



Towards Video-based Activated Muscle Group Estimation in the Wild

Kunyu Peng
Karlsruhe Institute of Technology
Karlsruhe, Germany
kunyu.peng@kit.edu

David Schneider
Karlsruhe Institute of Technology
Karlsruhe, Germany
david.schneider@kit.edu

Alina Roitberg
Stuttgart University
Stuttgart, Germany
alina.roitberg@ki.uni-stuttgart.de

Kailun Yang*
Hunan University
Changsha, China
kailun.yang@hnu.edu.cn

Jiaming Zhang
Karlsruhe Institute of Technology
Karlsruhe, Germany
jiaming.zhang@kit.edu

Chen Deng
Beijing Sport University
Beijing, China
dengchen@bsu.edu.cn

Kaiyu Zhang
Beijing Sport University
Beijing, China
kaiyuzhang@bsu.edu.cn

M. Saquib Sarfraz
Mercedes-Benz Tech Innovation
Stuttgart, Germany
saquib.sarfraz@kit.edu

Rainer Stiefelhagen
Karlsruhe Institute of Technology
Karlsruhe, Germany
rainer.stiefelhagen@kit.edu

Abstract

In this paper, we tackle the new task of video-based Activated Muscle Group Estimation (AMGE) aiming at identifying active muscle regions during physical activity in the wild. To this intent, we provide the MuscleMap dataset featuring >15K video clips with 135 different activities and 20 labeled muscle groups. This dataset opens the vistas to multiple video-based applications in sports and rehabilitation medicine under flexible environment constraints. The proposed MuscleMap dataset is constructed with YouTube videos, specifically targeting High-Intensity Interval Training (HIIT) physical exercise in the wild. To make the AMGE model applicable in real-life situations, it is crucial to ensure that the model can generalize well to numerous types of physical activities not present during training and involving new combinations of activated muscles. To achieve this, our benchmark also covers an evaluation setting where the model is exposed to activity types excluded from the training set. Our experiments reveal that the generalizability of existing architectures adapted for the AMGE task remains a challenge. Therefore, we also propose a new approach, TRANS^{M3}E, which employs a multi-modality feature fusion mechanism between both the video transformer model and the skeleton-based graph convolution model with novel cross-modal knowledge distillation executed on multi-classification tokens. The proposed method surpasses all popular video classification models when dealing with both, previously seen and new types of physical activities. The database and code can be found at <https://github.com/KPeng9510/MuscleMap>.

*Corresponding author: kailun.yang@hnu.edu.cn

CCS Concepts

• Computing methodologies → Activity recognition and understanding; Scene understanding.

Keywords

Activate muscle group estimation, video-based human activity understanding, video-based scene understanding.

ACM Reference Format:

Kunyu Peng, David Schneider, Alina Roitberg, Kailun Yang, Jiaming Zhang, Chen Deng, Kaiyu Zhang, M. Saquib Sarfraz, and Rainer Stiefelhagen. 2024. Towards Video-based Activated Muscle Group Estimation in the Wild. In *Proceedings of the 32nd ACM International Conference on Multimedia (MM '24)*, October 28-November 1, 2024, Melbourne, VIC, Australia. ACM, New York, NY, USA, 10 pages. <https://doi.org/10.1145/3664647.3680630>

1 Introduction

Human activity understanding is important as it enables the development of applications and systems that can enhance health-care, improve security, and optimize various aspects of daily life by automatically identifying and understanding human actions and behaviors [1, 28, 69]. Knowing which skeletal muscles of the human body are activated benefits human activity understanding, and sport and rehabilitation medicine from multiple perspectives and prevents inappropriate muscle usage which may cause physical injuries [18]. In health care, patients need to know how to conduct the exercise correctly to recover from surgery [33] or specific diseases [3], e.g., COVID-19 [57]. Knowledge about muscle activations allows for user-centric fitness applications providing insights for everyday users or professional athletes who need specially adapted training. The majority of existing work on Activated Muscle Group Estimation (AMGE) is based on wearable devices with electrode sensors [14]. Yet, many wearable devices are inconvenient and heavy [38], even harmful to health [4], and have limited usage time due to the battery [64]. A big strength of wearable devices is the high accuracy achieved through direct signal measurement from skin or muscle tissue. However, such exact bio-electrical changes are not required in a large number of medical recovery programs,



This work is licensed under a Creative Commons Attribution International 4.0 License.

MM '24, October 28-November 1, 2024, Melbourne, VIC, Australia
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0686-8/24/10
<https://doi.org/10.1145/3664647.3680630>

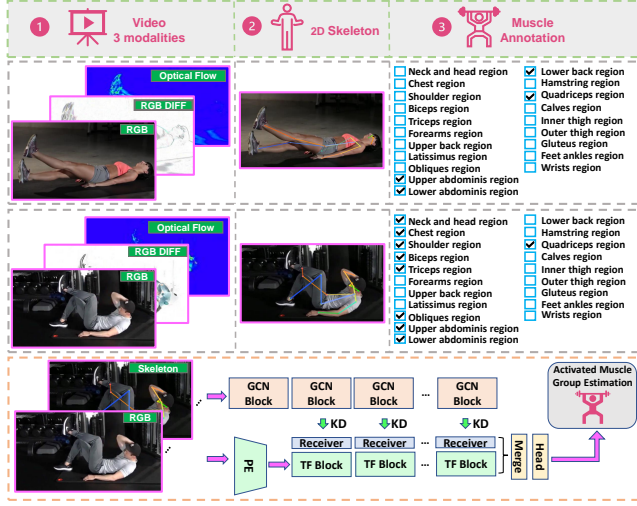


Figure 1: Overview of the proposed MuscleMap dataset (Top) and the TRANSM³E model (Bottom). Our dataset contains four data modalities, i.e., RGB, RGB difference (RGB Diff), optical flow, and 2D skeleton. PE and TF denote the patch embedding layer and the transformer block, respectively.

and knowing the binary activation status of the muscle as shown in Figure 1 is sufficient in many situations [44, 61, 78]. In contrast to wearable devices, most people have a video camera available at hand on their phone or laptop. Applying video-based AMGE on in-the-wild data collected by using smartphones or other widely available smart devices would allow for the application of such programs even without access to specialized hardware. Thereby, end-to-end video-based AMGE approaches are expected to be developed to prevent overburdens caused by wearable devices from both physical and psychological points of view. *Can modern deep learning algorithms relate fine-grained physical movements to individual muscles?* To answer this question, we tackle the barely researched task of video-based active muscle group estimation under an in-the-wild setting, which estimates muscle contraction during physical activities from video recordings without a restricted environment and background constraints.

Current research in video-based AMGE is limited by small-scale datasets and constrained data collection settings [12], where the data is often annotated with sensor signals and confined to restricted environments, covering only a limited range of actions. However, with the expansion of deep learning model capacities, there is a pressing need for larger datasets encompassing a wider variety of environments and activities. This expansion is vital for advancing the field of video-based AMGE within the research community.

In this work, we collect the first large-scale in-the-wild AMGE dataset from YouTube without environment constraints and give binary activation for different muscle regions by inquiring about sports field researchers. We created the MuscleMap dataset — a video-based dataset with 135 different exercises collected from

Table 1: A comparison among the statistics of the video-based datasets, where AR, AQA, and CE indicate activity recognition, activity quality assessment, and calorie consumption estimation.

Dataset	NumClips	Task	MultiLabel	NumActions
KTH [30]	599	AR	False	6
UCF101 [67]	13,320	AR	False	101
HMDB51 [32]	6,849	AR	False	51
ActivityNet [7]	28,108	AR	False	200
Kinetics400 [7]	429,256	AR	False	400
Video2Burn [51]	9,789	CE	False	72
MTL-AQA [48]	1,412	AQA	True	/
FineDive [79]	3,000	AQA	True	29
FineGym [65]	32,697	AQA	True	530
MiA [12]	15,000	AMGE	False	15
MuscleMap135 (Ours)	15,004	AMGE	True	135

YouTube considering in-the-wild videos. Each exercise type is annotated with one or multiple out of 20 different muscle group activations, as described in Table 1, which opens the door for video-based activated muscle group estimation in the wild task to the community. We annotate the dataset in a multi-label manner since human body movement is produced by the coordinated operation of diverse muscle regions. To acquire such annotations, we ask two senior researchers in the biomedical and sports research field to give the annotations.

We select various off-the-shelf Convolutional Neural Networks (CNNs)[8, 22], Graph Convolutional Networks (GCNs)[10, 34, 82], and transformer-based architectures [19, 37, 42] from the human activity recognition field, along with statistical methods, as baselines. Our proposed MuscleMap benchmark addresses a multi-label classification problem where each sample may have one to twenty labels. These models struggle with new activity types featuring new muscle combinations at test time, impacting AMGE generalizability. Skeleton-based models perform well on new activity types but not on known ones, whereas video-based models excel on known types but perform poorly on new ones. An approach that performs well on both known and new activity types is needed.

To tackle the aforementioned issue, we propose TRANSM³E, a cross-modality knowledge distillation and fusion architecture that combines RGB and skeleton data via a new classification tokens-based knowledge distillation and fusion mechanism. To achieve better extraction of underlying cues for AMGE, we propose and equip TRANSM³E with three essential novel components, i.e., *Multi-Classification Tokens (MCT)*, *Multi-Classification Tokens Knowledge Distillation (MCTKD)*, and *Multi-Classification Tokens Fusion (MCTF)*, atop the most competitive performing architecture MVITv2 [37] as the backbone. As it is fundamental to mine and predict the activities at the global level for AMGE, the proposed TRANSM³E, appearing as a transformer-based approach, is endowed with the capacity for long-term reasoning of visual transformers [73]. Since AMGE is a multi-label classification task, MCT is introduced, in view that using more classification tokens is expected to introduce more benefits toward finding informative cues. MCT also builds up the base for cross-modality MCT-level knowledge distillation.

Knowledge distillation [27] is leveraged for cross-modality knowledge transfer after the feature map reduction of the transformer

block to enable a more informative latent space learning for different modalities. Transferring cross-modality knowledge during training significantly benefits the model in finding out cross-modality informative cues for the AMGE task. However, we find that it is difficult to achieve the knowledge distillation between two models with obvious architecture differences, e.g., GCNs and video transformers, considering the alignment of the feature maps coming from different backbones and modalities to achieve the appropriate and effective knowledge distillation. Aside from the architectural differences, we examine that using late fusion to fuse the skeleton-based model and video-based model can not achieve a satisfactory performance due to the lack of alignment of the two different feature domains.

We propose a cross-modality MCT-level knowledge distillation scheme considering intermediate and final layer distillation by designing a specific knowledge distillation MCT for each modality. Alongside the classification MCT, another MCT executes knowledge distillation for each modality. Cross-modality knowledge distillation occurs only between the knowledge distillation MCTs from the two modalities, unlike existing works that use full embeddings or a single token at the final layer with a larger teacher [27, 39]. Our MCTKD mechanism integrates cross-modal knowledge into the main network, while MCTF merges the distilled knowledge MCT and classification MCT for final prediction of active muscle regions during human body motion. Combining these components, TRANS^{M3}E achieves state-of-the-art performance with superior generalizability compared to tested baselines. In summary, our contributions are listed as follows:

- We propose a new task of video-based Activated Muscle Group Estimation in the wild task with the aim of lowering the threshold of entry to muscle-activation-based health care and sports applications.
- We provide a new benchmark MuscleMap to propel research on the aforementioned task which includes the large-scale *MuscleMap* dataset. We also present baseline experiments for this benchmark, including CNN-, transformer-, and GCN-based approaches.
- We especially take the evaluation of the generalizability into consideration by constructing test and validation sets using new activities excluded during the training.
- We propose TRANS^{M3}E, targeting improving the AMGE generalizability towards new activity types. *Multi-classification Tokens* (MCT), *Multi-Classification Tokens Knowledge Distillation* (MCTKD) and *Multi-Classification Tokens Fusion* (MCTF) are used to formulate TRANS^{M3}E, which shows superior generalizability on new activities and introduces state-of-the-art results on the MuscleMap benchmark.

2 Related Work

Activate Muscle Group Estimation (AMGE) analysis is predominantly performed using electromyographic (EMG) data [5, 71] either with intramuscular (iEMG) or surface EMG sensors (sEMG). These methods use EMG data as input and detect activated muscle groups to achieve an understanding of the human body movement and the action, while we intend to infer muscle activations from body movements, therefore describing the opposite task. Chiquier *et al.* [12]

propose a video-based AMGE dataset by using the signal of the wearable devices as the annotation. Yet, the data collection setting and the environment are restricted. The scale of the introduced dataset is relatively small and it encompasses limited action types. In our work, we collect a large-scale dataset based on HIIT exercises on YouTube while delivering binary annotation for each muscle region. We reformulate it into a multi-label classification task, namely AMGE in the wild. The annotations are first derived from online resources and then checked and corrected by researchers in sports fields.

Activity Recognition is a dominating field within visual human motion analysis [1, 28] which was propelled by the advent of Convolutional Neural Networks (CNNs) with 2D-CNNs [24] in combination with recurrent neural networks (RNNS) [16] or different variations of 3D-CNNs [8, 22, 52]. More recently, transformer-based methods advanced over 3D-CNNs, especially with advanced pre-training methods and large datasets [37, 41, 42, 50]. Action Quality Assessment (AQA) [48, 70] and Visual Calory Estimation (VCE) [51] relate to our work since these methods likewise shift the question of research from *what?* to *how?* with the aim of detailed analysis of human motion. Multimodal data is a common strategy, e.g., by combining RGB video with audio [2, 49, 56], poses [62], optical flow [56], or temporal difference images [47]. Skeleton data is also commonly used as a modality for activity recognition on their own. Yan *et al.* [82] and follow-up research [54, 66, 77, 81] make use of GCNs, while competitive approaches leverage CNNs with special pre-processing methods [13, 17].

Knowledge distillation (KD) [27] became a common technique to reduce the size of a neural network while maintaining performance. In review [26], methods can be categorized to focus on knowledge distillation based on final network outputs (response-based) [29, 87], based on intermediate features (feature-based) [83, 86], or based on knowledge about the relations of data samples or features (relation-based) [9]. Recently, adaptations of distillation for transformer architectures gained attraction [35, 39]. Fusion strategies can be grouped into feature-fusion [55] and score fusion [31].

Multi-label classification methods allow for assigning more than a single class to a data sample. Common strategies include per-class binary classifiers with adapted loss functions to counter the imbalance problem [58], methods that make use of spatial knowledge [84, 85], methods that make use of knowledge about label relations [11, 68], or methods based on word embeddings [40, 80].

Datasets which combine visual data of the human body with muscle activation information are sparse and mainly limited to specific sub-regions of the human body, e.g., for hand gesture recognition [25]. In contrast, a large variety of full-body human activity recognition datasets were collected in recent years, which are labelled with high-level human activities [32, 53], fine grained human action segments [36, 88], or action quality annotations [65]. We leverage such datasets by extending them with muscle group activation labels.

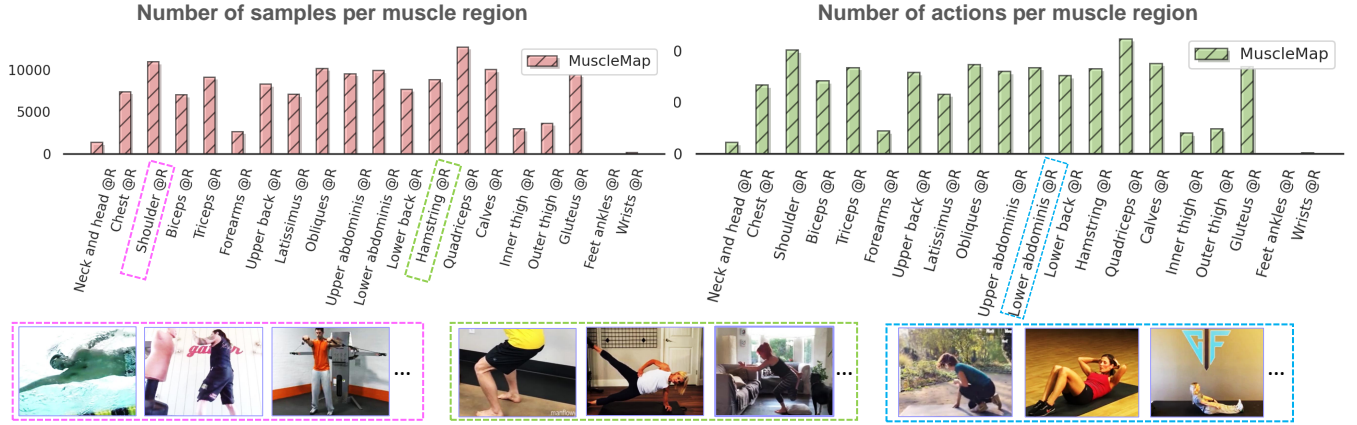


Figure 2: An overview of the number of samples and the number of activity types per muscle region (@R), depicted at the top left and the top right.

3 Benchmark

3.1 MuscleMap Dataset

With the new video-based active muscle group estimation in-the-wild task in mind, we collect the MuscleMap dataset by querying YouTube for the physical exercise video series. The collected dataset contains 135 activity types as well as 15,004 video clips and is competitive compared to other video-based datasets targeting fine-grained tasks, as shown in Table 1. Twenty activities are reserved for the validation and test splits of new activities, which are not included in the training set. MuscleMap targets physical exercise videos from fitness enthusiasts. High-Intensity Interval Training (HIIT) exercises are well suited for the AMGE in-the-wild task since they display a large range of motions that are designed to activate specific muscle groups and instructional videos provide high-quality examples of the displayed motion. The collected videos in our dataset are mostly near-person, which can benefit video-based muscle contribution understanding for the in-the-wild videos. To deliver diverse modalities, we provide RGB, RGB Diff, optical flow extracted by DenseFlow [75], and 2D skeleton data extracted through AlphaPose [20]. A small set of activities from the MuscleMap dataset is shown in the bottom part of Figure 2. In Table 1, MuscleMap is compared with existing human activity recognition, action quality assessment, calorie consumption datasets, and time series-wise muscle activation regression dataset. We ensured that all YouTube videos used were publicly available and complied with the platform’s terms of service.

3.2 Activated Muscle Group Annotation

We cluster skeletal muscles of the human body into 20 major muscle groups with binary activation as shown in the checkboxes in Figure 1. To ensure the quality of the annotation, we ask 2 researchers from the biomedical and sports fields to give the annotation for each activity by watching the video from the dataset. If the two biomedical and sports researchers fully agree with the AMGE annotation towards one activity, this activity is included in our dataset.

Both of the two annotators are senior researchers in the biomedical and sports fields.

3.3 Evaluation Protocol

To evaluate the generalizability of the leveraged approaches for the AMGE in-the-wild task, we formulate the **new val/test** and **known val/test**, where we use val and test to indicate the validation set and the test set, respectively. For MuscleMap, 20 of 135 activities are leveraged to formulate the **new val/test** set, which are *hollow hold, v-ups, calf raise hold, modified scissors, scissors, reverse crunches, march twists, hops on the spot, up and down planks, diamond push ups, running, plank jacks, archer push ups, front kicks, triceps dip hold, side plank rotation, raised leg push ups, reverse plank kicks, circle push ups, and shoulder taps*. The activity types for the **known test** and **known val** are the same as the activity types in the training set. The sample number for train, **new val**, **known val**, **new test**, **known test** sets are 7,069, 2,355, 1,599, 2,360, and 1,594. The performances are finally averaged for new and known sets (**mean test** and **mean val**). We randomly pick up half of the samples from each **new** activity type to construct the **new val** while the rest of the samples from the selected **new** activities are leveraged to construct the **new test**. After the training of the leveraged model, we test the performance of the trained model on **known/new** evaluation and **known/new** test sets, and then average the performance of **known** and **new** sets to get the averaged performance on evaluation and test sets by considering both **known** and **new** activities which are both important for the AMGE in-the-wild task.

3.4 Evaluation Metric

Mean averaged precision (mAP) is used as the evaluation metric for the AMGE in-the-wild task. We let $\mathbf{I} = \{I_i | i \in [1, \dots, N_I]\}$ denote the multi-hot annotation for the sample i and $\mathbf{y} = \{y_i | i \in [1, \dots, N_I]\}$ denote the prediction of the model for the given sample i . We first select the subset of \mathbf{y} and \mathbf{I} by calculating the mask through $\mathbf{m} = \text{where}(\mathbf{I} = 1)$. The corresponding subsets are thereby denoted as $\mathbf{y}[\mathbf{m}]$ and $\mathbf{I}[\mathbf{m}]$. Then we calculated the mean averaged precision score using the function and code from sklearn [6].

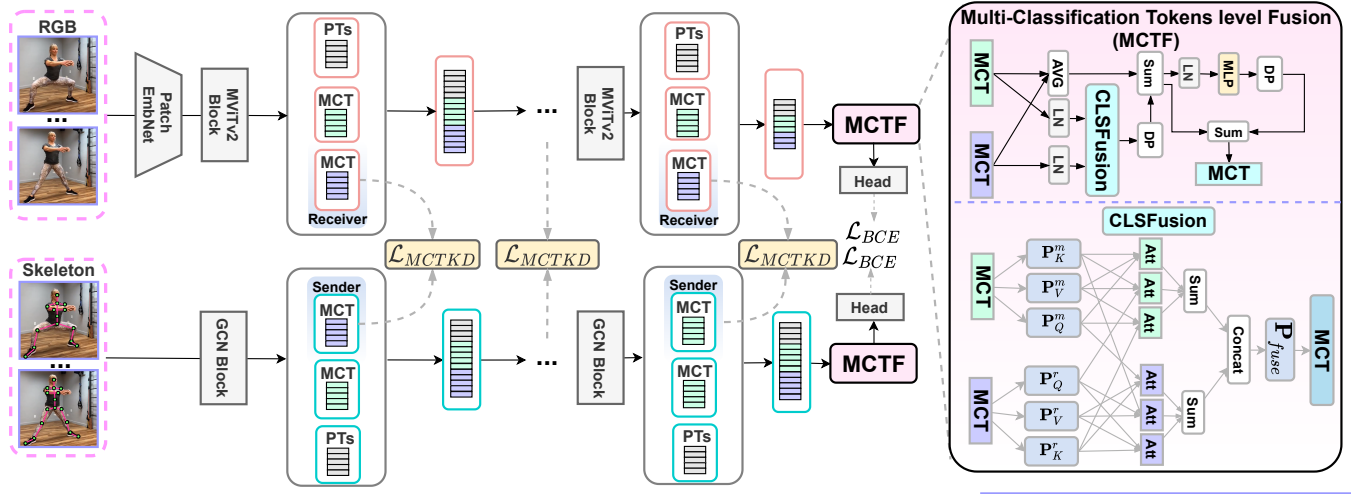


Figure 3: An overview of the proposed TRANSM³E architecture.

4 Architecture

4.1 Preliminaries of MVIT

TRANSM³E is based on MVITv2 [37]. The model architecture of TRANSM³E is shown in Figure 3. MVITv2 uses decomposed relative position embeddings and residual pooling connections to integrate shift-invariance and reduce computational complexity, while MVITv1 achieves downscaling by large strides on Keys (K) and Values (V).

4.2 Multi-Classification Tokens (MCT)

MCTs are used to harvest more informative components to achieve good generalizability for AMGE and to construct sender and receiver for cross-modality knowledge distillation in our work as shown in Figure 3. In our MCT setting, we directly use the final layer output of MCT and aggregate the MCT along the token dimension together with SoftMax to achieve multi-label classification.

Assuming the classification tokens of MCT to be referred to by $\{\text{cls}_j | j \in [1, \dots, C]\}$ and the flattened patch embeddings to be referred to as $\{\mathbf{p}_i | i \in [1, \dots, N_{\text{patches}}]\}$ for the given input video, where N_{patches} is the length of the patch sequence, the input of the first MVITv2 block is $[\text{cls}_1, \dots, \text{cls}_C, \mathbf{p}_1, \dots, \mathbf{p}_{N_{\text{patches}}}]$. The final prediction \mathbf{y} is computed through,

$$\mathbf{y} = \text{SoftMax}(\mathbf{P}_\alpha(\sum_{i=1}^C \text{cls}_i / C), \dim = -1), \quad (1)$$

where \mathbf{P}_α indicates a fully connected (FC) layer projecting the merged MCT to a single vector with the number of muscle regions as dimensionality. We make use of the same MCT settings for both the video-based backbone and the skeleton-based backbone according to Figure 4. After the first GCN block, the MCT for knowledge distillation and the MCT for classification are added to the model. We first flatten the spatial temporal nodes from the graph structure preserved by the GCN block. We use \mathbf{z}_{GCN}^* to denote the nodes of the constructed graph structure, cls_m^* to denote the MCT for classification, and cls_r^* to denote the MCT for knowledge distillation

regarding skeleton branch. We then concatenate all of these components along the node dimension and execute feature projection by using linear projection layer \mathbf{P}_m as follows,

$$\mathbf{z}_{GCN}^*, \text{cls}_m^*, \text{cls}_r^* = \text{Split}(\mathbf{P}_m(\text{Concat}(\mathbf{z}_{GCN}^*, \text{cls}_m^*, \text{cls}_r^*))). \quad (2)$$

Then we execute an internal knowledge merge from the nodes to the MCT for the classification, as follows,

$$\hat{\text{cls}}_m = \mathbf{P}_s(\mathbf{z}_{GCN}^*) + \text{cls}_m^*, \quad (3)$$

where \mathbf{P}_s denotes a FC layer. Finally, the node features, MCT for classification, and MCT for the knowledge distillation will be transferred to the next GCN block and the same procedure will be executed.

4.3 Multi-Classification Tokens Knowledge Distillation (MCTKD)

Multi-Classification Tokens Knowledge Distillation (MCTKD) is one of our main contributions. We are the first to introduce this technique, enabling knowledge distillation on multi-classification tokens between two structurally different backbones. Directly merging features from skeleton-based and video-based models underperforms due to structural and modality differences. To achieve effective feature fusion between these distinct architectures, we need a new solution. Our work explores knowledge distillation for feature space alignment from the MCT perspective, aiding cross-modality feature fusion for the AMGE in-the-wild task.

In the past, transformer-based knowledge distillation mainly focused on using intermediate full patch embeddings [45] or final classification token [72], while we propose knowledge distillation on the proposed MCT for both intermediate and final layers by using additional MCT for the knowledge distillation.

The underlying benefit of MCTKD is that the token number of the MCT is fixed, while knowledge distillation on the patch embeddings [21] may encounter the alignment issue when facing different modalities with different token sizes. Instead of directly

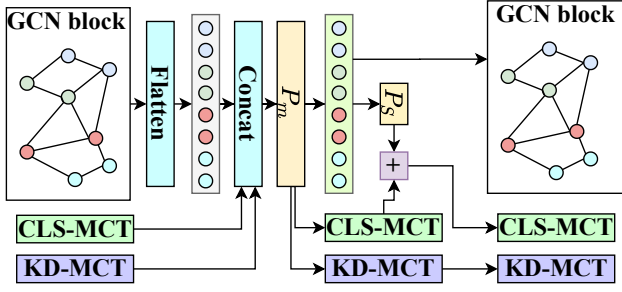


Figure 4: An overview of the modified GCN block with knowledge distillation MCT and classification MCT.

distilling knowledge from the MCT of an auxiliary modality towards the MCT of a major modality, knowledge distillation MCT is introduced to serve as a knowledge receiver. This approach avoids disruption on the MCT for classification for the major modality, *i.e.*, RGB video modality. The knowledge distillation MCT of the major modality branch is denoted as $\text{cls}_r = \{\text{cls}_{r,1}, \text{cls}_{r,2}, \dots, \text{cls}_{r,C}\}$ and the knowledge distillation MCT from the branch of auxiliary modality is indicated by $\text{cls}_s = \{\text{cls}_{s,1}, \text{cls}_{s,2}, \dots, \text{cls}_{s,C}\}$, MCTKD is achieved by applying KL-Divergence (KL-Div) loss after each feature map reduction block of MVitv2 on cls_r and cls_s :

$$L_{MCTKD,all} = \left(\sum_{i=1}^{N_B} \text{KL-Div}(\text{cls}_r^i, \text{cls}_s^i) \right) / N_B, \quad (4)$$

where N_B and $L_{MCTKD,all}$ refer to the block number and the sum of MCTKD losses. $L_{MCTKD,all}$ is combined equally with the binary cross entropy loss (L_{BCE}).

4.4 Multi-Classification Tokens Fusion (MCTF)

Multi-Classification Tokens Fusion (MCTF) is designed to fuse MCT for knowledge distillation and the MCT for classification as in Figure 3. We use cls_r to denote the knowledge distillation MCT, and cls_m denotes the classification MCT. \mathbf{K} , \mathbf{Q} , and \mathbf{V} for each MCT can be obtained through linear projections $\mathbf{P}_K^m, \mathbf{P}_Q^m, \mathbf{P}_V^m, \mathbf{P}_K^r, \mathbf{P}_Q^r$, and \mathbf{P}_V^r as follows,

$$\begin{aligned} \mathbf{K}_m, \mathbf{Q}_m, \mathbf{V}_m &= \mathbf{P}_K^m(\text{cls}_m), \mathbf{P}_Q^m(\text{cls}_m), \mathbf{P}_V^m(\text{cls}_m), \\ \mathbf{K}_r, \mathbf{Q}_r, \mathbf{V}_r &= \mathbf{P}_K^r(\text{cls}_r), \mathbf{P}_Q^r(\text{cls}_r), \mathbf{P}_V^r(\text{cls}_r). \end{aligned} \quad (5)$$

After obtaining the $\mathbf{Q}_{m/r}$, $\mathbf{K}_{m/r}$, and $\mathbf{V}_{m/r}$ from the MCT for classification and the MCT for the knowledge distillation, a mixed attention mechanism is calculated as follows,

$$\begin{aligned} \mathbf{A}_{mm}^m &= \mathbf{P}_{mm}(\text{DP}(\text{Att}(\mathbf{Q}_m, \mathbf{K}_m, \mathbf{V}_m))), \\ \mathbf{A}_{mr}^m &= \mathbf{P}_{mr}(\text{DP}(\text{Att}(\mathbf{Q}_m, \mathbf{K}_r, \mathbf{V}_m))), \\ \mathbf{A}_{rm}^m &= \mathbf{P}_{rm}(\text{DP}(\text{Att}(\mathbf{Q}_r, \mathbf{K}_m, \mathbf{V}_m))), \end{aligned} \quad (6)$$

where Att denotes the attention operation $\text{Att}(\mathbf{Q}_{m/r}, \mathbf{K}_{m/r}, \mathbf{V}_{m/r}) = \text{SoftMax}(\mathbf{Q}_{m/r} @ \mathbf{K}_{m/r}) * \mathbf{V}_{m/r}$ and DP indicates Dropout. The above equations provide attention considering different perspectives including self-attention \mathbf{A}_{mm}^m and two types of cross attention, *i.e.*, \mathbf{A}_{rm}^m and \mathbf{A}_{mr}^m which use the Queries from the MCT for the classification and the Keys from the MCT for knowledge distillation and vice versa. The same procedure is conducted for the knowledge distillation MCT to generate \mathbf{A}_{rr}^r , \mathbf{A}_{rm}^r , and \mathbf{A}_{mr}^r with DP by,

$$\begin{aligned} \mathbf{A}_{rr}^r &= \mathbf{P}_{rr}(\text{DP}(\text{Att}(\mathbf{Q}_r, \mathbf{K}_r, \mathbf{V}_r))), \\ \mathbf{A}_{rm}^r &= \mathbf{P}_{rm}(\text{DP}(\text{Att}(\mathbf{Q}_r, \mathbf{K}_m, \mathbf{V}_r))), \\ \mathbf{A}_{mr}^r &= \mathbf{P}_{mr}(\text{DP}(\text{Att}(\mathbf{Q}_m, \mathbf{K}_r, \mathbf{V}_r))). \end{aligned} \quad (7)$$

Then the attention is finalized as,

$$\mathbf{A}_m, \mathbf{A}_r = \text{Sum}(\mathbf{A}_{mm}^m, \mathbf{A}_{mr}^m, \mathbf{A}_{rm}^m), \text{Sum}(\mathbf{A}_{rr}^r, \mathbf{A}_{mr}^r, \mathbf{A}_{rm}^r). \quad (8)$$

The fused attention is thereby calculated through,

$$\mathbf{A}_f = \mathbf{P}_f(\text{Concat}(\mathbf{A}_m, \mathbf{A}_r)), \quad (9)$$

where \mathbf{P}_f denotes an FC layer. The whole procedure is indicated by,

$$\mathbf{A}_f = \text{CLS}_f(\text{LN}(\text{cls}_m), \text{LN}(\text{cls}_r)), \quad (10)$$

where LN demonstrates the layer normalization and CLS_f is the CLS-Fusion. Assuming we use cls_a to denote the average of MCT for classification and the MCT for knowledge distillation by $\text{cls}_a = (\text{cls}_m + \text{cls}_r)/2$, the final classification tokens are harvested by,

$$\begin{aligned} \text{cls}_f &= \text{cls}_a + \text{CLS}_f(\text{LN}(\text{cls}_r), \text{LN}(\text{cls}_m)), \\ \text{cls}_f &:= \text{cls}_a + \text{DP}(\mathbf{M}_\theta(\text{LN}(\text{cls}_f))), \end{aligned} \quad (11)$$

where \mathbf{M}_θ denotes a Multi-Layer Perception (MLP) based projection and DP denoted dropout operation. MCTKD and MCTF are added after N_{MCT} epochs of training of TRANS^{M3}E with only MCT, for both of the leveraged modalities and models. During the test phase, we make use of the average of the prediction results from the two branches as the final prediction.

5 Evaluation

5.1 Implementation Details

All the video models are pre-trained on ImageNet1K [15] using PyTorch 1.8.0 with four V100 GPUs. To reproduce TRANS^{M3}E, we first train MVitv2-S with only MCT for classification on RGB modality and HD-GCN with only MCT for classification on skeleton modality for 80 epochs and then train TRANS^{M3}E with all components for another 80 epochs. We use AdamW [43] with learning rate of $1e^{-4}$. The input video for *train*, *test*, and *val* is center cropped and rescaled as 224×224 with color jitter parameter as 0.4.

5.2 Analysis on the MuscleMap Benchmark

The results of different architectures on our benchmark are provided in Table 2. First, the approaches include *Random*, in which the muscle activation is predicted randomly, and *All Ones*, in which all the samples are predicted as using all the muscle regions. These two simple approaches are used to serve as statistic baselines. *Random* and *All Ones* show overall low performances with $<30\%$ mAP on all the evaluations. These statistical approaches are leveraged to make comparisons between deep-learning-based approaches to verify whether the model predicts muscle activation randomly or not. The skeleton-based approach, *e.g.*, HD-GCN [34], ST-GCN [82], and CTR-GCN [10], obviously outperform the statistic approaches and deliver promising performances when dealing with unseen activity types. Video-based approaches surpass statistic and skeleton baselines in terms of the AMGE of the known activities, where transformer-based approaches, *e.g.*, MVitv2 S/B [37] and VideoSwin S/B [42], and CNN-based approaches, *e.g.*, C2D [23], I3D [8], Slow [22], SlowFast [22], are leveraged. MVitv2-S shows good performance

Table 2: Experimental results on the MuscleMap benchmark.

Model	#PM	MuscleMap @ mAP					
		known val	new val	mean val	known test	new test	mean test
Random	0.0M	29.7	29.0	29.4	28.9	29.5	29.2
All Ones	0.0M	28.2	28.1	28.2	27.8	28.6	28.2
ST-GCN [82]	2.6M	90.4	63.5	77.0	90.5	63.3	76.9
CTR-GCN [10]	1.4M	93.7	62.2	78.0	93.6	61.7	77.7
HD-GCN [34]	0.8M	93.4	63.1	78.3	93.4	63.1	78.3
C2D (R50) [23]	23.5M	97.2	59.1	78.2	97.4	58.5	78.0
I3D (R50) [8]	20.4M	97.0	59.4	78.2	97.0	58.4	77.7
Slow (R50) [22]	24.3M	96.8	60.7	78.8	96.9	60.5	78.7
SlowFast (R50) [22]	25.3M	89.7	60.2	75.0	94.4	59.6	77.0
MViTv2-S [37]	34.2M	97.7	61.4	79.6	97.9	61.4	79.7
MViTv2-B [37]	51.2M	97.4	61.2	79.3	97.7	61.0	79.4
VideoSwin-S [42]	50.0M	92.6	58.8	75.7	92.4	58.8	75.6
VideoSwin-B [42]	88.0M	91.8	58.7	75.3	91.9	58.3	75.1
VideoMAEv2-B [74]	87.0M	97.1	62.8	80.0	97.5	61.7	79.6
Hiera-B [60]	52.0M	96.8	60.9	78.9	97.0	60.7	78.9
TransM³E (Ours)	55.4M	97.8	64.1	81.0	97.8	64.2	81.0

Table 3: Ablation for TransM³E on MuscleMap.

MCT	MCTKD	MCTF	known val	new val	mean val	known test	new test	mean test
✓	✓	✓	95.7	62.1	78.9	95.9	62.0	79.0
✓	✓	✓	95.4	62.4	78.9	95.6	62.1	78.9
✓	✓	✓	97.8	64.1	81.0	97.8	64.2	81.0

Table 4: Ablation of MCTKD on MuscleMap.

Method	known val	new val	mean val	known test	new test	mean test
FL-KD	96.5	63.0	79.8	96.4	63.4	79.9
DE-KD	95.9	63.9	79.9	96.6	63.9	80.3
SP-KD	97.5	63.0	80.3	96.7	63.1	79.9
FL-MCTKD	95.1	63.0	79.1	95.5	62.8	79.2
DE-MCTKD	95.1	63.3	79.2	95.2	63.4	79.3
SP-MCTKD	97.8	64.1	81.0	97.8	64.2	81.0

due to its ability to reason long-term information and its multi-scale pooling setting, achieving 79.6% and 79.7% for **mean val** and **mean test** on the MuscleMap dataset. However, skeleton-based approaches perform well on new activities but not on known ones due to the lack of visual appearance, while video-based approaches excel on known activities but not on new ones due to the sensitivity to the background changes. A good AMGE model should perform well in both scenarios.

To achieve this, we propose TransM³E, which combines the advantages of both skeleton-based and video-based approaches. It uses multi-classification tokens (MCT) for feature fusion and knowledge distillation, leveraging the top-performing backbones from both modalities: MViTv2-S and HD-GCN. TransM³E surpasses all the others by large margins. TransM³E is a transformer-based approach due to the capability for long-term reasoning of visual transformers [73] since the AMGE should consider the activities at the global level, which requires long-term information reasoning. TransM³E has 64.1%, 97.8%, 64.2%, and 81.0% mAP considering **new val**, **known val**, **new test**, and **known test** on our benchmark, while the generalizability to new activities is mostly highlighted. TransM³E outperforms MViTv2-S by 1.4% and 1.3% on the **mean val** and **mean test**, which especially works well for **new val** and **new test** as TransM³E surpasses MViTv2-S by 2.7% and 2.8%. During the experiments we observe that the obliques group

Table 5: Ablation for the MCTF on MuscleMap.

Method	known val	new val	mean val	known test	new test	mean test
Sum [59]	95.4	62.4	78.9	95.6	62.1	78.9
Multiplication [59]	94.5	62.8	78.7	94.7	62.8	78.8
SelfAttention [46]	97.4	62.9	80.2	97.6	62.8	80.2
CrossAttention [46]	94.9	63.7	79.3	95.1	63.5	79.3
MCTF (ours)	97.8	64.1	81.0	97.8	64.2	81.0

Table 6: Comparison of MMF/KD on MuscleMap.

Method	known val	new val	mean val	known test	new test	mean test
LateFusionSum [59]	80.6	59.8	70.2	80.1	60.0	70.1
LateFusionConcat [76]	83.5	60.8	72.2	83.3	61.2	72.3
LateFusionMul [59]	82.3	60.4	71.4	82.0	60.9	71.5
Ours	97.8	64.1	81.0	97.8	64.2	81.0

is hardest region to achieve AMGE. We also conduct per-label analysis towards sports with body weights and find that the AMGE performance of the motions with fitness equipment is higher than those with body weight.

5.3 Analysis of the Ablation Studies

Module ablation. The ablation study of MCT, MCTKD, and MCTF, is shown in Table 3, where we deliver the results for *w/o MCT*, *w/o MCTKD*, *w/o MCTF*, and *w/ all*. When we compare the results between *w/o MCT* and *w/ all*, we find that using MCT to enlarge the attributes prediction space can contribute performance improvements by 2.1%, 2.0%, 2.1%, 1.9%, 2.2%, and 2.0% in terms of **known val**, **new val**, **mean val**, **known test**, **new test**, and **mean test**. When comparing the results between *w/o MCTKD* and *w/ all*, we observe that leveraging MCTKD shows more benefits. When we compare the results between *w/o MCTF* and *w/ all*, we find that using MCTF to achieve the fusion between the information derived from the classification MCT and knowledge distillation MCT can bring performance improvements of 2.4%, 1.7%, 2.1%, 2.2%, 2.1%, and 2.1% in terms of the six aforementioned evaluations.

MCTKD ablation. We evaluate the effects of varying knowledge distillation location and the knowledge distillation on a single distillation token (KD) and MCT (MCTKD), where they are named differently, *i.e.*, KD/MCTKD at the final layer (FL-KD/MCTKD),

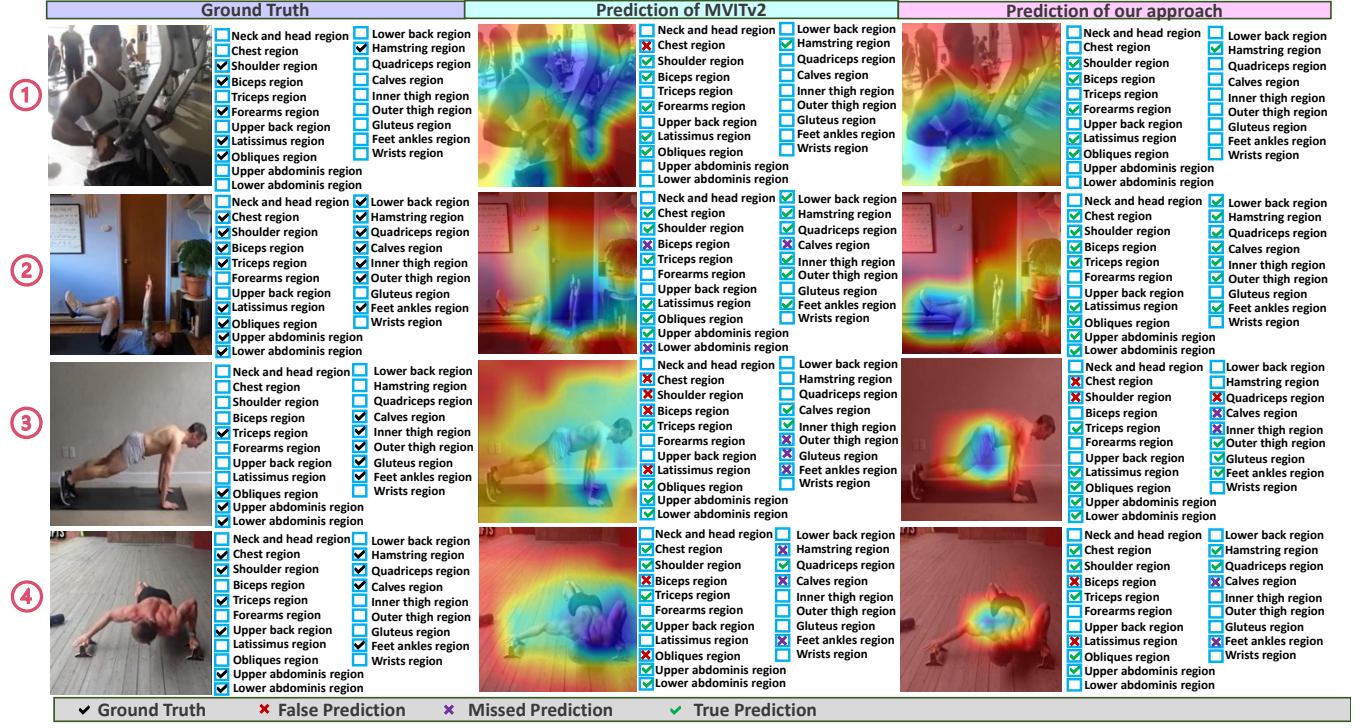


Figure 5: Qualitative results for the MVitv2-S [37] and TRANS³E. GradCam [63] visualization is given.

KD/MCTKD after token size reduction (SP-KD/MCTKD), or KD/MCTKD after each MVitv2 block (DE-KD/MCTKD), in Table 4. SP-KD and SP-MCTKD achieve the best performances for KD and MCTKD individually, demonstrating their superiority of using sparse knowledge distillation settings after the reduction of the feature map size, and SP-MCTKD outperforms SP-KD. Using MCTKD in sparse locations—specifically after each feature map size reduction—yields the best results on the MuscleMap benchmark. This enhances the aggregation of AMGE cues across modalities after pooling, facilitating effective knowledge distillation. SP-MCTKD achieves the best performance and is selected.

MCTF ablation. The ablations on MCTF for TRANS³E are presented in Table 5, where our approach is compared with existing fusion approaches, *e.g.*, *Sum*, *Multiplication*, *SelfAttention*, and *CrossAttention*. MCTF shows the best performance with 81.0% and 81.0% on **mean val** and **mean test**. The superiority of MCTF compared to other approaches, especially on generalizability, depends on using attention from a more diverse perspective.

5.4 Comparison with Other Fusion Approaches

Table 6 presents the comparison between TRANS³E and existing multi-modality fusion approaches, *i.e.*, *LateFuionSum*, *LateFusionConcat*, and *LateFusionMul*. We compare our proposed method to these conventional multi-modal fusion approaches to illustrate that the performance improvement of our approach is not solely delivered by using the feature fusion between the skeleton modality and the RGB video modality. Compared with the best performing baseline *LateFusionConcat*, our approach achieves a better performance.

5.5 Analysis of Qualitative Results

Qualitative results are shown in Figure 5, the label and GradCam [63] visualizations of MVitv2-S and TRANS³E are given from left to right. The true/missed/false prediction is marked as green checkmark/purple crossmark/red crossmark. Overall, our approach has more accurate predictions and fewer false and missed predictions for all the samples considering known activities, *i.e.*, ① and ② in Figure 5, and new activities, *e.g.*, ③ and ④, where ① and ② are correctly predicted by our model. TRANS³E concentrates mostly on the accurate body regions, *e.g.*, in sample ③ TRANS³E focuses on the leg and abdominis related region, while the focus of the MVitv2-S is distracted, which results in more false predictions of MVitv2-S. Due to the integration of the learned knowledge from both the video and skeleton modalities, our model can achieve a better focus.

6 Conclusion

In this paper, we propose the new task of video-based activated muscle group estimation in the wild. We contribute the first large-scale video-based AMGE dataset with in-the-wild videos and establish the MuscleMap benchmark using statistical baselines and existing video- and skeleton-based methods. Considering AMGE generalizability, we propose TRANS³E with multi-classification token distillation and fusion in a cross-modality manner to enhance generalization to new activity types. TRANS³E sets the state-of-the-art on the MuscleMap benchmark. Future works will explore missing modality AMGE and leverage shared encoder.

Acknowledgement

The project served to prepare the SFB 1574 Circular Factory for the Perpetual Product (project ID: 471687386), approved by the German Research Foundation (DFG, German Research Foundation) with a start date of April 1, 2024. This work was also partially supported in part by the SmartAge project sponsored by the Carl Zeiss Stiftung (P2019-01-003; 2021-2026). This work was performed on the HoreKa supercomputer funded by the Ministry of Science, Research and the Arts Baden-Wuerttemberg and by the Federal Ministry of Education and Research. The authors also acknowledge support by the state of Baden-Wuerttemberg through bwHPC and the German Research Foundation (DFG) through grant INST 35/1597-1 FUGG. This project is also supported by the National Key RD Program under Grant 2022YFB4701400.

References

- [1] Tasweer Ahmad, Lianwen Jin, Xin Zhang, Songxuan Lai, Guozhi Tang, and LuoJun Lin. 2021. Graph convolutional neural network for human action recognition: A comprehensive survey. *TAI* (2021).
- [2] Jean-Baptiste Alayrac et al. 2020. Self-supervised multimodal versatile networks. In *NeurIPS*.
- [3] Naji A Alibeji, Vahidreza Molazadeh, Frank Moore-Clingenpeel, and Nitin Sharma. 2018. A muscle synergy-inspired control design to coordinate functional electrical stimulation and a powered exoskeleton: Artificial generation of synergies to reduce input dimensionality. *CSM* (2018).
- [4] Yuliya Bababekova, Mark Rosenfield, Jennifer E Hue, and Rae R Huang. 2011. Font Size and Viewing Distance of Handheld Smart Phones. *Optometry and Vision Science* (2011).
- [5] Kelly R Berckmans, Birgit Castelein, Dorien Borms, Thierry Parlevliet, and Ann Cools. 2021. Rehabilitation exercises for dysfunction of the scapula: exploration of muscle activity using fine-wire EMG. *The American Journal of Sports Medicine* (2021).
- [6] Lars Buitinck et al. 2013. API design for machine learning software: experiences from the scikit-learn project. In *ECML W*.
- [7] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Nieves. 2015. ActivityNet: A large-scale video benchmark for human activity understanding. In *CVPR*.
- [8] Joao Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? A new model and the kinetics dataset. In *CVPR*.
- [9] Pengguang Chen, Shu Liu, Hengshuang Zhao, and Jiaya Jia. 2021. Distilling knowledge via knowledge review. In *CVPR*.
- [10] Yuxin Chen, Ziqi Zhang, Chunfeng Yuan, Bing Li, Ying Deng, and Weiming Hu. 2021. Channel-wise Topology Refinement Graph Convolution for Skeleton-Based Action Recognition. In *ICCV*.
- [11] Zhao-Min Chen, Xiu-Shen Wei, Peng Wang, and Yanwen Guo. 2019. Multi-label image recognition with graph convolutional networks. In *CVPR*.
- [12] Mia Chiquier and Carl Vondrick. 2023. Muscles in Action. In *ICCV*.
- [13] Vasileios Choutas, Philippe Weinzaepfel, Jérôme Revaud, and Cordelia Schmid. 2018. PoTion: Pose motion representation for action recognition. In *CVPR*.
- [14] Aijse Willem de Vries, Frank Krause, and Michiel Pieter de Looze. 2021. The effectivity of a passive arm support exoskeleton in reducing muscle activation and perceived exertion during plastering activities. *Ergonomics* (2021).
- [15] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *CVPR*.
- [16] Jeffrey Donahue et al. 2015. Long-term recurrent convolutional networks for visual recognition and description. In *CVPR*.
- [17] Haodong Duan, Yue Zhao, Kai Chen, Dahua Lin, and Bo Dai. 2022. Revisiting skeleton-based action recognition. In *CVPR*.
- [18] Rene D Estembar and Chih-Jung Huang. 2019. Essential occupational risk and health interventions for Taiwan's bus drivers. In *Proc. ICIEA*.
- [19] Haoqi Fan et al. 2021. Multiscale vision transformers. In *ICCV*.
- [20] Hao-Shu Fang et al. 2023. AlphaPose: Whole-body Regional Multi-Person Pose Estimation and Tracking in Real-Time. *TPAMI* (2023).
- [21] Zhiyuan Fang, Jianfeng Wang, Xiaowei Hu, Lijuan Wang, Yezhou Yang, and Zicheng Liu. 2021. Compressing visual-linguistic model via knowledge distillation. In *ICCV*.
- [22] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. 2019. Slow-Fast networks for video recognition. In *ICCV*.
- [23] Christoph Feichtenhofer, Axel Pinz, and Richard P. Wildes. 2017. Spatiotemporal multiplier networks for video action recognition. In *CVPR*.
- [24] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. 2016. Convolutional two-stream network fusion for video action recognition. In *CVPR*.
- [25] Qing Gao, Jinguo Liu, and Zhaojie Ju. 2021. Hand gesture recognition using multimodal data fusion and multiscale parallel convolutional neural network for human-robot interaction. *Expert Systems* (2021).
- [26] Jianping Gou, Baosheng Yu, Stephen J. Maybank, and Dacheng Tao. 2021. Knowledge distillation: A survey. *IJCV* (2021).
- [27] Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531* (2015).
- [28] Ankit Jain, Rajendra Akerkar, and Abhishek Srivastava. 2023. Privacy-Preserving Human Activity Recognition System for Assisted Living Environments. *TAI* (2023).
- [29] Xiao Jin et al. 2019. Knowledge distillation via route constrained optimization. In *ICCV*.
- [30] Soo-Min Kang and Richard P. Wildes. 2016. Review of action recognition and detection methods. *arXiv preprint arXiv:1610.06906* (2016).
- [31] Evangelos Kazakos, Arsha Nagrani, Andrew Zisserman, and Dima Damen. 2019. EPIC-Fusion: Audio-visual temporal binding for egocentric action recognition. In *ICCV*.
- [32] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. 2011. HMDB: A large video database for human motion recognition. In *ICCV*.
- [33] Todd A Kuiken, Laura A Miller, Robert D Lipschutz, Kathy A Stubblefield, and Gregory A Dumanian. 2006. Prosthetic command signals following targeted hyper-reinnervation nerve transfer surgery. In *Proc. EMBAC*.
- [34] Jungho Lee, Minhyeok Lee, Dogyoon Lee, and Sangyoon Lee. 2022. Hierarchically Decomposed Graph Convolutional Networks for Skeleton-Based Action Recognition. *arXiv preprint arXiv:2208.10741* (2022).
- [35] Yeonghyeon Lee, Kangwook Jang, Jahyun Goo, Youngmoon Jung, and Hoirin Kim. 2022. FitHuBERT: Going Thinner and Deeper for Knowledge Distillation of Speech Self-Supervised Learning. *arXiv preprint arXiv:2207.00555* (2022).
- [36] Ang Li, Meghana Thotakuri, David A. Ross, João Carreira, Alexander Vostrikov, and Andrew Zisserman. 2020. The AVA-Kinetics localized human actions video dataset. *arXiv preprint arXiv:2005.00214* (2020).
- [37] Yanghao Li et al. 2022. MVITv2: Improved Multiscale Vision Transformers for Classification and Detection. In *CVPR*.
- [38] Ting Liang and Yong J Yuan. 2016. Wearable medical monitoring systems based on wireless networks: A review. *IEEE Sensors Journal* (2016).
- [39] Ruiping Liu et al. 2022. TransKD: Transformer Knowledge Distillation for Efficient Semantic Segmentation. *arXiv preprint arXiv:2202.13393* (2022).
- [40] Shilong Liu, Lei Zhang, Xiao Yang, Hang Su, and Jun Zhu. 2021. Query2Label: A simple transformer way to multi-label classification. *arXiv preprint arXiv:2107.10834* (2021).
- [41] Ze Liu et al. 2022. Swin transformer V2: Scaling up capacity and resolution. In *CVPR*.
- [42] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. 2022. Video swin transformer. In *CVPR*.
- [43] Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *ICLR*.
- [44] Yue Lu, Yujie Wang, and Qian Lu. 2019. Effects of exercise on muscle fitness in dialysis patients: A systematic review and meta-analysis. *American Journal of Nephrology* (2019).
- [45] Alessio Monti, Angelo Porrello, Simone Calderara, Pasquale Coscia, Lamberto Ballan, and Rita Cucchiara. 2022. How many Observations are Enough? Knowledge Distillation for Trajectory Forecasting. In *CVPR*.
- [46] Arsha Nagrani, Shan Yang, Anurag Arnab, Areen Jansen, Cordelia Schmid, and Chen Sun. 2021. Attention bottlenecks for multimodal fusion. In *NeurIPS*.
- [47] Rameswar Panda et al. 2021. AdaMML: Adaptive multi-modal learning for efficient video recognition. In *ICCV*.
- [48] Paritosh Parmar and Brendan Tran Morris. 2019. What and how well you performed? A multitask learning approach to action quality assessment. In *CVPR*.
- [49] Mandela Patrick, Yuki Markus Asano, Ruth Fong, João F. Henriques, Geoffrey Zweig, and Andrea Vedaldi. 2020. Multi-modal self-supervision from generalized data transformations. *arXiv preprint arXiv:2003.04298* (2020).
- [50] Kunyu Peng, Jia Fu, Kailun Yang, Di Wen, Yufan Chen, Ruiping Liu, Junwei Zheng, Jiaming Zhang, M Saquib Sarfraz, Rainer Stiefelhofen, et al. 2024. Referring Atomic Video Action Recognition. *arXiv preprint arXiv:2407.01872* (2024).
- [51] Kunyu Peng, Alina Roitberg, Kailun Yang, Jiaming Zhang, and Rainer Stiefelhofen. 2022. Should I take a walk? Estimating Energy Expenditure from Video Data. In *CVPRW*.
- [52] Kunyu Peng, Alina Roitberg, Kailun Yang, Jiaming Zhang, and Rainer Stiefelhofen. 2022. TransDARC: Transformer-based Driver Activity Recognition with Latent Space Feature Calibration. In *IROS*.
- [53] Kunyu Peng, Alina Roitberg, Kailun Yang, Jiaming Zhang, and Rainer Stiefelhofen. 2023. Delving Deep into One-Shot Skeleton-based Action Recognition with Diverse Occlusions. *TMM* (2023).
- [54] Kunyu Peng, Cheng Yin, Junwei Zheng, Ruiping Liu, David Schneider, Jiaming Zhang, Kailun Yang, M Saquib Sarfraz, Rainer Stiefelhofen, and Alina Roitberg. 2024. Navigating open set scenarios for skeleton-based action recognition. In *AAAI*.

- [55] Cuong Pham, Linh Nguyen, Anh Nguyen, Ngoc Nguyen, and Van-Toi Nguyen. 2021. Combining skeleton and accelerometer data for human fine-grained activity recognition and abnormal behaviour detection with deep temporal convolutional networks. *Multimedia Tools and Applications* (2021).
- [56] A. J. Piergiovanni, Anelia Angelova, and Michael S. Ryoo. 2020. Evolving losses for unsupervised video representation learning. In *CVPR*.
- [57] Firstyani Imannisa Rahma, Rizki Mawan, Harianto Harianto, and Kusri. 2020. Nutrition and Lifestyle Recommendations for Patients Recovering from Covid-19 in Nusa Tenggara Barat Province. In *ICORIS*.
- [58] Tal Ridnik et al. 2021. Asymmetric loss for multi-label classification. In *ICCV*.
- [59] Alina Roitberg, Kunyu Peng, Zdravko Marinov, Constantin Seibold, David Schneider, and Rainer Stiefel. 2022. A Comparative Analysis of Decision-Level Fusion for Multimodal Driver Behaviour Understanding. In *IV*.
- [60] Chaitanya Ryali, Yuan-Ting Hu, Daniel Bolya, Chen Wei, Haoqi Fan, Po-Yao Huang, Vaibhav Aggarwal, Arkabandhu Chowdhury, Omid Poursaeed, Judy Hoffman, et al. 2023. Hiera: A hierarchical vision transformer without the bells-and-whistles. In *ICML*.
- [61] David Sanchez-Lorente, Ricard Navarro-Ripoll, Rudith Guzman, Jorge Moises, Elena Gimeno, Marc Boada, and Laureano Molins. 2018. Prehabilitation in thoracic surgery. *Journal of Thoracic Disease* (2018).
- [62] David Schneider, Saquib Sarfraz, Alina Roitberg, and Rainer Stiefel. 2022. Pose-Based Contrastive Learning for Domain Agnostic Activity Representations. In *CVPRW*.
- [63] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *ICCV*.
- [64] Suranga Seneviratne et al. 2017. A survey of wearable devices and challenges. *IEEE Communications Surveys & Tutorials* (2017).
- [65] Dian Shao, Yue Zhao, Bo Dai, and Dahua Lin. 2020. FineGym: A hierarchical video dataset for fine-grained action understanding. In *CVPR*.
- [66] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. 2020. Skeleton-based action recognition with multi-stream adaptive graph convolutional networks. *TIP* (2020).
- [67] Khuram Soomro, Amir Roshan Zamir, and Mubarak Shah. 2012. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402* (2012).
- [68] Vladislav Sovrasov. 2022. Combining Metric Learning and Attention Heads For Accurate and Efficient Multilabel Image Classification. *arXiv preprint arXiv:2209.06585* (2022).
- [69] Venkata Subbareddy K and Nirmala Devi L. 2023. View Invariant Spatio-Temporal Descriptor for Action Recognition From Skeleton Sequences. *TAI* (2023).
- [70] Yansong Tang et al. 2020. Uncertainty-aware score distribution learning for action quality assessment. In *CVPR*.
- [71] Mauricio Cagliari Tosin, Juliano Costa Machado, and Alexandre Balbinot. 2022. sEMG-based Upper Limb Movement Classifier: Current Scenario and Upcoming Challenges. *JAIR* (2022).
- [72] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. 2021. Training data-efficient image transformers & distillation through attention. In *ICML*.
- [73] Ashish Vaswani et al. 2017. Attention is all you need. In *NeurIPS*.
- [74] Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yan He, Yi Wang, Yali Wang, and Yu Qiao. 2023. VideoMAE V2: Scaling Video Masked Autoencoders With Dual Masking. In *CVPR*.
- [75] Shiguang Wang, Zhizhong Li, Yue Zhao, Yuanjun Xiong, Limin Wang, and Dahua Lin. 2020. Denseflow. <https://github.com/open-mmlab/denseflow>.
- [76] Ping-Cheng Wei, Kunyu Peng, Alina Roitberg, Kailun Yang, Jiaming Zhang, and Rainer Stiefel. 2022. Multi-modal Depression Estimation based on Sub-attentional Fusion. In *ECCV*.
- [77] Yiping Wei, Kunyu Peng, Alina Roitberg, Jiaming Zhang, Junwei Zheng, Ruiping Liu, Yufan Chen, Kailun Yang, and Rainer Stiefel. 2024. Elevating Skeleton-Based Action Recognition with Efficient Multi-Modality Self-Supervision. In *ICASSP*.
- [78] Felicity R. Williams, Annalisa Berzigotti, Janet M. Lord, Jennifer C. Lai, and Matthew J. Armstrong. 2019. impact of exercise on physical frailty in patients with chronic liver disease. *Alimentary Pharmacology & Therapeutics* (2019).
- [79] Jinglin Xu, Yongming Rao, Xumin Yu, Guangyi Chen, Jie Zhou, and Jiwen Lu. 2022. FineDiving: A Fine-grained Dataset for Procedure-aware Action Quality Assessment. In *CVPR*.
- [80] Shichao Xu, Yikang Li, Jenhao Hsiao, Chiuman Ho, and Zhu Qi. 2022. A Dual Modality Approach For (Zero-Shot) Multi-Label Classification. *arXiv preprint arXiv:2208.09562* (2022).
- [81] Yi Xu, Kunyu Peng, Di Wen, Ruiping Liu, Junwei Zheng, Yufan Chen, Jiaming Zhang, Alina Roitberg, Kailun Yang, and Rainer Stiefel. 2024. Skeleton-Based Human Action Recognition with Noisy Labels. *arXiv preprint arXiv:2403.09975* (2024).
- [82] Sijie Yan, Yuanjun Xiong, and Dahua Lin. 2018. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *AAAI*.
- [83] Zhendong Yang, Zhe Li, Mingqi Shao, Dachuan Shi, Zehuan Yuan, and Chun Yuan. 2022. Masked Generative Distillation. In *ECCV*.
- [84] Jin Ye, Junjun He, Xiaojiang Peng, Wenhao Wu, and Yu Qiao. 2020. Attention-driven dynamic graph convolutional network for multi-label image recognition. In *ECCV*.
- [85] Renchun You, Zhiyao Guo, Lei Cui, Xiang Long, Yingze Bao, and Shilei Wen. 2020. Cross-modality attention with semantic graph embedding for multi-label classification. In *AAAI*.
- [86] Zhengbo Zhang, Chunlun Zhou, and Zhigang Tu. 2022. Distilling Inter-Class Distance for Semantic Segmentation. In *IJCAI*.
- [87] Borui Zhao, Quan Cui, Renjie Song, Yiyu Qiu, and Jiajun Liang. 2022. Decoupled Knowledge Distillation. In *CVPR*.
- [88] Hang Zhao, Zhicheng Yan, Lorenzo Torresani, and Antonio Torralba. 2019. HACS: Human Action Clips and Segments Dataset for Recognition and Temporal Localization. In *ICCV*.