Supplementary Materials for
Making the Invisible Visible: Action Recognition Through Walls and Occlusions

Tianhong Li* Lijie Fan* Mingmin Zhao Yingcheng Liu Dina Katabi
MIT CSAIL

This supplementary material provides more detailed description of our data collection system, our implementation details and some ablation studies. Besides, we also provide a demo video to demonstrate the model and results of our proposed RF-Action. Please see the video for details.

1. Appendix A: Data Collection System

We collect our own dataset which we refer to as RF Multi-Modality Dataset (RF-MMD), including RF signals, multi-view videos, and the corresponding 3D skeletons.

We use a radio device to collect RF signals (orange box in Figure 1). Our device uses an FMCW radio with two antenna arrays, one horizontal and one vertical. The radio transmits an FMCW chirp with frequencies sweeping from 5.4 to 7.2 GHz. The raw RF signal is then processed using standard FMCW and antenna array equations to generate vertical and horizontal heatmaps, which are then synchronized with the camera frames. The frame rate is set to 30 heatmaps per second and the transmission power of our device is less than one millie Watt.

We implemented a wireless camera system consisting of 10 cameras to collect multi-view videos (blue boxes in Figure 1). We use a laptop as the camera system controller and 10 Raspberry Pi 3 single-board computers as the remote node to control each camera and synchronize with the master camera system controller. Our radio and cameras are synchronized using the network time protocol (NTP), whose synchronization error is typically less than 10ms.

To generate 3D skeletons for each person, our system first uses the video from each camera to generate 2D skeletons using AlphaPose [1]. However, since each camera may see different people at the same time, we need to associate the 2D skeletons from each camera to get 2D skeletons for the same person from different views. To tell whether a pair of 2D skeletons are from the same person, we compute the distance between each two 2D skeletons. Specifically, a 2D keypoint (e.g. head) determines one line in the 3D space where the 3D keypoint must lie on. Then if a pair of 2D keypoints corresponds to the same keypoint from same person, those two lines in 3D space will have small distance in 3D space. On the other hand, if the pair of 2D keypoints are from two different people, those two lines in 3D space will have a large distance.

Figure 1: The figure shows our data collection system. Orange box is our radio device to collect RF signals. Blue boxes show the camera system to collect video frames. RF signals and video frames are synchronized.

Based on this, we use the average distance between the 3D lines corresponding to various keypoints as the distance metric of two 2D skeletons, and use hierarchical clustering [5] to cluster 2D skeletons from the same person. Once the skeletons are associated, we can triangulate their keypoints to generate the corresponding 3D skeleton [2].

2. Appendix B: Implementation Details and Running Time

The skeleton generation network (section 4.1) is similar to the one used in [6]. It has a feature extraction network (12-layer ResNet), followed by a region proposal network (6-layer ResNet) and a pose estimation network (2-layer ResNet). The skeleton-to-action network (section 4.2) has 5 convolution layers with two-stream (special and temporal) as in HCN [3]. For the attention module, we use a convolu-

*Indicates equal contribution. Ordering determined by inverse alphabetical order.

tion layer with kernel size=1 to generate the masks, and we use a multi-head attention with 8 heads. Training is done with the Adam optimizer on 8 Nvidia TiTan Xp GPUs, with batch size=16 and learning rate=0.001.

The runtime of RF-based action detection is 29.8 FPS, and vision-based action detection is 3.0 FPS on a single Nvidia Titan Xp GPU. The runtime for vision-based inputs is slower because we have to run a pose estimation network on each view of the camera system then triangulate across views to get 3D skeletons.

3. Appendix C: Ablation Study

Baseline: We also compare our method with Graph Distillation [4]. Their mAP on the PKU-MMD dataset is 83.3, which is lower than ours, which is 93.3. Also, we adapt their method and test it on our RF-MMD dataset. This yields mAP of 78.1, which is also lower than our approach at 87.8. As for comparison with models for vision-only modality, the mAP of STA-LSTM, JCRRNN, and Skeleton boxes on the PKU-MMD dataset is 44.4, 32.5, and 54.8 respectively, which is significantly lower than ours.

Max Pooling vs. Multi-Proposal Module To show the effectiveness of our multi-proposal module, we compare max-pooling with our multi-proposal module, as shown in Table 1. The multi-proposal module demonstrates much better performance because max-pooling can predict only one action label during any time period, and cannot deal with scenarios where multiple people are doing actions and interacting simultaneously.

Methods	Action	Interaction	Total
Our Multi-Proposal	87.4	88.6	87.8
Max-pooling	66.7	63.2	65.5

Table 1: Comparison between Multi-Proposal and Max-pooling.

References

[1] H.-S. Fang, S. Xie, Y.-W. Tai, and C. Lu. Rmpe: Regional multi-person pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2334–2343, 2017.

[2] R. Hartley and A. Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003.

[3] C. Li, Q. Zhong, D. Xie, and S. Pu. Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation. *arXiv preprint arXiv:1804.06055*, 2018.

[4] Z. Luo, J.-T. Hsieh, L. Jiang, J. Carlos Niebles, and L. Fei-Fei. Graph distillation for action detection with privileged modalities. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 166–183, 2018.

[5] L. Rokach and O. Maimon. Clustering methods. In *Data mining and knowledge discovery handbook*, pages 321–352. Springer, 2005.

[6] M. Zhao, Y. Tian, H. Zhao, M. A. Alsheikh, T. Li, R. Hristov, Z. Kabelac, D. Katabi, and A. Torralba. Rf-based 3d skeletons. In *Proceedings of the 2018 Conference of the ACM Special Interest Group on Data Communication*, pages 267–281. ACM, 2018.