Supplementary Materials for *Supplementary Materials for Should I take a walk?* Estimating Energy Expenditure from Video Data

Kunyu Peng*, Alina Roitberg*, Kailun Yang, Jiaming Zhang, and Rainer Stiefelhagen
*denotes equal contribution
Karlsruhe Institute of Technology
firstname.lastname@kit.edu

1. Limitations

In addition to the summary of limitations mentioned at the end of our main paper, more details about the limitations of our approaches and proposed benchmarks will be given in this section. This work targets estimation of energy expenditure from videos. The benefits of such methods extend to multiple applications, such as supporting active and healthy lifestyle, *e.g.*, by tracking exercise routines [4] or monitoring the daily physical activity level for elderly care [2, 5, 8]. However, our work is not without limitations. Energy expenditure is a complex physiological process [3], and while bodily movement, (*i.e.* active muscled and intensities) are its primary drivers, there is a variety of the contributing factors, such as age, gender, weight and personal metabolic rate. Many of these factors are not considered in our work. For example, for simplicity, we derive energy annotations from medical compendiums assuming the weight of 150 *lb* (our study with the heart-rate based ground truth estimation is an exception, where age/gender/weight were taken into account). The ground truth values of our dataset are therefore only *approximate estimates*. Furthermore, as with most data-driven algorithms, our models may learn shortcuts and biases presenting in the data (in our cases oftentimes category- and context-related biases), which may cause a false sense of security. Direct caloriometry [7] or heart rate-based estimation [1] are more accurate ways to estimate caloric cost than visual models.

2. Broader impact

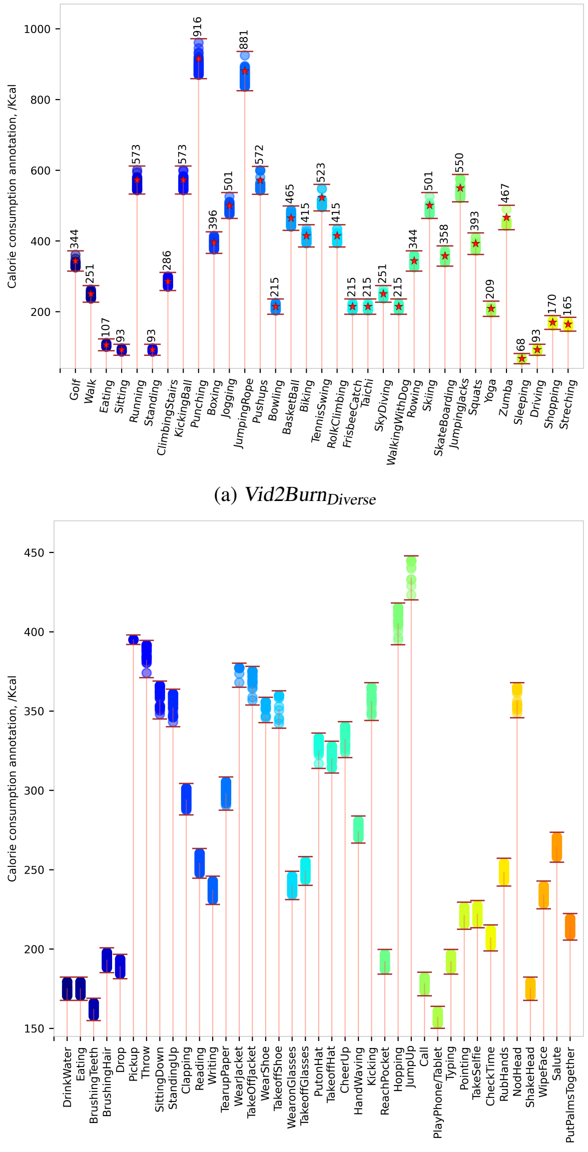
Our work introduces two video-based calorie consumption estimation benchmarks – *Vid2Burn*$_{Diverse}$ and *Vid2Burn*$_{ADL}$, together with several deep learning-based baselines targeting at end-to-end calorie consumption estimation. A wide range applications for health monitoring and human physical movement level prediction will directly benefit from this work. Moreover, since our work also tackle the generalization issue through evaluating the calorie estimation performance on the unseen activity types which can simulate the scenario for facing with out-of-distribution samples. The baselines leveraged in our work show a certain performance difference between the evaluations of known and unknown action types, indicating that offensive predictions, biased content and possible misclassifications can result in false sense of security while it still points out a valuable future research direction to us for further investigation. To allow future work constructed based on our benchmarks and baselines, we will make our code, models, and data publicly available.

3. License of existing asserts

Since we use multiple public datasets and online resources to form the video dataset and annotation set, we have carefully cited the related works for these leveraged datasets and marked the website link of the online resources in the corresponding footnote in our paper.

4. Human subjects data collection clarification

In *Vid2Burn*$_{ADL}$ dataset, we collect the heart rate, body weight and age data from 4 subjects to improve the accuracy of our calorie consumption annotation. The collected data is only leveraged to generate the global calorie consumption annotation which is highly aggregated and can not directly identify a specific person. The data and annotation are all anonymous. During the data collection procedure, each subject is well instructed to collect the heart rate data through wrist band (MIBAND 4) which can't bring any negative impact to the human body. From the dataset *Vid2Burn* which will be published soon, no person data is involved since all the data are highly aggregated. All participants are voluntary and signed a data collection agreement. We did not place the signed form for voluntary data collection in the supplementary materials in order to ensure anonymous submission.



(a) *Vid2Burn*$_{Diverse}$

(b) *Vid2Burn*$_{ADL}$

Figure 1. An overview of the calorie consumption annotation for *Vid2Burn* dataset. The boundaries of the fluctuations for each category is marked as brown lines, which define the sample-wise calorie consumption annotation based on the intensity of skeleton movement.

5. Supplementary for *Vid2Burn* dataset

5.1. Comparison between *Vid2Burn* and other human energy expenditure datasets

In order to further clarify the strengthens of our proposed benchmarks, we make a comparison between the proposed two benchmarks, – *Vid2Burn*$_{Diverse}$ and *Vid2Burn*$_{ADL}$, and the other two existed video-based benchmarks which are Standford-ECM [6] and Sphere [10]. The camera setting for our proposed benchmarks and Sphere are all fixed-position while Standford-ECM leveraged egocentric perspective requiring the camera to be mounted on a wearable device which limits the comfort of the user and requires contact, if application has been taken into consideration. Concerning the action numbers, our *Vid2Burn* contains in total 72 kinds of activities which contains simultaneously high- and low-intensity activities, together with >9K video clips which is much larger compared with the other two datasets offering more possibility to achieve deep learning based end-to-end calorie consumption estimation. In addition, our benchmarks provide sample-wise calorie consumption annotation which is more precise compared with the other datasets that only provide category-level human energy expenditure annotations. We also provide the description of the two proposed benchmarks in Figures 3 and 4 to show part of the label-sample pair with sample-wise calorie consumption annotation for each benchmark. There are 33 and 39 label-sample pairs for *Vid2Burn*$_{Diverse}$ and *Vid2Burn*$_{ADL}$ separately.

5.2. Supplementary for Vid2Burn-ADL dataset

First, a detailed introduction of sample numbers under each activity type, indicated by the number of samples on each histogram, and the corresponding category-wise annotation, denoted by the number on each image, are introduced in Fig. 2. The sample numbers for different actions show a balanced distribution with minimum sample number as 122 and maximum sample number as 159. Second, the statistic analysis for *Vid2Burn*$_{ADL}$ dataset is shown in Fig. 1b. Similar to *Vid2Burn*$_{Diverse}$ dataset, we use 39 categories coming from NTU RGBD [9] dataset to construct the *Vid2Burn*$_{ADL}$ dataset, where the color dot indicates the sample-wise calorie consumption annotation. Compared with *Vid2Burn*$_{Diverse}$ introduced in Fig. 1a, *Vid2Burn*$_{ADL}$ shows relatively lower movement intensity. In Fig. 1b we can find that for sample-wise calorie consumption annotation, there is an overlapping for fluctuated calorie consumption ranges among different action types. Finally, we will give a detail description for the heart rate collection procedure. During the data collection process, each participant needs to wear the wrist band and monitor the heart rate. For a specific action, participants were asked to repeat the action for two minutes and maintain the same action frequency as the original video (randomly selected for each action category leveraged in our work based on NTU RGBD [9] dataset) to obtain stable heart rate data. The interval between each action is carefully selected to ensure that the heart rate has returned to the rest state heart rate based on measurement.

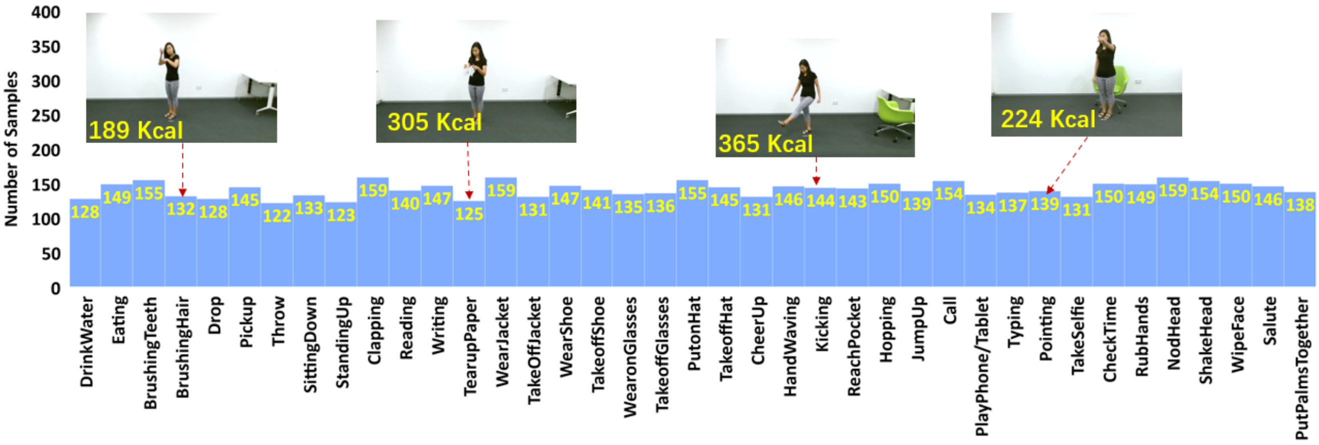


Figure 2. An overview of the dataset structure for *Video2Burn$_{ADL}$* dataset. The number in each histogram indicates the number of samples for the corresponding category and the number on each cluster of image represents the category-wise calorie consumption ground truth.

Datasets	Ours-DVS	Ours-ADL	Standford-ECM	Sphere
Modality	Video	Video	Video	video
Setting	Fixed-	Fixed-	Ego-	Fixed-
#Actions	33	39	24	11
#Clips	4260	5529	113	20
Unit	Calorie	Calorie	MET	MET
Label	s /c	s	c	c

Table 1. A comparison among datasets for human energy expenditure prediction, where c indicates category-wise annotation and s indicates sample-wise annotation. Ours-DVS indicates the *Vid2Burn$_{Diverse}$* benchmark and Ours-ADL indicates the *Vid2Burn$_{ADL}$* benchmark

Method	Known activity types			New activity types		
	MAE	SPC	NLL	MAE	SPC	NLL
Skeleton (Diverse)	330.7	-	-	665.1	-	-
Skeleton (ADL)	78.1	-	-	95.2	-	-
SF-AVR (Diverse)	57.9	49.41	6.08	134.0	42.20	9.02
SF-AVR (ADL)	20.1	68.61	5.57	36.4	72.30	6.69

Table 2. A comparison between deep learning-based approach (*I3D-AVR*) and skeleton-based forward computation approach (the same with the skeleton-based calorie consumption annotation generation procedure) on *Vid2Burn$_{Diverse}$* and *Vid2Burn$_{ADL}$* benchmarks.

6. Supplementary for the experiments and analyses

First, more details about the category-wise performance for calorie consumption estimation on the *Vid2Burn$_{Diverse}$* benchmark using category-wise annotation for supervision is represented in Table 5. Second, we provide additional comparison between deep learning-based and pure skeleton-based forward computation for calorie consumption prediction. Finally, we provide an additional ablation studies for different σ when generating soft label for supervision.

6.1. Comparison between deep learning-based approaches and skeleton-based computations

Since one of the annotation source for calorie consumption estimation is skeleton data, the performance of directly using skeleton to compute calorie consumption is interesting to be researched. We thereby conduct experiments on the two proposed benchmarks with sample-wise annotations between deep learning-based approaches and pure skeleton-based forward calculation. According to the experimental results introduced by Table 2, pure skeleton based forward calculation shows a performance difference by 272.8 kcal and 531.1 kcal on the known and unknown action types on the *Vid2Burn$_{Diverse}$* dataset and the performance difference on the *Vid2Burn$_{ADL}$* dataset for the known and unknown action types are 58.1 kcal and 58.8 kcal compared with *I3D-AVR* illustrating the outstanding performance of the deep learning-based approaches for video-based calorie consumption estimation.

6.2. Ablation studies on different standard error for soft label generation

In order to investigate the influence brought by different σ when generating soft label for calorie consumption prediction, we conduct corresponding ablation studies shown in Table 4 using *I3D-AVR* approach on the *Vid2Burn$_{Diverse}$* dataset under category-wise supervision and choose σ as 5, 15, 25 and 50 kcal separately. According to the experimental results, choosing σ as 15 shows the best performance on known activity types and 50 shows the best performance on unknown activity types in the MAE metric.

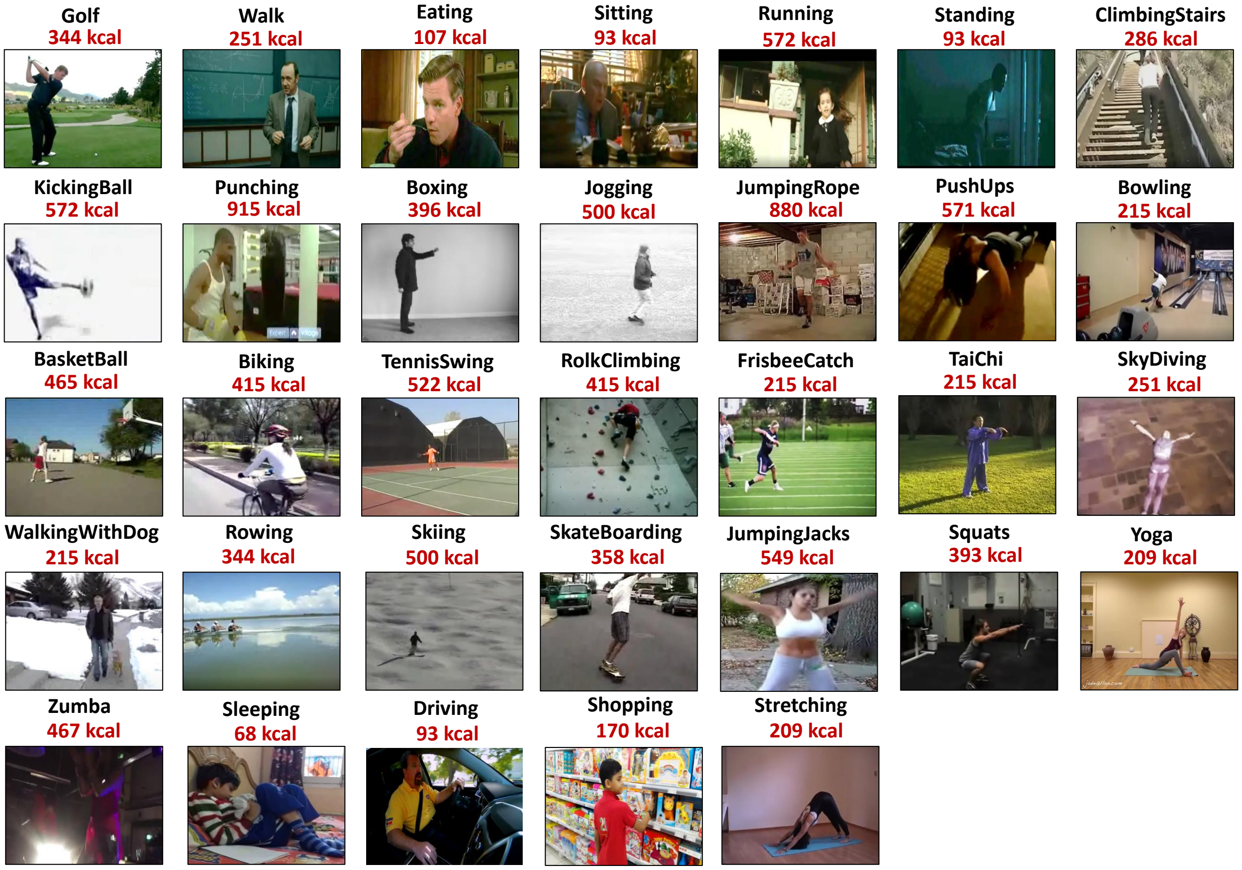


Figure 3. An overview of the calorie consumption annotation for *Vid2Burn*$_{Diverse}$ dataset for all 33 leveraged action types (sample-wise annotation). We mark the corresponding calorie consumption annotation under each category name for each selected sample.

Method	Known activity types			New activity types		
	MAE	SPC	NLL	MAE	SPC	NLL
Video	43.2	72.82	6.19	56.0	71.08	**6.49**
I3D-AVR (TFS)	102.8	43.08	6.67	69.5	52.69	6.65
I3D-AVR	22.9	72.97	5.66	39.6	70.45	6.38

Table 3. Experiments on *Vid2Burn*$_{ADL}$ with sample-wise label, where *Video* denotes the model *I3D-AVR* only fine-tuned on the calorie consumption estimation head consisted mainly of fc layers, *I3D-AVR (TFS)* denotes the *I3D-AVR* model while training from scratch.

STD	Known activity types			New activity types		
	MAE	SPC	NLL	MAE	SPC	NLL
50	35.9	85.99	6.64	183.2	56.97	8.25
25	32.5	71.60	6.07	229.7	40.44	9.56
15	29.3	60.74	5.63	194.5	34.16	10.61
5	38.9	32.25	4.70	228.7	7.61	16.32

Table 4. Ablation studies by adjusting the σ for label softing on *Vid2Burn*$_{Diverse}$ using category-wise label supervision.

6.3. Further illustration of the calorie consumption estimation ability

When digging deeper into the direction of the deep learning-based calorie consumption estimation, the relationship between action recognition and calorie consumption estimation is interesting to be investigated, especially for the question about whether there is only a lookup relationship between calorie consumption estimation and action recognition or not. First if looking into labels, we have sample-wise label differing among the samples inside same action type according to different human body movement intensity, which makes sure that it will not be a simple lookup relationship. According to Fig. 1b, there are calorie consumption range overlapping among different action types. Second we conduct several ablation studies listed in Table 3 to support our argument. If our models predict lookup relationship between calorie consumption and action classes, the performance of the model only fine-tuning the fc layers should be higher than the perfor-

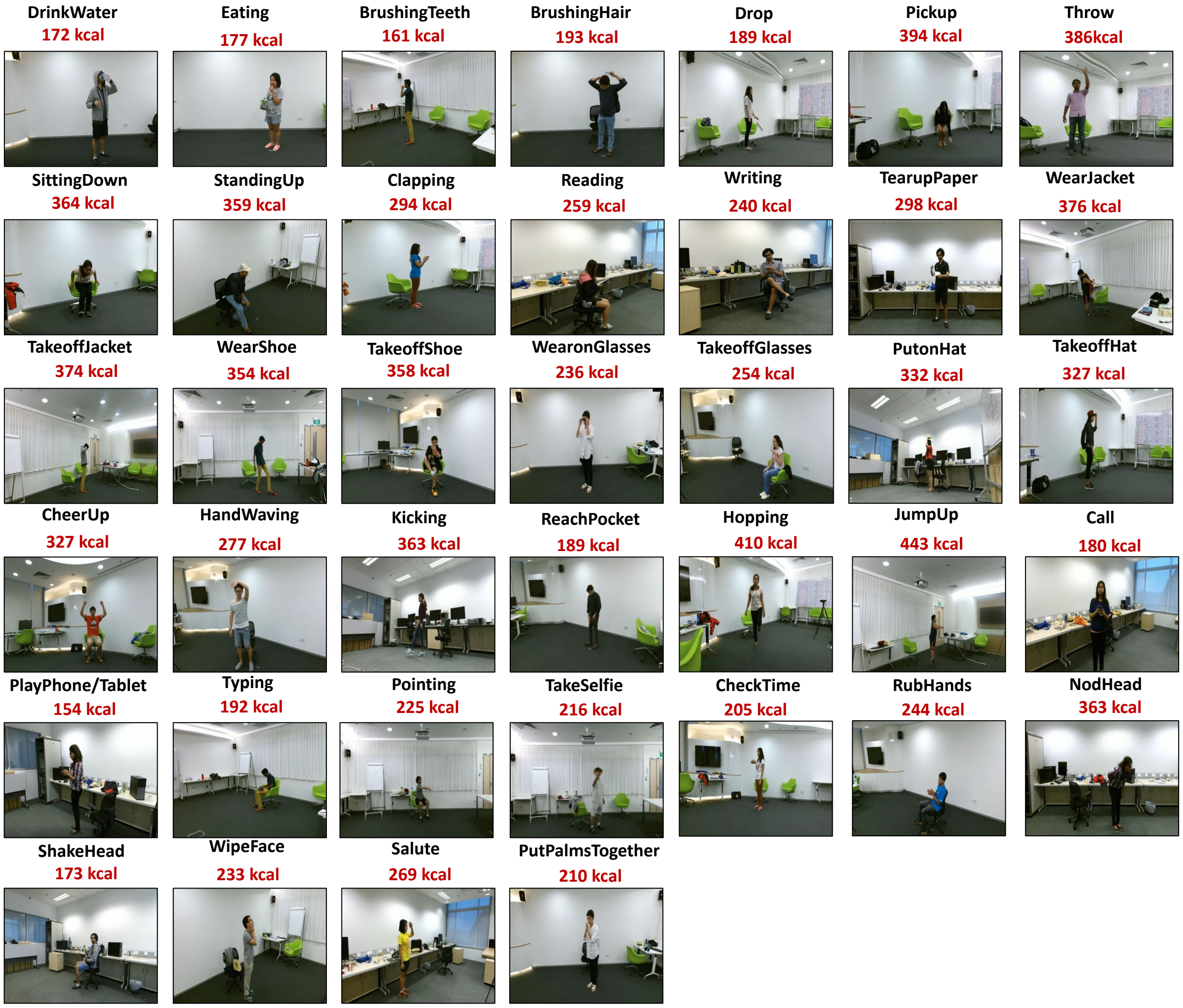


Figure 4. An overview of the calorie consumption annotation for *Vid2Burn*$_{ADL}$ dataset for all 39 leveraged action types (sample-wise annotation). We mark the corresponding calorie consumption annotation under each category name for the selected sample.

Method	Known action types (five common classes)										Unknown action types (Three common classes)					
	Sitting		Running		Climbing		KickBall		Punching		Yoga		Sleeping		Shopping	
	SPC	MAE	SPC	MAE	SPC	MAE	SPC	MAE	SPC	MAE	SPC	MAE	SPC	MAE	SPC	MAE
ST-GCN	9.41	245.6	-13.51	310.9	14.81	225.6	-9.71	554.2	10.93	367.5	19.65	230.5	-14.38	370.2	-0.24	322.3
I3D-AVR	11.09	176.8	7.69	82.6	13.96	81.8	16.51	216.3	46.79	1.1	13.10	191.9	1.05	273.2	1.92	190.5
SF-AVR	22.02	88.6	11.76	43.8	24.19	57.7	15.22	202.0	30.08	0.9	20.18	101.9	3.37	205.6	3.43	174.4
R(2+1)D-AVR	21.18	124.6	12.88	288.0	25.03	74.0	-0.02	31.8	32.70	6.3	15.20	217.9	10.97	314.2	4.68	271.1
R3D-AVR	18.14	166.5	6.14	102.3	21.13	135.5	2.87	246.4	38.84	0.2	17.82	97.4	2.71	168.3	1.17	108.5
I3D-LSTM	14.60	74.7	6.09	147.7	14.85	99.9	10.15	250.0	48.97	0.6	12.04	217.6	0.14	287.2	4.15	280.0
SF-LSTM	15.02	182.0	12.59	9.6	15.79	86.7	8.69	161.4	45.81	7.6	19.49	133.6	0.26	373.7	3.91	266.8
R(2+1)D-LSTM	8.97	218.5	11.22	37.1	10.47	98.4	9.86	347.1	45.54	5.3	18.26	181.8	5.71	152.6	8.53	134.4
R3D-LSTM	8.78	128.3	6.54	262.1	8.82	83.8	6.48	204.5	41.82	3.4	15.66	223.2	-0.04	340.9	0.02	370.1

Table 5. Experimental results for human calorie consumption estimation for the selected action categories on the *Vid2Burn*$_{Diverse}$ dataset supervised with category-wise annotation.

mance of our approach. Since video classes are highly dependent on action classes, we conduct experiment by freezing weights of pretrained video-based backbone while only adjusting weights of fully-connected layers as *Video* in Table 3, where MAE of *Video* for both known- and unknown-action types evaluation are all worse than *I3D-AVR*. We also test train-from-scratch for the *I3D-AVR* baseline denoted as *I3D-AVR (TFS)* which shows the worst performance when compared with others, illustrating that pretraining is important. Through the above analyses it can be seen that the relationship between human action and calorie consumption prediction is not a simple lookup relationship and also pretraining is essential.

6.4. Supplementary for implementation details

In addition to the mentioned implementation details in our paper, our model is built based on PyTorch toolbox. Since we leverage temporal sliding window to aggregate features along time axis, the temporal overlapping of the sliding window for I3D, R3D, R(2+1)D backbones are chosen as 6 frames while the temporal overlapping for SlowFast is chosen as 16 frames since it requires larger temporal window length (32 frames) compared with the others (16 frames). For the $\textit{Vid2Burn}_{ADL}$ dataset, the estimation head has 500 channels output as the maximum calorie consumption estimation range is set as 500 kcal together with resolution 1 kcal. For the $\textit{Vid2Burn}_{Diverse}$ dataset, the channel number of the final output is 1000.

References

[1] Barbara E. Ainsworth, William L. Haskell, Stephen D Herrmann, Nathanael Meckes, David R. Bassett, Catrine Tudor-Locke, Jennifer L. Greer, Jesse Vezina, Melicia C. Whitt-Glover, and Arthur S. Leon. 2011 compendium of physical activities: a second update of codes and MET values. *Medicine and Science in Sports and Exercise*, 2011. 1

[2] Muhammad Awais, Lorenzo Chiari, Espen Alexander F. Ihlen, Jorunn L. Helbostad, and Luca Palmerini. Physical activity classification for elderly people in free-living conditions. *IEEE Journal of Biomedical and Health Informatics*, 2018. 1

[3] Carl J. Caspersen, Kenneth E. Powell, and Gregory M. Christenson. Physical activity, exercise, and physical fitness: definitions and distinctions for health-related research. *Public Health Reports*, 1985. 1

[4] Ara Jo, Bryan D. Coronel, Courtney E. Coakes, and Arch G. Mainous III. Is there a benefit to patients using wearable devices such as fitbit or health apps on mobiles? A systematic review. *The American Journal of Medicine*, 2019. 1

[5] Tae Hee Jo, Jae Hoon Ma, and Seung Hyun Cha. Elderly perception on the internet of things-based integrated smart-home system. *Sensors*, 2021. 1

[6] Katsuyuki Nakamura, Serena Yeung, Alexandre Alahi, and Li Fei-Fei. Jointly learning energy expenditures and activities using egocentric multimodal signals. In *CVPR*, 2017. 2

[7] A. C. Pinheiro Volp, F. C. de Oliveira, R. Duarte Moreira Alves, E. A. Esteves, and Josefina Bressan. Energy expenditure: components and evaluation methods. *Nutricion Hospitalaria*, 2011. 1

[8] Blaine Reeder, Jane Chung, Kate Lyden, Joshua Winters, and Catherine M. Jankowski. Older women's perceptions of wearable and smart home activity sensors. *Informatics for Health and Social Care*, 2020. 1

[9] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. NTU RGB+D: A large scale dataset for 3D human activity analysis. In *CVPR*, 2016. 2

[10] Lili Tao, Tilo Burghardt, Majid Mirmehdi, Dima Damen, Ashley Cooper, Massimo Camplani, Sion Hannuna, Adeline Paiement, and Ian Craddock. Energy expenditure estimation using visual and inertial sensors. *IET Computer Vision*, 2018. 2