Egocentric Audio-Visual Object Localization
(Supplementary Materials)

A. Video Demo

In our video demo, we illustrate our motivation and method. Next, we show (a) localization results on the Epic Sounding Object dataset, and (b) on the Ego4D dataset.

B. Details of Our Framework

Our framework consists of a visual branch and an audio branch. We start with elaborating on the audio branch and then move to the visual one.

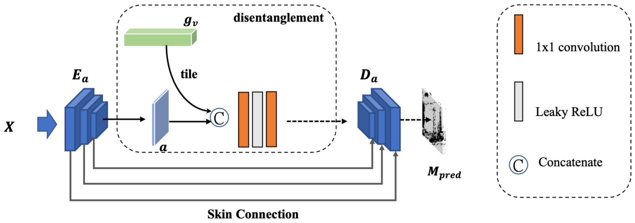


Figure 8. Details of audio branch, including an encoder E_a, a disentanglement network f and a decoder D_a.

Audio Encoder E_a and Decoder D_a. To extract the audio feature from input spectrogram X and predict the separation mask, our audio branch contains an encoder E_a and a decoder D_a as shown in Fig. 8. In practice, we design the encoder and decoder in a U-NET style architecture, i.e., skin connections are enabled. The encoder-decoder network consists of five convolution layers and five up-convolution layers, and all layers adopt a 4x4 kernel with stride 2. A BatchNorm (BN) layer and a ReLU activation layer are appended after each convolution/up-convolution layer. For the last layer in D_a, we use a Sigmoid layer instead of the ReLU and remove the BN layer to output the mask. Skip connections allow the information flow from layer i in E_a to layer $n - i$ in D_a, where $n = 5$ is the total number of layers.

Disentanglement Network. The disentanglement network consists of two 1x1 convolution layers to obtain the visually indicated audio features. Given the audio input features a with a size of 4x4x512, we first tile the visual feature $g_v \in \mathbb{R}^{512}$ by 4x4 times along the Time and Frequency axes to match the size of a. Then we concatenate audio and visual features along the channel dimension. The concatenated features will go through two 1x1 convolution layers with Leaky ReLU in between. The output feature maps are also of size 4x4x512.

In the following, we will explain some details about our visual branch.

Visual Encoder E_v. The visual branch includes a visual encoder E_v to extract the features in the beginning. It takes $T = 5$ frames of dimension 224x224x3 as inputs and outputs 5 feature maps of dimension 28x28x512. We use a pre-trained Dilated ResNet to implement E_v.

Geometry-Aware Temporal Modeling Module. Given frames I_i and I_j and their corresponding features v_i and v_j, we first estimate the homography transformation $\mathcal{H}_{ji}$ between frames. The homography is a 3x3 matrix:

$$\mathcal{H}_{ji} = \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{bmatrix} \tag{1}$$

I_j can thus be warped to I_{ji} as $I_{ji} = \mathcal{H}_{ji} \otimes I_j$. Instead of applying homography at the image level, we use it in the feature space to warp the features. If the homography estimation fails due to the poor image condition or drastic viewpoint change, $\mathcal{H}_{ji}$ will be replaced with an identity matrix:

$$\mathcal{H}_{ji} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{2}$$

Therefore, our GATM will degrade to a vanilla temporal modeling approach but still can leverage the temporal contexts. The temporal context aggregation is implemented as single-head attention at the time dimension.

C. Analysis on Audio Disentanglement

To further investigate the issue of out-of-view sounds in egocentric videos, we experimented with analyzing the performance of our model when dealing with audio mixtures. We compare two different models. One is trained with disentanglement loss $\mathcal{L}_{ids}$ while the other does not. During inference, we generate two different audio inputs: one is by mixing the original audio $s^{(1)}$ with an audio clip $s^{(2)}$

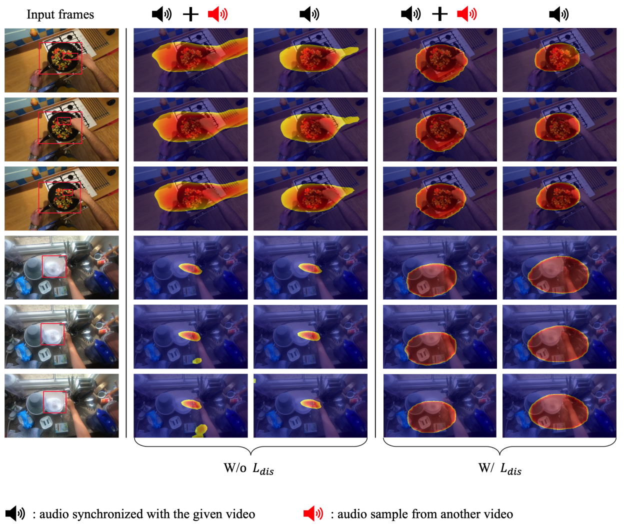

Input frames
W/o L_{dis}
W/ L_{dis}
: audio synchronized with the given video
: audio sample from another video

Figure 9. Visualizations on localization results generated from the model without (column 2-3) and with (column 4-5) audio disentanglement. Our model trained with L_{dis} can yield consistent results when dealing with and without visually irrelevant audio in the input.

Model	Input	@0.2	@0.3	@0.4
w/o L_{dis}	$s^{(1)} + s^{(2)}$	38.15	18.60	10.42
w/ L_{dis}	$s^{(1)} + s^{(2)}$	38.70	19.42	10.51
w/o L_{dis}	$s^{(1)}$	38.21	18.67	10.42
w/ L_{dis}	$s^{(1)}$	38.71	19.42	10.51

Table 4. Analysis of our model's performance by controlling the out-of-view sounds. We compare two models (trained w/ and w/o L_{dis}) and report results over CIoU@{0.2, 0.3, 0.4}.

sampled from another video in the dataset, while the other one takes the original audio $s^{(1)}$ as input. The corresponding results are shown in Table 4. We found that with audio feature disentanglement, our model achieves stable performance for both mixed and original audio inputs. Visualizations in Fig. 9 also demonstrate the effectiveness of disentanglement training.

D. More Experiments

In this section, we provide extra ablations and hyper-parameter experiments to add to the thoroughness of the evaluation.

Baseline	Baseline + SL	Baseline + L_{dis}
27.41	28.83	32.96

Table 5. Analysis of the efficacy of Soft Localization (SL) and disentanglement loss L_{dis}. We report results on CIoU@0.2 metric.

Detailed ablations. We present a more detailed ablation to complement Table 3 in the main paper. Specifically, we add Soft Localization or L_{dis} to the baseline model before adding the geometry-aware temporal modeling module. Results are reported in Table 5. The SL module can slightly increase the performance to 28.83 (5.2%↑). When the disentanglement module is added to the baseline model without GATM, the performance is boosted to 32.96 (20.2%↑), which validates the effectiveness of audio disentanglement for solving out-of-view sounds problem.

$\lambda = 1$	$\lambda = 5$	$\lambda = 10$
30.85	32.96	32.10

Table 6. Hyper-parameter evaluation on the coefficient that controls the impact of disentanglement loss L_{dis}. The results are reported on model "*Baseline + SL*".

Hyper-parameter λ. In Table 6, we evaluate the effect of λ that is used to balance the losses. When λ is small, the model is not sufficiently trained to remove the visually unrelated contents from the audio representation, yielding inferior results. However, as λ becomes larger ($\lambda = 10$), the training objective will focus more on separation instead of accurate localization. Therefore, we select $\lambda = 5$ empirically for our main model.

$T = 3$	$T = 5$	$T = 7$
36.46	37.38	35.34

Table 7. Hyper-parameter evaluation on the number of frames T. The results are reported on model "*Baseline + GATM*".

Hyper-parameter T. The number of frames T used to aggregate the temporal context is important and could further demonstrate the usefulness of the geometry-aware temporal modeling module. In Table 7, we show quantitative results with various frame numbers. We can find that aggregating temporal information is significant as the performance boost from $T = 3$ to $T = 5$. When the frame number becomes even larger, the difficulty in effectively aligning visual frame features increases as greater viewpoint changes happen. Therefore, we choose $T = 5$ in our main experiments.

E. More Visualizations

We visualize more localization results for examples in the Epic Sounding Object and Ego4D [2] datasets (shown in Fig. 10). The figure shows that our framework can correctly localize various sounding objects, e.g., pan, spoon, box, vacuum cleaner, scissor, car and etc. Both indoor and outdoor scenarios are covered. An example of people watch-

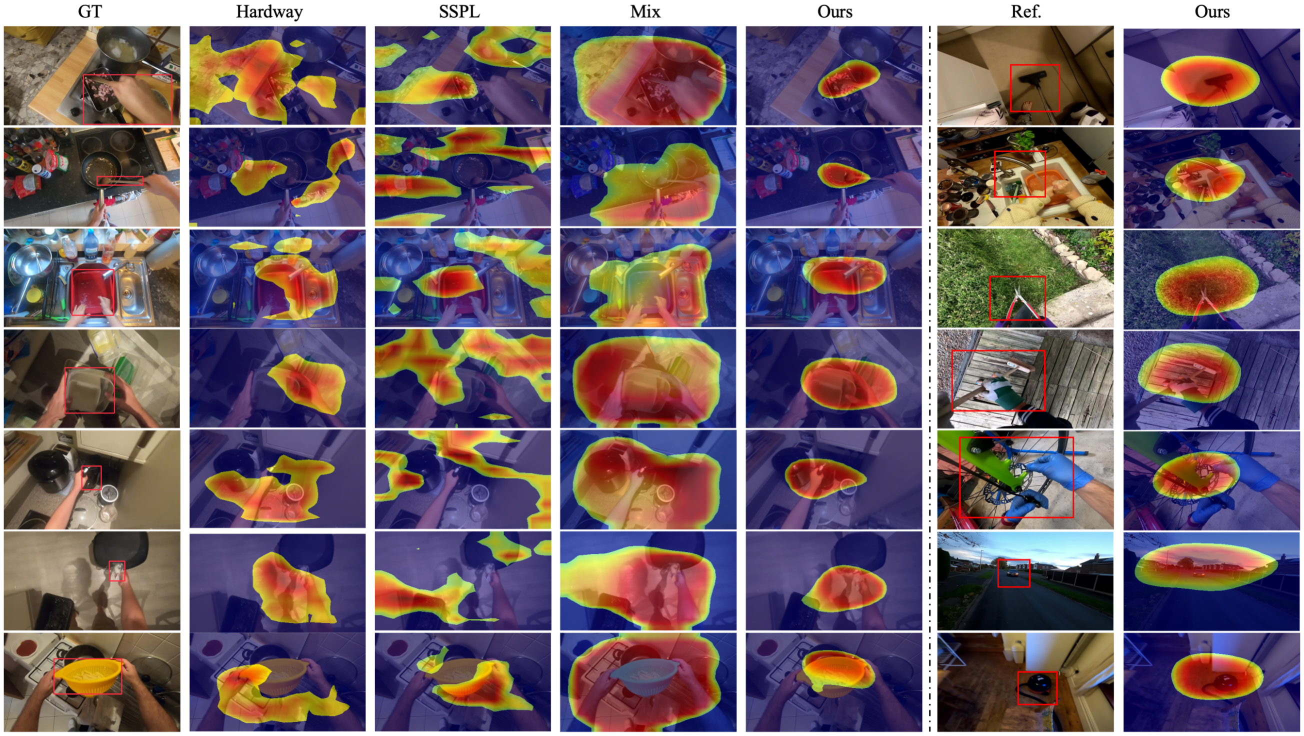


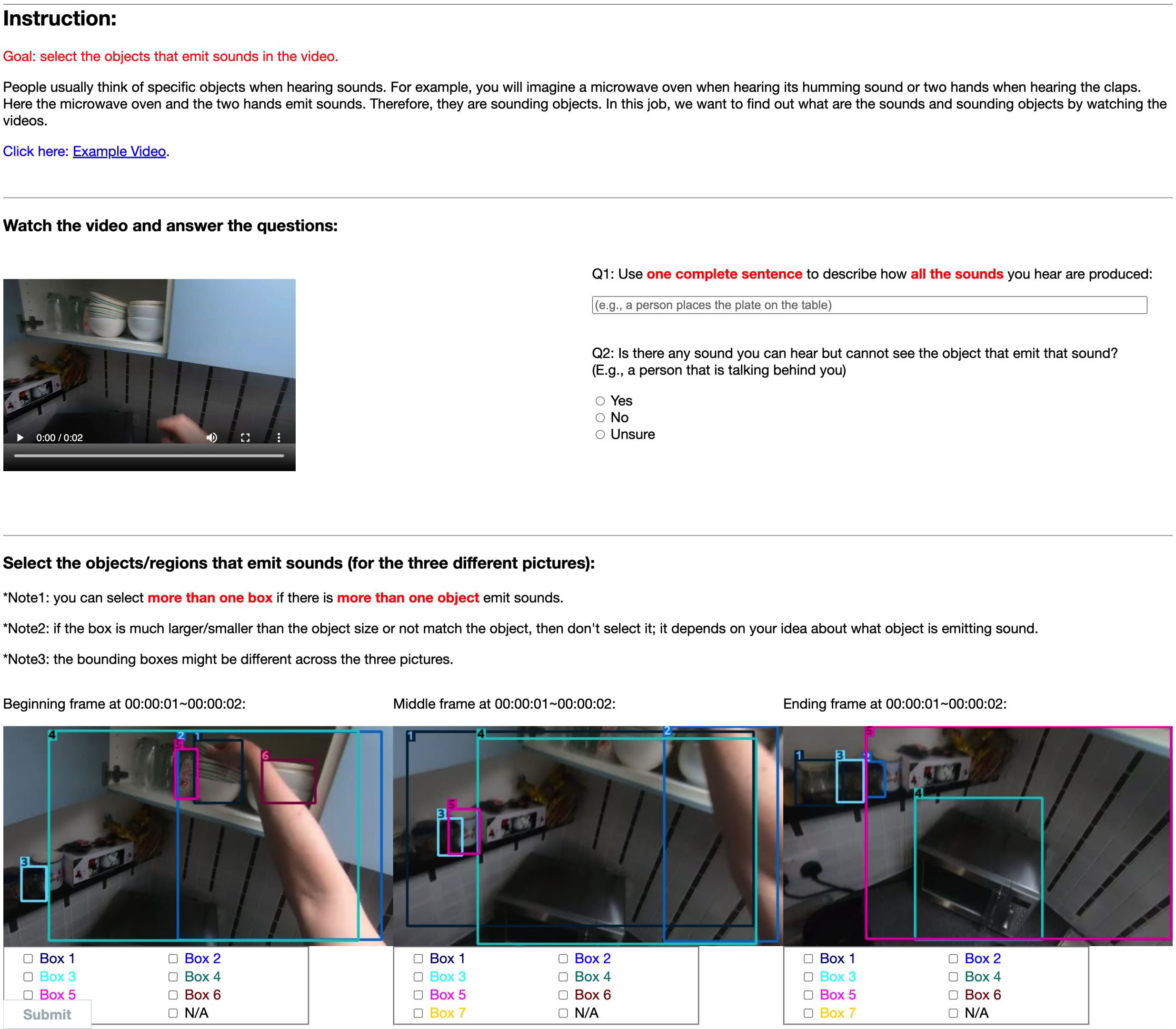
Figure 10. More visualization results of comparative methods and our approach on both Epic Sounding Object dataset (column 1-5) and Ego4D [2] dataset (column 6-7). The corresponding sounding objects are shown in red boxes (columns 1 and 6). Our method can produce more precise localization results and generalize to diverse daily scenarios.

ing independent activity is also shown (Row 6 for Ego4D results).

F. Epic Sounding Object Dataset

Amazon Mechanic Turk Annotation Collection. Annotating sounding objects in egocentric videos is challenging. Sounds are often correlated with human-object interactions, and sounding objects are sometimes occluded or under severe deformation due to frequent viewpoint changes. Therefore, annotating sounding objects automatically is difficult. To this end, we follow a semi-automatic labeling process by first generating bounding boxes for potential sounding objects. We use a Mask R-CNN object detector [3] trained on MS-COCO [5] and a hand-objects detector [6] that pretrained with 42K egocentric images [1, 4, 7] to produce bounding boxes. Second, we manually annotate the sounding objects. Due to the above-described difficulty, people may have different opinions about what objects are emitting sounds. To reduce this uncertainty, we ask three or more people to annotate the same video and apply a voting process to the annotations.

Specifically, we take advantage of the Amazon Mechanic Turk to label the sounding objects. We develop an interface (as shown in Fig. 11) to support this process. First, the annotator is required to watch the video to ensure that sounding objects correspond to the sounds. Then the annotator will answer the following questions: (1) *Use one complete sentence to describe how all the sounds you hear are produced.* Sound sometimes is ambiguous to annotate in bounding boxes when it is made by the interaction between objects, e.g., putting down the dish on the table. Whether "only dish" or "both dish and table" are the sounding object is hard to determine. In this case, a language description is suitable to handle the ambiguity; (2) *Whether the video contains sound you can hear but cannot see.* The answer hints at whether out-of-view sounds exist in the video. In egocentric videos, out-of-view sounds might be created due to the limited FoV. Therefore, we include this question to provide statistical analysis about the out-of-view sound problem; (3) by watching the video, the annotators are required to select the bounding boxes that correspond to the objects that emit sounds. The goal is to select the bounding box that humans recognize as a sounding object. All the collected answers will be passed to a voting process to determine the final annotations for each video (see Fig. 5 in the main paper).



Instruction:

Goal: select the objects that emit sounds in the video.

People usually think of specific objects when hearing sounds. For example, you will imagine a microwave oven when hearing its humming sound or two hands when hearing the claps. Here the microwave oven and the two hands emit sounds. Therefore, they are sounding objects. In this job, we want to find out what are the sounds and sounding objects by watching the videos.

Click here: Example Video.

Watch the video and answer the questions:

Q1: Use **one complete sentence** to describe how **all the sounds** you hear are produced:

(e.g., a person places the plate on the table)

Q2: Is there any sound you can hear but cannot see the object that emit that sound?
(E.g., a person that is talking behind you)

○ Yes
○ No
○ Unsure

Select the objects/regions that emit sounds (for the three different pictures):

*Note1: you can select **more than one box** if there is **more than one object** emit sounds.

*Note2: if the box is much larger/smaller than the object size or not match the object, then don't select it; it depends on your idea about what object is emitting sound.

*Note3: the bounding boxes might be different across the three pictures.

Beginning frame at 00:00:01~00:00:02:

□ Box 1 □ Box 2
□ Box 3 □ Box 4
□ Box 5 □ Box 6
Submit □ N/A

Middle frame at 00:00:01~00:00:02:

□ Box 1 □ Box 2
□ Box 3 □ Box 4
□ Box 5 □ Box 6
□ Box 7 □ N/A

Ending frame at 00:00:01~00:00:02:

□ Box 1 □ Box 2
□ Box 3 □ Box 4
□ Box 5 □ Box 6
□ Box 7 □ N/A

Figure 11. Example of our annotation interface. We ask every Amazon Mechanic Turk worker to watch the video first. Then, they are required to answer two questions and annotate the sounding objects correspondingly. The rules of selecting sounding objects ensure the quality of annotations. Besides, we conduct a voting process to obtain precise annotations for each video.

G. Potential Applications

Our work has the potential to facilitate a range of applications, which are described below:

Audio-Visual Episodic Memory. As egocentric video records what and where of an individual's daily life experience, it would be interesting to build an intelligent AR assistant to search the object that has been presented in the past. Previously, the episodic memory task could take an image or language query as input to localize the object. The correlation between visual objects and sound is less explored. Our egocentric audio-visual object localization task can extend episodic memory with auditory sense. As shown in Fig. 12 (a), people can ask "where did I use the vacuum cleaner?", and the AR assistant can give the answer by feeding an audio query (a vacuum cleaner audio clip from the web) to the localization network. Therefore, the vacuum cleaner can be found in the video.

Audio-Visual Object State. In egocentric research, it is important to know the state of objects that humans interact with. The recognized object's state helps understand the wearer's actions. Interestingly, human-object interaction often makes a sound. Therefore, localizing objects by sounds provides a new angle in recognizing an object's state. For example, in Fig. 12 (b), both the laptop and the pot make sounds. By feeding the audio and the frame to the localization model, the laptop and the pot are localized correspondingly, which indicates that both objects are in

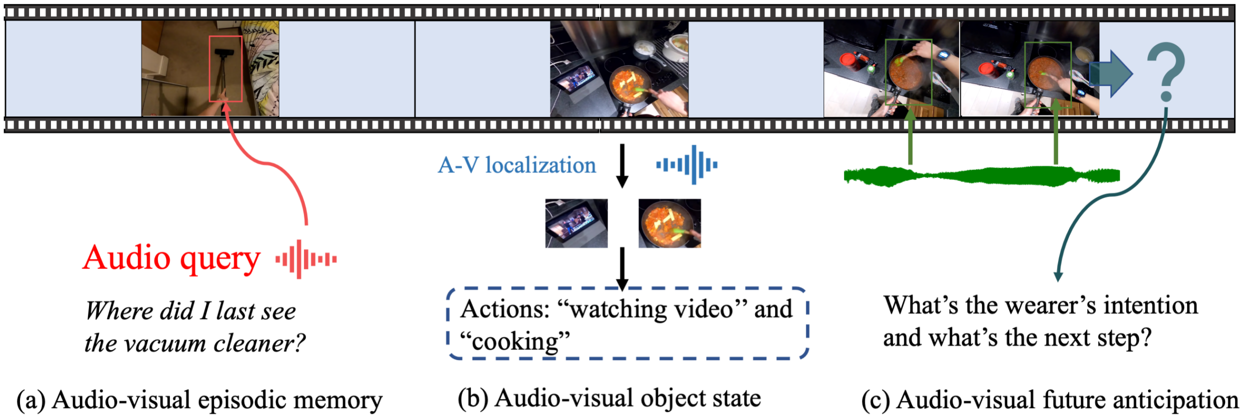

A-V localization
Audio query
Where did I last see the vacuum cleaner?
Actions: "watching video" and "cooking"
What's the wearer's intention and what's the next step?
(a) Audio-visual episodic memory
(b) Audio-visual object state
(c) Audio-visual future anticipation

Figure 12. An overview of potential applications following egocentric audio-visual object localization, including (a) audio-visual episodic memory, (b) audio-visual object state, and (c) audio-visual future anticipation. These applications target understanding the *past*, *current*, and *future* of the wearer's experience. Such capability is enabled by taking fine-grained audio-visual perception.

a "working" state. Consequently, the "watch video" and "cooking food using a pot" actions are easily detected.

Audio-Visual Future Anticipation. As the audio-visual object state task indicate the human activity at the "current" moment by utilizing the sounding object results, it is natural to predict the future by analyzing the most recent audio-visual clips. The sound may change continuously as the object's state changes. In Fig. 12 (c), when the cooking is completed, the frying pot sound will be different and hence indicate the wearer's future action: the wearer may move the pot to the table or pour food into a bowl. Thus, by analyzing the audio-visual object state changes, the future of the wearer can be anticipated.

References

[1] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Evangelos Kazakos, Jian Ma, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Rescaling egocentric vision. *arXiv preprint arXiv:2006.13256*, 2020. 3

[2] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. *arXiv preprint arXiv:2110.07058*, 2021. 2, 3

[3] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 3

[4] Yin Li, Zhefan Ye, and James M Rehg. Delving into egocentric actions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 287–295, 2015. 3

[5] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 3

[6] Dandan Shan, Jiaqi Geng, Michelle Shu, and David F Fouhey. Understanding human hands in contact at internet scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9869–9878, 2020. 3

[7] Gunnar A Sigurdsson, Abhinav Gupta, Cordelia Schmid, Ali Farhadi, and Karteek Alahari. Charades-ego: A large-scale dataset of paired third and first person videos. *arXiv preprint arXiv:1804.09626*, 2018. 3