

Supplementary Materials: Towards Activated Muscle Group Estimation in the Wild

Anonymous Authors

1 SOCIETY IMPACT AND LIMITATIONS

In our work, a new dataset targeting at the AMGE is collected based on YouTube videos, termed as MuscleMap135. We build up MuscleMap benchmark for the AMGE by using statistic baselines and existing video-based approaches including both video-based and skeleton-based methods, while the three aforementioned datasets are all considered. Through the experiments we find that the generalizability targeting AMGE on new activities is not satisfied for the existing activity recognition approaches. In order to tackle this issue, we propose a new cross modality knowledge distillation approach named as TRANSM³E while using MVITv2-S [6] as its basic backbone. The proposed approach alleviates the generalization problem in a certain degree, however there is still large space for further improvement and future research. The AMGE performance gap between the known activities and new activities illustrates that our model has potential to give offensive predictions, misclassifications and biased content which may cause false prediction resulting in the negative social impact. The dataset and code will be released publicly.

Limitations. The annotations of MuscleMap135 are created for each video clip instead of being created for each frame and the label is binary without giving the different levels of muscle activations. In addition, there is still a clear gap between the performance of known and new categories. While our method has enhanced the generalization capacity, there remains room for future improvement. **Additional clarification of the submission.** We notice that the title in the system is slightly different from the title in the submission (where video-based is removed in our submission). We will make changes in the system on the final version if it is accepted.

2 MORE DETAILS OF THE DATASET

The muscle regions where the number of sources are bigger than the threshold are chosen as activated muscle region. We can see that no obvious deviation could be found in AMGE annotation. We annotate the commonly leveraged human body muscles in the daily life into 20 muscle regions according to the suggestion of the experts, *i.e.*, neck and head region, chest region, shoulder region, biceps region, triceps region, forearms region, upper back region, latissimus region, obliques region, upper abdominis region, lower abdominis region, lower back region, hamstring region, quadriceps region, calves region, inner thigh region, outer thigh region, gluteus region, feet ankles region, and wrists region. We rearrange occipitofrontalis, temporoparietalis, levator labii superioris, masticatorii, sternocleidomastoideus as neck and head muscle region; pectoralis major as chest region; deltoideus as shoulder region; biceps brachii as biceps region; triceps brachii as triceps region; flexor carpi radialis, palmaris longus, abductor pollicis longus as forearm region; trapezius as upper back region; latissimus dorsi as latissimus region; external oblique, serratus anterior as obliques region; rectus abdominis, quadratus lumborum as upper abdominis region;

transversus abdominis, pyramidalis as lower abdominis region; erector spinae as lower back region; biceps femoris, semimembranosus, semitendinosus as hamstring region; rectus femoris, vastus medialis as quadriceps region; gastrocnemius, soleus as calves region; adductor longus, sartorius, gracilis as inner thigh region; iliotibial tract as outer thigh region; gluteus maximus as gluteus region; peroneus longus and brevis, extensor digitorum longus, flexor hallucis longus, flexor digitorum longus, peroneus tertius, tibialis posterior as feet ankles region; extensor pollicis, 1st dorsal interosseous, pronator quadratus as wrists region.

3 FURTHER IMPLEMENTATION DETAILS

For our TRANSM³E, we use 16 MVIT-S blocks and choose the number of heads as 1. The feature dimension of the patch embedding net is 96 while using 3D CNN and choosing the patch kernel as {3, 7, 7}, patch stride kernel as {2, 4, 4} and patch padding as {1, 3, 3}. The MLP ratio for the feature extraction block is 4.0, QKV bias is chosen as True and the path dropout rate is chosen as 0.2. The dimensions of the tokens and number of heads are multiplied by 2 after the 1-st, 3-th, and 14-th blocks. The pooling kernel of QKV is chosen as {3, 3, 3}, the adaptive pooling stride of KV is chosen as {1, 8, 8} while the stride for the pooling on Q is chosen as {1, 2, 2} for the 1-st, 3-th, and 14-th block. For the rest of the blocks among 0~15-th blocks, the stride for the pooling on Q is chosen as {1, 1, 1}. Regarding the MCTF, we choose the head number as 1, the QK scale number as 0.8, the dropout for attention as 0.0, and the dropout rate of the path as 0.2. The input embeddings of the MCTF have 768 channels while the intermediate embeddings of the MCTF structure have the same number of channels as the input of MCTF. All the hyperparameters are chosen according to the performance measured on the validation set.

4 BASELINE METHODS

Video classification approaches, *e.g.*, I3D [1], SlowFast [4], and MVITv2 [6], skeleton approaches, *i.e.*, ST-GCN [8], CTR-GCN [2], and HD-GCN [5], and statistic calculations, *e.g.*, randomly guess (Random), are selected as baselines to formulate our MuscleMap benchmark on the proposed new dataset to achieve AMGE in-the-wild. Statistic calculation-based approaches serve for performance verification considering the question regarding whether the prediction of the model is random or not. Skeleton-based approaches are selected since they directly take the geometric relationship of the human body into consideration without disrupting information from the background. Considering video-based approaches, transformer-based models, *i.e.*, MVITv2 and VideoSwin, and Convolutional Neural Network (CNN) based models, *i.e.*, C2D, I3D, Slow, and SlowFast, are leveraged. Transformers are expected to have better performance compared with CNNs due to their excellent long-term reasoning ability [7], which is also verified in the experiments conducted on the MuscleMap benchmark.

Table 1: Results for different modalities on the MuscleMap benchmark.

Modality	known val	new val	mean val	known test	new test	mean test
Optical Flow	72.7	59.8	66.3	69.7	57.7	63.7
RGB Difference	96.8	60.3	78.6	97.5	59.8	78.7
RGB	98.5	62.1	80.3	98.6	60.7	79.7

5 MORE DETAILS OF THE MCTKD

Since we introduced the ablation regarding MCTKD in our main paper with experimental results, only more details regarding the KD format and position will be introduced in this section. In order to make it clearer for understanding, we illustrate more details regarding the KD/MCTKD position in Figure 1 to give a detailed clarification. For the MCTKD related approaches, we use the MCTKD as depicted by (d), where the KD is executed between the knowledge receiver MCTs of the main modality and the sender MCTs of the auxiliary modality. For all the other basic KD-based approaches, we use the format as depicted by (c), where the KD is executed between the MCTs of the main modality and the MCTs of the auxiliary modality, regarded as conventional KD. All the experiments are executed with MCTs while without MCTF aggregation. We simply average the MCTs for all the experiments in this ablation. Regarding the sparse format as depicted in (a), the knowledge of the auxiliary modality is only transferred after the size reduction of the pooling layer denoted as DownSampling (DS) in Figure 1 and after the final layer. Only SparseMCTKD and DenseMCTKD are depicted since the SparseKD and DenseKD use the same position settings. SparseKD/MCTKD aims at reducing the KD/MCTKD calculation by selecting the most important intermediate layers to transfer the knowledge. After each pooling layer which has size reduction, the informative cues will be highlighted, which makes the corresponding changes of the tokens from auxiliary modality necessary to be integrated through KD/MCTKD. We choose the position after the pooling with size reduction to do the KD/MCTKD on the intermediate layer. DenseKD/MCTKD is designed to transfer the knowledge directly after each transformer block in order to leverage the knowledge from the other modality thoroughly. We make use of both KD positions to conduct comparison and select the most appropriate method to build the MCTKD in our final model.

6 ANALYSIS OF DIFFERENT MODALITIES

We systematically search for the best-performing primary modality considering the video data and present the results in Table 1. We deliver the experimental results on MViT2-S architecture with MCT pre-trained with ImageNet1K [3] for *Optical Flow*, *RGB Difference*, and *RGB* modalities. We observe that the RGB modality outperforms the other modalities due to its informative temporal-spatial appearance cues which contributes to good AMGE results. We thereby choose the RGB modality as the primary modality to conduct the research and hope that the provided other modalities can enable future research for the multi-modal AMGE.

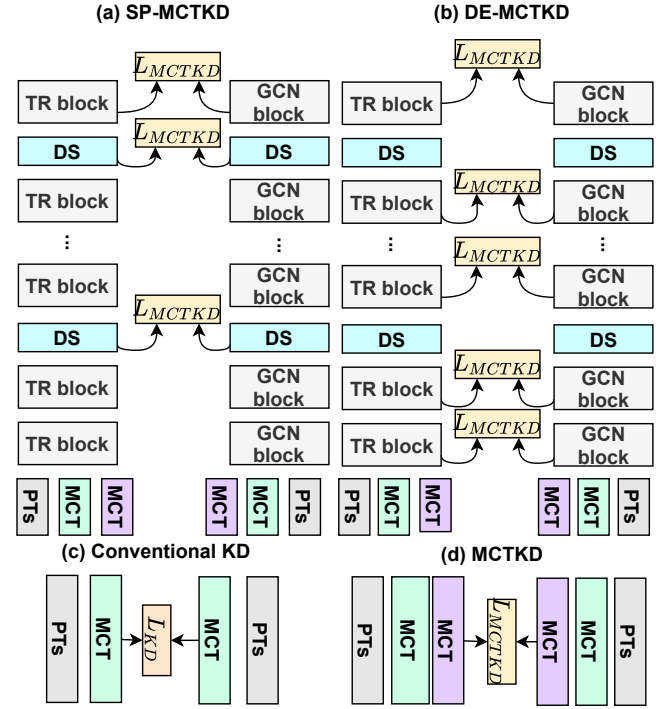


Figure 1: An overview of the details regarding our ablation study for the MCTKD position and format, where (a) we execute MCTKD after the downsampling of the pooling layer and after the final transformer block to formulate sparse MCTKD, named as SP-MCTKD, (b) we leverage the MCTKD after each transformer block (TR Block) to formulate the dense MCTKD, named as DE-MCTKD, (c) indicates the conventional knowledge distillation (w/o knowledge distillation MCT), and (d) indicates the MCTKD we leveraged.

REFERENCES

- [1] Joao Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? A new model and the kinetics dataset. In *CVPR*.
- [2] 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*.
- [3] 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*.
- [4] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. 2019. Slow-Fast networks for video recognition. In *ICCV*.
- [5] 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).
- [6] Yanghao Li et al. 2022. MViT2: Improved Multiscale Vision Transformers for Classification and Detection. In *CVPR*.
- [7] Ashish Vaswani et al. 2017. Attention is all you need. In *NeurIPS*.
- [8] Sijie Yan, Yuanjun Xiong, and Dahua Lin. 2018. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *AAAI*.