Camera Pose Estimation and Localization with Active Audio Sensing

Karren Yang[2⋆] Michael Firman[1] Eric Brachmann[1] Clément Godard[3⋆]

[1]Niantic [2]MIT [3]Google

Abstract. In this work, we show how to estimate a device's position and orientation indoors by echolocation, i.e., by interpreting the echoes of an audio signal that the device itself emits. Established visual localization methods rely on the device's camera and yield excellent accuracy if unique visual features are in view and depicted clearly. We argue that audio sensing can offer complementary information to vision for device localization, since audio is invariant to adverse visual conditions and can reveal scene information beyond a camera's field of view. We first propose a strategy for learning an audio representation that captures the scene geometry around a device using supervision transfer from vision. Subsequently, we leverage this audio representation to complement vision in three device localization tasks: relative pose estimation, place recognition, and absolute pose regression. Our proposed methods outperform state-of-the-art vision models on new audio-visual benchmarks for the Replica and Matterport3D datasets.

1 Introduction

Audio signals are rich with information about the scenes around us. As humans, we can often identify objects based on the sounds they make, and we can also localize objects based on the direction of their sounds. Beyond passively listening to sounds, animals such as bats and dolphins, as well as some individuals who are visually impaired, use *echolocation* (*i.e., active audio sensing*) to sense the spatial layout of their surroundings; they actively emit sounds that bounce off major surfaces, creating audio echoes that convey structural properties such as scene geometry and surface material [33, 26].

A growing body of research has proposed active audio sensing for vision tasks such as room geometry estimation [26], depth estimation [33, 65, 24], and floor-plan estimation [68]. Inspired by these pioneering works, we ask:

Can we train a machine to "hear" where it is in an indoor scene?

Figure 1(a) illustrates the problem setting: a device consisting of a camera with a co-registered microphone emits a sound and records the echoes from the surrounding indoor scene. Our goal is to leverage these audio echoes, either alone or in conjunction with the camera's image, to perform the three classic camera

⋆ Work done while at Niantic, during Karren's internship.

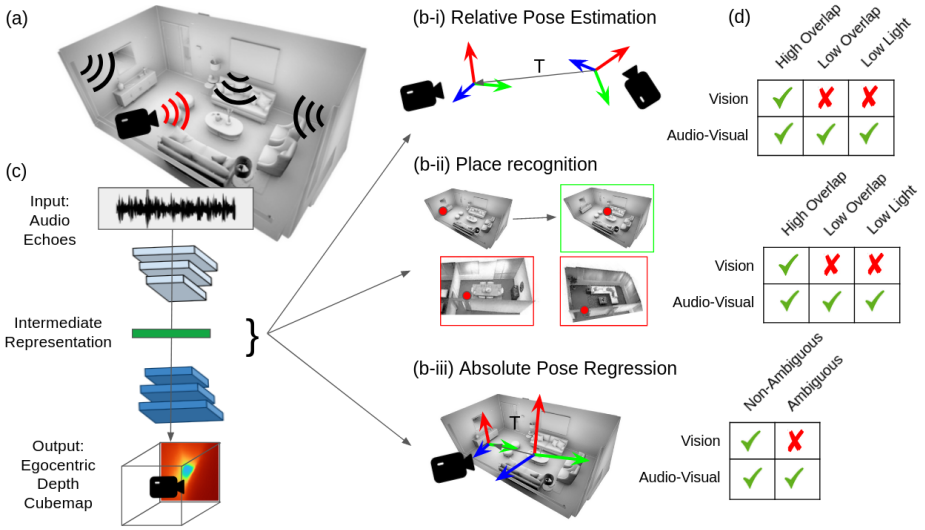

Fig. 1. Indoor device localization with active audio sensing. (a) Problem setting. A device consisting of a co-located microphone and camera generates sound (red) that bounces off major surfaces to create echoes (black). We leverage these audio echoes to perform the audio-visual device localization tasks proposed in (b). (b-i). Relative pose estimation. Audio-visual from two devices are used to estimate their relative transformation (e.g., rotation and translation). (b-ii) Place recognition. Audio-visual input from the device is used to retrieve nearby locations using a database of reference captures. (b-iii) Absolute pose regression. Audio-visual input from the device is used to estimate its global position and orientation with respect to the scene. (c) To learn audio features that capture the full geometry of the device's surroundings, we propose a pretraining task that distills an egocentric visual cube map [37] into the audio representation. (d) Audio sensing improves performance over established vision baselines.

localization tasks shown in Figure 1(b): (i) relative pose estimation, (ii) place recognition, and (iii) absolute pose regression. From this point on, we refer to these tasks as *device localization* tasks, since they involve both a camera and a microphone– a reasonable assumption for most applications in AR/VR [101, 16] and robotics [25, 52]. See the figure caption for an overview of each task.

While device localization tasks are conventionally tackled with only camera images, audio offers two key advantages. First, audio echoes reflect off surfaces beyond a camera's field of view, capturing more scene information than just what can be seen in an image. Second, audio signals are invariant to adverse visual conditions such as low lighting and occlusions. Our idea is that these attributes of audio sensing can enable us to solve cases of device localization that are generally challenging for vision.

Interestingly, we find that directly optimizing models to use audio inputs for device localization is not an optimal strategy. Unlike vision, the high-dimensional input representation of audio does not explicitly depict the scene geometry around the device, which is useful for localization. To overcome this challenge,

we propose a pretraining framework that distills visual information of the surroundings, represented as an egocentric depth cubemap [37], into the audio representation. As shown in Figure 1(c), as pretraining, we task a model with reconstructing an egocentric view captured at one of six possible orientations from the microphone. In this way, we learn useful spatial audio features through the natural co-occurrence of audio-visual data [107, 58], without the need for manually annotating surfaces in the scene (e.g., using a floorplan [68]).

Subsequently, we integrate these audio features into audio-visual methods for the three device localization tasks shown in Figure 1(b). Since these tasks have not previously been done with audio sensing, we introduce new benchmarks on the Replica [85] and Matterport3D [17] datasets. Integrating our audio features with established vision baselines achieves superior results across all tasks. Importantly, audio sensing enables us to solve cases that are challenging for vision, as summarized in Figure 1(d).

To summarize our main **contributions**: 1) We propose a pretraining framework for extracting features from audio echo recordings that are useful for device localization; 2) We introduce audio sensing to three classic visual localization tasks, that are conventionally tackled using only camera images: relative pose estimation, place recognition, and absolute pose regression, and achieve superior results on all three tasks; 3) We propose novel audio-visual benchmarks for these tasks, building on publicly available datasets and simulation platforms.

To our knowledge, our work is the first to extend classic camera localization tasks to the audio-visual domain. Our code and pretrained models are available at `https://github.com/nianticlabs/audio-localization`.

1.1 Related Work

Audio for Spatial Sensing. Existing research has leveraged audio to sense locations of surfaces or objects in a scene. Echolocation has previously been used to compute the shape of a convex polyhedral room [26], to predict the shape of an object around a corner [53], to predict distances to surfaces [84, 28, 100], to estimate frontal depth maps [24, 33, 65], and to reconstruct floor plans [68]. Other prior work leverages passive rather than active audio (i.e., audio naturally emitted by sound sources in the scene) for spatial scene understanding. These generally focus on localizing the source of the object producing the sound, for example, predicting the direction of sound arrival [66], localizing multiple sound sources using SVD [93], highlighting sound sources in a video [40, 31, 47, 62], drawing bounding boxes around moving vehicles [32] and performing semantic object detection on street scenes [98]. Recent work in robotics even uses ambient audio in a scene to estimate distances to walls [23]. Different from all of these works, we use echolocation to estimate the surround depth of a scene in pretraining, with the ultimate goal of performing device localization.

Audio-Visual Scene Perception. Audio and visual signals often occur together in a scene and offer useful joint information for performing tasks such as action recognition [105, 55, 103, 44, 35, 106] and object labeling [110]. The co-occurrence of these signals enables self-supervised representation learning [72, 61,

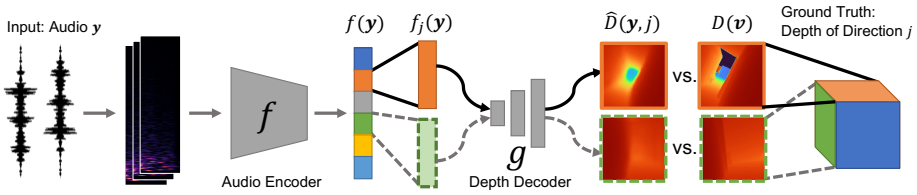

Fig. 2. Pretraining Task. We distill an egocentric visual depth map into an audio representation for downstream localization tasks. During training, we provide audio-visual samples where the camera is rotated with respect to the microphone, and we teach the model to reconstruct this egocentric view from audio. See text for details.

64, 2, 62, 107, 58, 60], audio synthesis [63, 83] or spatialization [34, 59] corresponding to a visual scene, and navigation [19, 21, 20, 18]. Inspired by these works, our pretraining task leverages co-occurring audio-visual signals to learn useful audio features for device localization.

Visual place recognition. Determining a device's location based on a captured image is a device localization problem that can be formulated as an image retrieval task, where the objective is to retrieve a database image taken from the same place as the query, rather an image that looks similar [36]. Prior works have proposed performing a nearest-neighbor search on global visual descriptors [43, 1, 3, 94] and/or performing matching between local visual features [54, 109, 92, 38]. Ongoing challenges in visual place recognition include the need for visual overlap between query and database images [56] as well as the need for invariance to different visual appearance conditions [36] such as lighting. Different from prior work, here we propose to augment visual place recognition with audio sensing to overcome these challenges and demonstrate our method on new audio-visual benchmarks for place recognition.

Relative Camera Pose Estimation. Relative pose estimation is used to localize one device with respect to another by predicting the relative transformation between them, usually based on a pair of images. The most prevalent methods are feature-matching methods that use a pose solver integrated within a RANSAC framework [69], with state-of-the-art approaches using learned methods for feature detection [27, 71, 96, 6], matching [74, 86] and robust model fitting [108, 70, 11, 87]. Deep learning methods use convolutional neural networks to directly regress the transformation from a pair of images [108, 57, 29, 67], including a recent work that frames regression as a classification problem [22]. An ongoing challenge in relative pose estimation is handling wide rotation cases where there is limited visual overlap between the image pair [15]. Different from prior work, here we augment relative camera pose estimation with audio sensing and demonstrate our method on new audio-visual benchmarks for relative pose estimation.

Absolute Camera Pose Estimation. Absolute pose estimation infers the camera position and orientation based on a single query frame relative to a pre-scanned environment. Traditional methods match sparse features of the query to a full 3D reconstruction of the scene, and solve for the pose [50, 51, 88, 76, 79, 89, 77, 78]. Recent iterations of this classic formula utilize learned components

	Frontal Camera FoV						Overall					
	RMS↓	**REL↓**	**Log10↓**	**A1 ↑**	**A2↑**	**A3↑**	**RMS↓**	**REL↓**	**Log10↓**	**A1↑**	**A2↑**	**A3↑**
Replica dataset												
Echo2Depth [65]	0.583	0.443	0.143	0.603	0.765	0.851	-	-	-	-	-	-
Ours	**0.474**	**0.360**	**0.121**	**0.677**	**0.817**	**0.884**	**0.501**	**0.371**	**0.130**	**0.643**	**0.797**	**0.874**
Matterport3D dataset												
Echo2Depth [65]	1.166	0.351	0.133	0.571	0.756	0.847	-	-	-	-	-	-
Ours	**1.118**	**0.341**	**0.125**	**0.597**	**0.773**	**0.860**	**0.994**	**0.327**	**0.114**	**0.638**	**0.795**	**0.872**

Table 1. Performance on depth pretraining task. Our framework outperforms the state-of-the-art Echo2Depth [65]; see text for details.

for some of the steps, particularly for image retrieval, feature extraction and feature matching [73, 42, 90, 91, 75]. Scene coordinate regression dispenses with the need for discrete feature matching by regressing image-to-scene correspondences directly, via random forests [82, 97, 8] or neural networks [7, 9, 11, 12]. Absolute pose regression networks predict poses in a single forward pass, and avoid any potentially costly geometric optimization altogether [46, 45, 102, 13, 81]. Finally, relative pose regression can be coupled with image retrieval to infer absolute poses [48, 4, 111, 95, 104]. One of the major challenges in absolute pose estimation is to handle scene ambiguities, such as feature-less areas or repeating structures. These are more likely to appear in larger scenes. Difficult visual conditions, such as low lighting, can also create ambiguous images. Different strategies exist to cope with ambiguities in absolute pose estimation, such as avoiding full-scale reconstructions [79], using global image context to resolve local ambiguities [10, 49] or modeling uncertainty to make multi-modal predictions [14]. Orthogonal to these strategies, we demonstrate that active audio sensing effectively helps disambiguates a query by providing a surround view of the environment.

2 Spatial Audio Representation Learning Framework

The main objective of this work is to evaluate the capabilities of audio sensing for classic visual localization tasks. However, a key challenge is wrangling the high-dimensional audio input into a meaningful form. While camera images explicitly display the spatial configuration of a scene, spatial cues in audio echoes are reflected in subtler differences in signal arrival times and levels [33].

To overcome this challenge, we propose to first learn an audio representation that captures the spatial configuration of the scene around the device. Our pretraining framework exploits the natural co-occurrence between visual and audio signals to distill visual information of the device's surroundings, which we represent as an egocentric depth cubemap [37], into the audio representation. We hypothesize that such a representation will be helpful for device localization.

Method. Let $\mathbf{y}$ denote audio signal and $\mathbf{v}$ denote the visual frame. As shown in Figure 2, we provide our model with audio-visual samples where the camera is rotated in one of six orientations with respect to the microphone. These orientations (denoted by j) correspond to the faces of an egocentric cube map. The audio encoder f extracts a feature embedding for the full scene, and the depth decoder g uses the j-th subset of this embedding to reconstruct a depth map of

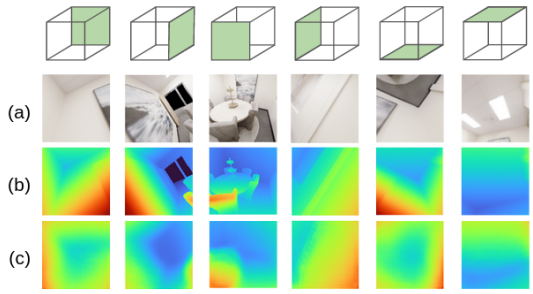

Fig. 3. Scene Geometry from Audio. Our pretraining task distills scene geometry into an audio representation. (a) Camera views corresponding to an egocentric cubemap. (b) Ground truth depth maps for these views. (c) Depth decoded from our audio representation.

the camera view. We train the model by minimizing the log-loss [41, 65],

$$\mathcal{L}(\mathbf{y}, \mathbf{v}, j) := \frac{1}{WH} \sum_{p=1}^{W} \sum_{q=1}^{H} \log(1 + |D(\mathbf{v})_{[p,q]} - g(f_j(\mathbf{y}), j)_{[p,q]}|), \tag{1}$$

where $D(\mathbf{v})$ is the target depth map derived from visual frame $\mathbf{v}$, and (W, H) is the size of the depth map. While supervision is only provided for one face of the cube map at a time, over the course of training on many samples, the model learns to capture the full surround depth of the device's surroundings.

Dataset. We train our framework on the Replica [85] and Matterport3D [17] datasets. Replica contains 18 indoor scenes of hotels, apartments, rooms and offices. Matterport3D contains 85 indoor scenes, most of which are large, multi-room homes. Following previous work [20, 33, 65, 68], audio-visual data for these scenes is obtained by simulating echo responses using the SoundSpaces platform [20] and rendering the corresponding camera view using the Habitat platform [80]. SoundSpaces simulates acoustics by pre-computing an impulse response (IR) for each source-receiver location pair on a dense grid with 0.5m and 1m spatial resolution for Replica and Matterport3D respectively. To simulate an echo recorded by a microphone with a specific location and orientation, the IR with both source and receiver at this grid location is selected, rotated to the desired orientation, and convolved with a 3ms audio chirp (20Hz-20kHz frequency sweep) [33]. We follow previously defined scene splits for these datasets for training, validation and testing [33, 65]. Devices are placed at all possible grid locations within each scene, with random azimuth and elevation angles.

Task Performance. To assess whether our model extracts meaningful scene information from audio echoes, we compare to the SOTA Echo2Depth [65]. Table 1 shows that we significantly outperform their model across all evaluation metrics. Note that their approach only predicts the depth for the frontal camera, whereas our approach predicts omnidirectional scene geometry. Training Echos2Depth separately on six faces of the cube with the same representation size performs worse than our model, in part due to the sharing of weights in our encoder and decoder (see Supplemental Material for ablations). Figure 3 shows a qualitative result of the scene geometry captured by our audio representation.

3 Relative Pose Estimation

Having learned audio features that capture 3D scene geometry, we now show how these signals can help localize a device. We start with relative pose estimation: given inputs from two nearby devices, predict their relative transformation. Visual methods match features between two camera images, but they have difficulty handling cases with low overlap between images [22]. Audio sensing can help as it captures spatial cues beyond the camera's field of view.

3.1 Proposed Models with Audio Sensing

Let $(\mathbf{y}_1, \mathbf{v}_1)$ and $(\mathbf{y}_2, \mathbf{v}_2)$ denote the audio-visual inputs for the two devices, and let (R, t) denote their 3×3 relative rotation matrix and 3D translation vector. As in previous work, we take t to be a normalized direction without scale.

Audio regression model. We first propose a model that regresses relative pose directly from audio. The audio signals $\mathbf{y}_1, \mathbf{y}_2$ are passed through our pretrained feature extractor f. We concatenate the features and pass them through a shallow multi-layer perceptron (MLP) to produce three vectors: $(\hat{r}_x, \hat{r}_y, \hat{t}_a)$. We use a partial Gram-Schmidt projection to obtain a rotation matrix $\hat{R}_a$ from $\hat{r}_x, \hat{r}_y$ [112], and train the MLP to minimize the mean-squared error between the predicted and ground truth rotation matrices [112], as well as a direction loss [22] given by the negative cosine similarity between the two translation vectors, i.e.

$$\mathcal{L}_R(\hat{R}_a, R) := ||\hat{R}_a - R||^2 \quad \text{and} \quad \mathcal{L}_t(\hat{t}_a, t) := -\frac{\hat{t}_a^T t}{||\hat{t}_a||||t||}.$$

The full loss is given by $\mathcal{L}_{audio}(\hat{R}_a, \hat{t}_a, R, t) := \beta\mathcal{L}_R(\hat{R}_a, R) + \mathcal{L}_t(\hat{t}_a, t)$, where hyperpameter $\beta > 0$ weighs the relative importance between the losses.

Audio-visual regression model. Some existing visual methods tackle relative pose by regressing pose from images directly using a Siamese architecture [108, 57, 29, 67]. We augment this approach with audio sensing. The audio signals $\mathbf{y}_1, \mathbf{y}_2$ are passed through our pretrained feature extractor f, and the images $\mathbf{v}_1, \mathbf{v}_2$ are passed through a deep residual network [39]. Similar to the audio regression model, the audio-visual features are concatenated and passed through a shallow MLP to predict pose. The model is trained to minimize $\mathcal{L}_{audio}$.

Visual feature matching + audio. State-of-the-art visual methods such as Superglue [74] match local features between two images and then predict relative pose via essential matrix estimation within a RANSAC loop [30]. To incorporate audio sensing into such methods, we propose a mixture-of-experts (MoE) type model, in which a gating function decides whether to use the audio expert (audio regression model) or the visual expert (SuperGlue). An intuitive gating function to use is the output of the visual matching: if visual matching produces a pose $(\hat{R}_v, \hat{t}_v)$, then there is likely overlap between the images, and we should use this result; otherwise, we use the audio expert's prediction $(\hat{R}_a, \hat{t}_a)$.

Visual feature matching + audio + (learned) gating. Since visual matching does not necessarily produce a better result than audio, we also propose a learned gating function that assigns an expert based on the predicted poses.

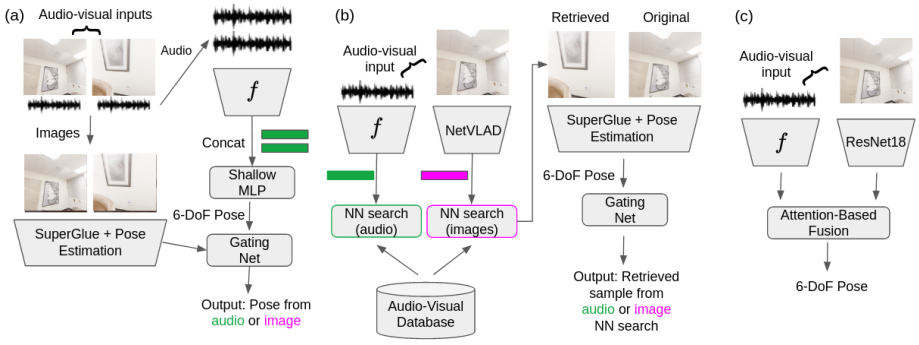

Fig. 4. Audio-visual methods for device localization. Proposed models for (a) relative pose estimation, (b) place recognition, and (c) absolute pose regression.

Concretely, the learned gating function is a neural network that takes as input the predicted poses from both streams $(\hat{R}_a, \hat{t}_a, \hat{R}_v, \hat{t}_v)$ and outputs a vector $z \in [0,1]^4, \sum_i z_i = 1$ indicating the composition of the final prediction. Each entry of z gives the probability that one modality will outperform the other for estimating R or t. We train the gating network to minimize the cross-entropy loss between z and z^*, a one-hot vector indicating the optimal combination of expert outputs. See Figure 4(a) for a schematic of this full model.

3.2 Evaluation

Benchmarks. Since there are no datasets for relative pose estimation with audio-visual data, we introduce new benchmarks on the Replica [85] and Matterport3D [17] scenes. For training, validation, and test scenes, we use the same splits as our pretraining task. We sample audio-visual inputs from adjacent navigable points on the scene grid with random azimuth and elevation angles. We consider three evaluation scenarios: standard baseline cases (<90° rotation) that are typically studied in the vision literature where cameras have considerable visual overlap; extreme wide baseline cases (>90° rotation) where cameras have very limited visual overlap; and low-lighting cases. We evaluate methods on median angular error [22], as well as accuracy at a 20° cutoff [74].

Baselines. We compare our audio-visual models to the established vision models that they build upon. The first is a visual regression baseline that uses a Siamese architecture to regress pose from two images, as in [108, 57, 29, 67]. The second is SuperGlue, a SOTA visual matching method for relative pose estimation [74]; we use the pretrained indoor model released by the authors. To assess the pretraining task, we compare our audio model to one trained from scratch.

Results. Table 2 shows quantitative results for all methods.

Audio-visual vs. Vision-only. Audio sensing improves the performance of both the visual regression and SOTA visual matching methods across all metrics and evaluation settings. This validates the benefit of audio for visual positioning.

		Standard Baseline (Relative Rot. <90°)			Wide Baseline (Relative Rot. >90°)			Low Lighting (Dark Image)		
		t↓	R↓	Acc.↑	t↓	R↓	Acc.↑	t↓	R↓	Acc.↑
	Replica dataset									
	Visual regression	35.0	15.0	21.0	37.4	20.0	17.3	57.3	99.1	0.7
	Visual matching [74]	34.1	12.2	39.9	55.0*	108.4*	2.0	47.6*	14.3*	1.6
Ours	Audio only (scratch)	30.2	17.9	21.9	28.8	19.5	22.3	30.2	17.9	21.9
Ours	Audio only (pretrained)	21.6	10.5	38.1	23.0	10.9	36.8	21.6	10.5	38.1
Ours	Audio-visual regression	33.0	14.3	24.1	35.6	18.4	19.5	58.5	91.6	0.9
Ours	Visual matching + Audio	19.0	7.3	47.7	27.3	13.3	31.5	23.3	10.8	35.3
Ours	Visual matching + Audio + Gating	**13.8**	**6.5**	**55.4**	**22.6**	**10.6**	**37.6**	**21.5**	**10.3**	**38.4**
	Matterport3D dataset									
	Visual regression	44.5	37.1	6.6	45.8	56.5	3.9	53.5	110.8	0.1
	Visual matching [74]	19.0	9.5	49.2	55.8*	113.09*	2.1	50.1*	31.1*	0.6
Ours	Audio only (scratch)	35.0	24.7	13.8	35.1	25.7	13.2	35.0	24.7	13.8
Ours	Audio only (pretrained)	31.2	20.5	17.7	31.7	22.8	15.9	31.2	20.5	17.7
Ours	Audio-visual regression	41.4	37.0	8.0	43.3	51.3	5.4	48.1	99.9	1.0
Ours	Visual matching + Audio	13.5	7.7	52.3	36.7	30.1	14.1	32.0	20.8	17.4
Ours	Visual matching + Audio + Gating	**11.7**	**7.2**	**54.4**	**31.2**	**22.5**	**17.2**	**31.1**	**20.3**	**18.0**

Table 2. Audio sensing improves relative pose estimation. Our best audio-visual method, which combines our audio feature representation with visual matching and a learned gating network, outperforms other methods including SOTA visual matching [74] on both the Replica and Matterport3D datasets. Audio-visual methods that outperform both visual baselines are underlined. *Visual matching fails to find a match on most test images; median values are computed from valid matches only.

As expected, adding audio is most helpful to the wide-baseline and low-lighting cases, achieving large gains over the visual matching baseline.

Figure 5 shows a qualitative result of visual matching + audio + gating. In Fig. 5(a), the two devices have large relative rotation and there is low overlap between their images. As a result, visual matching performs poorly, as shown in Fig. 5(b). Our gating function chooses the audio expert to produce robust results, as shown in Fig. 5(c).

Does pretraining help? Our pretraining task significantly boosts the performance of the audio model over one trained from scratch. This validates our hypothesis that learning 3D scene geometry is beneficial for device localization.

Audio-Visual vs. Audio-only. In the wide-baseline and low-lighting cases, visual matching + audio does worse than the audio only model. This is due to false matches found by SuperGlue. The learned gating function correctly disambiguates many of these false matches, yielding improved scores. Interestingly, while the audio-visual regression model improves over visual regression, it does not improve over the audio only model. A different representation of the pose outputs from the visual and audio streams, e.g., using the classification framework of [22] rather than regression, may yield superior fusion results.

4 Place Recognition

Given a device capture, place recognition aims to determine its rough location by retrieving a similar capture from a reference database. Visual place recognition

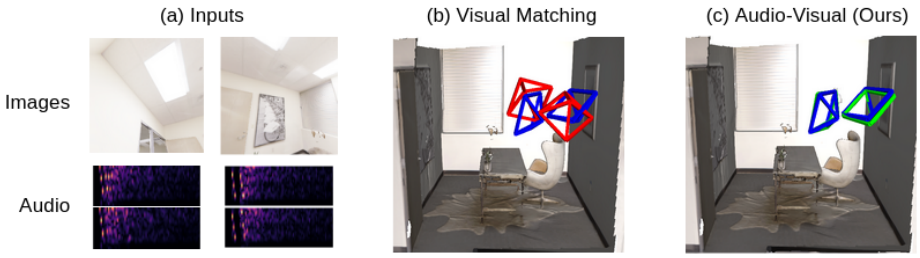

Fig. 5. Combining audio and vision for relative pose estimation. (a) Audio-visual inputs from two devices. (b) Due to the low overlap between images, visual matching preforms poorly. (c) In this case, our audio-visual chooses the pose predicted by the audio expert to make a robust prediction. Blue – ground truth. Red – visual prediction. Green – our prediction.

typically involves performing retrieval on camera inputs, but it performs poorly in situations with low overlap between query and database images [56]. Audio sensing can help by providing spatial cues beyond the camera's field of view. We let $(\mathbf{y}_q, \mathbf{v}_q)$ denote the query audio-visual input captured at position c_q.

4.1 Proposed Methods with Audio Sensing

Audio Descriptor. We first propose to learn an audio descriptor from the output of pretrained feature extractor f. We do this by appending a shallow MLP to the end of f, which we train using the triplet margin loss [5]:

$$\mathcal{L}_{\text{triplet}}(\mathbf{y}, \mathbf{y}_+, \mathbf{y}_-) := \max\{||\mathbf{y} - \mathbf{y}_+|| - ||\mathbf{y} - \mathbf{y}_-|| + m, 0\}, \tag{2}$$

where $\mathbf{y}$ denotes the anchor, $\mathbf{y}_+$ is a spatially-neighboring audio sample, $\mathbf{y}_-$ is a non-neighboring audio sample, and $m > 0$ represents the margin. While f captures 3D scene geometry, this additional training with triplet loss enforces that close distances in the output of the MLP reflect close distances in physical space. To perform place recognition, we compare $f^*(\mathbf{y}_q)$ to the descriptors in the reference database and perform an exact nearest neighbor search based on Euclidean distance to retrieve a sample located at $c_{\text{NN}(\mathbf{y}_q)}$.

Visual descriptor + audio. Visual place recognition commonly relies on state-of-the-art networks, such as NetVLAD [1], to produce visual descriptors for retrieval. To incorporate audio sensing into such methods, we propose a mixture-of-experts (MoE) type model, in which a gating function decides whether to use the audio expert (audio descriptor) or the visual expert (NetVLAD). Since our goal is to use the audio stream when the visual match between the query image and reference images is poor, we propose an intuitive gating function based on a validation step for the visual retrieval result: if local feature matching (i.e., using SuperGlue [74]) between $\mathbf{v}_q$ and the retrieved image $\mathbf{v}_{\text{NN}(\mathbf{v}_q)}$ predicts a positive result, then we use the location $c_{\text{NN}(\mathbf{v}_q)}$ corresponding to this result; otherwise, we use the audio expert's prediction $c_{\text{NN}(\mathbf{y}_q)}$.

		Overall All Queries			High Overlap Subset of Queries			Low Overlap Subset of Queries			Low Lighting All Queries		
		R@1	R@5	Rank	R@1	R@5	Rank	R@1	R@5	Rank	R@1	R@5	Rank
	Replica dataset												
	Visual descriptors [1]	0.59	0.76	4.0	0.91	**0.98**	2.3	0.38	0.61	5.0	0.18	0.34	5.0
Ours	Audio only (scratch)	0.59	0.68	4.5	0.57	0.66	5.0	0.60	0.70	**1.5**	0.59	0.68	4.0
Ours	Audio only (pretrained)	0.65	0.74	2.5	0.69	0.78	4.0	**0.64**	0.72	**1.5**	**0.65**	**0.74**	**1.0**
Ours	Visual descriptors + Audio	0.67	0.81	2.5	**0.92**	**0.98**	**1.8**	0.51	0.69	4.0	0.63	0.73	3.0
Ours	Visual descriptors + Audio + Gating	**0.71**	**0.83**	**1.5**	**0.92**	**0.98**	2.0	0.58	**0.73**	2.5	0.64	**0.74**	2.0
	Matterport3D dataset												
	Visual descriptors [1]	0.44	0.66	4.3	0.90	**0.97**	2.4	0.39	0.63	4.3	0.12	0.26	5.0
Ours	Audio only (scratch)	0.41	0.54	4.3	0.42	0.54	4.8	0.41	0.54	4.3	0.41	0.54	3.3
Ours	Audio only (pretrained)	0.49	0.62	3.3	0.50	0.63	4.3	0.49	0.62	3.0	**0.49**	**0.62**	2.0
Ours	Visual descriptors + Audio	0.54	**0.71**	2.3	**0.91**	**0.97**	**1.5**	0.50	0.68	2.3	**0.49**	**0.62**	2.9
Ours	Visual descriptors + Audio + Gating	**0.55**	**0.71**	**1.0**	0.90	**0.97**	2.1	**0.51**	**0.69**	**1.3**	**0.49**	**0.62**	**1.9**

Table 3. Audio sensing improves visual place recognition. Our models that combine our audio feature representation with visual descriptors outperform SOTA image retrieval with NetVLAD descriptors [1] on both the Replica and Matterport3D datasets. Audio-visual methods that outperform the visual baseline are underlined. Results are averaged over scenes; rank refers to average rank over scenes in the dataset.

Visual descriptor + audio + (learned) gating. In addition to a fixed gating function, we also propose a learned function for our MoE model. The learned gating function is a shallow MLP that takes the match predicted by Superglue [74], represented as relative pose, and outputs a scalar value $z \in [0, 1]$ indicating whether to use the position retrieved by vision or audio. The gating function is trained to minimize the binary cross-entropy loss between z and z^*, which indicates whether the retrieved result from vision is better than audio.

4.2 Evaluation

Benchmarks. Since there are no datasets for place recognition with audio-visual data, we introduce new benchmarks on two scenes from the Replica dataset and and four scenes from the Matterport3D dataset that were held out from the pretraining task. Reference and query audio-visual data are sampled from distinct locations on the scene grid. For reference samples, we place devices within 90° and 45° of the first cardinal direction (N/0°) for Replica and Matterport3D respectively. We consider two evaluation scenarios: high overlap cases where the query devices are oriented in the same range of rotations as the reference devices; low overlap cases where query devices are oriented outside of this range; and low-lighting cases. To evaluate methods, we use the Recall@k metric [36].

Baselines. We compare our audio-visual methods to SOTA image retrieval using NetVLAD [1]. To determine the usefulness of the pretraining task, we compare our audio descriptors to those trained from scratch.

Results. Table 3 shows quantitative results for all methods.

Audio-visual vs. Vision-only. Audio-visual results that outperform vision are underlined in the table. We find that including audio sensing improves performance over NetVLAD image retrieval across almost all metrics and evaluation scenarios. This validates the benefit of audio for indoor visual place recognition.

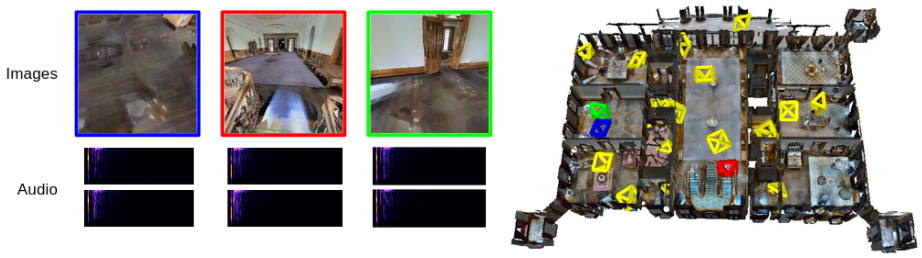

Fig. 6. Combining audio and vision for place recognition. The query audio-visual sample is highlighted in blue and its position is depicted by the blue frustum. Since there is no overlap between reference frustums (subset shown in yellow) and the query, using visual descriptors results in incorrect retrieval (red). Our audio-visual method selects the audio expert to make a correct retrieval (green). Note that frustum rotations do not matter for accuracy, only its position.

As expected, the addition of audio benefits low-overlap queries and low-lighting cases the most, bringing large gains over the visual baseline. Figure 6 shows a qualitative result of our visual descriptor + audio + gating model. Since there is no visual overlap between the query sample (blue) and the images in the reference database, NetVLAD retrieval returns an incorrect result (red). The audio-visual model chooses the audio expert to make a correct retrieval (green). *Does pretraining help?* Our pretraining task significantly boosts the performance of the audio descriptors over those trained from scratch. This supports our hypothesis that learning 3D scene geometry from audio can help device localization. *Audio-Visual vs. Audio-only.* Note that in the low-overlap cases and low-lighting cases, the visual descriptor + audio model does worse than the model with audio alone. This is because the query images have very little visual overlap with the reference, so many positive matches for these images are false matches. The learned gating function manages to correctly disambiguate many false matches, yielding improved scores. Audio-only shows slightly better performance on low-overlap and low-lighting queries on the smaller Replica scenes, while visual descriptor + audio + gating performs better on the larger Matterport3D scenes.

5 Absolute Pose Regression

Absolute pose estimation involves estimating the position and rotation of a device with respect to a known 3D environment. Many recent direct regression approaches [48, 4, 111, 95, 104] as well as some local feature-based approaches [73, 42] combine retrieval with relative pose estimation. Our approaches for integrating audio sensing into place recognition and relative pose estimation could directly be used in those pipelines to tackle challenging cases with low visual overlap. Here, we focus instead on regression methods and use audio to tackle a separate challenge: scene ambiguities that cause images to match with different parts of the same scene. Recent work on absolute pose regression has focused on

		Overall All Queries				**Low Ambiguity** Normal Light				**High Ambiguity** Low Light			
		Position		Rotation		Position		Rotation		Position		Rotation	
		Error	Rank	Error	Rank	Error	Rank	Error	Rank	Error	Rank	Error	Rank
	Replica dataset												
	PoseNet [46]	0.53	1.75	9.4	2.0	**0.43**	**1.25**	7.3	2.25	0.74	2.0	13.7	2.0
Ours	Audio only	2.49	4.0	27.7	4.0	2.27	4.0	27.4	4.0	2.93	4.0	28.3	3.5
Ours	Audio-Visual (scratch)	1.88	3.0	19.3	3.0	1.52	3.0	12.5	2.75	2.45	3.0	27.2	3.5
Ours	Audio-Visual	**0.52**	**1.25**	**6.9**	**1.0**	0.46	1.75	**5.4**	**1.0**	**0.64**	**1.0**	**9.9**	**1.0**
	Matterport3D dataset												
	PoseNet [46]	2.17	2.0	21.0	2.0	1.41	2.0	**13.5**	1.75	3.73	2.0	36.5	2.0
Ours	Audio only	9.60	4.0	74.9	4.0	9.82	4.0	76.2	4.0	9.14	4.0	72.2	3.75
Ours	Audio-Visual (scratch)	3.34	3.0	41.8	3.0	2.31	3.0	30.3	3.0	5.46	3.0	64.9	3.25
Ours	Audio-Visual	**1.86**	**1.0**	**19.4**	**1.0**	**1.31**	**1.0**	13.6	**1.25**	**3.01**	**1.0**	**31.2**	**1.0**

Table 4. Audio sensing improves absolute pose regression. Our audio-visual fusion model, which combines our audio features with visual features, outperforms the established vision baseline on both the Replica and Matterport3D datasets. Results are averaged over scenes; rank refers to average rank over scenes in the dataset.

modeling distributions of poses to handle this problem [14]. As an orthogonal solution, we propose to use audio sensing to disambiguate between regions of the scene that appear visually similar.

Model. We augment an established absolute pose regression network, PoseNet [46], with our audio features. Let $(\mathbf{y}, \mathbf{v})$ denote the audio-visual capture of the device, and let (R, t) denote the device's pose. We use a deep residual network [13] to extract visual features from $\mathbf{v}$, and we use our pretrained audio feature extractor to obtain audio features from $\mathbf{y}$. The features are fused using a self-attention module [99] and a shallow MLP produces three vectors: $(\hat{r}_x, \hat{r}_y, \hat{t})$. A partial Gram-Schmidt projection is used to obtain a rotation matrix $\hat{R}$ from $\hat{r}_x, \hat{r}_y$ [112]. We train the weights of our network to minimize the mean-squared error between the predicted and ground truth poses, i.e., $\mathcal{L}_{\text{audio}}(\hat{R}, \hat{t}, R, t) := \beta||\hat{R} - R||^2 + ||\hat{t} - t||^2$, where $\beta > 0$ is a hyperparameter that weights the relative importance between the rotation and translation errors.

5.1 Evaluation

Benchmarks. There are no datasets for absolute pose estimation with audio-visual data, so we introduce new benchmarks on two scenes from Replica [85] and four scenes from Matterport3D [17] that are held out from pretraining. Training and test samples are obtained from distinct locations on the scene grid, at random azimuth and elevation angles. To introduce more ambiguity, we reduce the brightness of specific regions (approximately 30% of the floorplan area).

Baselines. We compare our audio-visual regression network to PoseNet [46], the established visual regression approach that we build upon. To assess the usefulness of the pretraining task, we also perform an ablation of our model that trains the audio network from scratch. Note that we also experimented with DSAC* [12] and ESAC [10], state-of-the-art scene coordinate regression models. However, we found DSAC* to perform poorly on these large, ambiguous

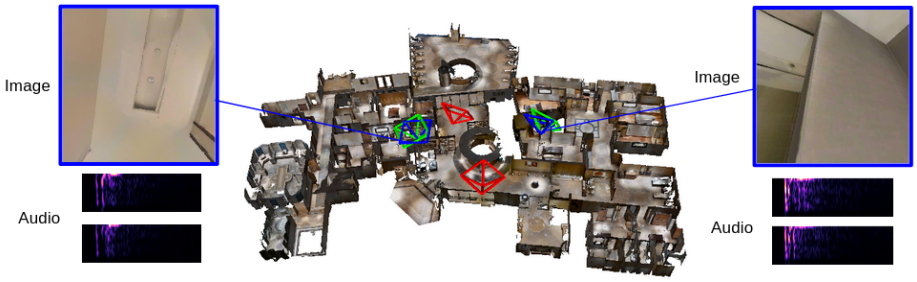

Fig. 7. Combining audio and vision for absolute pose regression. The input images observe ambiguous views of the scene (blue frustums). This result in poor performance on the part of the visual model (red frustums), whereas the audio-visual model uses audio to disambiguate the position of the device (green frustums).

datasets, and we found ESAC to consume an unreasonable amount of training time to cover each scene with dozens of expert networks.

Results. Table 4 shows quantitative results for our approach. The audio-visual model provides a boost over the established visual baseline. This is already evident in portions of the scene with regular illumination, but particularly prominent in the portions of the scene with poor illumination (and greater visual ambiguity). Figure 7 provides a qualitative example of how audio sensing benefits the vision model for absolute pose regression in a large Matterport3D scene: the input images observe ambiguous views of the scene (blue). This result in poor performance on the part of the visual model (red), whereas the audio-visual model uses audio to disambiguate the position of the device (green).

Does pretraining help? Using our pretrained audio feature extractor significantly boosts the performance of the audio-visual model. When we train the model from scratch, without our pretrained audio feature extractor, we find that it performs significantly worse– surprisingly, even worse than the vision-only model. [103] provides an explanation: multimodal models do not necessarily outperform unimodal models, since modalities trained from scratch may generalize and overfit at different rates. This further validates the use of our pretrained audio features.

6 Discussion

Overall, we present an exciting new research direction that leverages active audio sensing for classic camera localization tasks. Our experiments show that integrating our scene-aware audio features into established vision models improves performance across relative pose estimation, place recognition, and absolute pose regression. We hope that our work inspires further research in this direction, including collection of real-world audio-visual datasets of indoor scenes with ground truth poses. While we focus on improving specific vision models in this work, our insights on using audio sensing are not limited to these architectures and could be combined with other task-specific advances in the literature.

References

1. Arandjelovic, R., Gronat, P., Torii, A., Pajdla, T., Sivic, J.: NetVLAD: CNN architecture for weakly supervised place recognition. In: CVPR (2016)
2. Arandjelovic, R., Zisserman, A.: Look, listen and learn. In: ICCV (2017)
3. Babenko, A., Lempitsky, V.: Aggregating local deep features for image retrieval. In: ICCV (2015)
4. Balntas, V., Li, S., Prisacariu, V.: RelocNet: Continuous Metric Learning Relocalisation using Neural Nets. In: The European Conference on Computer Vision (ECCV) (September 2018)
5. Balntas, V., Riba, E., Ponsa, D., Mikolajczyk, K.: Learning local feature descriptors with triplets and shallow convolutional neural networks. In: BMVC (2016)
6. Bhowmik, A., Gumhold, S., Rother, C., Brachmann, E.: Reinforced feature points: Optimizing feature detection and description for a high-level task. In: CVPR (June 2020)
7. Brachmann, E., Krull, A., Nowozin, S., Shotton, J., Michel, F., Gumhold, S., Rother, C.: DSAC - Differentiable RANSAC for Camera Localization. In: CVPR (2017)
8. Brachmann, E., Michel, F., Krull, A., Yang, M.Y., Gumhold, S., Rother, C.: Uncertainty-driven 6d pose estimation of objects and scenes from a single rgb image. In: CVPR (2016)
9. Brachmann, E., Rother, C.: Learning Less is More - 6D Camera Localization via 3D Surface Regression. In: CVPR (2018)
10. Brachmann, E., Rother, C.: Expert sample consensus applied to camera relocalization. In: ICCV (2019)
11. Brachmann, E., Rother, C.: Neural-guided RANSAC: Learning where to sample model hypotheses. In: ICCV (2019)
12. Brachmann, E., Rother, C.: Visual camera re-localization from RGB and RGB-D images using DSAC. TPAMI (2021)
13. Brahmbhatt, S., Gu, J., Kim, K., Hays, J., Kautz, J.: Geometry-aware learning of maps for camera localization. In: CVPR (2018)
14. Bui, M., Birdal, T., Deng, H., Albarqouni, S., Guibas, L., Ilic, S., Navab, N.: 6d camera relocalization in ambiguous scenes via continuous multimodal inference. In: ECCV (2020)
15. Cai, R., Hariharan, B., Snavely, N., Averbuch-Elor, H.: Extreme rotation estimation using dense correlation volumes. In: CVPR (2021)
16. Castle, R., Klein, G., Murray, D.W.: Video-rate localization in multiple maps for wearable augmented reality. In: 2008 12th IEEE International Symposium on Wearable Computers. pp. 15–22. IEEE (2008)
17. Chang, A., Dai, A., Funkhouser, T., Halber, M., Niessner, M., Savva, M., Song, S., Zeng, A., Zhang, Y.: Matterport3D: Learning from RGB-D data in indoor environments. In: 3DV (2017)
18. Chen, C., Al-Halah, Z., Grauman, K.: Semantic audio-visual navigation. In: CVPR (2021)
19. Chen, C., Jain, U., Schissler, C., Gari, S.V.A., Al-Halah, Z., Ithapu, V.K., Robinson, P., Grauman, K.: Audio-visual embodied navigation. environment **97**, 103 (2019)
20. Chen, C., Jain, U., Schissler, C., Gari, S.V.A., Al-Halah, Z., Ithapu, V.K., Robinson, P., Grauman, K.: SoundSpaces: Audio-visual navigation in 3D environments. In: ECCV (2020)

21. Chen, C., Majumder, S., Al-Halah, Z., Gao, R., Ramakrishnan, S.K., Grauman, K.: Learning to set waypoints for audio-visual navigation. arXiv preprint arXiv:2008.09622 (2020)
22. Chen, K., Snavely, N., Makadia, A.: Wide-baseline relative camera pose estimation with directional learning. In: CVPR (2021)
23. Chen, Z., Hu, X., Owens, A.: Structure from silence: Learning scene structure from ambient sound. arXiv preprint arXiv:2111.05846 (2021)
24. Christensen, J.H., Hornauer, S., Stella, X.Y.: Batvision: Learning to see 3d spatial layout with two ears. In: ICRA (2020)
25. Debski, A., Grajewski, W., Zaborowski, W., Turek, W.: Open-source localization device for indoor mobile robots. Procedia Computer Science **76**, 139–146 (2015)
26. Dokmanić, I., Parhizkar, R., Walther, A., Lu, Y.M., Vetterli, M.: Acoustic echoes reveal room shape. Proceedings of the National Academy of Sciences **110**(30), 12186–12191 (2013)
27. Dusmanu, M., Rocco, I., Pajdla, T., Pollefeys, M., Sivic, J., Torii, A., Sattler, T.: D2-net: A trainable cnn for joint detection and description of local features. arXiv preprint arXiv:1905.03561 (2019)
28. Eliakim, I., Cohen, Z., Kosa, G., Yovel, Y.: A fully autonomous terrestrial bat-like acoustic robot. PLoS computational biology **14**(9), e1006406 (2018)
29. En, S., Lechervy, A., Jurie, F.: Rpnet: An end-to-end network for relative camera pose estimation. In: Proceedings of the European Conference on Computer Vision (ECCV) Workshops (2018)
30. Fischler, M.A., Bolles, R.C.: Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Communications of the ACM **24**(6), 381–395 (1981)
31. Fisher III, J.W., Darrell, T., Freeman, W., Viola, P.: Learning joint statistical models for audio-visual fusion and segregation. NeurIPS (2000)
32. Gan, C., Zhao, H., Chen, P., Cox, D., Torralba, A.: Self-supervised moving vehicle tracking with stereo sound. In: ICCV (2019)
33. Gao, R., Chen, C., Al-Halah, Z., Schissler, C., Grauman, K.: VisualEchoes: Spatial image representation learning through echolocation. In: ECCV (2020)
34. Gao, R., Grauman, K.: 2.5D visual sound. In: CVPR (2019)
35. Gao, R., Oh, T.H., Grauman, K., Torresani, L.: Listen to look: Action recognition by previewing audio. In: CVPR (2020)
36. Garg, S., Fischer, T., Milford, M.: Where is your place, visual place recognition? IJCAI (2021)
37. Greene, N.: Environment mapping and other applications of world projections. IEEE computer graphics and Applications **6**(11), 21–29 (1986)
38. Hausler, S., Garg, S., Xu, M., Milford, M., Fischer, T.: Patch-NetVLAD: Multi-scale fusion of locally-global descriptors for place recognition. In: CVPR (2021)
39. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
40. Hershey, J., Movellan, J.: Audio vision: Using audio-visual synchrony to locate sounds. NeurIPS (1999)
41. Hu, J., Ozay, M., Zhang, Y., Okatani, T.: Revisiting single image depth estimation: Toward higher resolution maps with accurate object boundaries. In: WACV (2019)
42. Humenberger, M., Cabon, Y., Guerin, N., Morat, J., Revaud, J., Rerole, P., Pion, N., de Souza, C., Leroy, V., Csurka, G.: Robust Image Retrieval-based Visual Localization using Kapture. arXiv:2007.13867 (2020)

43. Jégou, H., Douze, M., Schmid, C., Pérez, P.: Aggregating local descriptors into a compact image representation. In: CVPR (2010)
44. Kazakos, E., Nagrani, A., Zisserman, A., Damen, D.: Epic-fusion: Audio-visual temporal binding for egocentric action recognition. In: ICCV (2019)
45. Kendall, A., Cipolla, R.: Geometric loss functions for camera pose regression with deep learning. In: CVPR (2017)
46. Kendall, A., Grimes, M., Cipolla, R.: PoseNet: A Convolutional Network for Real-Time 6-DOF Camera Relocalization. In: ICCV (2015)
47. Kidron, E., Schechner, Y.Y., Elad, M.: Pixels that sound. In: CVPR (2005)
48. Laskar, Z., Melekhov, I., Kalia, S., Kannala, J.: Camera relocalization by computing pairwise relative poses using convolutional neural network. In: ICCV Workshops (2017)
49. Li, X., Wang, S., Zhao, Y., Verbeek, J., Kannala, J.: Hierarchical scene coordinate classification and regression for visual localization. In: CVPR (2020)
50. Li, Y., Snavely, N., Huttenlocher, D.P.: Location recognition using prioritized feature matching. In: ECCV (2010)
51. Li, Y., Snavely, N., Huttenlocher, D.P., Fua, P.: Worldwide pose estimation using 3D point clouds. In: ECCV (2012)
52. Lim, H., Sinha, S.N., Cohen, M.F., Uyttendaele, M.: Real-time image-based 6-dof localization in large-scale environments. In: CVPR (2012)
53. Lindell, D.B., Wetzstein, G., Koltun, V.: Acoustic non-line-of-sight imaging. In: CVPR (2019)
54. Liu, D., Cui, Y., Yan, L., Mousas, C., Yang, B., Chen, Y.: Densernet: Weakly supervised visual localization using multi-scale feature aggregation. In: Proceedings of the AAAI Conference on Artificial Intelligence (2021)
55. Long, X., Gan, C., De Melo, G., Wu, J., Liu, X., Wen, S.: Attention clusters: Purely attention based local feature integration for video classification. In: CVPR (2018)
56. Masone, C., Caputo, B.: A survey on deep visual place recognition. IEEE Access **9**, 19516–19547 (2021)
57. Melekhov, I., Ylioinas, J., Kannala, J., Rahtu, E.: Relative camera pose estimation using convolutional neural networks. In: International Conference on Advanced Concepts for Intelligent Vision Systems. pp. 675–687. Springer (2017)
58. Morgado, P., Li, Y., Nvasconcelos, N.: Learning representations from audio-visual spatial alignment. NeurIPS **33**, 4733–4744 (2020)
59. Morgado, P., Nvasconcelos, N., Langlois, T., Wang, O.: Self-supervised generation of spatial audio for 360 video. NeurIPS **31** (2018)
60. Morgado, P., Vasconcelos, N., Misra, I.: Audio-visual instance discrimination with cross-modal agreement. In: CVPR (2021)
61. Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., Ng, A.Y.: Multimodal deep learning. In: ICML (2011)
62. Owens, A., Efros, A.A.: Audio-visual scene analysis with self-supervised multisensory features. In: ECCV. pp. 631–648 (2018)
63. Owens, A., Isola, P., McDermott, J., Torralba, A., Adelson, E.H., Freeman, W.T.: Visually indicated sounds. In: CVPR (2016)
64. Owens, A., Wu, J., McDermott, J.H., Freeman, W.T., Torralba, A.: Ambient sound provides supervision for visual learning. In: ECCV (2016)
65. Parida, K.K., Srivastava, S., Sharma, G.: Beyond image to depth: Improving depth prediction using echoes. In: CVPR (2021)

66. Politis, A., Mesaros, A., Adavanne, S., Heittola, T., Virtanen, T.: Overview and evaluation of sound event localization and detection in dcase 2019. IEEE/ACM Transactions on Audio, Speech, and Language Processing **29**, 684–698 (2020)
67. Poursaeed, O., Yang, G., Prakash, A., Fang, Q., Jiang, H., Hariharan, B., Belongie, S.: Deep fundamental matrix estimation without correspondences. In: Proceedings of the European Conference on Computer Vision (ECCV) Workshops (2018)
68. Purushwalkam, S., Gari, S.V.A., Ithapu, V.K., Schissler, C., Robinson, P., Gupta, A., Grauman, K.: Audio-visual floorplan reconstruction. In: ICCV (2021)
69. Raguram, R., Frahm, J.M., Pollefeys, M.: A comparative analysis of ransac techniques leading to adaptive real-time random sample consensus. In: European conference on computer vision. pp. 500–513. Springer (2008)
70. Ranftl, R., Koltun, V.: Deep fundamental matrix estimation. In: ECCV (2018)
71. Revaud, J., Weinzaepfel, P., de Souza, C.R., Humenberger, M.: R2D2: repeatable and reliable detector and descriptor. In: NeurIPS (2019)
72. de Sa, V.R.: Learning classification with unlabeled data. NeurIPS (1994)
73. Sarlin, P.E., Cadena, C., Siegwart, R., Dymczyk, M.: From Coarse to Fine: Robust Hierarchical Localization at Large Scale. In: CVPR (2019)
74. Sarlin, P.E., DeTone, D., Malisiewicz, T., Rabinovich, A.: Superglue: Learning feature matching with graph neural networks. In: CVPR (2020)
75. Sarlin, P.E., Unagar, A., Larsson, M., Germain, H., Toft, C., Larsson, V., Pollefeys, M., Lepetit, V., Hammarstrand, L., Kahl, F., Sattler, T.: Back to the Feature: Learning robust camera localization from pixels to pose. In: CVPR (2021), `https://arxiv.org/abs/2103.09213`
76. Sattler, T., Havlena, M., Radenovic, F., Schindler, K., Pollefeys, M.: Hyperpoints and fine vocabularies for large-scale location recognition. In: ICCV (2015)
77. Sattler, T., Leibe, B., Kobbelt, L.: Improving Image-Based Localization by Active Correspondence Search. In: ECCV (2012)
78. Sattler, T., Leibe, B., Kobbelt, L.: Efficient & Effective Prioritized Matching for Large-Scale Image-Based Localization. PAMI (2017)
79. Sattler, T., Torii, A., Sivic, J., Pollefeys, M., Taira, H., Okutomi, M., Pajdla, T.: Are Large-Scale 3D Models Really Necessary for Accurate Visual Localization? In: CVPR (2017)
80. Savva, M., Kadian, A., Maksymets, O., Zhao, Y., Wijmans, E., Jain, B., Straub, J., Liu, J., Koltun, V., Malik, J., Parikh, D., Batra, D.: Habitat: A Platform for Embodied AI Research. In: ICCV (2019)
81. Shavit, Y., Ferens, R., Keller, Y.: Learning multi-scene absolute pose regression with transformers. In: ICCV. pp. 2733–2742 (October 2021)
82. Shotton, J., Glocker, B., Zach, C., Izadi, S., Criminisi, A., Fitzgibbon, A.: Scene Coordinate Regression Forests for Camera Relocalization in RGB-D Images. In: CVPR (2013)
83. Singh, N., Mentch, J., Ng, J., Beveridge, M., Drori, I.: Image2reverb: Cross-modal reverb impulse response synthesis. In: ICCV (2021)
84. Sohl-Dickstein, J., Teng, S., Gaub, B.M., Rodgers, C.C., Li, C., DeWeese, M.R., Harper, N.S.: A device for human ultrasonic echolocation. IEEE Transactions on Biomedical Engineering **62**(6), 1526–1534 (2015)
85. Straub, J., Whelan, T., Ma, L., Chen, Y., Wijmans, E., Green, S., Engel, J.J., Mur-Artal, R., Ren, C., Verma, S., Clarkson, A., Yan, M., Budge, B., Yan, Y., Pan, X., Yon, J., Zou, Y., Leon, K., Carter, N., Briales, J., Gillingham, T., Mueggler, E., Pesqueira, L., Savva, M., Batra, D., Strasdat, H.M., Nardi, R.D., Goesele,

M., Lovegrove, S., Newcombe, R.: The Replica dataset: A digital replica of indoor spaces. arXiv preprint arXiv:1906.05797 (2019)
86. Sun, J., Shen, Z., Wang, Y., Bao, H., Zhou, X.: LoFTR: Detector-free local feature matching with transformers. CVPR (2021)
87. Sun, W., Jiang, W., Trulls, E., Tagliasacchi, A., Yi, K.M.: ACNe: Attentive context normalization for robust permutation-equivariant learning. In: CVPR (June 2020)
88. Svarm, L., Enqvist, O., Oskarsson, M., Kahl, F.: Accurate localization and pose estimation for large 3D models. In: CVPR (2014)
89. Svärm, L., Enqvist, O., Kahl, F., Oskarsson, M.: City-scale localization for cameras with known vertical direction. TPAMI (2017)
90. Taira, H., Okutomi, M., Sattler, T., Cimpoi, M., Pollefeys, M., Sivic, J., Pajdla, T., Torii, A.: InLoc: Indoor visual localization with dense matching and view synthesis. In: CVPR (2018)
91. Taira, H., Okutomi, M., Sattler, T., Cimpoi, M., Pollefeys, M., Sivic, J., Pajdla, T., Torii, A.: InLoc: Indoor Visual Localization with Dense Matching and View Synthesis. TPAMI (2021)
92. Taubner, F., Tschopp, F., Novkovic, T., Siegwart, R., Furrer, F.: Lcd–line clustering and description for place recognition. In: 2020 International Conference on 3D Vision (3DV) (2020)
93. Thrun, S.: Affine structure from sound. NeurIPS (2005)
94. Torii, A., Arandjelovic, R., Sivic, J., Okutomi, M., Pajdla, T.: 24/7 place recognition by view synthesis. In: CVPR (2015)
95. Türkoğlu, M.Ö., Brachmann, E., Schindler, K., Brostow, G., Monszpart, A.: Visual Camera Re-Localization Using Graph Neural Networks and Relative Pose Supervision. In: 3DV. IEEE (2021)
96. Tyszkiewicz, M., Fua, P., Trulls, E.: Disk: Learning local features with policy gradient. In: NeurIPS (2020)
97. Valentin, J., Nießner, M., Shotton, J., Fitzgibbon, A., Izadi, S., Torr, P.: Exploiting Uncertainty in Regression Forests for Accurate Camera Relocalization. In: CVPR (2015)
98. Vasudevan, A.B., Dai, D., Gool, L.V.: Semantic object prediction and spatial sound super-resolution with binaural sounds. In: ECCV (2020)
99. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. NeurIPS **30** (2017)
100. Villalpando, A.P., Schillaci, G., Hafner, V.V., Guzmán, B.L.: Ego-noise predictions for echolocation in wheeled robots. In: ALIFE 2019: The 2019 Conference on Artificial Life. pp. 567–573. MIT Press (2019)
101. Wagner, D., Reitmayr, G., Mulloni, A., Drummond, T., Schmalstieg, D.: Real-time detection and tracking for augmented reality on mobile phones. IEEE transactions on visualization and computer graphics **16**(3), 355–368 (2009)
102. Walch, F., Hazirbas, C., Leal-Taixé, L., Sattler, T., Hilsenbeck, S., Cremers, D.: Image-Based Localization Using LSTMs for Structured Feature Correlation. In: ICCV (2017)
103. Wang, W., Tran, D., Feiszli, M.: What makes training multi-modal classification networks hard? In: CVPR (2020)
104. Winkelbauer, D., Denninger, M., Triebel, R.: Learning to localize in new environments from synthetic training data. In: ICRA (2021)
105. Wu, Z., Jiang, Y.G., Wang, X., Ye, H., Xue, X.: Multi-stream multi-class fusion of deep networks for video classification. In: Proceedings of the 24th ACM international conference on Multimedia. pp. 791–800 (2016)

106. Yang, K., Lin, W.Y., Barman, M., Condessa, F., Kolter, Z.: Defending multimodal fusion models against single-source adversaries. In: CVPR (2021)
107. Yang, K., Russell, B., Salamon, J.: Telling left from right: Learning spatial correspondence of sight and sound. In: CVPR (2020)
108. Yi, K.M., Trulls, E., Ono, Y., Lepetit, V., Salzmann, M., Fua, P.: Learning to find good correspondences. In: CVPR (2018)
109. Yue, H., Miao, J., Yu, Y., Chen, W., Wen, C.: Robust loop closure detection based on bag of superpoints and graph verification. In: IROS (2019)
110. Zhang, Z., Wu, J., Li, Q., Huang, Z., Traer, J., McDermott, J.H., Tenenbaum, J.B., Freeman, W.T.: Generative modeling of audible shapes for object perception. In: ICCV (2017)
111. Zhou, Q., Sattler, T., Pollefeys, M., Leal-Taixé, L.: To Learn or Not to Learn: Visual Localization from Essential Matrices. In: ICRA (2019)
112. Zhou, Y., Barnes, C., Lu, J., Yang, J., Li, H.: On the continuity of rotation representations in neural networks. In: CVPR (2019)