Supplementary Material: Sound and Visual Representation Learning with Multiple Pretraining Tasks

Arun Balajee Vasudevan[1], Dengxin Dai[2], Luc Van Gool[1,3]
ETH Zurich[1] MPI for Informatics[2] KU Leuven [3]
{arunv,vangool}@vision.ee.ethz.ch, ddai@mpi-inf.mpg.de

1. SSL tasks of Binaural sounds

Spatial alignment. We train the SSL model for Spatial alignment to predict the angular difference *i.e.,* rotation angle R. We employ cross entropy loss between the predicted and actual angular difference as follows:

$$L_{\text{CE}} = \mathop{\mathbb{E}}_{i \in \mathcal{D}} \left[CE(h(v_{iR}, a_i), \Delta a_g) \right], \tag{1}$$

where $h(v_{iR}, a_i)$ is a prediction head followed by a softmax layer to predict the angle $\hat{R}$. v_{iR} and a_i refer to video features of rotated video segment v_i and sound features respectively. R is the groundtruth angular difference due to the rotation on $360°$ video segment v_i. CE denotes the cross entropy loss and $\mathcal{D}$ represents the set of all the $360°$ video segments of all videos in the batch.

Foreground alignment. Coming to the loss, we define it as:

$$L_{\text{FA}} = -\mathop{\mathbb{E}}_{i \in \mathcal{D}} \left[log \frac{exp(v_p \cdot a_i/\tau)}{exp(v_p \cdot a_i/\tau) + \sum_{v_n \in P_i} exp(v_n \cdot a_i/\tau)} \right] \tag{2}$$

where v_p and v_n are video feature vectors from aligned (positive) and misaligned (negative) video segments respectively, with respect to sound segment feature a_i, and τ is a temperature hyperparameter. P_i represents the set of misaligned video features for a_i while $\mathcal{D}$ denotes the set of all the $360°$ video segments of all videos in the batch.

Temporal gap prediction. As discussed in the main paper, let the temporal gap between given two sound slices is normalized as δ_m while the network predicts the temporal gap as $\hat{\delta}_m$. We train the model to minimize a huber loss [5]:

$$L_{\text{huber}}(\delta_m, \hat{\delta}_m) = \begin{cases} \frac{1}{2}(\delta_m - \hat{\delta}_m)^2, & \text{if } |\delta_m - \hat{\delta}_m| \leq \epsilon \\ \epsilon|\delta_m - \hat{\delta}_m| - \frac{1}{2}\epsilon^2, & \text{otherwise} \end{cases}$$

$$L_{gap} = \mathop{\mathbb{E}}_{m \in \mathcal{D}_p} \left[L_{\text{huber}}(\delta_m, \hat{\delta}_m) \right]$$

$L_{huber}(\delta_m, \hat{\delta}_m)$ between the groundtruth and the predicted temporal gap of the m^{th} sound slice pair. We set ϵ to be 1.5. $\mathcal{D}_p$ denotes the set of all sound slice pairs in in a batch and L_{huber} computes the loss for the batch.

2. Multi-SSL models

Figure 1 shows model diagrams of other approaches of Multi-SSL in addition to Figure 2 of the main paper. Here, approaches such as *ProgressiveNet, Euclidean dist:* and *Contrastive dist:* are shown in the figure.

3. Dataset details

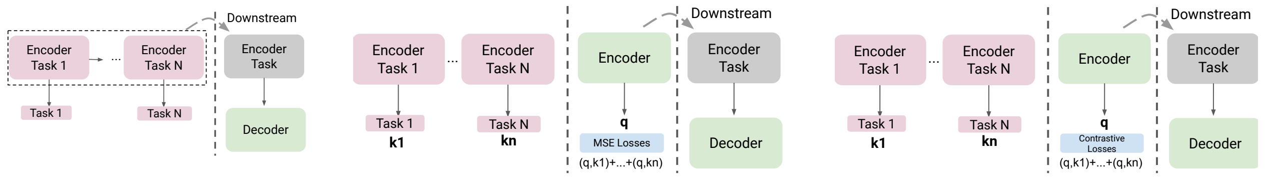
OmniAudio dataset [7]. The dataset consists of 360° audio-visual recording and it is done at 165 locations. For each location, the data is recorded for around 5-7 minutes. Thus, the dataset consists of 165 city traffic videos and audios with an average length of 6.49 minutes, totalling 15 hours. The raw video-audio data is post-processed into 2 second segments, resulting in 64250 video clips. The videos contain numerous sound-making objects such as cars, trams, motorcycles, pedestrians, buses and trucks.

PASCAL VOC dataset [1]. PASCAL VOC 2007 is a dataset for image recognition using 20 object classes. The test split of VOC07 has 2.5K images and 15K annotations for evaluation. Th train and validation set has 16.5K images and 47K annotations. The dataset is used for image classification.

MS COCO dataset [6]. For detection and segmentation in Table 4 we use COCO2017, which contains 80 labeled objects with segmentation masks and boxes. The training set contains $118K$ images with 850K annotations, and the validation set $5K$ images with $36K$ annotations.

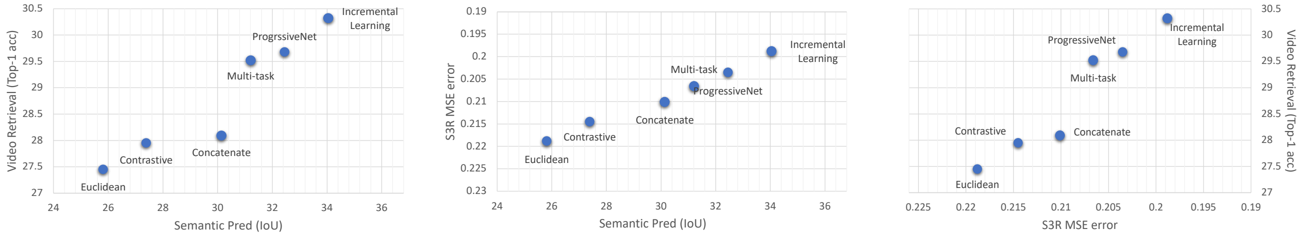
4. More experimental details

Sound networks. For all the experiments, we collect training and testing samples of 2-second video segments and a pair of binaural sound channels from OmniAudio dataset [7]. We preprocess sound samples following techniques from



(a) ProgressiveNet (b) Euclidean dist: (c) Contrastive dist:

Figure 1: Different approaches of Multi-SSL in addition to Figure 2 of the main paper. Left side of each subfigure indicates the Multi-SSL pretraining and right-most side depicts the downstream task training and evaluation. Middle parts of (b) and (c) denotes an encoder being trained from the pretrained encoders using MSE loss and contrastive loss respectively. Gray blocks denote frozen part when it trained for the downstream task.



(a) Semantic Pred vs VR (b) Semantic Pred vs S^{3}R (c) S^{3}R vs VR

Figure 2: This presents the trade-off between semantic prediction, video retrieval accuracy and S^{3}R performance of different Multi-SSL approaches (blue) that we tabulated in Table 2(b) of the main paper. All the methods are trained and evaluated on OmniAudio dataset.

[2, 7]. We keep the audio samples at 96kHz and their amplitude is normalized to a desired RMS level, which we set to 0.1 for all the audio channels. For normalization, we compute mean RMS values of amplitude over the entire dataset separately for each channel. An Short-time fourier transform is applied to the normalized sound waveform, with a window size of 512 (5.3ms), hop length of 160 (1.6ms) resulting a Time-Frequency representation of size of 257 × 601 pixels, which is our sound format.

Vision networks. We closely follow MoCo-v2 [3] for ImageNet pretraining. We ues ResNet50 [4] as the backbone as stated in the main paper. L2 normalized feature vector obtained in MoCov2 or DenseCL represents a query or key. For both the global and dense contrastive learning, the dictionary size is set to 65536. The momentum is set to 0.999. We adopt SGD as the optimizer and set its weight decay and momentum to 0.0001 and 0.9. We optimize each model on 8 GPUs with a cosine learning rate decay schedule and use a mini-batch size of 256. We train for 200 epochs, a total of 1 million iterations. The data augmentation pipeline consists of random resized cropping of 224 × 224 pixels with random color jittering, grayscale conversion, gaussian blurring and random horizontal flip.

SSL	Downstream tasks		
	SP↑	S^{3}R↓	VR↑
$\mathcal{A}$	15.32	**0.2105**	9.13
$\mathcal{B}_1$	23.82	0.2477	27.01
$\mathcal{B}_2$	**24.33**	0.2501	**27.35**
$\mathcal{C}$	16.85	0.2931	20.44

Table 1: Sound representations from SSL tasks are evaluated on 3 downstream tasks. $\mathcal{A}$:Spatial alignment, $\mathcal{B}_1$&$\mathcal{B}_2$: Foreground alignment a) and b) parts, $\mathcal{C}$: Temporal gap prediction. $\mathcal{B}_1$ in green is added to the Table 2(a) of the main paper.

5. Additional Experimental results

Foreground Alignment results. Section 3.1 details about SSL task of *foreground alignment* with two approaches. The objectives of the corresponding approaches are: a) Features of the masked foreground objects are learned to align with sounds, b) aligning motion flow features with sound features. We tabulate in Table 1 as $\mathcal{B}_1$ and $\mathcal{B}_2$. We note that the performance of $\mathcal{B}_1$ and $\mathcal{B}_2$ in the three tasks are quite comparable. In the main paper, we proceed further experiments with $\mathcal{B}_2$ since its performance is slightly better than $\mathcal{B}_1$. Comparable performances from $\mathcal{B}_1$ and $\mathcal{B}_2$ indicate that sound representations learn quite similar features from motion flow features and masked foreground object feature representations. This can be because that most of the objects in the scene are in motion and in addition, the recorded sounds are predominantly from the motion of objects around sensor setup (used in OmniAudio dataset).

Multi-SSL tradeoff results. Figure 2 depicts the tradeoff between semantic prediction, video retrieval accuracy and S^{3}R performance. We plot the results of different Multi-SSL approaches Euclidean, Contrastive dist:, Concatenate, Multitask, ProgressiveNet, and Incremental Learning. We observe in Figure 2 that Incremental learning outperforms all other Multi-SSL approaches in sound representation learning. This is in line with the Figure 1(b) of the main paper where we show results on image representation learning.

References

[1] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010.

[2] Ruohan Gao and Kristen Grauman. 2.5 d visual sound. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 324–333, 2019.

[3] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020.

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[5] Peter J Huber. Robust estimation of a location parameter. In *Breakthroughs in statistics*, pages 492–518. Springer, 1992.

[6] 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.

[7] Arun Balajee Vasudevan, Dengxin Dai, and Luc Van Gool. Semantic object prediction and spatial sound super-resolution with binaural sounds. In *European Conference on Computer Vision*, pages 638–655. Springer, 2020.