ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-6654-0540-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICASSP43922.2022.9746752

CROSS-MODAL KNOWLEDGE DISTILLATION FOR VISION-TO-SENSOR ACTION RECOGNITION

Jianyuan Ni[1] *Raunak Sarbajna*[2] *Yang Liu*[3] *Anne H.H. Ngu*[1] *Yan Yan*[4]

[1] Texas State University, USA
[2] University of Houston, USA
[3] Sun Yat-sen University, China
[4] Illinois Institute of Technology, USA

ABSTRACT

Human activity recognition (HAR) based on multi-modal approach has been recently shown to improve the accuracy performance of HAR. However, restricted computational resources associated with wearable devices, *i.e.,* smartwatch, failed to directly support such advanced methods. To tackle this issue, this study introduces an end-to-end Vision-to-Sensor Knowledge Distillation (VSKD) framework. In this VSKD framework, only time-series data, *i.e.,* accelerometer data, is needed from wearable devices during the testing phase. Therefore, this framework will not only reduce the computational demands on wearable devices, but also produce a learning model that closely matches the performance of the computational expensive multi-modal approach. In order to retain the local temporal relationship and facilitate visual deep learning models, we first convert time-series data to two-dimensional images by applying the Gramian Angular Field (GAF) based encoding method. We adopted multi-scale TRN with BN-Inception and ResNet18 as the teacher and student network in this study, respectively. A novel loss function, named Distance and Angle-wised Semantic Knowledge loss (DASK), is proposed to mitigate the modality variations between the vision and the sensor domain. Extensive experimental results on UTD-MHAD, MMAct, and Berkeley-MHAD datasets demonstrate the competitiveness of the proposed VSKD model which can be deployed on wearable devices.

***Index Terms*—** Cross-modal knowledge distillation, Vision-to-sensor, Human activity recognition.

1. INTRODUCTION

Human Activity Recognition (HAR), *i.e.,* perceiving and recognizing human actions, is crucial for real-time applications, such as healthcare and human-robot interaction [1]. Vision-based methods have dominated the HAR community because video data contains rich appearance information [2]. However, video-based HAR is intrinsically restricted in various occlusion and illumination conditions similar to the human vision limitations. Meanwhile, utilizing time-series data, *i.e.,* accelerometer data, from wearable devices is another typical way of identifying the HAR problem. But sensor-based HAR approaches are difficult to achieve reliable performance compared to video modality due to the constraint of single context information [3, 4]. For instance, previous work indicated that deep learning method using the accelerometer data from a wrist-worn watch for fall detection only achieved 86% accuracy due to the fact that it is difficult to differentiate various wrist movements when someone falls [4]. Multi-modal HAR systems can solve such problems by utilizing the complementary information acquired from different modalities. For example, vision-based approaches could provide global motion features while sensor-based methods can give 3D information about local body movement [5]. Nevertheless, limited resources associated with wearable devices, such as CPU and memory storage, cannot support such powerful and advanced multi-modal systems. In order to overcome such issues, the technique of cross-modal transfer, *i.e.,* knowledge distillation (KD), that needs only one modality input during the testing phase to reach the performance close to the combination of multi-modal data during the training phase has been proposed [6]. Using this approach, we can transfer the knowledge from vision to sensor domain by reducing computation resource demand, but also eventually boost the performance of HAR using wearable devices.

In this paper, we propose an end-to-end Vision-to-Sensor Knowledge Distillation (VSKD) for HAR problem. The overview of the proposed method is shown in Figure 1. First, we adopted the Gramian Angular Field (GAF) to encode the accelerometer data to an image representation while preserving the temporal information from the data [7]. Next, we trained the teacher networks with video stream inputs with the standard cross-entropy loss function. The KD process of accelerometer data was accomplished by using the new loss function, named Distance and Angle-wised Semantic Knowledge loss (DASK). Overall, the contributions of this paper are summarized as follows: 1) To the best of our knowledge, this is the first study conducting the knowledge distillation

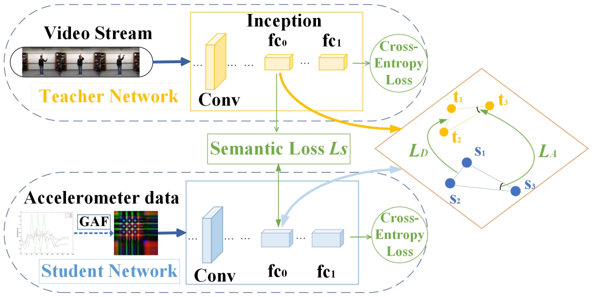

Fig. 1. Schematic overview of the proposed VSKD method.

(KD) model from the video-to-sensor domain. In this VSKD model, a student network with input of wearable sensor data, *i.e.,* accelerometer data, learns the compensatory information from the teacher network with input of video streams. 2) We proposed a novel loss function (DASK), which is utilized to alleviate the modality gap between the teacher and student network. 3) We demonstrated the effectiveness of the proposed VSKD method on three public datasets.

2. RELATED WORK

HAR has been an active research field due to its wide application in various areas [1, 2, 8]. Despite the fact that video modality contains rich RGB information, video modality is subject to various viewpoints or illumination conditions which affects its effectiveness. HAR studies with time-series data, *i.e.,* accelerometer data, from wearable devices are growing rapidly [4, 9, 10]. Although those works demonstrated the feasibility of sensor-based HAR approaches, they cannot achieve reliable performance due to the noisy data or sensor variations [3]. By aggregating the advantages of various data modalities, a multi-modal approach can ultimately provide a robust and accurate HAR method. However, the limited computation capabilities of a low-cost wearable devices prevent the complexity of multi-modal methods that can be deployed on the device directly. In order to build lightweight and accurate HAR models, the knowledge distillation approach has emerged to build a student model with less computational overhead and yet can retain similar accuracy performance as the teacher model [6]. Kong *et al.* [11] proposed a multi-modal attention distillation method to model video-based HAR with the instructive side information from inertial sensor modality. Similarly, Liu *et al.* [5] introduced a multi-modal KD method where the knowledge from multiple sensor data were adaptively transferred to video domain. Even though those works provide promising results on HAR with the multi-modal approach, no multi-modal KD work has yet been proposed where the time-series data is used as the student model. Using the reversed approach will improve the accuracy performance of a sensor-based HAR, but also reduce the computational resource demand making it viable to run the model on the wearable devices directly.

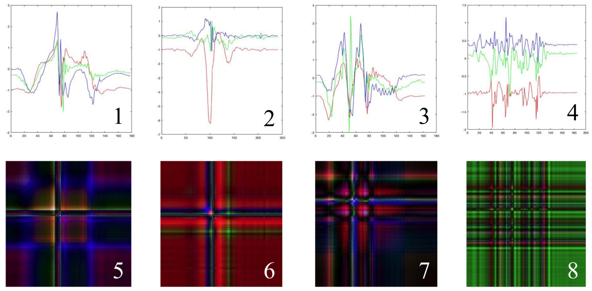

Fig. 2. Selected sensor (top) and their corresponding GAF images (bottom) in UTD-MHAD [12] : (1) basketball shooting; (2) bowling; (3) knock on door and (4) walking.

3. METHODOLOGY

3.1. Virtual Image Generation

Inspired by [7], we encodes the accelerometer data to image representation first. In short, we denote one of the three axial accelerometer data (for example, x coordinate) as $\mathbf{X} = \{x_1, x_2, ..., x_n\}$ and normalize it into $\hat{\mathbf{X}}$ among interval [-1, 1]. The normalized $\hat{\mathbf{X}}$ was then encoded into the polar coordinate (θ, γ) from the normalised amplitude and the radius from the time t, as represented in Eq.1:

$$g\left(\hat{x_i}, t_i\right) = \left[\theta_i, r_i\right] \quad \text{where} \quad \begin{cases} \theta_i = arccos(\hat{x_i}), x_i \in \hat{\mathbf{X}} \\ r_i = t_i \end{cases} \tag{1}$$

After this transformation, the correlation coefficient between vectors can be easily calculated using the trigonometric sum between points [7]. The tri-axial sensor data with the size of n can be assembled as an image representation $\mathbf{P} = (\mathbf{G_x}, \mathbf{G_y}, \mathbf{G_z})$ of size $n \times n \times 3$. Selected examples of sensor and their corresponding GAF images in UTD-MHAD [12] are shown in Figure 2.

3.2. DASK Loss

Hinton *et al.* [6] proposed a KD method which compress knowledge from larger model (i,e. *teacher*) into a smaller model (i,e. *student*), while retaining decent accuracy performance. Given a teacher model T_k and a student model S_k, the soft-target $\tilde{y}^T$ produced by the teacher model is considered as high-level knowledge. The loss of KD when training student model can be defined as:

$$\mathcal{L}_{KD} = \mathcal{L}_C(y, y^S) + \alpha \mathcal{L}_K(\tilde{y}^T, \tilde{y}^S) \tag{2}$$

$$\mathcal{L}_K = \frac{1}{m} \sum_{k=0}^{m} KL(\frac{P^{T_k}}{T}, \frac{P^{S_k}}{T}) \tag{3}$$

where y and y^S refer to the predicted labels and class probability for the student network in this study, respectively. $\tilde{y}^S$ is the soft target generated by the student model. Here $\mathcal{L}_C$ is the typical cross-entropy loss and $\mathcal{L}_K$ is the Kullback-Leibler (KL) divergence, while P^{T_k} is the class probability for the teacher network and P^{S_k} is the class probability for the student network. T represents the temperature controlling the distribution of the provability and we use $T = 4$ in this study according to [6].

KD methods [6, 13] assume the knowledge as a learned mapping from inputs to outputs, which means the outputs themselves contain some relative information from inputs. Therefore, in order to minimize the modality gap between the vision and the sensor domain, we focus on information transfer. More specifically, not only do we to conduct the VSKD based on individual predicted outputs, we also need to consider the structural relation, such as the distance and the angle information, as well as the semantic information among those two modalities that share the same action activity. Therefore, given a pair of training examples, the distance-wise function ψ_D tries to minimize the Euclidean distance between teacher and student examples. μ is a normalization factor for distance and l_δ is Huber loss. The distance-wise distillation loss L_D, which tries to penalize the distance differences between teacher and student outputs is defined as:

$$\psi_D(t_i, t_j) = \frac{1}{\mu}\|(t_i - t_j)\|_2 \tag{4}$$

$$\mathcal{L}_D = \sum_{(x_i, x_j)\in X^2} l_\delta(\psi_D(t_i, t_j), \psi_D(s_i, s_j)) \tag{5}$$

Similarly, given three training examples, the angle-wise function ψ_A tries to minimize the angle between teacher and student examples. The angle-wise distillation loss L_A which tries to transfer the angle-relation information among teacher and students outputs is defined as:

$$\begin{aligned} &\psi_A(t_i, t_j, t_k) = \cos\angle t_j t_j t_k = \langle e^{ij}, e^{kj}\rangle \\ \text{where} \quad &e^{ij} = \frac{t_i - t_j}{\|(t_i - t_j)\|_2}, e^{kj} = \frac{t_k - t_j}{\|(t_k - t_j)\|_2} \end{aligned} \tag{6}$$

$$\mathcal{L}_A = \sum_{(x_i, x_j, x_k)\in X^2} l_\delta(\psi_A(t_i, t_j, t_k), \psi_A(s_i, s_j, s_k)) \tag{7}$$

In addition, since multi-modal data have the same semantic content, semantic loss is defined as:

$$\mathcal{L}_S = \frac{1}{m}\sum_{k=1}^{m}(\|H^S - H^T\|)_2^2 \tag{8}$$

where H^S and H^T represents the feature of fc0 layer from both student and teacher models. To keep H^S and H^T spatial dimensions same, we add one more fc layer (fc0) before its original fc layer (fc1) in the ResNet18 shown in Figure 1.

In summary, we use the original KD loss L_{K_D} and augment it to include the distance and the angle-wised distillation loss L_D, L_A as well as the semantic loss L_S, to train the student network and the final DASK loss for the student model is defined as follow:

$$\mathcal{L}_T^S = L_{KD} + \beta(L_D + L_A) + \gamma L_S \tag{9}$$

where α, β, γ are the tunable hyperparameters to balance the loss terms for the student network.

Method	Testing Modality	Accuracy (%)
Singh *et al.* [15]	Acc. + Gyro.	91.40
Ahmad and Khan [9]	Acc. + Gyro.	95.80
Wei *et al.* [16]	Acc. + Gyro.	90.30
Chen *et al.* [17]	Acc. + Gyro.	96.70
Garcia-Ceja *et al.* [18]	Acc.	90.20
Student baseline	Acc.	94.87
VSKD model	**Acc.**	**96.97 (2.1 ↑)**

Table 1. Comparison result on accuracy performance of UTD-MHAD. The number in parenthesis means increased accuracy over the student baseline. Acc. denotes accelerometer and Gyro. denotes gyroscope.

Method	Testing Modality	Accuracy (%)	F1 score (%)
Das *et al.* [19]	Acc.(Six locations)	88.90	88.80
Student Baseline	Acc. (Left Wrist)	89.09	89.27
VSKD model	**Acc. (Left Wrist)**	**90.18 (1.09 ↑)**	**91.58**
Student Baseline	Acc. (Right Wrist)	86.54	88.33
VSKD model	Acc. (Right Wrist)	87.64 (1.10 ↑)	89.90
Student Baseline	Acc. (Left Hip)	83.27	84.45
VSKD model	Acc. (Left Hip)	83.82 (0.55 ↑)	83.91
Student Baseline	Acc. (Right Hip)	81.09	81.45
VSKD model	Acc. (Right Hip)	82.55 (1.46 ↑)	82.99
Student Baseline	Acc. (Left Ankle)	65.09	64.73
VSKD model	Acc. (Left Ankle)	65.82 (0.73 ↑)	66.90
Student Baseline	Acc. (Right Ankle)	62.91	62.73
VSKD model	Acc. (Right Ankle)	64.36 (1.45 ↑)	63.60

Table 2. Comparison result on accuracy and F1 performance of Berkeley-MHAD. The number in parenthesis means increased accuracy and F1 score over the student baseline.

4. EXPERIMENTS

4.1. Dataset

In this study, three benchmark datasets were selected due to their multi-modal data forms: *MMAct* [11], *UTD-MHAD* [12] and *Berkeley-MHAD* [14]. Specially, we use video streams as the teacher modality input and accelerometer data as the student modality input.

4.2. Experimental Settings

For the teacher network, we used multi-scale TRN [20] with BN-Inception pre-trained on ImageNet due to its balance between the number of parameters and efficiency. In the teacher network, we set the dropout ratio as 0.5 to reduce the effect of over-fitting. The number of segments is set as 8 for Berkeley-MHAD and UTD-MHAD, while 3 for the MMAct. For the student model, we used ResNet18 as the backbone. All the experiments were performed on four Nvidia GeForce GTX 1080 Ti GPUs using PyTorch. We employed the classification accuracy and F-measure as the evaluation metric to compare the performance of the VSKD model with: 1) a student baseline model (ResNet18); 2) other work in which time-series data were applied.

4.3. Experimental Results

The comparison results of three datasets are shown in Table 1, 2, and 3, respectively. In Table 1, the proposed VSKD model performs better than all the previous comparable models. We make an improvement in the testing accuracy of 6.77% compared to the accelerometer view method which extracted 16 features from accelerometer signals for classification [18]. The VSKD model achieved 2.1% higher in accuracy performance, compared to just the student model alone. This result sheds light on incorporating video modality for improving sensor-based HAR. It is worth noting that the proposed VSKD model even performs better as compared to the methods where the accelerometer and gyroscope data were used for testing [15, 9, 16, 17]. These results demonstrated that accelerometer data in the VSKD model can significantly learn knowledge from video streams, thus make an improvement in testing accuracy by 0.37%-6.77%. In Table 2, the proposed VSKD model trained with vision and sensor modality can outperform all the student baseline models. Even though gray-scale video streams on the Berkeley-MHAD dataset lack color information which may degrade the knowledge transfer process, the improvements ranged from 0.55% to 1.46% can still be obtained by the additional support of multi-modal modalities. Also, the VSKD model tested with the left wrist accelerometer data performs better compared to the previous study where accelerometer data from six locations were used [19]. This might be due to the fact that Berkeley-MHAD dataset includes activities more related to hand activity, such as waving hands, clapping, and throwing. In Table 3, while accelerometer data from the phone is the only modality in the testing phase, the method achieves better F-score performance compared to [11, 21] in which either video streams or accelerometer data from phone and watch was used in the testing phase. We also note that the VSKD method trained with accelerometer data from the watch performs worse than the one with accelerometer data from phone. This result was consistent with previous works which showed arm movements introduce additional variability giving rise to a degradation in HAR [22]. Another reason is because MMAct dataset includes activities more related to leg activity, such as sitting, kicking, and jumping.

Method	Testing Modality	Cross Subject (%)	Cross Session (%)
Kong *et al.* [11]	Acc.(Watch+Phone)	62.67	70.53
Kong *et al.* [21]	RGB video	65.10	62.80
Student baseline	Acc. (Phone)	55.44	61.38
VSKD model	**Acc. (Phone)**	**65.83 (10.39 ↑)**	**73.64 (12.26 ↑)**
Student baseline	Acc. (Watch)	46.83	20.63
VSKD model	Acc. (Watch)	60.14 (13.31 ↑)	40.82 (20.19 ↑)

Table 3. Comparison result on F-measurement performance of MMAct. The number in parenthesis means increased F1 score over the student baseline.

Method	Testing Modality	Accuracy (%)	F1 score (%)
ST [6]	Acc.	96.04	96.15
AT [23]	Acc.	96.03	95.80
SP [24]	Acc.	95.80	95.57
DASK-VGG16	Acc.	95.34	95.69
DASK-ResNet18	Acc.	**96.97**	**96.38**
ASK (W/O D)-ResNet18	Acc.	96.73	96.27
DSK (W/O A)-ResNet18	Acc.	96.51	95.80
SK (W/O D and A)-ResNet18	Acc.	96.50	96.06
DAK (W/O S)-ResNet18	Acc.	95.80	96.04

Table 4. Ablation study of accuracy and F1 score performance on UTD-MHAD dataset. W/O denotes Without. D denotes the distance-wise loss L_D. A denotes the angle-wise loss L_A. S denotes the semantic distillation loss L_S.

4.4. Ablation Study

To evaluate the contribution of the proposed DASK loss function, we compare the DASK function with previous KD methods [6, 23, 24]. For those methods, we use the shared codes, and the parameters are selected according to the default setting. As shown in Table 4, the proposed DASK loss function performs better than all of the previous comparable KD loss functions, proving that both structural relation and semantic information are critical for time-series data in a KD process. Also, angle-wised loss L_A contributes more (0.22%) to accuracy improvement as compared to distance-wised loss L_D, indicating time-series data are more valuable in giving 3D information about local body movement. Furthermore, compared to structural relation information, semantic loss L_S contributes more to accuracy improvement (0.70%), which highlights the critical role of semantic information on sensor-based HAR. The proposed VSKD model with ResNet18 as the student baseline performed better (an accuracy of 96.97%) as compared to the VSKD model where VGG16 was used as the student baseline (95.34%). This happens because different convolutional layers in ResNet tend to learn different types of features regarding the input [25], therefore, ResNet18 model is more effective in capturing these features compared to VGG16 model.

5. CONCLUSION

In this paper, we propose an end-to-end Vision-to-Sensor Knowledge Distillation (VSKD) model, which not only improves the sensor-based HAR performance, but also reduces the computational resource demand during the testing phase. We also propose a novel loss function (DASK), which highlights the importance of structural relation and semantic information for bridging the modality gap between vision and sensor domain. Extensive experimental results on UTD-MHAD, MMAct, and Berkeley-MHAD datasets demonstrate the competitiveness of the proposed VSKD model.

6. ACKNOWLEDGEMENTS

This research is supported by NSF SCH-2123749 Collaborative Research. This article solely reflects the opinions and conclusions of its authors and not the funding agents.

7. REFERENCES

[1] Michalis Vrigkas, Christophoros Nikou, and Ioannis A Kakadiaris, "A review of human activity recognition methods," *Frontiers in Robotics and AI*, vol. 2, pp. 28, 2015.

[2] Zehua Sun, Jun Liu, Qiuhong Ke, Hossein Rahmani, Mohammed Bennamoun, and Gang Wang, "Human action recognition from various data modalities: A review," *arXiv preprint arXiv:2012.11866*, 2020.

[3] Haodong Guo, Ling Chen, Liangying Peng, and Gencai Chen, "Wearable sensor based multimodal human activity recognition exploiting the diversity of classifier ensemble," in *UbiComp*, 2016, pp. 1112–1123.

[4] Taylor R Mauldin, Marc E Canby, Vangelis Metsis, Anne HH Ngu, and Coralys Cubero Rivera, "Smartfall: A smartwatch-based fall detection system using deep learning," *Sensors*, vol. 18, no. 10, pp. 3363, 2018.

[5] Yang Liu, Keze Wang, Guanbin Li, and Liang Lin, "Semantics-aware adaptive knowledge distillation for sensor-to-vision action recognition," *TIP*, 2021.

[6] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[7] Feri Setiawan, Bernardo Nugroho Yahya, and Seok-Lyong Lee, "Deep activity recognition on imaging sensor data," *Electronics Letters*, vol. 55, no. 17, pp. 928–931, 2019.

[8] Yuheng Wang, Haipeng Liu, Kening Cui, Anfu Zhou, Wensheng Li, and Huadong Ma, "m-activity: Accurate and real-time human activity recognition via millimeter wave radar," in *ICASSP*, 2021, pp. 8298–8302.

[9] Zeeshan Ahmad and Naimul Mefraz Khan, "Multidomain multimodal fusion for human action recognition using inertial sensors," in *BigMM*, 2019, pp. 429–434.

[10] Katerina Karagiannaki, Athanasia Panousopoulou, and Panagiotis Tsakalides, "An online feature selection architecture for human activity recognition," in *ICASSP*, 2017, pp. 2522–2526.

[11] Quan Kong, Ziming Wu, Ziwei Deng, Martin Klinkigt, Bin Tong, and Tomokazu Murakami, "Mmact: A large-scale dataset for cross modal human action understanding," in *ICCV*, 2019, pp. 8658–8667.

[12] Chen Chen, Roozbeh Jafari, and Nasser Kehtarnavaz, "Utd-mhad: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor," in *ICIP*, 2015, pp. 168–172.

[13] Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho, "Relational knowledge distillation," in *CVPR*, 2019, pp. 3967–3976.

[14] Ferda Ofli, Rizwan Chaudhry, Gregorij Kurillo, René Vidal, and Ruzena Bajcsy, "Berkeley mhad: A comprehensive multimodal human action database," in *WACV*, 2013, pp. 53–60.

[15] Satya P Singh, Madan Kumar Sharma, Aimé Lay-Ekuakille, Deepak Gangwar, and Sukrit Gupta, "Deep convlstm with self-attention for human activity decoding using wearable sensors," *IEEE Sensors Journal*, vol. 21, no. 6, pp. 8575–8582, 2020.

[16] Haoran Wei, Roozbeh Jafari, and Nasser Kehtarnavaz, "Fusion of video and inertial sensing for deep learning–based human action recognition," *Sensors*, vol. 19, no. 17, pp. 3680, 2019.

[17] Chen Chen, Roozbeh Jafari, and Nasser Kehtarnavaz, "Fusion of depth, skeleton, and inertial data for human action recognition," in *ICASSP*, 2016, pp. 2712–2716.

[18] Enrique Garcia-Ceja, Carlos E Galván-Tejada, and Ramon Brena, "Multi-view stacking for activity recognition with sound and accelerometer data," *Information Fusion*, vol. 40, pp. 45–56, 2018.

[19] Avigyan Das, Pritam Sil, Pawan Kumar Singh, Vikrant Bhateja, and Ram Sarkar, "Mmhar-ensemnet: A multimodal human activity recognition model," *IEEE Sensors Journal*, vol. 21, no. 10, pp. 11569–11576, 2020.

[20] Bolei Zhou, Alex Andonian, Aude Oliva, and Antonio Torralba, "Temporal relational reasoning in videos," in *ECCV*, 2018, pp. 803–818.

[21] Quan Kong, Wenpeng Wei, Ziwei Deng, Tomoaki Yoshinaga, and Tomokazu Murakami, "Cycle-contrast for self-supervised video representation learning," *arXiv preprint arXiv:2010.14810*, 2020.

[22] Rubén San-Segundo, Henrik Blunck, José Moreno-Pimentel, Allan Stisen, and Manuel Gil-Martín, "Robust human activity recognition using smartwatches and smartphones," *Engineering Applications of Artificial Intelligence*, vol. 72, pp. 190–202, 2018.

[23] Sergey Zagoruyko and Nikos Komodakis, "Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer," *ICLR*, 2017.

[24] Frederick Tung and Greg Mori, "Similarity-preserving knowledge distillation," in *ICCV*, 2019, pp. 1365–1374.

[25] Aditya Prakash, Kashyap Chitta, and Andreas Geiger, "Multi-modal fusion transformer for end-to-end autonomous driving," in *CVPR*, 2021, pp. 7077–7087.