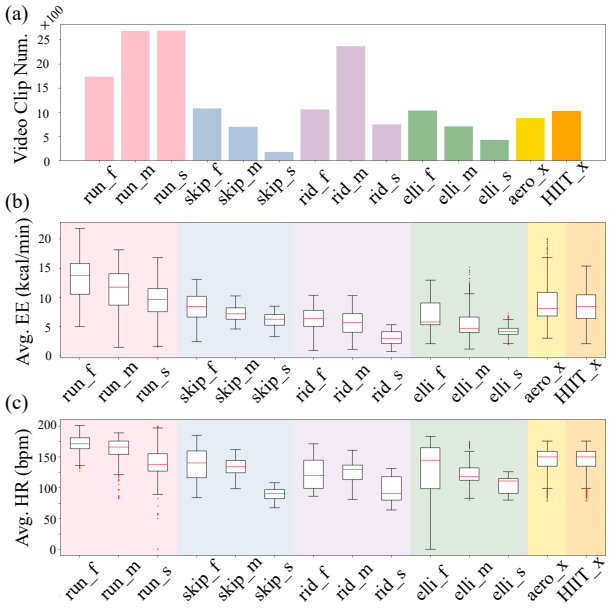

Fig. 2: Statistics of E3V-K5 dataset. (a) The number of video clips. (b) Average energy expenditure (EE) for each class. (c) Average heart rate (HR) for each class.

significant dispersion in heart rate within the same class, especially in *elliptical_f*, demonstrating that heart rate has obvious individual differences. Based on the above findings, it is difficult to estimate EE accurately by heart rate measurement alone.

3.3 Cross-Subject Data Split

The application of energy expenditure estimation based on video requires the model to have a strong generalization on individuals not seen in the training set. In order to evaluate the generalization of the model, we divided the E3V-K5 dataset into training, validation, and test sets according to the subjects. Specifically, we randomly divide the 36 subjects in a roughly 6:2:2 ratio, assigning 22 subjects to the training set, with 7 subjects each in the validation and test sets. Accordingly, the number of video clips in the training, validation, and test sets are 10,049 and 3,234 and 3,243, respectively. This cross-subject data split ensures that the subjects used for evaluating the model's performance are not seen during the model training process, which allows for an effective assessment of the model's generalization ability. Besides, the original complete exercise videos and their corresponding EE labels are still preserved in E3V-K5 dataset. Future researchers can use these videos to investigate the relationship between EE and accumulated training time.

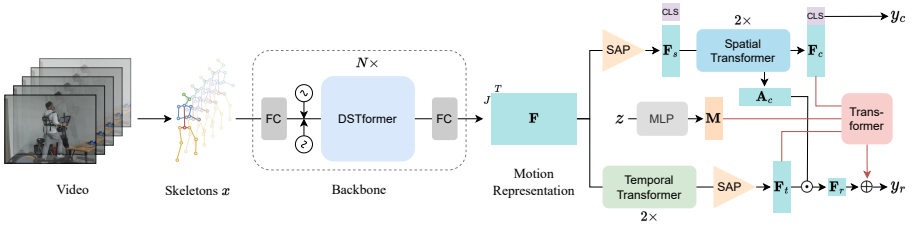

Fig. 3: Framework of E3SFormer. The human skeleton sequence x is extracted using a pose estimator from the video and then fed into a backbone to obtain motion representation $\mathbf{F}$. It is then sent to an action recognition branch (upper) and an energy estimation regression branch (lower). The category-related joint-specific attention $\mathbf{A}_c$ from the action recognition branch is transferred to the energy estimation regression branch to boost its performance. The multi-modal data z are used for more personalized energy estimation estimation.

4 E3SFormer: Energy Expenditure Estimation Skeleton Transformer

Accurately estimating energy expenditure requires fine-grained analysis of video, which is a computationally intensive task. Traditional video understanding methods usually sample a small number of frames in each video [6, 23, 60, 66], which is not adequate for predicting precise energy expenditure of human motion. If we input all frames of a video clip into these methods, the GPU memory usage and inference time will be excessive, making it unfavorable for practical applications. Furthermore, irrelevant stuff and background in the video may affect the prediction of EE. Therefore, we adopt a human skeleton-based method to accurately estimate EE on our E3V-K5 dataset and reduce computational cost and inference time.

In this section, we introduce our proposed E3SFormer in detail. The overall procedure is illustrated in Figure 3, including a backbone and two branches for action recognition and EE regression, respectively. The entire network is based on the Transformer architecture. The backbone uses a spatio-temporal fusion for extracting spatial and temporal features of an inputted human skeleton sequence. After that, the features will be fed into the two different branches for different tasks simultaneously. The attention of each joint in the action recognition branch is transferred to the EE regression branch to facilitate precise EE regression.

4.1 Spatio-Temporal Motion Feature Extraction

The key component of the backbone is a Dual-stream Spatio-temporal Transformer (DSTformer) block. One DSTformer block consists of two different branches. The first branch initially performs a Transformer along the spatial (joint) dimension, followed by a Transformer on the temporal dimension. The second branch switches the order of these two Transformers. The result of these two branches is

fused through adaptive weights produced by an attention regressor. Each branch of DSTformer has the capability of modeling comprehensive spatio-temporal information, and different branches are interested in different spatio-temporal aspects. The fusion operation can dynamically balance the results of these two branches.

Specifically, we define the input skeleton sequence as $x \in \mathbb{R}^{T\times J\times C_{\text{in}}}$, where T is the temporal sequence length, J is the number of body joints, and C_{in} is the channel number of input. Specifically, $C_{\text{in}} = 3$ in here, the first and second channels are the x-coordinate and y-coordinate of body joints respectively, and the third channel is the visibility confidence of each joint offered by the pose estimation method [20]. The skeleton sequence x is projected to a high-dimensional feature $\mathbf{F}^0 \in \mathbb{R}^{T\times J\times C}$, and concatenated with a pretrained spatial position encoding $\mathbf{P}_{\text{S}} \in \mathbb{R}^{1\times J\times C}$ and a temporal position encoding $\mathbf{P}_{\text{T}} \in \mathbb{R}^{T\times 1\times C}$. Then the input feature is fed into the backbone that contains N DSTformer blocks to get the motion representation $\mathbf{F} \in \mathbb{R}^{T\times J\times C}$. C denotes the channel of features used in the backbone and thereafter branches. The obtained motion representation $\mathbf{F}$ is then fed into two transformer branches for both action recognition and energy expenditure regression.

4.2 Spatial-based Action Recognition

For the action recognition branch, we first use a Self Attention Pooling (SAP) layer to squeeze the temporal dimension T of $\mathbf{F}$, which is defined as follows:

$$\text{SAP}(\mathbf{F}_j) = \sum_{t=1}^{T} \frac{\exp(\text{FC}(\mathbf{F}_j^t))}{\sum_{t'=1}^{T}\exp(\text{FC}(\mathbf{F}_j^{t'}))} \cdot \mathbf{F}_j^t, \tag{1}$$

where $\mathbf{F}_j$ is the slice of $\mathbf{F}$ along the joint dimension, FC is a Fully Connected layer. The result of this SAP layer is denoted as $\mathbf{F}_s \in \mathbb{R}_{J\times C}$, which is concatenated with a class token (CLS) and fed into a two Spatial Transformer (ST) layer to model the relation shape among the joints. The ST aims to perform Transformer operation along the joint dimension, the key component of which is the Multi-Head Self-Attention (MHSA). First, the query $\mathbf{Q}^i$, key $\mathbf{K}^i$, and value $\mathbf{V}^i$ of head $i \in [1, h]$ is calculated as follows:

$$\mathbf{Q}^i = \mathbf{F}_s\mathbf{W}_Q^i,\ \mathbf{K}^i = \mathbf{F}_s\mathbf{W}_K^i,\ \mathbf{V}^i = \mathbf{F}_s\mathbf{W}_V^i, \tag{2}$$

where $\mathbf{W}_Q^i$, $\mathbf{W}_K^i$, and $\mathbf{W}_V^i$ are projection layers of head i. Then, we calculate the attention matrices as follows:

$$\mathbf{A}^i = \text{softmax}(\frac{\mathbf{Q}^i(\mathbf{K}^i)^\top}{\sqrt{d_K}}), \tag{3}$$

where d_K is the feature dimension of $\mathbf{K}^i$. After that, the output of the MHSA is defined as:

$$\text{MHSA}(\mathbf{F}_s) = [\mathbf{A}^1\mathbf{V}^1; \ldots; \mathbf{A}^h\mathbf{V}^h; \ldots]\mathbf{W}_O, \tag{4}$$

where $\mathbf{W}_O$ is a output projection. Residual connection is used to the MHSA result, which is fed into a multi-layer perceptron (MLP), and followed by a residual connection. The Pre-LayerNorm trick is used for both MHSA and MLP.

4.3 Joint-Specific Attention for Enhanced Energy Expenditure Regression

Every token in the action recognition branch leverages its query to calculate the similarity of all keys to form the attention matrix, representing which tokens should be concerned. The CLS token is used to classify action, so in our intuition, which joints are important for a certain action category can be represented by the attention of the CLS token. Therefore, the average of multi-head attention of CLS token in the second ST, which is termed as category-related joint-specific attention $\mathbf{A}_c \in \mathbb{R}^J$, is used to signify the importance.

For the EE prediction branch, there are two Temporal Transformer (TT) layers followed by a SAP layer. The only difference between ST and TT is that TT is performed along the temporal dimension of each joint. The result can be denoted as $\mathbf{F}_t \in \mathbb{R}^{J\times C}$. To gain the enhanced representation for regression, we use $\mathbf{A}_c$ as a weight to calculate a weighted sum of $\mathbf{F}_t$ along the joint dimension, resulting in $\mathbf{F}_r \in \mathbb{R}^C$. For the integration of multi-modal data z including heart rate and physical attributes, an MLP is used to extract feature $\mathbf{M}$ of them. Then, it is concatenated with the $\mathbf{F}_t$ and the result of action recognition branch $\mathbf{F}_c \in \mathbb{R}^{J\times C}$ without CLS token and fed into a Transformer layer. The result as well as $\mathbf{F}_r$ is used to regress EE.

We use the Cross-Entropy Loss L_c to train the action recognition branch, together with L1 Loss L_r to train the EE regression branch. The overall loss function is as follows:

$$L = L_r + \alpha L_c, \tag{5}$$

where α is a hyperparameter.

5 Experiments

5.1 Experiment Setup

Comparison Methods. We compare the proposed E3SFormer with the following skeleton-based action recognition frameworks on the E3V-K5 dataset: ST-GCN [84], AAGCN [59], MS-G3D [43], CTR-GCN [8], ST-GCN++ [16], DG-STGCN [15], and PoseConv3D [17]. Among them, the last method is based on CNN, while the other methods are based on GCN. We modify the output channel of the last Linear layer originally for classification to 1 for EE regression. Besides, We alter the input channel of PoseConv3D [17] from J to 3, and use the sequence of RGB video frames as input to simply compare the performance of the skeleton-based approach and video-based approach. The altered framework is designated as RGBConv3D.

Training Details. We use the pretrained weight of MotionBERT [87] to initialize the backbone. However, the length of pretrained temporal position encoding $\mathbf{P}_\mathrm{T}$ is insufficient for our fine-grained task that has a quite long skeleton sequence. Therefore, we perform linear interpolation on the T dimension of $\mathbf{P}_\mathrm{T}$ from the original number to a longer number to accommodate longer input

sequences. Our model and comparison models are implemented by PyTorch and optimized by Lion [7] optimizer with a learning rate of 10^{-4}, weight decay of 5×10^{-4}, and cosine annealing as the learning rate decay schedule. We train all the settings for 50 epochs with a batch size of 16, except the two CNN-based models, PoseConv3D and RGBConv3D. Considering the larger GPU memory usage of these two models, we set the batch size of these two models to 8. For all the skeleton-based models, the joint coordinates are normalized to the range of $[-1, 1]$. The random horizontal flipping is applied as the data augmentation.

Evaluation Metrics. We adopt L1 Loss to train every model for EE regression, which is also known as Mean Absolute Error (MAE). In addition to MAE, we also use Mean Relative Error (MRE), Pearson Correlation Coefficient (PCC), and Coefficient of Determination (R^2) as evaluation metrics for the model. The MRE, PCC, and R^2 are calculated as follows:

$$\text{MRE} = \frac{1}{N}\sum_{i=1}^{N}\frac{|y_i - \hat{y}_i|}{y_i}, \tag{6}$$

$$\text{PCC} = \frac{\sum_{i=1}^{N}(y_i - \overline{y})(\hat{y}_i - \overline{\hat{y}})}{\sqrt{\sum_{i=1}^{N}(y_i - \overline{y})^2}\sqrt{\sum_{i=1}^{N}(\hat{y}_i - \overline{\hat{y}})^2}}, \tag{7}$$

$$\text{R}^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \overline{y})^2}, \tag{8}$$

where y_i and $\hat{y}_i$ denote the actual value and predicted value respectively, while $\overline{y}$ and $\overline{\hat{y}}$ denote the mean of the actual values and the predicted values respectively, and N denotes the number of samples.

5.2 Results of Energy Expenditure Estimation

Pure Skeleton Results. As shown in Table 2 (top), when we only leverage the human skeleton sequence as input, our proposed E3SFormer surpasses all comparison methods on most evaluation metrics, except for R^2. But the R^2 still ranks second among all the methods, and very close to the first (0.5118 compared to 0.5175). These results demonstrate the effectiveness of our method.

The PoseConv3D [17] ranks first on R^2 and performs relatively better compared to other GCN-based methods on other evaluation metrics, exhibiting the superior capability to extract fine-grained features in our task. We conjecture that this is because the issue of over-smoothing in GCNs results in a diminished ability to express fine-grained motion features in the deeper layers of the network. Accurate estimation of EE, however, requires precise capture of the displacement of each joint to measure muscle contractions, a capability where CNNs excel.

Despite being a CNN, the RGBConv3D performs much worse compared to PoseConv3D [17] and other GCN-based methods. The main reason, in our opinion, is that the inputs of RGBConv3D are RGB video clips that contain irrelevant objects, other people, and various backgrounds, which may disturb the

Table 2: Energy expenditure regression results on E3V-K5 dataset. The **w/o MM** and **w/ MM** denote without and with heart rate and physical attributes as multi-modal data, respectively. We use the percentage form of MRE for a clear presentation. The ↓ indicates the lower the better, and the ↑ indicates the higher the better.

	Metric\Method	ST-GCN	AAGCN	MS-G3D	CTR-GCN	ST-GCN++	DG-STGCN	PoseConv3D	RGBConv3D	**E3SFormer** (Ours)
w/o MM	MRE (%) ↓	36.42	37.00	47.67	34.56	35.98	34.28	33.03	42.93	**28.81**
	MAE ↓	2.1939	2.2050	2.7804	2.3978	2.2023	2.0796	2.0670	2.5408	**2.0304**
	PCC ↑	0.6632	0.6686	0.5523	0.6798	0.6944	0.7397	0.7232	0.5186	**0.7528**
	R^2 ↑	0.3722	0.4037	0.1325	0.3412	0.4317	0.5054	**0.5175**	0.2663	0.5118
w/ MM	MRE (%) ↓	23.06	21.47	21.86	20.11	22.37	20.72	21.52	28.83	**18.01**
	MAE ↓	1.4895	1.4176	1.4862	1.3016	1.5122	1.3874	1.3939	1.7382	**1.2490**
	PCC ↑	0.8637	0.8701	0.8570	0.8967	0.8640	0.8667	0.8976	0.8988	**0.9157**
	R^2 ↑	0.7169	0.7425	0.7257	0.7909	0.7265	0.7360	0.7861	0.7048	**0.7953**

prediction of EE. In contrast, PoseConv3D [17] renders the joint coordinates to the video space as the input of CNN, focusing on human body movement while disregarding the influence of background factors.

Multi-modal Input Results. Based on the fact that different individuals will have varying energy expenditures when engaging in the same type and intensity of exercise, using only video clips or skeleton sequences to accurately predict EE is inadequate. More personalized data are required for this purpose. Therefore, when augmented with heart rate and physical attributes, the model performances are much better than without these multi-modal data, shown in Table 2 (bottom).

For all the comparison methods, we leverage a three-layer MLP to extract a feature of heart rate and physical attributes of each input sample. The hidden layers and output layer of the MLP have the same number of channels as the output channels of each backbone in these methods. The extracted attribute feature is concatenated with the backbone feature, and fed into a fully connected layer to predict EE.

With the help of multi-modal data, the performances of all methods improved significantly. Our method ranks first on all of the evaluation metrics, owing to a meticulously designed architecture. The gap between CNN-based methods and GCN-based methods becomes less pronounced. The PoseConv3D [17] does not stand out on the evaluation metrics representing prediction accuracy (MRE and MAE), but performs well on the evaluation metrics related to correlation (PCC and R^2). The PCC of RGBConv3D is quite high while the R^2 is relatively lower, which is related to the worst performance on MRE and MAE, showing a high correlation but low prediction accuracy. The incorporation of multi-modal data boosts the prediction accuracy of all methods. However, according to the two analyses above, due to the structural advantages of CNNs, CNN-based methods exhibit better predictive correlation.

5.3 Ablation Study

Table 3 shows the ablation study that we conducted. The left half of the table is experiments using only heart rate (HR) and physical attributes (Attr) to predict

Table 3: Ablation study of our method. "Formula" denotes using a predefined set of formulas to calculate Energy Expenditure based on heart rate and physical attributes. "AR" is an abbreviation for Action Recognition.

Metric\Ablation	Formula	Only HR	Only Attr	HR+Attr	w/o MM w/o AR	w/o MM w/ AR	w/ MM w/o AR	w/ MM w/ AR
MRE (%) ↓	65.02	32.75	58.41	25.78	39.22	28.81	24.88	18.01
MAE ↓	3.5047	2.7904	3.0140	1.5276	2.1071	2.0304	1.5105	1.2490
PCC ↑	0.7767	0.7871	0.5812	0.8712	0.7155	0.7528	0.8560	0.9157
R^2 ↑	-0.1493	0.0268	0.0416	0.7297	0.4704	0.5118	0.7035	0.7953

EE. The Formula [61] is given by the *American College of Sports Medicine* to estimate EE based on these data. The parameters of the formula differ for males and females. For males, the formula is as follows:

$$\mathrm{EE} = \frac{(0.6309 \times HR + 0.1988 \times W + 0.2017 \times A - 55.0969)}{4.184},$$

while for females, the formula is:

$$\mathrm{EE} = \frac{(0.4472 \times HR + 0.1263 \times W + 0.074 \times A - 20.4022)}{4.184},$$

where EE denotes the energy expenditure ($kcal/min$), HR, W, and A denote heart rate, weight, and age, respectively. The rest three columns are the experiments using a three-layer MLP to predict EE according to the specified data. The channel number of the hidden layers is 512. It is shown that the neural networks are more appropriate than the predefined formula for this task. Both using only heart rate and using only physical attributes are not sufficient to produce an acceptable result, indicating that EE is related to a combination of both, rather than either one alone. The right half of Table 3 is the ablation study of E3SFormer's action recognition branch with the category-related joint-specific attention. The "w/o AR" refers to replacing the joint-specific attention with average pooling for averaging regression outputs. The results show that without the joint-specific attention, the performance will degenerate, demonstrating the importance of it.

6 Conclusion

We curate E3V-K5, an authentic benchmark for energy expenditure estimation based on video. A total of 16,526 videos are included in this dataset, labeled with the COSMED K5 calorimeter to gain the authentic energy expenditure of people. Additionally, it includes the heart rate and physical attributes of each subject. The data volume, label diversity, and authenticity of E3V-K5 surpass all previous video datasets for energy expenditure estimation. Moreover, we propose the E3SFormer that utilizes human skeleton data from videos to regress energy expenditure. Comprehensive experiments exhibit the challenging nature of E3V-K5 and the effectiveness of E3SFormer, aiming to inspire further research based on this benchmark.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (U20B2066), Zhejiang Province "JianBingLingYan+" Research and Development Plan (No.2024C01021), and the Fundamental Research Funds for the Central Universities (226-2024-00058).

References

1. Ainslie, P.N., Reilly, T., Westerterp, K.R.: Estimating human energy expenditure. Sports Medicine **33**, 683–698 (2003)
2. Altini, M., Casale, P., Penders, J.F., Amft, O.: Personalization of energy expenditure estimation in free living using topic models. IEEE Journal of Biomedical and Health Informatics **19**(5), 1577–1586 (2015)
3. Alves, A.J., Viana, J.L., Cavalcante, S., Oliveira, N.L., Duarte, J.A., Mota, J., Oliveira, J.C., Ribeiro, F.: Physical activity in primary and secondary prevention of cardiovascular disease: Overview updated. World Journal of Cardiology **8**, 575 – 583 (2016)
4. Blake, O.M., Wakeling, J.M.: Estimating changes in metabolic power from emg. Springerplus **2**(1), 1–7 (2013)
5. Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: Realtime multi-person 2d pose estimation using part affinity fields. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (July 2017)
6. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (July 2017)
7. Chen, X., Liang, C., Huang, D., Real, E., Wang, K., Liu, Y., Pham, H., Dong, X., Luong, T., Hsieh, C.J., et al.: Symbolic discovery of optimization algorithms. arXiv preprint arXiv:2302.06675 (2023)
8. Chen, Y., Zhang, Z., Yuan, C., Li, B., Deng, Y., Hu, W.: Channel-wise topology refinement graph convolution for skeleton-based action recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13359–13368 (2021)
9. Cheng, K., Zhang, Y., Cao, C., Shi, L., Cheng, J., Lu, H.: Decoupling gcn with dropgraph module for skeleton-based action recognition. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIV 16. pp. 536–553. Springer (2020)
10. Clark, C.C.T., Barnes, C.M., Stratton, G., McNarry, M.A., Mackintosh, K.A., Summers, H.D.: A review of emerging analytical techniques for objective physical activity measurement in humans. Sports Medicine **47**, 439–447 (2016)
11. Crouter, S.E., Clowers, K.G., Bassett, D.: A novel method for using accelerometer data to predict energy expenditure. Journal of Applied Physiology **100 4**, 1324–31 (2006)
12. Crouter, S.E., LaMunion, S.R., Hibbing, P.R., Kaplan, A., Bassett, D.: Accuracy of the cosmed k5 portable calorimeter. PLoS ONE **14** (2019)
13. Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., Darrell, T.: Long-term recurrent convolutional networks for visual recognition and description. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (June 2015)

14. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
15. Duan, H., Wang, J., Chen, K., Lin, D.: Dg-stgcn: dynamic spatial-temporal modeling for skeleton-based action recognition. arXiv preprint arXiv:2210.05895 (2022)
16. Duan, H., Wang, J., Chen, K., Lin, D.: Pyskl: Towards good practices for skeleton action recognition. In: Proceedings of the 30th ACM International Conference on Multimedia. pp. 7351–7354 (2022)
17. Duan, H., Zhao, Y., Chen, K., Lin, D., Dai, B.: Revisiting skeleton-based action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2969–2978 (2022)
18. Durstine, J.L., Grandjean, P.W., Cox, C.A., Thompson, P.D.: Lipids, lipoproteins, and exercise. Journal of Cardiopulmonary Rehabilitation and Prevention **22**(6), 385–398 (2002)
19. Edgcomb, A., Vahid, F.: Estimating daily energy expenditure from video for assistive monitoring. In: 2013 IEEE International Conference on Healthcare Informatics. pp. 184–191. IEEE (2013)
20. Fang, H.S., Li, J., Tang, H., Xu, C., Zhu, H., Xiu, Y., Li, Y.L., Lu, C.: Alphapose: Whole-body regional multi-person pose estimation and tracking in real-time. IEEE Transactions on Pattern Analysis and Machine Intelligence (2022)
21. Fang, Y., Wang, W., Xie, B., Sun, Q., Wu, L., Wang, X., Huang, T., Wang, X., Cao, Y.: Eva: Exploring the limits of masked visual representation learning at scale. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19358–19369 (June 2023)
22. Feichtenhofer, C.: X3d: Expanding architectures for efficient video recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (June 2020)
23. Feichtenhofer, C., Fan, H., Malik, J., He, K.: Slowfast networks for video recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (October 2019)
24. Feichtenhofer, C., Li, Y., He, K., et al.: Masked autoencoders as spatiotemporal learners. Advances in Neural Information Processing Systems **35**, 35946–35958 (2022)
25. Feichtenhofer, C., Pinz, A., Wildes, R.P.: Spatiotemporal multiplier networks for video action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (July 2017)
26. Feichtenhofer, C., Pinz, A., Zisserman, A.: Convolutional two-stream network fusion for video action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (June 2016)
27. Ferguson, B.: Acsm's guidelines for exercise testing and prescription 9th ed. 2014. Journal of the Canadian Chiropractic Association **58**, 328–328 (2014)
28. Guo, D., Li, K., Hu, B., Zhang, Y., Wang, M.: Benchmarking micro-action recognition: Dataset, method, and application. IEEE Transactions on Circuits and Systems for Video Technology (2024)
29. Guo, S., Xiong, Z., Zhong, Y., Wang, L., Guo, X., Han, B., Huang, W.: Cross-architecture self-supervised video representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19270–19279 (June 2022)
30. Hamasaki, H.: Daily physical activity and type 2 diabetes: A review. World Journal of Diabetes **7 12**, 243–51 (2016)

31. Hand, G.A., Shook, R.P., O'Connor, D.P., Kindred, M.M., Schumacher, S.M., Drenowatz, C., Paluch, A.E., Burgess, S., Blundell, J.E., Blair, S.N.: The effect of exercise training on total daily energy expenditure and body composition in weight-stable adults: A randomized, controlled trial. Journal of physical activity & health pp. 1–8 (2020)
32. Hughson, R.L., Tschakovsky, M.E., Houston, M.E.: Regulation of oxygen consumption at the onset of exercise. Exercise and sport sciences reviews **29**(3), 129–133 (2001)
33. Ji, S., Xu, W., Yang, M., Yu, K.: 3d convolutional neural networks for human action recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **35**(1), 221–231 (2012)
34. Kaiyala, K.J., Ramsay, D.S.: Direct animal calorimetry, the underused gold standard for quantifying the fire of life. Comparative Biochemistry and Physiology. Part A, Molecular & Integrative Physiology **158 3**, 252–64 (2011)
35. Kalkwarf, H.J., Haas, J.D., Belko, A.Z., Roach, R.C., Roe, D.A.: Accuracy of heart-rate monitoring and activity diaries for estimating energy expenditure. The American journal of Clinical Nutrition **49 1**, 37–43 (1989)
36. Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., Fei-Fei, L.: Large-scale video classification with convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (June 2014)
37. Kirkham, A.A., Davis, M.K.: Exercise prevention of cardiovascular disease in breast cancer survivors. Journal of Oncology **2015** (2015)
38. Korban, M., Li, X.: Ddgcn: A dynamic directed graph convolutional network for action recognition. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16. pp. 761–776. Springer (2020)
39. Li, B., Li, X., Zhang, Z., Wu, F.: Spatio-temporal graph routing for skeleton-based action recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 8561–8568 (2019)
40. Li, K., Wang, Y., Li, Y., Wang, Y., He, Y., Wang, L., Qiao, Y.: Unmasked teacher: Towards training-efficient video foundation models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 19948–19960 (October 2023)
41. Lin, J., Gan, C., Han, S.: Tsm: Temporal shift module for efficient video understanding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (October 2019)
42. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. pp. 740–755. Springer (2014)
43. Liu, Z., Zhang, H., Chen, Z., Wang, Z., Ouyang, W.: Disentangling and unifying graph convolutions for skeleton-based action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 143–152 (2020)
44. Lyden, K., Kozey, S.L., Staudenmeyer, J.W., Freedson, P.S.: A comprehensive evaluation of commonly used accelerometer energy expenditure and met prediction equations. European Journal of Applied Physiology **111**, 187–201 (2011)
45. Macdonald, T.: Preventing chronic diseases: A vital investment. Journal of The Royal Society for The Promotion of Health **126**, 95 (2006)

46. Masullo, A., Burghardt, T., Damen, D., Hannuna, S., Ponce-López, V., Mirmehdi, M.: Calorinet: From silhouettes to calorie estimation in private environments. arXiv preprint arXiv:1806.08152 (2018)
47. Murdoch, D.R.: High life: A history of high altitude physiology and medicine. British Medical Journal **318**, 1631 (1999)
48. Nakamura, K., Yeung, S., Alahi, A., Fei-Fei, L.: Jointly learning energy expenditures and activities using egocentric multimodal signals. 2017 IEEE Conference on Computer Vision and Pattern Recognition pp. 6817–6826 (2017)
49. Neely, F.G.: Biomechanical risk factors for exercise-related lower limb injuries. Sports Medicine **26**, 395–413 (1998)
50. Pang, J., Qiu, L., Li, X., Chen, H., Li, Q., Darrell, T., Yu, F.: Quasi-dense similarity learning for multiple object tracking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 164–173 (2021)
51. Peng, K., Roitberg, A., Yang, K., Zhang, J., Stiefelhagen, R.: Should i take a walk? estimating energy expenditure from video data. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops pp. 2074–2084 (2022)
52. Perrett, T., Masullo, A., Damen, D., Burghardt, T., Craddock, I., Mirmehdi, M., et al.: Personalized energy expenditure estimation: Visual sensing approach with deep learning. JMIR Formative Research **6**(9), e33606 (2022)
53. Pham, H.H., Khoudour, L., Crouzil, A., Zegers, P., Velastin, S.A.: Video-based human action recognition using deep learning: a review. arXiv preprint arXiv:2208.03775 (2022)
54. Qing, Z., Zhang, S., Huang, Z., Wang, X., Wang, Y., Lv, Y., Gao, C., Sang, N.: Mar: Masked autoencoders for efficient action recognition. IEEE Transactions on Multimedia (2023)
55. Ryali, C., Hu, Y.T., Bolya, D., Wei, C., Fan, H., Huang, P.Y., Aggarwal, V., Chowdhury, A., Poursaeed, O., Hoffman, J., et al.: Hiera: A hierarchical vision transformer without the bells-and-whistles. arXiv preprint arXiv:2306.00989 (2023)
56. Shao, S., Zhao, Z., Li, B., Xiao, T., Yu, G., Zhang, X., Sun, J.: Crowdhuman: A benchmark for detecting human in a crowd. arXiv preprint arXiv:1805.00123 (2018)
57. Shcherbina, A., Mattsson, C.M., Waggott, D., Salisbury, H., Christle, J.W., Hastie, T.J., Wheeler, M.T., Ashley, E.A.: Accuracy in wrist-worn, sensor-based measurements of heart rate and energy expenditure in a diverse cohort. Journal of Personalized Medicine **7** (2016)
58. Shi, L., Zhang, Y., Cheng, J., Lu, H.: Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12026–12035 (2019)
59. Shi, L., Zhang, Y., Cheng, J., Lu, H.: Skeleton-based action recognition with multi-stream adaptive graph convolutional networks. IEEE Transactions on Image Processing **29**, 9532–9545 (2020)
60. Simonyan, K., Zisserman, A.: Two-stream convolutional networks for action recognition in videos. Advances in Neural Information Processing Systems **27** (2014)
61. of Sports Medicine, A.C.: Guidelines for exercise testing and prescription. Williams & Wilkins (1991)
62. Tan, H., Lei, J., Wolf, T., Bansal, M.: Vimpac: Video pre-training via masked token prediction and contrastive learning. arXiv preprint arXiv:2106.11250 (2021)
63. Tao, L., Burghardt, T., Mirmehdi, M., Damen, D., Cooper, A., Hannuna, S.L., Camplani, M., Paiement, A., Craddock, I.: Calorie counter: Rgb-depth visual estimation of energy expenditure at home. In: 2016 Asian Conference on Computer Vision (2016)

64. Tikkanen, O., Kärkkäinen, S., Haakana, P., Kallinen, M., Pullinen, T., Finni, T.: Emg, heart rate, and accelerometer as estimators of energy expenditure in locomotion. Medicine and Science in Sports and Exercise **46 9**, 1831–9 (2014)
65. Tong, Z., Song, Y., Wang, J., Wang, L.: Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In: Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A. (eds.) Advances in Neural Information Processing Systems. vol. 35, pp. 10078–10093. Curran Associates, Inc. (2022)
66. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3d convolutional networks. In: Proceedings of the IEEE International Conference on Computer Vision (December 2015)
67. Tran, D., Wang, H., Torresani, L., Feiszli, M.: Video classification with channel-separated convolutional networks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (October 2019)
68. Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., Paluri, M.: A closer look at spatiotemporal convolutions for action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (June 2018)
69. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in Neural Information Processing Systems **30** (2017)
70. Villar, R., Beltrame, T., Hughson, R.L.: Validation of the hexoskin wearable vest during lying, sitting, standing, and walking activities. Applied Physiology, Nutrition, and Metabolism **40 10**, 1019–24 (2015)
71. Wang, L., Koniusz, P.: 3mformer: Multi-order multi-mode transformer for skeletal action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5620–5631 (June 2023)
72. Wang, L., Xiong, Y., Wang, Z., Qiao, Y.: Towards good practices for very deep two-stream convnets. arXiv preprint arXiv:1507.02159 (2015)
73. Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., Van Gool, L.: Temporal segment networks: Towards good practices for deep action recognition. In: European Conference on Computer Vision. pp. 20–36. Springer (2016)
74. Wang, R., Chen, D., Wu, Z., Chen, Y., Dai, X., Liu, M., Jiang, Y.G., Zhou, L., Yuan, L.: Bevt: Bert pretraining of video transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14733–14743 (June 2022)
75. Wang, R., Chen, D., Wu, Z., Chen, Y., Dai, X., Liu, M., Yuan, L., Jiang, Y.G.: Masked video distillation: Rethinking masked feature modeling for self-supervised video representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6312–6322 (June 2023)
76. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (June 2018)
77. Wang, Y., Li, K., Li, Y., He, Y., Huang, B., Zhao, Z., Zhang, H., Xu, J., Liu, Y., Wang, Z., et al.: Internvideo: General video foundation models via generative and discriminative learning. arXiv preprint arXiv:2212.03191 (2022)
78. Wang, Y., Long, M., Wang, J., Yu, P.S.: Spatiotemporal pyramid network for video action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (July 2017)
79. Wei, C., Fan, H., Xie, S., Wu, C.Y., Yuille, A., Feichtenhofer, C.: Masked feature prediction for self-supervised visual pre-training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14668–14678 (June 2022)

80. Welk, G.J.: Physical activity assessments for health-related research (2002)
81. Westerterp, K.R.: Doubly labelled water assessment of energy expenditure: principle, practice, and promise. European Journal of Applied Physiology **117**, 1277 – 1285 (2017)
82. White, L.E., DeBlois, J.P., Barreira, T.V.: Reliability analysis of the cosmed k5 portable metabolic system. Medicine & Science in Sports & Exercise (2019)
83. Williams, G.L., Li, S., Pathirana, P.N.: Preliminary investigation of energy comparation between gyroscope, electromyography and vo2 wearable sensors. 2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society pp. 4963–4966 (2016)
84. Yan, S., Xiong, Y., Lin, D.: Spatial temporal graph convolutional networks for skeleton-based action recognition. In: Proceedings of the AAAI conference on artificial intelligence. vol. 32 (2018)
85. Yue-Hei Ng, J., Hausknecht, M., Vijayanarasimhan, S., Vinyals, O., Monga, R., Toderici, G.: Beyond short snippets: Deep networks for video classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (June 2015)
86. Zhou, B., Andonian, A., Oliva, A., Torralba, A.: Temporal relational reasoning in videos. In: Proceedings of the European Conference on Computer Vision (September 2018)
87. Zhu, W., Ma, X., Liu, Z., Liu, L., Wu, W., Wang, Y.: Motionbert: A unified perspective on learning human motion representations. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2023)