Novel-View Acoustic Synthesis

Changan Chen[1,3] Alexander Richard[2] Roman Shapovalov[3] Vamsi Krishna Ithapu[2]
Natalia Neverova[3] Kristen Grauman[1,3] Andrea Vedaldi[3]
[1]University of Texas at Austin [2]Reality Labs Research at Meta [3]FAIR, Meta AI

8. Supplementary Material

In this supplementary material, we provide additional details about:

1. Supplementary video for qualitative assessment of our model's performance.
2. Replay-NVAS dataset details (referenced in Sec. 4.1 of the main paper).
3. SoundSpaces-NVAS dataset details.
4. Implementation and training details (referenced in Sec. 5.6).
5. Baseline details (referenced in Sec. 6).

8.1. Supplementary Video

This video includes examples for the Replay-NVAS dataset and the SoundSpaces-NVAS dataset as well our model's prediction on both datasets. Listen with a headphone for the spatial sound.

8.2. Replay-NVAS Dataset Details

Multi-view camera calibration. We estimate camera poses with COLMAP [2] Structure-from-Motion (SfM) framework on each scene separately. Each scene is filmed with 8 static DSLR cameras and 3 wearable GoPro cameras (the latter are not used in our acoustic synthesis experiments). We first run SfM on the segments of the GoPro recordings where the wearers move significantly; followed by registration of the static camera frames to the model and a final round of bundle adjustment where we enforce constant relative poses between static camera frames taken at the same timestamp. This two-stage procedure greatly reduces the scale of the problem by making SfM focus first on the most diverse part of the data. Upon feature extraction stage, we cull the local features belonging to potentially dynamic object categories (such as people or animals) as detected by Detectron2 instance segmentation [4]. We then exploit the stationarity of DSLRs by picking a medoid camera pose among the frames filmed by each camera. Finally, we rotate and scale the coordinate system so that Z axis is pointing roughly upwards and scale the scene so that distances between cameras match the approximate field measurements in centimeters. Fig. 5 plots all camera coordinates and orientations projected to XY plane.

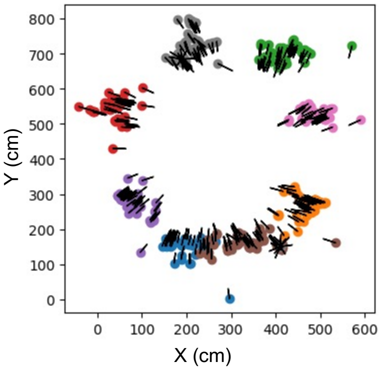

Figure 5. Coordinates of static cameras estimated by COLMAP, aggregated over scenes. The same colours correspond to the same camera position estimated from different scenes; they do not collapse to a point since the cameras could be moved between recordings.

Training data construction. We temporally align different DSLR videos using the clapper sound that is prominent in the waveform, which gives us synchronized multi-view audio-visual data. However, this data is not directly usable for training because some of it is noisy (e.g., people frequently talking over each other) or silent, which leads to additional learning challenges for the model. Thus, we design an automatic process for filtering out noisy clips. More specifically, we first extract all one-second audio clips of all videos and obtain the corresponding near-range audio clips and bounding boxes for each speaker. As described in Sec. 5.2, we select the active speaker based on the maximum energy of near-range audio with $\Delta t = 0.2$. For a one-second video clip, we obtain 5 candidate bounding boxes. We choose a threshold $\delta\%$ and only keep clips where more than $\delta\%$ of the bounding boxes belong to the same person. We set δ to 80. In this way, we keep clips where there is only one dominant speaker talking, and this speaker's bounding box is used as the localization feature V_L.

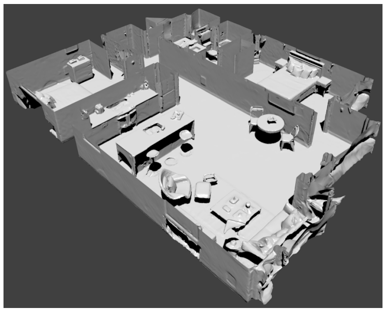

Figure 6. Environment mesh.

8.3. SoundSpaces-NVAS Dataset Details

For the single environment experiment, we use an apartment environment from the Gibson dataset [5][1]. Fig. 6 shows the mesh of the environment (the ceiling is removed). For the novel environment experiment, we use the public train/val/test splits.

For all images, we render with a resolution of 256 × 256 and a field of view of 120 degrees. We render binaural audio at a sample rate of 16000.

8.4. Implementation and Training Details

All audio clips during training are one second long with a sample rate of 16000. The shape of A_S and A_T is thus 2×16000. The audio encoder is a conv1d layer that encodes audio from 2 channel (binaural) to latent features of 64 channels, i.e., A_F^k is of shape 64×16000. For acoustic synthesis, we have $M = 30$ gated multi-modal fusion layers, which are equally divided into 3 blocks. In each block, the dilation of the dilated conv1d increases exponentially with base 3. The kernel size for each dilated conv1d is also 3. Both the skip and residual layers are conv1d layers with kernel size 1. The decoder network is a conv1d layer that encodes the latent audio features from 64 channels back to 2 channels.

The image resolutions are downsampled to 216×384 and 256×256 for Replay-NVAS (downsampled) and SoundSpaces-NVAS respectively. After being processed by a cond1d layer and flattened, the output visual feature V_F is of size 672 for Replay-NVAS and 512 for SoundSpaces-NVAS. The fusion layer consists of two fully connected layers with the first output dimension being 512 and the second being 256.

We train all models for 1000 epochs on the SoundSpaces-NVAS dataset and for 600 epochs on the Replay-NVAS dataset with a learning rate of 0.001. We evaluate the checkpoint with the lowest validation loss on the test set.

8.5. Baseline Details

For the Digital Signal Processing (DSP) baseline, we use the head-related transfer function (HRTF) measured by a KEMAR Dummy-Head Binaural Microphone. We apply a Wiener filter [3] to estimate the inverse HRTF. We adjust the gain of the HRTF by performing a binary search on the validation dataset and selecting the best gain value for testing. For the VAM [1] baseline, we take the original model from the paper, and we make minimal modifications by concatenating the visual feature with the target viewpoint pose P_T. We train the model with the same hyper-parameters described in the paper until convergence on both datasets.

References

[1] Changan Chen, Ruohan Gao, Paul Calamia, and Kristen Grauman. Visual acoustic matching. In *CVPR*, 2022. 2

[2] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proc. CVPR*, 2016. 1

[3] Norbert Wiener. Extrapolation, interpolation, and smoothing of stationary time series. *Report of the Services 19, Research Project DIC-6037 MIT*, 1942. 2

[4] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. `https://github.com/facebookresearch/detectron2`, 2019. 1

[5] Fei Xia, Amir R. Zamir, Zhi-Yang He, Alexander Sax, Jitendra Malik, and Silvio Savarese. Gibson Env: real-world perception for embodied agents. In *Computer Vision and Pattern Recognition (CVPR), 2018 IEEE Conference on*. IEEE, 2018. 2

---

[1] `http://gibsonenv.stanford.edu/models/?id=Oyens`