Robust Video Activity Recognition Augmented by Network Traffic in Connected Environments

ANONYMOUS AUTHOR(S)

[Placeholder.] Activity recognition using video data is widely adopted for elder care, safety/security surveillance, and home automation in home. However, it suffers a lot of operational failures (camera angles, blockages, lighting conditions, and etc.) . In order to solve this, we propose to use network traffic to augment video and add robustness, correct mistakes in in-home activity recognition. Network traffic is lightweight and uniformly exists in any home with a router. We not only propose intermediate fusion of both video-based skeleton features and network traffic features, but also a grey-box model augmentation framework that uses a white-box network traffic model to targetedly augment on and explain any black-box video model. We are able to show the robustness of network features in emulated errors, and the robustness of both sources combined, as well as the robustness when trained with simulated losses.

1 INTRODUCTION

In-home activity recognition can be used for elder care, health monitoring, surveillance, home automation, etc.

The most mature way to do in-home activity recognition is to use Video and side channels (vibration sensors, power lines, networks, and etc.) TODO: A bunch of citations to be put here.

Video Activity Recognition (AR) is well studied and has a lot of pre-trained models or solutions. But it is not robust, i.e., it suffers from a lot of operational failures (camera angles, blockages, lighting conditions, home settings, and etc.) and statistical failures (demonstration, user).

So VAR needs to be paired with one of the side channels. The characteristics of video for AR are (1) it is universal; (2)it is independent of device models; (3) it is very dependent on the environment; (4) it is less accurate at the beginning, but more robust to data losses. Network traffic have the characteristics that directly compliment video. (1) it is universal; (2)it is independent of the environments; (3) it is independent of the environment; (4) it is more accurate at the beginning, but more robust to data losses. TODO: admit one of our limitations is that 15% is not the only data that we needed

TODO: Why network is robust? independent of physical environment, universal measurement, collected in a lab env

TODO: Why not sensor based approach to augment it? Because sensors are not readily available.

Contributions:

- How do we combine? Featurization.
- Robustness evaluation in different settings.
- Algorithm for paired observation collection. TODO: zero paired observation (soft voting), some paired observation (stacking), to all paired observations (intermediate)

venn diagram, the opposite of it

Video: generalizable across devices, dependent on physical environment; generalizable across physical environments, dependent on device model.

2 BACKGROUND

2.1 Related Work

Takeaway: We present previous works on activity recognition, showing the deltas of our approach.

Video-based activity recognition. 1. Pose/Keypoint Estimation OpenPose [6, 7, 19, 22] is a 2D human pose estimation project that represents a person in 135 keypoints. VideoPose3D [17] is a two-step method that does 3D human pose estimation after 2D keypoint detection.

	Universal interface	Environment dependence	Device dependence	Accuracy	Literature
Video	✓	✓		✓	
Network traffic	✓		✓	✓	
Vibration sensors			✓	✓	
Powerline	✓		✓		
Wireless	✓	✓			

Table 1. *Activity recognition sources.TODO: or a venn diagram*

2. Human Action Categorization 2.1. Algorithms/Models SlowFast [12]

Sensor-based activity recognition. [15, 24]

Network-based activity privacy leakage. [1, 2]

Data / feature Augmentation.

ML Robustness. Adversarial ML on Video-based activity recognition [21].

Multi-modal learning and sensor fusion. The world is naturally multi-modal and we have data generated from diverse sources. Multi-model learning builds machine learning models that integrates data sources from multiple modalities, so we can take advantage of complementary information from different sources [4].

Survey papers: [3, 18] Early fusion (feature level) vs. Late fusion (decision level)

[23] fuses audiovisual data for human action recognition. TODO: ted: more details...

Ensemble learning. Ensemble learning combines multiple machine learning models in a way that achieves better accuracy than all composing models [11]. Common ensemble methods include bagging, boosting and stacking. Bagging generates multiple bootstrap sample sets from existing data and learns a small model for each set of these samples. They are then aggregated to get a model with lower variance. Examples of such aggregation include soft voting, hard voting and random forests [5]. Boosting also tries to aggregate multiple weak models, but it focuses on sequentially fitting the most difficult (inaccurate) samples from previous runs. Boosting is usually used to reduce bias and cannot run in parallel. Examples of boosting include AdaBoost [13], XGBoost [10]. Stacking, on the other hand, combines multiple weak models by training a meta-model. Since this step requires extra data that has not been used by any weak models, one interesting hyperparameter is how to split the original dataset.

Related public datasets. Takeaway: what's missing / why do we collect our own data? Existing video activity recognition datasets (such as Kinetics [8, 9, 14, 20] and AVA [16]) were collected from various environments annotated with hundreds of classes. TODO: ted: more reasoning... To the best of our knowledge, there has been no activity recognition datasets involving both video and network traffic modalities.

2.2 Taxonomy of Smart Home Devices and Activities

Takeaway: We taxonomize the smart home devices and related activites by the proximity of interactions and different signal sources.

Activities. Table 2

TODO: List of activities and classes

3 DATASET

Takeaway: We introduce how the dataset are collected: showing the components of testbed, the data collection pipeline, and describe the formations of dataset.

Proximity	Signal sources	Devices
Physical (mechanical, touch) interactions	Network + Video	Smart appliances
Nearby (audio) interactions	Network + Audio	Smart agent, surveillance camera
Remote (app) interactions	Network	Controlled devices

Table 2. *Data collection for testbed in IoT Lab.*

3.1 Testbed

Takeaway: We introduce the IoT Lab and all the IoT devices included in it, showing the diversity of devices.

IoT Lab. We collect data and conduct experiments in our IoT Lab. It is a smart home broadband network environment within an experimental laboratory. We leverage Single Board Computers (SBC, e.g. NVIDIA Jetson nano and rasperry Pi 4) as the home network gateway. They also act as a compute interface to monitor network traffic, internet performance, security and privacy. Various data science and machine learning techniques are applied on these SBCs to uncover valuable insight. Through a dedicated single WiFi connection, powered by Comcast, we connect all smart devices.

Device	Location	Data collected
Router	1 in the lab	Network traffic
Stationary Cameras	2 in the lab	Video (from two angles)
Microphone	1 in the lab	Audio data

Table 3. *Data collection for testbed in IoT Lab.*

As stated in Table 3, to provide a great coordination of ground truths, multiple sources of data are collected within the IoT Lab. These data collectors include in-lab recorder and on-body sensors. (1) an NVIDIA Jetson nano acting as home gateway, which records every single packet from smart devices in both inbound and outbound directions; (2) two IP cameras connected with the Jeston nano. They provide simultaneous video and audio from different perspectives, which ensures human activities and interactions with all smart devices are presented; (3) an additional high-quality [brand name?] microphone to capture voice commands, TV noises, and ambient sound of human activities, so we know when things were triggered.

IoT devices. What devices we have in the lab. The IoT Lab is augmented by seven different types of IoT devices listed in Table. 4: computer, mobile device, wearable, home automation (Philips light bulbs, TP-link smart plug), smart agent (Google Home, Amazon Echo), surveillance (Nest camera, Wyze camera, Ring camer), smart appliances (Samsung smart fridge, Samsung smart TV, Samsung smart stove,Samsung dishwasher, Bose wireless speaker).

3.2 Data collection pipeline

Takeaway: We describe the Web controlled collection pipeline and the collection process.

Justify why network, video and audio: data collection using in a unified way. don't have to deal with data collections across multiple different devices, apps, platforms, sensors, and interfaces. More deployable and generic.

Labeling. Web-based data collection interface. Participant ID, repitition, activity, device, label, timestamp.

Types	Devices
Computer	personal laptop
Mobile devices	Android / iOS smart phone
Wearable	Fossil watch
Smart appliances	Samsung smart fridge, Samsung smart TV, Samsung smart stove, Samsung dishwasher, Bose wireless speaker
Smart agents	Google Home, Amazon Echo
Home automation	Philips light bulbs, TP-link smart plug
Surveillance	Nest camera, Wyze camera, Ring camera

Table 4. *Types and devices in the IoT Lab.*

3.3 Dataset descriptions

Takeaway: We describe the collection of activities and devices and why we make such choices.

TODO: Activities we collected. Show a table

4 ACTIVITY RECOGNITION USING SINGLE SOURCES

4.1 Video Activity Recognition

Takeaway: We describe Classic and SOTA video-based activity recognition methods, including feature selection, network structure, and metrics. Although Video is good for activity recognition because:

- + Lots of pre-trained models
- + Direct observation of activity

It faces severe problems in:

- - Occlusion (Requires LoS)
- - Sensitivity (drift) to environment changes (angle, lighting)
- - Adversarial attacks

4.1.1 Methodology. **SlowFast pretrained model.**

Time series Facebook poses for video.

4.1.2 Operational and statistical failures. Takeaway: We arugue that video classifiers are often harmed by failures during operations, environmental / contextual changes cause the classifications to fail, also by statistical differences across users and demonstrations.

Operational failures. Occlusion, angle, lighting, Adversarial attacks.

Operational failures. User, demonstration

4.2 Why Can Network Traffic Augment?

TODO: If Network Traffic is so good, why not just use it as the main source? Because there's no large pretrained models and it's quite costly?

Network can augment Video because:

- + Robustness under different conditions (generalizability in emulated network conditions)
- + Slower pace to drift (?) Based on per-demonstration differences
- + Don't require LoS
- + Universal In-home interface

But network faces the following problems that need to be addressed

- - Lead and lag
- - No direct observation of activity (strong labels for a side channel)
- - Diverse network traffic pattern mapped to a single activity (e.g., touch screen on fridge). Certain activity could not be easily detected by network.
- - Remote control vs nearby interaction (need to at least discuss it)
- - Multiple people.
- - Limitation of IoT devices.

4.2.1 *Methodology.*

Takeaway: We describe the NetML activity recognition methods, including feature selection, network structure, and metrics.

Assumptions: Device & ip matching, needs more retrains to generalize across devices: collect more data, provide technical solutions to adapt to new environment

Device-ip mapping relationship (let users provide them)

NetML for network.

Tree-structure model management. For every device and activity, we can readily collect the traffic of them in a lab environment.

Robustness under Operational Failures Takeaway: We show with evidence that the results even with failures during operations, like environmental / contextual changes. The results kept robust to under realistic emulation of jitter, latency, packet loss, and packet duplications.

Emulation settings: Tools are TCPReplay + NetEM + tcpdump, on Ubuntu 20.04 in a VM.

Robustness under Statistical Failures Takeaway: We argue the transferability of network traffic to user and demonstration.

Split by user.

Split by demonstration.

4.2.2 *Why is Network Not the Main Source?*

Network traffic does not have big pretrained models.

There are environments that do not have network traffic.

Certain activity could not be easily detected by network.

No direct observation of activity.

Can not handle multiple people (? Maybe we don't mention it at all.)

4.3 Single Source Classification Results

Takeaway: We show the results of classification on 4 devices and multiple activities from both network and video classifiers.

5 ACTIVITY RECOGNITION USING HETEROGENEOUS SOURCES

5.1 Framework

Takeaway: We introduce the overview of our framework, including the goals of this multi-modal approach, two augmentation methods, and their use cases.

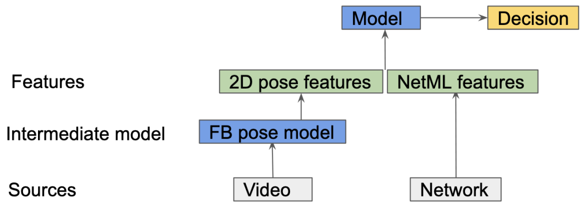


Fig. 1. *Intermediate fusion of video pose features and NetML Stats features.*

Goals: (1) As we degrade the quality of one input, you don't completely lose all accuracy. (2) We can choose the source with more reliability by inspecting data quality

5.2 All Paired Observations: Intermediate Combination

Takeaway: We describe the assumptions required by intermediate augmentation and its details.

Assumption. Video model and network model are trained together.

Methodology. Intermediate augmentation, to concatenate pose features with netml features.

5.3 No Paired Observations: Model Soft Voting

Assumption. In a given environment, video model and network model already exist, and they are trained separately.

TODO: need the experiments that show some paired observation helped.

5.4 Some Paired Observations: Grey-box Model Combination

Takeaway: We describe the assumptions required by grey-box augmentation and the idea behind it and the details of this methodology.

Assumption. A video model already exists in an environment (Because there are many pretrained video models). And network models are trained in a lab setting.

Idea. Use a meta-learner to combine two existing models, and use the information generated from them to smartly collect a minimum set of data to train the meta-learner.

Methodology. A smart(er) way to do late fusion. (1) Use uncertainty scoring techniques to diagnose weak points / under-training regions of a given black-box video model. (2) Based on the information of the diagnosis, augment a subset of devices / activity using white-box network model.

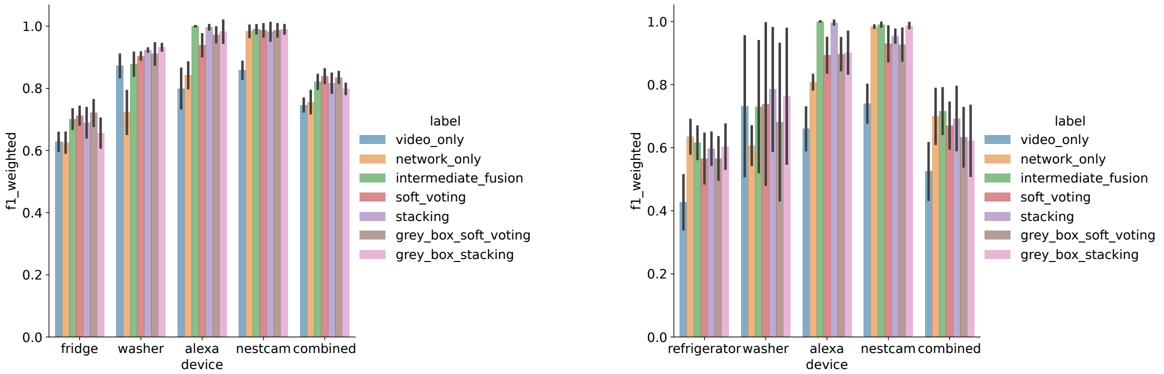
Advantages.

- Can both use the power of pre-trained video models and network traffic
- Targetedly augment device / activity on meta-learner.
- Faster saturation and higher accuracy.

6 EVALUATION

6.1 Combination Methods Effectiveness

TODO: Confusion matrix on 14/12 classes?



(a) *Tested on equal splits.*

(b) *Tested on unseen users.*

Fig. 2. *Weighted F1 Score for Different devices.*

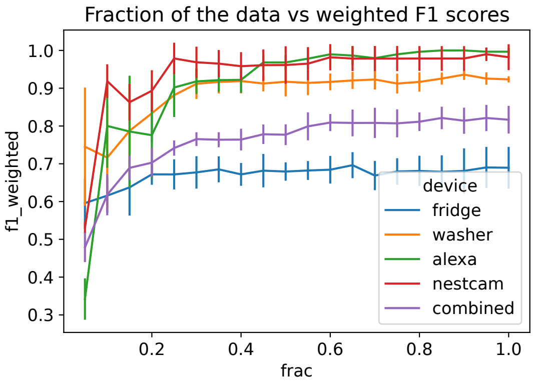


Fig. 3. *Fraction of the data vs weighted F1 scores.*

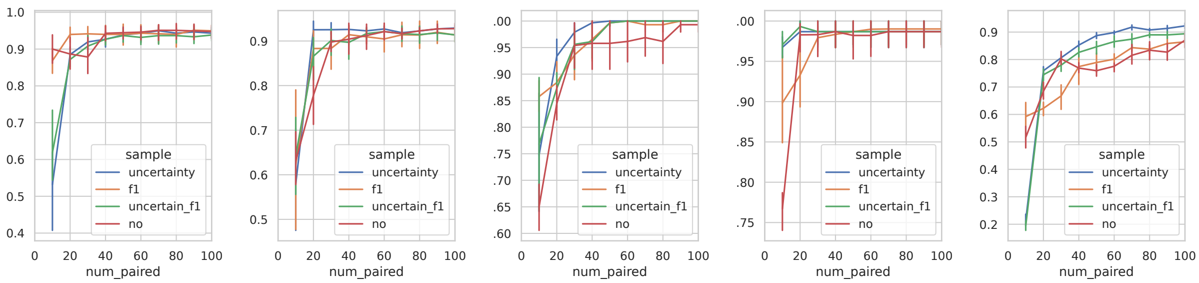
6.2 Smart Stacking Effectiveness

6.3 Robustness Test under Synthetic Data Losses

Takeaway: We demonstrate that using heterogeneous data sources to recognize activity is more effective and robust than any single source, under different scenarios of data losses. And the augmentation methods are better than baselines like bagging and boosting.

Settings.

- Element random loss: crop out N times of 1-sec data randomly.
- Block random loss: crop out roughly same successive timestamp (randomly) for pcaps and videos.
- Element frequent loss in row: crop out N/2 times of 2-sec data randomly, randomly crop out additional 1-sec data if N is odd.
- Successive element loss: crop out random successive timestamp (N-sec).



(a) *Fridge.* **(b)** *Washer.* **(c)** *Alexa.* **(d)** *Nestcam.* **(e)** *Combined.*

Fig. 4. *The weighted F1 score changes when the number of paired observation varies for different classifiers. TODO: equivalant to how many human hours*

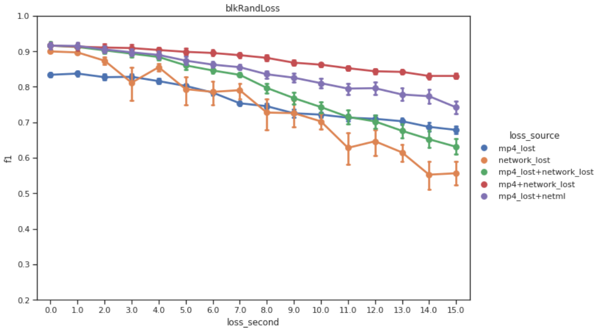


Fig. 5. *Intermediate fusion of video pose features and NetML Stats features.*

Due to its bursty nature, network traffic is more sensitive to losses.
TODO: Rerun on all classes

6.4 Robustness Test under Realistic Failures

Takeaway: We demonstrate that using heterogeneous data sources to recognize activity is more effective and robust than any single source, under a 3D axis of realistic failures. And the augmentation methods are better than baselines like bagging and boosting.

3D axis of failures on network and video. Synthetic and Realistic failures.

3 axis and experiment design:

- How clean the video could be? (e.g. line of sight, angle of camera) Easy: line of sight, Middle: some obstruction, Hard: NLoS.
- How clean the network traffic (e.g. NAT or not NATed, sampling rate, pcap or flow-level summary) Easy: pcap, not NATed, Middle: pcap, NATed, Hard: flow-level summary and NATed Jitter, packet losses (model packet loss)
- Complexity of task (e.g. in-kitchen activity recognition) Stateful or stateless activities, Easy: device on/off, Middle: take sth out of fridge, Hard: undecided.

Device	Video model				Network model			
	equal_mean	equal_std	user_mean	user_std	equal_mean	equal_std	user_mean	user_std
fridge	62.79%	3.21%	42.69%	9.55%	62.52%	3.57%	63.45%	5.96%
washer	87.21%	4.03%	73.12%	27.12%	72.26%	7.68%	60.65%	7.45%
alexa	79.93%	7.12%	65.98%	7.80%	84.19%	4.60%	80.76%	2.64%
nestcam	85.83%	3.01%	73.94%	6.88%	98.35%	2.01%	98.38%	0.39%
combined	74.63%	2.28%	52.47%	10.02%	75.57%	4.00%	69.92%	9.71%

Table 5. *Weighted F1 scores tested on equal split vs on unseen users, for single-source models.*

TODO: Consider boosting and bagging as baselines.

Specific confusing cases.

e.g. poses of fridge vs dryer

7 DISCUSSION

7.1 User influence

Split by user.

7.2 Capacity boundary of each source

Video easier, network harder. Multiple people.

Network easier, Video harder. Device.

7.3 Limitation

7.4 Future work

Activity tracking to provide information for home designers about a better home layout. Activity recognition for elder care, fell down, leave stove / heater / fridge on. Not only in a smart home, but also other connected environments like smart factory, smart building, smart city

8 CONCLUSION

REFERENCES

[1] Abbas Acar, Hossein Fereidooni, Tigist Abera, Amit Kumar Sikder, Markus Miettinen, Hidayet Aksu, Mauro Conti, Ahmad-Reza Sadeghi, and Selcuk Uluagac. 2020. Peek-a-boo: I see your smart home activities, even encrypted!. In *Proceedings of the 13th ACM Conference on Security and Privacy in Wireless and Mobile Networks*. 207–218.

[2] Noah Apthorpe, Dillon Reisman, Srikanth Sundaresan, Arvind Narayanan, and Nick Feamster. 2017. Spying on the smart home: Privacy attacks and defenses on encrypted iot traffic. *arXiv preprint arXiv:1708.05044* (2017).

[3] Pradeep K Atrey, M Anwar Hossain, Abdulmotaleb El Saddik, and Mohan S Kankanhalli. 2010. Multimodal fusion for multimedia analysis: a survey. *Multimedia systems* 16, 6 (2010), 345–379.

[4] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2018. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence* 41, 2 (2018), 423–443.

[5] Leo Breiman. 2001. Random Forests. *Machine Learning* 45, 1 (2001), 5–32. https://doi.org/10.1023/A:1010933404324

[6] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. 2019. OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2019).

[7] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2017. Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. In *CVPR*.

[8] João Carreira, Eric Noland, Andras Banki-Horvath, Chloe Hillier, and Andrew Zisserman. 2018. A Short Note about Kinetics-600. *CoRR* abs/1808.01340 (2018). arXiv:1808.01340 http://arxiv.org/abs/1808.01340

[9] João Carreira, Eric Noland, Chloe Hillier, and Andrew Zisserman. 2019. A Short Note on the Kinetics-700 Human Action Dataset. *CoRR* abs/1907.06987 (2019). arXiv:1907.06987 http://arxiv.org/abs/1907.06987

[10] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Francisco, California, USA) *(KDD '16)*. Association for Computing Machinery, New York, NY, USA, 785–794. https://doi.org/10.1145/2939672.2939785

[11] Thomas G. Dietterich. 2000. Ensemble Methods in Machine Learning. In *Multiple Classifier Systems, First International Workshop, MCS 2000, Cagliari, Italy, June 21-23, 2000, Proceedings (Lecture Notes in Computer Science, Vol. 1857)*, Josef Kittler and Fabio Roli (Eds.). Springer, 1–15. https://doi.org/10.1007/3-540-45014-9_1

[12] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. 2019. SlowFast Networks for Video Recognition. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*. IEEE, 6201–6210. https://doi.org/10.1109/ICCV.2019.00630

[13] Yoav Freund and Robert E Schapire. 1997. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *J. Comput. System Sci.* 55, 1 (1997), 119–139. https://doi.org/10.1006/jcss.1997.1504

[14] Will Kay, João Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. 2017. The Kinetics Human Action Video Dataset. *CoRR* abs/1705.06950 (2017). arXiv:1705.06950 http://arxiv.org/abs/1705.06950

[15] Gierad Laput, Yang Zhang, and Chris Harrison. 2017. Synthetic sensors: Towards general-purpose sensing. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 3986–3999.

[16] Ang Li, Meghana Thotakuri, David A. Ross, João Carreira, Alexander Vostrikov, and Andrew Zisserman. 2020. The AVA-Kinetics Localized Human Actions Video Dataset. *CoRR* abs/2005.00214 (2020). arXiv:2005.00214 https://arxiv.org/abs/2005.00214

[17] Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 2019. 3D human pose estimation in video with temporal convolutions and semi-supervised training. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

[18] Shankar T Shivappa, Mohan Manubhai Trivedi, and Bhaskar D Rao. 2010. Audiovisual information fusion in human–computer interfaces and intelligent environments: A survey. *Proc. IEEE* 98, 10 (2010), 1692–1715.

[19] Tomas Simon, Hanbyul Joo, Iain Matthews, and Yaser Sheikh. 2017. Hand Keypoint Detection in Single Images using Multiview Bootstrapping. In *CVPR*.

[20] Lucas Smaira, João Carreira, Eric Noland, Ellen Clancy, Amy Wu, and Andrew Zisserman. 2020. A Short Note on the Kinetics-700-2020 Human Action Dataset. *CoRR* abs/2010.10864 (2020). arXiv:2010.10864 https://arxiv.org/abs/2010.10864

[21] He Wang, Feixiang He, Zhexi Peng, Tianjia Shao, Yong-Liang Yang, Kun Zhou, and David Hogg. 2021. Understanding the robustness of skeleton-based action recognition under adversarial attack. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14656–14665.

[22] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. 2016. Convolutional pose machines. In *CVPR*.

[23] Qiuxia Wu, Zhiyong Wang, Feiqi Deng, Zheru Chi, and David Dagan Feng. 2013. Realistic Human Action Recognition With Multimodal Feature Selection and Fusion. *IEEE Trans. Syst. Man Cybern. Syst.* 43, 4 (2013), 875–885. https://doi.org/10.1109/TSMCA.2012.2226575

[24] Shibo Zhang, Yaxuan Li, Shen Zhang, Farzad Shahabi, Stephen Xia, Yu Deng, and Nabil Alshurafa. 2022. Deep Learning in Human Activity Recognition with Wearable Sensors: A Review on Advances. *Sensors* 22, 4 (2022), 1476.