Listen to Look: Action Recognition by Previewing Audio (Supplementary Materials)

Ruohan Gao[1,2] Tae-Hyun Oh[2] Kristen Grauman[1,2] Lorenzo Torresani[2]
[1]The University of Texas at Austin [2]Facebook AI Research
rhgao@cs.utexas.edu, {taehyun,grauman,torresani}@fb.com

The supplementary materials for [3] consist of:

A. Supplementary video showing useful moments selected by our method for a few video examples.

B. Dataset details.

C. Implementation details.

D. Single-modality architecture of IMGAUD-SKIMMING.

E. Clips where audio helps the most/least in distillation.

F. Trade-off between efficiency and accuracy on Mini-Sports1M.

G. Ablation study.

H. Additional qualitative results.

A. Supplementary Video

In our supplementary video, we show examples of (a) the *visually* useful moments selected by our method using the *visual* modality versus those obtained by uniform sampling, and (b) the *acoustically* useful moments selected by our method using the *audio* modality versus those obtained by uniform sampling.

From these examples, we can see that the selected moments by our method are more indicative of the corresponding action. The visually useful moments capture the key frames of representative object and scene configurations, such as the frames of a person on the bike for the action mountain biking, different stages of the cake for the action making a cake, and the key body pose for the action throwing discus; The acoustically useful moments capture the key audio segments, such as the sound of the barbell hitting the ground for the action barbell snatch, the sound of the buzzer indicating contact for the action doing fencing, and the car engine sound for the action stock car racing.

B. Dataset Details

We use a total of 4 datasets for evaluation: Kinetics-Sounds [1], UCF-101 [6], ActivityNet [2], and Mini-Sports1M [4]:

- **Kinetics-Sounds** is a subset of Kinetics and consists of only action classes that are potentially recognizable both visually and aurally. It is assembled by [1] and consists of 34 classes. However, 3 classes were removed from the original Kinetics dataset. Therefore, we use the remaining 31 classes in our experiments. The 31 action classes are: blowing nose, blowing out candles, bowling, chopping wood, dribbling basketball, laughing, mowing lawn, playing accordion, playing bagpipes, playing bass guitar, playing clarinet, playing drums, playing guitar, playing harmonica, playing keyboard, playing organ, playing piano, playing saxophone, playing trombone, playing trumpet, playing violin, playing xylophone, ripping paper, shoveling snow, shuffling cards, singing, stomping grapes, tap dancing, tapping guitar, tapping pen, and tickling.
- **UCF-101** is a dataset of about 13K short trimmed clips of 101 human actions. We use the official training/validation/testing splits (split1) in our experiments.
- **ActivityNet** contains videos of various lengths with average duration of 117 seconds. We use the latest release (version 1.3), which consists of around 20K videos of 200 classes. We use the official training/validation/testing splits in our experiments.
- **Mini-Sports1M** is a subset of Sports1M dataset containing an equal number of videos for each class. It is assembled by us to facilitate comparisons of video-level action recognition following [5]. We only take videos of length 2-5 mins, and randomly sample 30 videos for each class for training, and 10 videos for each class for testing.

Figure 1: Single-modality architecture of the IMGAUD-SKIMMING network. At each time step, it takes the indexed feature for the current time step as well as the previous hidden state and cell output as input, and produces the current hidden state and cell output. The hidden state for the current time step is used to make predictions about the next moment to focus on in the untrimmed video through the querying operation illustrated in Fig. 4 in the main paper. The average-pooled features of all selected time steps is used to make the final prediction of the action in the video.

C. Implementation Details

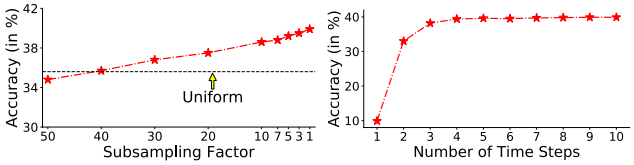
Both our IMGAUD2VID network and IMGAUD-SKIMMING network are implemented in PyTorch. For IMGAUD2VID distillation, we use a R(2+1)D-18 video recognition model [7] pre-trained on Kinetics as the teacher model. It takes 16 frames of size 112×112 as input and produces a video descriptor of 512 dimensions. The 16 frames are taken by sampling every other frame from the raw video at 30 fps, so they roughly span 1 second. The image network for the student model is a ResNet-18 network that takes the starting RGB frame of size 112×112 as input. The audio network is the same as the image network except that we change the first convolution layer to take a 1-channel audio-spectrogram of size 101×40 as input. We also change the average pooling layer after all ResNet blocks to pool only over the temporal dimension. For all experiments, we subsample the audio at 16kHz, and the input audio sample is 1s long. STFT is computed using a Hann window size of 400 and a hop length of 160, producing a 101×40 mel-spectrogram audio representation with 40 mel filterbanks. Both streams generate an output of 256 dimensions and thus the concatenated representations yield an image-audio embedding of 512 dimensions. The network is trained using an SGD optimizer with weight decay of 1×10^{-4} and Nesterov momentum of 0.9. The starting learning rate is set to 1×10^{-3}. The network is trained for 60 epochs with a batch size of 16, and the learning rate decays by 10 times at 30th epoch and 50th epoch.

For IMGAUD-SKIMMING, we use a one-layer LSTM with 1,024 hidden units. $\texttt{Query}(\cdot)$ is a two-layer MLP network that maps the LSTM hidden state of dimension 1,024 to a query vector of dimension 512. The hidden layer of the MLP has 1,024 dimension and is followed by batchnorm and ReLU. $\texttt{Key}(\cdot)$ is a linear layer that maps indexing features of 512 dimensions to indexing keys of the same dimensionality. We implement it using a conv 1x1 layer followed by batchnorm and ReLU. We sample an image-audio pair every 16 frames for each video, and extract the corresponding image and audio features by using the student models. We use $T = 10$ time steps during training, and use the mean image and audio feature vectors as input for the first time step. We mask out the selected index at each time step to encourage diversity of the selected moments. The masking operation is performed independently for the visual and audio modalities. The network is trained using an SGD optimizer with weight decay of 1×10^{-4} and Nesterov momentum of 0.9. The starting learning rate is set to 1×10^{-2}. The network is trained for 25 epochs with a batch size of 256 and the learning rate decays by 10 times at the 15th and the 20th epoch.

D. Single-Modality Architecture of IMGAUD-SKIMMING

Figure 1 illustrates the single modality version of our IMGAUD-SKIMMING network, where we only use a single modality for indexing and recognition. It is in the same spirit of Fig. 3 in the main paper, but here we only use the visual modality. This single-modality version of our method was mainly tested for compatibility with existing methods to compare in Fig. 7 in the main paper. The feature sequences can be any visual features extracted from a visual classifier, *e.g.*, ResNet-101 features and R(2+1)D-152 features as used in Fig. 7 and Table 2 in the main paper. At the t-th time step, the LSTM cell takes the *indexed* feature $\tilde{\mathbf{z}}_\mathbf{t}$ as input, as well as the previous hidden state $\mathbf{h}_{t-1}$ and the previous cell output $\mathbf{c}_{t-1}$ as input, and produces the current hidden state $\mathbf{h}_t$ and the cell output $\mathbf{c}_t$. To fetch the indexed features $\tilde{\mathbf{z}}_t$ from the feature sequences, the same indexing operation illustrated in Fig. 4 in the main paper is used but only for the visual modality.

Figure 2: Top-ranked/bottom-ranked clips where audio helps the most/least for our IMGAUD2VID distillation. The top-ranked clips (first row) belong to classes: grinding meat, jumpstyle dancing, playing cymbals, playing bagpipes, wrestling and welding; The bottom-ranked clips (second row) belong to classes: answering questions, bee keeping, clay pottery making, getting a haircut, tossing coin and extinguishing fire.



(a) Feature interpolation (b) Early stopping

Figure 3: Trade-off between efficiency and accuracy when using sparse indexing features or early stop on Mini-Sports1M. Uniform denotes the UNIFORM baseline in Table 1 in the main paper.

Our framework also has the flexibility to use different features for indexing and recognition. We can use cheaper features as indexing features (*e.g.*, MobileNetv2 features used in Fig. 7 in main) and more powerful (also often more expensive) features as recognition features (*e.g.*, R(2+1)D-152 features used in Table 2 in main). At inference time, we query the indexing features to obtain the weight vector $\mathbf{w}$ and get the selected index by $\arg\max(\mathbf{w})$. Then we use the recognition classifier to perform predictions for these selected moments, and average their prediction results as the final prediction. Note that when using the same features for indexing and recognition, we use the aggregated features for recognition, as discussed in the last ablation study in Sec. G.

E. Clips Where Audio Helps the Most/Least in Distillation

In Sec. 4.1 of the main paper, we perform an experiment to compute the $\mathcal{L}_1$ distance of the video descriptor hallucinated by our IMGAUD2VID distillation and the image-based distillation to the ground-truth video descriptor in order to identify the cases where the audio modality is particularly beneficial. As shown in Fig. 2, the top-ranked clips (first row) for which we best match the ground-truth tend to be dynamic scenes that have informative audio information, *e.g.*, grinding meat, jumpstyle dancing, playing cymbals, playing bagpipes, wrestling and welding. The bottom-ranked clips (second row) tend to be clips where the audio either contains just silence, narration, and background music, or are too difficult to perceive, *e.g.*, answering questions, bee keeping, clay pottery making, getting a haircut, tossing coin and extinguishing fire.

F. Trade-off between efficiency and accuracy on Mini-Sports1M

Similarly to Fig. 6 in the main paper, here we show the trade-off between efficiency and accuracy on Mini-Sports1M. Figure 3a shows the recognition results when using different subsampling factors for indexing features. We can see that the recognition remains robust to even aggressive subsampling of the indexing features. Figure 3b shows the results when stopping at different time steps. We can see that the first three steps yield sufficient cues for recognition. This suggests that we can stop around the third step with negligible accuracy loss.

G. Ablation Study

We perform three ablation studies in this section:

1. In Table 1 of the main paper, we have shown our results when using both the visual and audio modalities. To demonstrate the gain of selecting acoustically useful moments, we evaluate an ablated version of our method which uses only the visual features for indexing but both modalities for recognition. Namely, we only query the image indexing features to get weight vector $\mathbf{w}_t^{\mathbf{I}}$. For the next time step, we use the aggregated image feature and directly use the audio feature indexed by $\arg\max(\mathbf{w}_t^{\mathbf{I}})$ as the input to the fusion net-

work $\Psi(\cdot)$. Namely,

$$\begin{aligned} \tilde{\mathbf{z}}^{\mathbf{I}}_{t+1} &= \textstyle\sum_{j=1}^{N} w_j \mathbf{z}^{\mathbf{I}}_j, \quad & w_{j\in\{1,\cdots,N\}} \in \mathbb{R}_+; \\ \tilde{\mathbf{z}}^{\mathbf{A}}_{t+1} &= \mathbf{z}^{\mathbf{A}}_j, & j = \arg\max(\mathbf{w}^{\mathbf{I}}_t). \end{aligned} \tag{1}$$

The results we obtain are as follows. We can see that additionally leveraging audio based indexing can introduce an additional 1.1 accuracy gain for ActivityNet and a 0.7 gain for Mini-Sports1M.

	visual indexing	audio-visual indexing
ActivityNet	70.0	71.1
Mini-Sports1M	39.2	39.9

2. For our method in the main paper, we predict two query vectors to query the image indexing features and audio indexing features separately to find the visually and acoustically useful moments, respectively. As an alternative, we can also leverage image-audio features extracted from the image-audio network as the indexing and recognition features directly using the single modality version of our IMGAUD-SKIMMING network illustrated in Fig. 1. The results are shown as follows. We can see that separately querying the two modalities leads to 0.6 accuracy gain for ActivityNet and a 1.0 gain for Mini-Sports1M compared to using visual-only indexing features.

	single query	separate query
ActivityNet	70.5	71.1
Mini-Sports1M	38.9	39.9

3. During inference, we use the aggregated features at each time step, as done at training time. Based on the previous ablation study, we can also directly use the image-audio feature indexed by $\arg\max(\mathbf{w})$ at each time step, instead of aggregating the sequence of indexing image-audio features using our soft indexing mechanism. The results are shown as follows. We can see that the predicted frames indexed by $\arg\max(\mathbf{w})$ are already sufficient for recognition, demonstrating that our system truly learns to select useful moments in untrimmed videos. The feature aggregation step enabled by soft indexing can lead to an additional 0.4 accuracy gain for ActivityNet and a 0.5 gain for Mini-Sports1M.

	unaggregated	aggregated
ActivityNet	70.1	70.5
Mini-Sports1M	38.4	38.9

H. Additional Qualitative Results

Figure 4 shows some additional qualitative results of the frames selected by our method using the visual modality versus those obtained by uniform sampling. It can be noticed that the frames chosen by our method are much more informative of the action in the video compared to those uniformly sampled.

References

[1] R. Arandjelovic and A. Zisserman. Look, listen and learn. In *ICCV*, 2017.

[2] F. Caba Heilbron, V. Escorcia, B. Ghanem, and J. Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *CVPR*, 2015.

[3] R. Gao, T.-H. Oh, K. Grauman, and L. Torresani. Listen to look: Action recognition by previewing audio. In *CVPR*, 2020.

[4] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In *CVPR*, 2014.

[5] B. Korbar, D. Tran, and L. Torresani. Scsampler: Sampling salient clips from video for efficient action recognition. In *ICCV*, 2019.

[6] K. Soomro, A. R. Zamir, and M. Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

[7] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri. A closer look at spatiotemporal convolutions for action recognition. In *CVPR*, 2018.

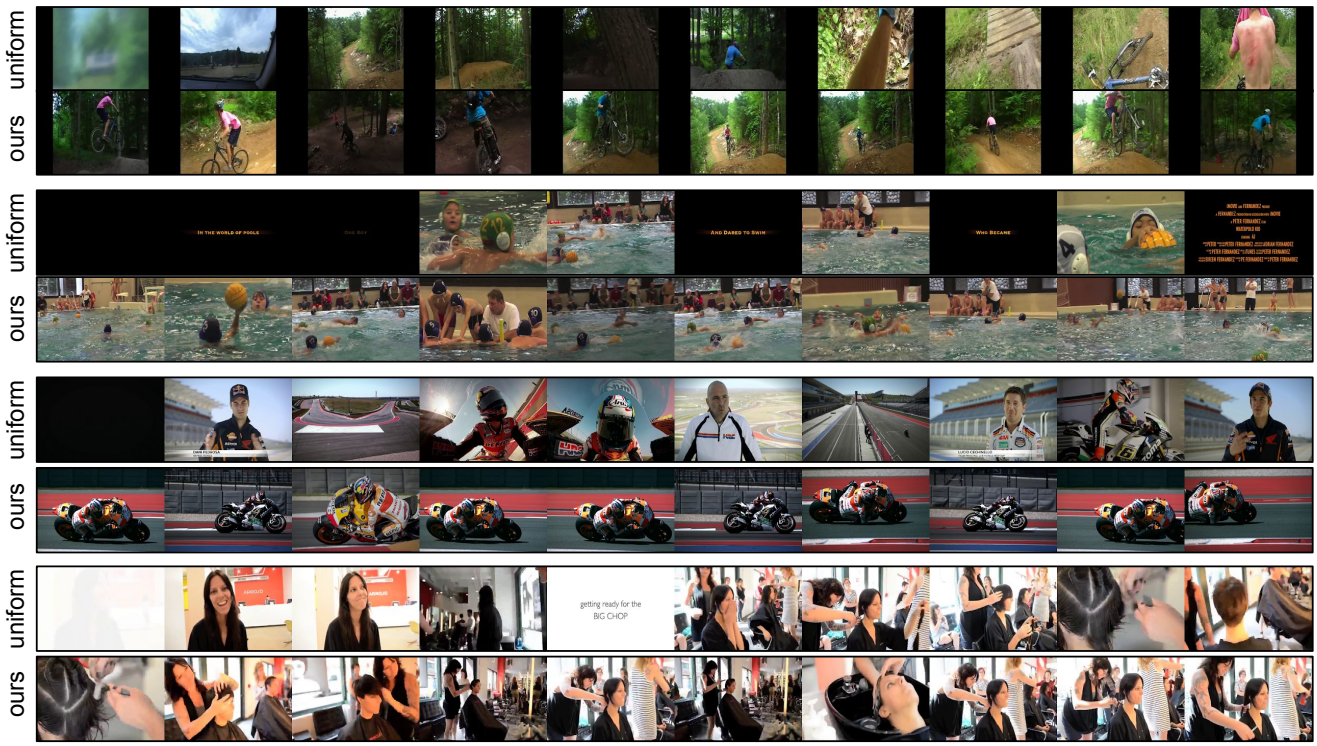


Figure 4: Qualitative examples of 10 uniformly selected moments (odd rows) and the first 10 visually useful moments selected by our method (even rows) for four untrimmed videos of the following actions: downhill mountain biking, playing waterpolo, doing motocross, and getting a haircut. The frames selected by our method are more indicative of the corresponding action.