VisualEchoes: Spatial Image Representation Learning through Echolocation

Ruohan Gao[1,3], Changan Chen[1,3], Ziad Al-Halah[1],
Carl Schissler[2], Kristen Grauman[1,3]

[1]The University of Texas at Austin, [2]Facebook Reality Lab, [3]Facebook AI Research
{rhgao,changan,ziad,grauman}@cs.utexas.edu, carl.schissler@fb.com

Abstract. Several animal species (e.g., bats, dolphins, and whales) and even visually impaired humans have the remarkable ability to perform echolocation: a biological sonar used to perceive spatial layout and locate objects in the world. We explore the spatial cues contained in echoes and how they can benefit vision tasks that require spatial reasoning. First we capture echo responses in photo-realistic 3D indoor scene environments. Then we propose a novel interaction-based representation learning framework that learns useful *visual* features via echolocation. We show that the learned image features are useful for multiple downstream vision tasks requiring spatial reasoning—monocular depth estimation, surface normal estimation, and visual navigation—with results comparable or even better than heavily supervised pre-training. Our work opens a new path for representation learning for embodied agents, where supervision comes from interacting with the physical world.

1 Introduction

The perceptual and cognitive abilities of embodied agents are inextricably tied to their physical being. We perceive and act in the world by making use of all our senses—especially looking and listening. We see our surroundings to avoid obstacles, listen to the running water tap to navigate to the kitchen, and infer how far away the bus is once we hear it approaching.

By using *two* ears, we perceive spatial sound. Not only can we identify the sound-emitting object (e.g., the revving engine corresponds to a bus), but also we can determine that object's location, based on the time difference between when the sound reaches each ear (Interaural Time Difference, ITD) and the difference in sound level as it enters each ear (Interaural Level Difference, ILD). Critically, even beyond objects, audio is also rich with information about the *environment* itself. The sounds we receive are a function of the geometric structure of the space around us and the materials of its major surfaces [5]. In fact, some animals capitalize on these cues by using *echolocation*—actively emitting sounds to perceive the 3D spatial layout of their surroundings [68].

We propose to learn image representations from echoes. Motivated by how animals and blind people obtain spatial information from echo responses, first

we explore to what extent the echoes of chirps generated in a scanned 3D environment are predictive of the depth in the scene. Then, we introduce VisualEchoes, a novel image representation learning method based on echolocation. Given a first-person RGB view and an echo audio waveform, our model is trained to predict the correct camera orientation at which the agent would receive those echoes. In this way, the representation is forced to capture the alignment between the sound reflections and the (visually observed) surfaces in the environment. At test time, we observe only pixels—no audio. Our learned VisualEchoes encoder better reveals the 3D spatial cues embedded in the pixels, as we demonstrate in three downstream tasks.

Our approach offers a new way to learn image representations without manual supervision by *interacting* with the environment. In pursuit of this high-level goal there is exciting—though limited—prior work that learns visual features by touching objects [59,63,2,62] or moving in a space [45,1,27]. Unlike mainstream "self-supervised" feature learning work that crafts pretext tasks for large static repositories of human-taken images or video (e.g., colorization [88], jigsaw puzzles [57], audio-visual correspondence [50,6]), in *interaction-based feature learning* an embodied agent[1] performs physical actions in the world that dynamically influence its own first-person observations and possibly the environment itself. Both paths have certain advantages: while conventional self-supervised learning can capitalize on massive static datasets of human-taken photos, interaction-based learning allows an agent to "learn by acting" with rich multi-modal sensing. This has the advantage of learning features adaptable to new environments. Unlike any prior work, we explore feature learning from echoes.

Our contributions are threefold: 1) We explore the spatial cues contained in echoes, analyzing how they inform depth prediction; 2) We propose VisualEchoes, a novel interaction-based feature learning framework that uses echoes to learn an image representation and does not require audio at test time; 3) We successfully validate the learned spatial representation for the fundamental downstream vision tasks of monocular depth prediction, surface normal estimation, and visual navigation, with results comparable to or even outperforming heavily supervised pre-training baselines.

2 Related Work

Auditory Scene Analysis using Echoes: Previous work shows that using echo responses only, one can predict 2D [5] or 3D [14] room geometry and object shape [22]. Additionally, echoes can complement vision, especially when vision-based depth estimates are not reliable, e.g., on transparent windows or featureless walls [49,86]. In dynamic environments, autonomous robots can leverage echoes for obstacle avoidance [77] or mapping and navigation [17] using a bat-like echolocation model. Concurrently with our work, a low-cost audio system called BatVision is used to predict depth maps purely from echo responses [12]. Our work explores a novel direction for auditory scene analysis by employing

[1] person, robot, or simulated robot

echoes for spatial visual feature learning, and unlike prior work, the resulting features are applicable in the absence of any audio.

Self-Supervised Image Representation Learning: Self-supervised image feature learning methods leverage structured information within the data itself to generate labels for representation learning [69,38]. To this end, many "pretext" tasks have been explored—for example, predicting the rotation applied to an input image [35,1], discriminating image instances [19], colorizing images [52,88], solving a jigsaw puzzle from image patches [57], predicting unseen views of 3D objects [44], or multi-task learning using synthetic imagery [66]. Temporal information in videos also permits self-supervised tasks, for example, by predicting whether a frame sequence is in the correct order [55,20] or ensuring visual coherence of tracked objects [82,31,43]. Whereas these methods aim to learn features generically useful for recognition, our objective is to learn features generically useful for spatial estimation tasks. Accordingly, our echolocation objective is well-aligned with our target family of spatial tasks (depth, surfaces, navigation), consistent with findings that task similarity is important for positive transfer [87]. Furthermore, unlike any of the above, rather than learn from massive repositories of human-taken photos, the proposed approach learns from interactions with the scene via echolocation.

Feature Learning by Interaction: Limited prior work explores feature learning through interaction. Unlike the self-supervised methods discussed above, this line of work fosters agents that learn from their own observations in the world, which can be critical for adapting to new environments and to realize truly "bottom-up" learning by experience. Existing methods explore touch and motion interactions. In [59], objects are struck with a drumstick to facilitate learning material properties when they sound. In [63], the trajectory of a ball bouncing off surfaces facilitates learning physical scene properties. In [62,2], a robot learns object properties by poking or grasping at objects. In [27], a drone learns not to crash after attempting many crashes. In [45,1], an agent tracks its egomotion in concert with its visual stream to facilitate learning visual categories. In contrast, our idea is to learn visual features by *emitting audio* to acoustically interact with the scene. Our work offers a new perspective on interaction-based feature learning and has the advantages of not disrupting the scene physically and being ubiquitously available, i.e., reaching all surrounding surfaces.

Audio-Visual Learning: Inspiring recent work integrates sound and vision in joint learning frameworks that synthesize sounds for video [59,92], spatialize monaural sounds from video [29,56], separate sound sources [58,18,28,89,30,24], perform cross-modal feature learning [8,60], track audio-visual targets [34,9,3,26], segment objects with multi-channel audio [42], direct embodied agents to navigate in indoor environments [11,25], recognize actions in videos [32,48], and localize pixels associated with sounds in video frames [75,71,7,40]. None of the prior methods pursues echoes for visual learning. Furthermore, whereas nearly all existing audio-visual methods operate in a passive manner, observing incidental sounds within a video, in our approach the system learns by actively emitting sound—a form of interaction with the physical environment.

Monocular Depth Estimation To improve monocular depth estimation, recent methods focus on improving neural network architectures [23] or graphical models [81,53,84], employing multi-scale feature fusion and multi-task learning [15,41], leveraging motion cues from successive frames [76], or transfer learning [47]. However, these approaches rely on depth-labeled data that can be expensive to obtain. Hence, recent approaches leverage scenes' spatial and temporal structure to self-supervise depth estimation, by using the camera motion between pairs of images [33,36] or frames [91,80,37,46], or consistency cues between depth and features like surface normals [85] or optical flow [65]. Unlike any of these existing methods, we show that audio in the form of an echo response can be effectively used to recover depth, and we develop a novel feature learning method that benefits a purely visual representation (no audio) at test time.

3 Approach

Our goals are to show that echoes convey spatial information, to learn visual representations by echolocation, and to leverage the learned representations for downstream tasks. In the following, we first describe how we simulate echoes in 3D environments (Sec. 3.1). Then we perform a case study to demonstrate how echoes can benefit monocular depth prediction (Sec. 3.2). Next, we present VisualEchoes, our interaction-based feature learning formulation to learn image representations (Sec. 3.3). Finally, we exploit the learned visual representation for monocular depth, surface normal prediction, and visual navigation (Sec. 3.4).

3.1 Echolocation Simulation

Our echolocation simulation is based on recent work on audio-visual navigation [11], which builds a realistic acoustic simulation on top of the Habitat [70] platform and Replica environments [73]. Habitat [70] is an open-source 3D simulator that supports efficient RGB, depth, and semantic rendering for multiple datasets [73,10,83]. Replica is a dataset of 18 apartment, hotel, office, and room scenes with 3D meshes and high definition range (HDR) textures and renderable reflector information. The platform in [11] simulates acoustics by pre-computing room impulse responses (RIR) between all pairs of possible source and receiver locations, using a form of audio ray-tracing [79]. An RIR is a transfer function between the sound source and the sound microphone, and it is influenced by the room geometry, materials, and the sound source location [51]. The sound received at the listener location is computed by convolving the appropriate RIR with the waveform of the source sound.

We use the binaural RIRs for all Replica environments to generate echoes for our approach. As the source audio "chirp" we use a sweep signal from 20Hz-20kHz (the human-audible range) within a duration of 3ms. While technically any emitted sound could provide some echo signal from which to learn, our design (1) intentionally provides the response for a wide range of frequencies and (2) does so in a short period of time to avoid overlap between echoes and

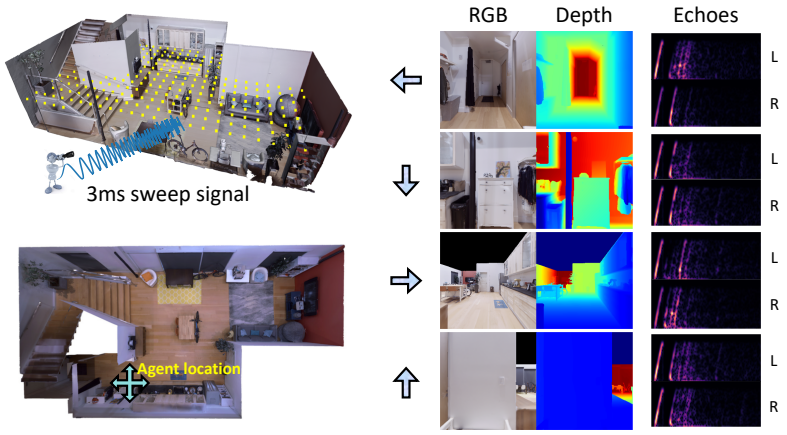

Fig. 1: Echolocation simulation in real-world scanned environments. During training, the agent goes to the densely sampled locations marked with yellow dots. The left bottom figure illustrates the top-down view of one Replica scene where the agent's location is marked. The agent actively emits 3 ms omnidirectional sweep signals to get echo responses from the room. The right column shows the corresponding RGB and depth of the agent's view as well as the echoes received in the left and right ears when the agent faces each of the four directions.

direct sounds. We place the source at the *same* location as the receiver and convolve the RIR for this source-receiver pair with the sweep signal. In this way, we compute the echo responses that would be received at the agent's microphone locations. We place the agents at all navigable points on the grid (every 0.5m [11]) and orient the agent in four cardinal directions (0°, 90°, 180°, 270°) so that the rendered egocentric views (RGB and depth) and echoes capture room geometry from different locations and orientations.

Fig. 1 illustrates how we perform echolocation for one scene environment. The agent goes to the densely sampled navigable locations marked with yellow dots and faces four orientations at each location. It actively emits omnidirectional chirp signals and records the echo responses received when facing each direction. Note that the spectrograms of the sounds received at the left (L) and right (R) ears reveal that the agent first receives the direct sound (strong bright curves), and then receives different echoes for the left and right microphones due to ITD, ILD, and pinnae reflections. The subtle difference in the two spectrograms conveys cues about the spatial configuration of the environment, as can be observed in the last column of Fig. 1.

3.2 Case Study: Spatial Cues in Echoes

With the synchronized egocentric views and echo responses in hand, we now conduct a case study to investigate the spatial cues contained in echo responses

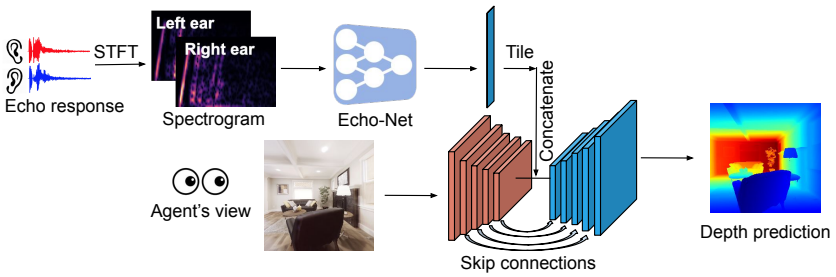

Fig. 2: Our RGB+Echo2Depth network takes the echo responses and the corresponding egocentric RGB view as input, and performs joint audio-visual analysis to predict the depth map for the input image. The injected echo response provides additional cues of the spatial layout of the scene. Note: in later sections we define networks that do not have access to the audio stream at test time.

in these realistic indoor 3D environments. We have two questions: (1) can we directly predict depth maps purely from echoes? and (2) can we use echoes to augment monocular depth estimation from RGB? Answering these questions will inform our ultimate goal of devising a interaction-supervised visual feature learning approach leveraging echoes only at training time (Sec. 3.3). Furthermore, it can shed light on the extent to which low-cost audio sensors can replace depth sensors, which would be especially useful for navigation robots under severe bandwidth or sensing constraints, e.g., nano drones [61,54].

Note that these two goals are orthogonal to that of prior work performing depth prediction from a single view [16,53,84,23,41]. Whereas they focus on developing sophisticated loss functions and architectures, here we explore how an agent *actively interacting with the scene acoustically* may improve its depth predictions. Our findings can thus complement existing monocular depth models.

We devise an RGB+Echo2Depth network (and its simplified variants using only RGB or echo) to test the settings of interest. The RGB+Echo2Depth network predicts a depth map based on the agent's egocentric RGB input and the echo response it receives when it emits a chirp standing at that position and orientation in the 3D environment. The core model is a multi-modal U-Net [67]; see Fig. 2. To directly measure the spatial cues contained in echoes alone, we also test a variant called Echo2Depth. Instead of performing upsampling based on the audio-visual representation, this model drops the RGB input, reshapes the audio feature, and directly upsamples from the audio representation. Similarly, to measure the cues contained in the RGB alone, a variant called RGB2Depth drops the echoes and predicts the depth map purely based on the visual features. The RGB2Depth model represents existing monocular depth prediction approaches that predict depth from a single RGB image, in the context of the same architecture design as RGB+Echo2Depth to allow apples-to-apples calibration of our findings. We use RGB images of spatial dimension 128×128. See Supp. for network details and loss functions used to train the three models.

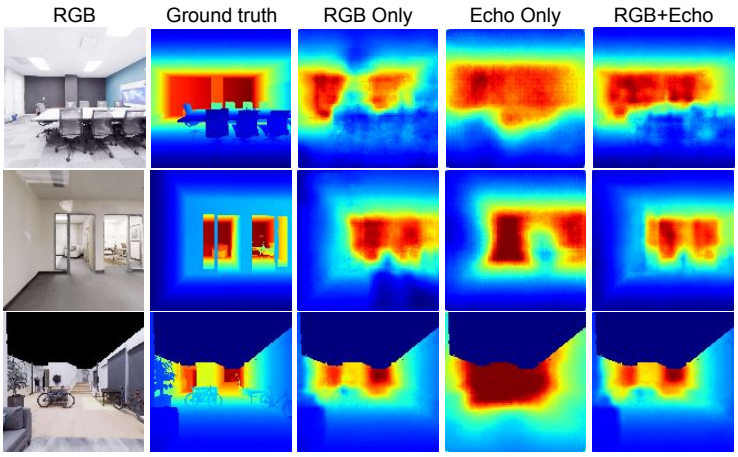

Fig. 3: Qualitative results of our case study on monocular depth estimation in unseen environments using echoes. Together with the quantitative results (Tab. 1), these examples show that echoes contain useful spatial cues that inform a visual spatial task. For example, in row 1, the RGB+Echo model better infers the depth of the column on the back wall, whereas the RGB-Only model mistakenly infers the strong contours to indicate a much closer surface. The last row shows a typical failure case (see text). See Supp. for more examples.

	RMS ↓	REL ↓	log 10 ↓	$\delta < 1.25$ ↑	$\delta < 1.25^2$ ↑	$\delta < 1.25^3$ ↑
Average	1.070	0.791	0.230	0.235	0.509	0.750
Echo2Depth	0.713	0.347	0.134	0.580	0.772	0.868
RGB2Depth	0.374	0.202	0.076	0.749	0.883	0.945
RGB+Echo2Depth	**0.346**	**0.172**	**0.068**	**0.798**	**0.905**	**0.950**

Table 1: Case study depth prediction results. ↓ lower better, ↑ higher better.

Table 1 shows the quantitative results of predicting depth from only echoes, only RGB, or their combination. We evaluate on a heldout set of three Replica environments (comprising 1,464 total views) with standard metrics: root mean squared error (RMS), mean relative error (REL), mean $\log 10$ error ($\log 10$), and thresholded accuracy [41,16]. We can see that depth prediction is possible purely from echoes. Augmenting traditional single-view depth estimation with echoes (bottom row) achieves the best performance by leveraging the additional acoustic spatial cues. Echoes alone are naturally weaker than RGB alone, yet still better than the simple Average baseline that predicts the average depth values in all training data.

Fig. 3 shows qualitative examples. It is clear that echo responses indeed contain cues of the spatial layout; the depth map captures the rough room layout, especially its large surfaces. When combined with RGB, the predictions are more

accurate. The last row shows a typical failure case, where the echoes alone cannot capture the depth as well due to far away surfaces with weaker echo signals.

3.3 VisualEchoes Spatial Representation Learning Framework

Having established the scope for inferring depth from echoes, we now present our VisualEchoes model to leverage echoes for visual representation learning. We stress that our approach assumes audio/echoes are available only during training; at test time, an RGB image alone is the input.

The key insight of our approach is that the echoes and visual input should be consistent. This is because both are functions of the same latent variable—the 3D shape of the environment surrounding the agent's co-located camera and microphones. We implement this idea by training a network to predict their correct association.

In particular, as described in Sec. 3.1, at any position in the scene, we suppose the agent can face four orientations, i.e., at an azimuth angle of 0°, 90°, 180°, and 270°. When the agent emits the sweep signal (chirp) at a certain position, it will hear different echo responses when it faces each different orientation. If the agent correctly interprets the spatial layout of the current view from *visual* information, it should be able to tell whether that visual input is congruous with the echo response it hears. Furthermore, and more subtly, to the extent the agent implicitly learns about probable views surrounding its current egocentric field of view (e.g., what the view just to its right may look like given the context of what it sees in front of it), it should be able to tell which direction the received echo *would* be congruous with, if not the current view.

We introduce a representation learning network to capture this insight. See Fig. 4. The visual stream takes the agent's current RGB view as input, and the audio stream takes the echo response received from one of the four orientations—not necessarily the one that coincides with the visual stream orientation. The fusion layer fuses the audio and visual information to generate an audio-visual feature of dimension D. A final fully-connected layer is used to make the final prediction among four classes. See Supp. and Sec. 4 for architecture details. The four classes are defined as follows:

$\uparrow$: The echo is received from the same orientation as the agent's current view.
$\rightarrow$: The echo is received from the orientation if the agent turns right by 90°.
$\downarrow$: The echo is received from the orientation opposite the agent's current view.
$\leftarrow$: The echo is received from the orientation if the agent turns left by 90°.

The network is trained with cross-entropy loss. Note that although the emitted source signal is always the same (3 ms *omnidirectional* sweep signal, cf. Sec. 3.1), the agent hears different echoes when facing the four directions because of the shape of the ears and the head shadowing effect modeled in the binaural head-related transfer function (HRTF). Since the classes above are defined relative to the agent's current view, it can only tell the orientation for which it is receiving the echoes if it can correctly interpret the 3D spatial layout within the RGB

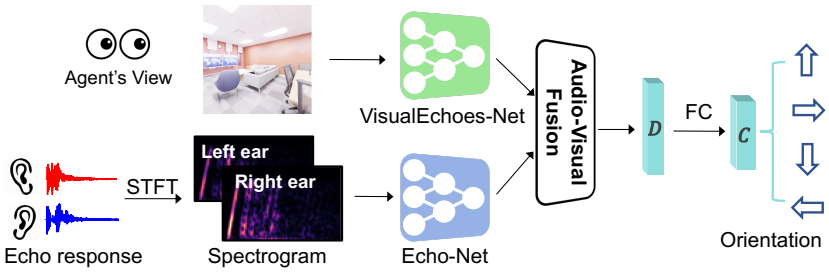

Fig. 4: Our VisualEchoes network takes the agent's current RGB view as visual input, and the echo responses from one of the four orientations as audio input. The goal is to predict the orientation at which the agent would receive the input echoes based on analyzing the spatial layout in the image. After training with RGB and echoes, the VisualEchoes-Net is a pre-trained encoder ready to extract spatially enriched features from novel RGB images, as we validate with multiple downstream tasks (cf. Sec. 3.4).

input. In this way, the agent's aural interaction with the scene enhances spatial feature learning for the visual stream.

The proposed idea generalizes trivially to use more than four discrete orientations—and even arbitrary orientations if we were to use regression rather than classification. The choice of four is simply based on the sound simulations available in existing data [11], though we anticipate it is a good granularity to capture the major directions around the agent. Our training paradigm requires the representation to discern mismatches between the image and echo using echoes generated from the same physical position on the ground plane but different orientations. This is in line with our interactive embodied agent motivation, where an agent can look ahead, then turn and hear echoes from another orientation at the same place in the environment, and learn their (dis)association. In fact, ecological psychologists report that humans can perform more accurate echolocation when moving, supporting the rationale of our design [74,68]. Furthermore, our design ensures the mismatches are "hard" examples useful for learning spatial features because the audio-visual data at offset views will naturally be related to one another (as opposed to views or echoes from an unrelated environment).

3.4 Downstream Tasks for the Learned Spatial Representation

Having introduced our VisualEchoes feature learning framework, next we describe how we repurpose the learned visual representation for three fundamental downstream tasks that require spatial reasoning: monocular depth prediction, surface normal estimation, and visual navigation. For each task, we adopt strong models from the literature and swap in our pre-trained encoder VisualEchoes-Net for the RGB input.

Monocular depth prediction: We explore how our echo-based pre-training can benefit performance for traditional monocular depth prediction. Note that

unlike the case study in Sec. 3.2, in this case there are no echo inputs at test time, only RGB. To evaluate the quality of our learned representation, we adopt a strong recent approach for monocular depth prediction [41] consisting of several novel loss functions and a multi-scale network architecture that is based on a backbone network. We pre-train ResNet-50 [39] using VisualEchoes and use it as the backbone for comparison with [41].

Surface normal estimation: We also evaluate the learned spatial representation to predict surface normals from a single image, another fundamental mid-level vision task that requires spatial understanding of the geometry of the surfaces [21]. We adopt the the state-of-the-art pyramid scene parsing network PSPNet architecture [90] for surface normal prediction, again swapping in our pre-trained VisualEchoes network for the RGB feature backbone.

Visual navigation: Finally, we validate on an embodied visual navigation task. In this task, the agent receives a sequence of RGB images as input and a point goal defined by a displacement vector relative to the starting position of the agent [4]. The agent is spawned at random locations and must navigate to the target location quickly and accurately. This entails reasoning about 3D spatial configurations to avoid obstacles and find the shortest path. We adopt a state-of-the-art reinforcement learning-based PointGoal visual navigation model [70]. It consists of a three-layer convolutional network and a fully-connected layer to extract visual feature from the RGB images. We pre-train its visual network using VisualEchoes, then train the full network end to end.

While other architectures are certainly possible for each task, our choices are based on both on the methods' effectiveness in practice, their wide use in the literature, and code availability. Our contribution is feature learning from echoes as a pre-training mechanism for spatial tasks, which is orthogonal to advances on architectures for each individual task. In fact, a key message of our results is that the VisualEchoes-Net encoder boosts multiple spatial tasks, under multiple different architectures, and on multiple datasets.

4 Experiments

We present experiments to validate VisualEchoes for three tasks and three datasets (Replica [73], NYU-V2 [72], and DIODE [78]). The goal is to examine the impact of our features compared to either learning features for that task from scratch or learning features with manual semantic supervision. See Supp. for details of the three datasets.

Implementation Details: All networks are implemented in PyTorch. For the echoes, we use the first 60 ms, which allows most of the room echo responses following the 3 ms chirp to be received. We use an audio sampling rate of 44.1 kHz. STFT is computed using a Hann window of length 64, hop length of 16, and FFT size of 512. The audio-visual fusion layer (see Fig. 4) concatenates the visual and audio feature, and then uses a fully-connected layer to reduce the feature dimension to $D = 128$. See Supp. for details of the network architectures and optimization hyperparameters.

		RMS ↓	REL ↓	log 10 ↓	$\delta < 1.25$ ↑	$\delta < 1.25^2$ ↑	$\delta < 1.25^3$ ↑
Sup	ImageNet Pre-trained	0.356	0.203	0.076	0.748	0.891	0.948
Sup	MIT Indoor Scene Pre-trained	0.334	0.196	0.072	0.770	0.897	0.950
Unsup	Scratch	0.360	0.214	0.078	0.747	0.879	0.940
Unsup	VisualEchoes (Ours)	**0.332**	**0.195**	**0.070**	**0.773**	**0.899**	**0.951**

(a) Replica

		RMS ↓	REL ↓	log 10 ↓	$\delta < 1.25$ ↑	$\delta < 1.25^2$ ↑	$\delta < 1.25^3$ ↑
Sup	ImageNet Pre-trained	0.812	0.249	0.102	0.589	0.855	0.955
Sup	MIT Indoor Scene Pre-trained	0.776	0.239	0.098	0.610	0.869	0.959
Unsup	Scratch	0.818	0.252	0.103	0.586	0.853	0.950
Unsup	VisualEchoes (Ours)	**0.797**	**0.246**	**0.100**	**0.600**	**0.863**	**0.956**

(b) NYU-V2

		RMS ↓	REL ↓	log 10 ↓	$\delta < 1.25$ ↑	$\delta < 1.25^2$ ↑	$\delta < 1.25^3$ ↑
Sup	ImageNet Pre-trained	2.250	0.453	0.199	0.336	0.591	0.766
Sup	MIT Indoor Scene Pre-trained	2.218	0.424	0.198	0.363	0.632	0.776
Unsup	Scratch	2.352	0.481	0.214	0.321	0.581	0.742
Unsup	VisualEchoes (Ours)	**2.223**	**0.430**	**0.198**	**0.340**	**0.610**	**0.769**

(c) DIODE

Table 2: Depth prediction results on the Replica, NYU-V2, and DIODE datasets. We use the RGB2Depth network from Sec. 3.2 for all models. Our VisualEchoes pre-training transfers well, consistently predicting depth better than the model trained from scratch. Furthermore, it is even competitive with the supervised models, whether they are pre-trained for ImageNet or MIT Indoor Scenes (1M/16K manually labeled images). ↓ lower better, ↑ higher better. (Un)sup = (un)supervised. We boldface the best unsupervised method.

Evaluation Metrics: We report standard metrics for the downstream tasks. 1) *Monocular Depth Prediction:* RMS, REL, and others as defined above, following [41,16]. 2) *Surface Normal Estimation:* mean and median of the angle distance and the percentage of good pixels (i.e., the fraction of pixels with cosine distance to ground-truth less than t) with $t = 11.25°, 22.5°, 30°$, following [21]. 3) *Visual Navigation:* success rate normalized by inverse path length (SPL), the distance to the goal at the end of the episode, and the distance to the goal normalized by the trajectory length, following [4].

4.1 Transferring VisualEchoes Features for RGB2Depth

Having confirmed echoes reveal spatial cues in Sec. 3.2, we now examine the effectiveness of VisualEchoes, our learned representation. Our model achieves 66% test accuracy on the orientation prediction task, while chance performance is only 25%; this shows learning the visual-echo consistency task itself is possible.

First, we use the same RGB2Depth network from our case study in Sec. 3.2 as a testbed to demonstrate the learned spatial features can be successfully transferred to other domains. Instead of randomly initializing the RGB2Depth UNet encoder, we initialize with an encoder 1) pre-trained for our visual-echo

	RMS ↓	REL ↓	log 10 ↓	$\delta < 1.25$ ↑	$\delta < 1.25^2$ ↑	$\delta < 1.25^3$ ↑
Scratch	0.360	0.214	0.078	0.747	0.879	0.940
SimpleVisualEchoes	0.340	0.198	0.073	0.763	0.892	0.948
BinaryMatching	0.345	0.199	0.074	0.760	0.889	0.944
VisualEchoes (Ours)	**0.332**	**0.195**	**0.070**	**0.773**	**0.899**	**0.951**

Table 3: Ablation study on Replica. See Supp. for results on NYU-V2 and Diode.

consistency task, 2) pre-trained for image classification using ImageNet [13], or 3) pre-trained for scene classification using the MIT Indoor Scene dataset [64]. Throughout, aside from the standard ImageNet pre-training baseline, we also include MIT Indoor Scenes pre-training, in case it strengthens the baseline due to its domain alignment with the indoor scenes in Replica, DIODE, and NYU-2.

Table 2 shows the results on all three datasets: Replica, NYU-V2, and DIODE. The model initialized with our pre-trained VisualEchoes network achieves much better performance compared to the model trained from scratch. Moreover, it even outperforms the supervised model pre-trained on scene classification in some cases. The ImageNet pre-trained model performs much worse; we suspect that the UNet encoder does not have sufficient capacity to handle ImageNet classification, and also the ImageNet domain is much different than indoor scene environments. This result accentuates that task similarity promotes positive transfer [87]: our unsupervised spatial pre-training task is more powerful for depth inference than a supervised semantic category pre-training task. See Supp. for low-shot experiments varying the amount of training data.

We also perform an ablation study to demonstrate that the design of our spatial representation learning framework is essential and effective. We compare with the following two variants: SimpleVisualEchoes, which simplifies our orientation prediction task to two classes; and BinaryMatching, which mimics prior work [6] that leverages the correspondence between images and audio as supervision by training a network to decide if the echo and RGB are from the same environment. As shown in Table 3, our method performs much better than both baselines. See Supp. for details.

4.2 Evaluating on Downstream Tasks

Next we evaluate the impact of our learned VisualEchoes representation on all three downstream tasks introduced in Sec. 3.4.

Monocular depth prediction: Table 4a shows the results.[2] All methods use the same settings as [41], where they evaluate and report results on NYU-V2. We use the authors' publicly available code[3] and use ResNet-50 as the encoder. See Supp. for details. With this apples-to-apples comparison, the difference in performance can be attributed to whether/how the encoder is pre-trained. Although

[2] We evaluate on NYU-V2, the most widely used dataset for the task of single view depth prediction and surface normal estimation. The authors's code [41,38] is tailored to this dataset.

[3] https://github.com/JunjH/Revisiting_Single_Depth_Estimation

		RMS ↓	REL ↓	log 10 ↓	$\delta < 1.25$ ↑	$\delta < 1.25^2$ ↑	$\delta < 1.25^3$ ↑
Sup	ImageNet Pre-trained [41]	**0.555**	**0.126**	**0.054**	**0.843**	**0.968**	**0.991**
Sup	MIT Indoor Scene Pre-trained	0.711	0.180	0.075	0.730	0.925	0.979
Unsup	Scratch	0.804	0.209	0.086	0.676	0.897	0.967
Unsup	VisualEchoes (Ours)	0.683	0.165	0.069	0.762	0.934	0.981

(a) Depth prediction results on NYU-V2.

		Mean Dist. ↓	Median Dist. ↓	$t < 11.25°$ ↑	$t < 22.5°$ ↑	$t < 30°$ ↑
Sup	ImageNet Pre-trained	26.4	17.1	36.1	59.2	68.5
Sup	MIT Indoor Scene Pre-trained	25.2	17.5	36.5	57.8	67.2
Unsup	Scratch	26.3	16.1	37.9	60.6	69.0
Unsup	VisualEchoes (Ours)	**22.9**	**14.1**	**42.7**	**64.1**	**72.4**

(b) Surface normal estimation results on NYU-V2. The results for the ImageNet Pre-trained baseline and the Scratch baseline are directly quoted from [38].

		SPL ↑	Distance to Goal ↓	Normalized Distance to Goal ↓
Sup	ImageNet Pre-trained	0.833	0.663	0.081
Sup	MIT Indoor Scene Pre-trained	0.798	1.05	0.124
Unsup	Scratch	0.830	0.728	0.096
Unsup	VisualEchoes (Ours)	**0.856**	**0.476**	**0.061**

(c) Visual navigation performance in unseen Replica environments.

Table 4: Results for three downstream tasks. ↓ lower better, ↑ higher better.

our VisualEchoes features are learned from Replica, they transfer reasonably well to NYU-V2, outperforming models trained from scratch by a large margin. This result is important because it shows that despite training with simulated audio, our model generalizes to real-world test images. Our features also compare favorably to supervised models trained with heavy supervision.

Surface normal estimation: Table 4b shows the results. We follow the same setting as [38] and we use the authors' publicly available code.[4] Our model performs much better even compared to the ImageNet-supervised pre-trained model, demonstrating that our interaction-based feature learning framework via echoes makes the learned features more useful for 3D geometric tasks.

Visual navigation: Table 4c shows the results. By pre-training the visual network, VisualEchoes equips the embodied agents with a better sense of room geometry and allows them to learn faster (see Supp. for training curves). Notably, the agent also ends much closer to the goal. We suspect it can better gauge the distance because of our VisualEchoes pre-training. Models pre-trained for classification on MIT Indoor Scenes perform more poorly than Scratch; again, this suggests features useful for recognition may not be optimal for a spatial task like point goal navigation.

This series of results on three tasks consistently shows the promise of our VisualEchoes features. We see that learning from echoes translates into a strengthened *visual* encoding. Importantly, while it is always an option to train multiple representations entirely from scratch to support each given task, our

[4] https://github.com/facebookresearch/fair_self_supervision_benchmark

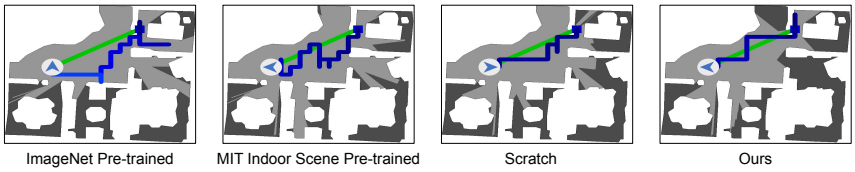

Fig. 5: Qualitative examples of visual navigation trajectories on top-down maps. Blue square and arrow denote agent's starting and ending positions, respectively. The green path indicates the shortest geodesic path to the goal, and the agent's path is in dark blue. Agent path color fades from dark blue to light blue as time goes by. Note, the agent sees a sequence of egocentric views, not the map.

results are encouraging since they show the *same* fundamental interaction-based pre-training is versatile across multiple tasks.

4.3 Qualitative Results

Fig. 5 shows example navigation trajectories on top-down maps. Our visual-echo consistency pre-training task allows the agent to better interpret the room's spatial layout to find the goal more quickly than the baselines. See Supp. for qualitative results on depth estimation and surface normal examples. Initializing with our pre-trained VisualEchoes network leads to much more accurate depth prediction and surface normal estimates compared to no pre-training, demonstrating the usefulness of the learned spatial features.

5 Conclusions and Future Work

We presented an approach to learn spatial image representations via echolocation. We performed an in-depth study on the spatial cues contained in echoes and how they can inform single-view depth estimation. We showed that the learned spatial features can benefit three downstream vision tasks. Our work opens a new path for interaction-based representation learning for embodied agents and demonstrates the potential of learning spatial visual representations even with a limited amount of multisensory data.

While our current implementation learns from audio rendered in a simulator, the results show that the learned spatial features already benefit transfer to vision-only tasks in real photos outside of the scanned environments (e.g., the NYU-V2 [72] and DIODE [78] images), indicating the realism of what our system learned. Nonetheless, it will be interesting future work to capture the echoes on a real robot. We are also interested in pursuing these ideas within a sequential model, such that the agent could actively decide when to emit chirps and what type of chirps to emit to get the most informative echo responses.

Acknowledgements: UT Austin is supported in part by DARPA Lifelong Learning Machines and ONR PECASE. RG is supported by Google PhD Fellowship and Adobe Research Fellowship.

References

1. Agrawal, P., Carreira, J., Malik, J.: Learning to see by moving. In: ICCV (2015)
2. Agrawal, P., Nair, A.V., Abbeel, P., Malik, J., Levine, S.: Learning to poke by poking: Experiential learning of intuitive physics. In: NeurIPS (2016)
3. Alameda-Pineda, X., Staiano, J., Subramanian, R., Batrinca, L., Ricci, E., Lepri, B., Lanz, O., Sebe, N.: Salsa: A novel dataset for multimodal group behavior analysis. TPAMI (2015)
4. Anderson, P., Chang, A., Chaplot, D.S., Dosovitskiy, A., Gupta, S., Koltun, V., Kosecka, J., Malik, J., Mottaghi, R., Savva, M., et al.: On evaluation of embodied navigation agents. arXiv preprint arXiv:1807.06757 (2018)
5. Antonacci, F., Filos, J., Thomas, M.R., Habets, E.A., Sarti, A., Naylor, P.A., Tubaro, S.: Inference of room geometry from acoustic impulse responses. IEEE Transactions on Audio, Speech, and Language Processing (2012)
6. Arandjelovic, R., Zisserman, A.: Look, listen and learn. In: ICCV (2017)
7. Arandjelović, R., Zisserman, A.: Objects that sound. In: ECCV (2018)
8. Aytar, Y., Vondrick, C., Torralba, A.: Soundnet: Learning sound representations from unlabeled video. In: NeurIPS (2016)
9. Ban, Y., Li, X., Alameda-Pineda, X., Girin, L., Horaud, R.: Accounting for room acoustics in audio-visual multi-speaker tracking. In: ICASSP (2018)
10. 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. 3DV (2017)
11. Chen, C., Jain, U., Schissler, C., Gari, S.V.A., Al-Halah, Z., Ithapu, V.K., Robinson, P., Grauman, K.: Audio-visual embodied navigation. In: ECCV (2020)
12. Christensen, J., Hornauer, S., Yu, S.: Batvision - learning to see 3d spatial layout with two ears. In: ICRA (2020)
13. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: CVPR (2009)
14. Dokmanić, I., Parhizkar, R., Walther, A., Lu, Y.M., Vetterli, M.: Acoustic echoes reveal room shape. Proceedings of the National Academy of Sciences (2013)
15. Eigen, D., Fergus, R.: Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In: ICCV (2015)
16. Eigen, D., Puhrsch, C., Fergus, R.: Depth map prediction from a single image using a multi-scale deep network. In: NeurIPS (2014)
17. Eliakim, I., Cohen, Z., Kosa, G., Yovel, Y.: A fully autonomous terrestrial bat-like acoustic robot. PLoS computational biology (2018)
18. Ephrat, A., Mosseri, I., Lang, O., Dekel, T., Wilson, K., Hassidim, A., Freeman, W.T., Rubinstein, M.: Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation. In: SIGGRAPH (2018)
19. Feng, Z., Xu, C., Tao, D.: Self-supervised representation learning by rotation feature decoupling. In: CVPR (2019)
20. Fernando, B., Bilen, H., Gavves, E., Gould, S.: Self-supervised video representation learning with odd-one-out networks. In: CVPR (2017)
21. Fouhey, D.F., Gupta, A., Hebert, M.: Data-driven 3d primitives for single image understanding. In: ICCV (2013)
22. Frank, N., Wolf, L., Olshansky, D., Boonman, A., Yovel, Y.: Comparing vision-based to sonar-based 3d reconstruction. ICCP (2020)
23. Fu, H., Gong, M., Wang, C., Batmanghelich, K., Tao, D.: Deep ordinal regression network for monocular depth estimation. In: CVPR (2018)

24. Gan, C., Huang, D., Zhao, H., Tenenbaum, J.B., Torralba, A.: Music gesture for visual sound separation. In: CVPR (2020)
25. Gan, C., Zhang, Y., Wu, J., Gong, B., Tenenbaum, J.B.: Look, listen, and act: Towards audio-visual embodied navigation. In: ICRA (2020)
26. Gan, C., Zhao, H., Chen, P., Cox, D., Torralba, A.: Self-supervised moving vehicle tracking with stereo sound. In: ICCV (2019)
27. Gandhi, D., Pinto, L., Gupta, A.: Learning to fly by crashing. In: IROS (2017)
28. Gao, R., Feris, R., Grauman, K.: Learning to separate object sounds by watching unlabeled video. In: ECCV (2018)
29. Gao, R., Grauman, K.: 2.5d visual sound. In: CVPR (2019)
30. Gao, R., Grauman, K.: Co-separating sounds of visual objects. In: ICCV (2019)
31. Gao, R., Jayaraman, D., Grauman, K.: Object-centric representation learning from unlabeled videos. In: ACCV (2016)
32. Gao, R., Oh, T.H., Grauman, K., Torresani, L.: Listen to look: Action recognition by previewing audio. In: CVPR (2020)
33. Garg, R., BG, V.K., Carneiro, G., Reid, I.: Unsupervised cnn for single view depth estimation: Geometry to the rescue. In: ECCV (2016)
34. Gebru, I.D., Ba, S., Evangelidis, G., Horaud, R.: Tracking the active speaker based on a joint audio-visual observation model. In: ICCV Workshops (2015)
35. Gidaris, S., Singh, P., Komodakis, N.: Unsupervised representation learning by predicting image rotations. In: ICLR (2018)
36. Godard, C., Mac Aodha, O., Brostow, G.J.: Unsupervised monocular depth estimation with left-right consistency. In: CVPR (2017)
37. Godard, C., Mac Aodha, O., Firman, M., Brostow, G.J.: Digging into self-supervised monocular depth estimation. In: ICCV (2019)
38. Goyal, P., Mahajan, D., Gupta, A., Misra, I.: Scaling and benchmarking self-supervised visual representation learning. In: ICCV (2019)
39. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
40. Hershey, J.R., Movellan, J.R.: Audio vision: Using audio-visual synchrony to locate sounds. In: NeurIPS (2000)
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. Irie, G., Ostrek, M., Wang, H., Kameoka, H., Kimura, A., Kawanishi, T., Kashino, K.: Seeing through sounds: Predicting visual semantic segmentation results from multichannel audio signals. In: ICASSP (2019)
43. Jayaraman, D., Grauman, K.: Slow and steady feature analysis: Higher order temporal coherence in video. In: CVPR (2016)
44. Jayaraman, D., Gao, R., Grauman, K.: Shapecodes: self-supervised feature learning by lifting views to viewgrids. In: ECCV (2018)
45. Jayaraman, D., Grauman, K.: Learning image representations equivariant to ego-motion. In: ICCV (2015)
46. Jiang, H., Larsson, G., Maire Greg Shakhnarovich, M., Learned-Miller, E.: Self-supervised relative depth learning for urban scene understanding. In: ECCV (2018)
47. Karsch, K., Liu, C., Kang, S.B.: Depth transfer: Depth extraction from video using non-parametric sampling. TPAMI (2014)
48. Kazakos, E., Nagrani, A., Zisserman, A., Damen, D.: Epic-fusion: Audio-visual temporal binding for egocentric action recognition. In: ICCV (2019)
49. Kim, H., Remaggi, L., Jackson, P.J., Fazi, F.M., Hilton, A.: 3d room geometry reconstruction using audio-visual sensors. In: 3DV (2017)

50. Korbar, B., Tran, D., Torresani, L.: Co-training of audio and video representations from self-supervised temporal synchronization. In: NeurIPS (2018)
51. Kuttruff, H.: Room Acoustics. CRC Press (2017)
52. Larsson, G., Maire, M., Shakhnarovich, G.: Colorization as a proxy task for visual understanding. In: CVPR (2017)
53. Liu, F., Shen, C., Lin, G.: Deep convolutional neural fields for depth estimation from a single image. In: CVPR (2015)
54. McGuire, K., De Wagter, C., Tuyls, K., Kappen, H., de Croon, G.: Minimal navigation solution for a swarm of tiny flying robots to explore an unknown environment. Science Robotics (2019)
55. Misra, I., Zitnick, C.L., Hebert, M.: Shuffle and learn: unsupervised learning using temporal order verification. In: ECCV (2016)
56. Morgado, P., Vasconcelos, N., Langlois, T., Wang, O.: Self-supervised generation of spatial audio for 360° video. In: NeurIPS (2018)
57. Noroozi, M., Favaro, P.: Unsupervised learning of visual representations by solving jigsaw puzzles. In: ECCV (2016)
58. Owens, A., Efros, A.A.: Audio-visual scene analysis with self-supervised multisensory features. In: ECCV (2018)
59. Owens, A., Isola, P., McDermott, J., Torralba, A., Adelson, E.H., Freeman, W.T.: Visually indicated sounds. In: CVPR (2016)
60. Owens, A., Wu, J., McDermott, J.H., Freeman, W.T., Torralba, A.: Ambient sound provides supervision for visual learning. In: ECCV (2016)
61. Palossi, D., Loquercio, A., Conti, F., Flamand, E., Scaramuzza, D., Benini, L.: A 64-mw dnn-based visual navigation engine for autonomous nano-drones. IEEE Internet of Things Journal (2019)
62. Pinto, L., Gupta, A.: Supersizing self-supervision: Learning to grasp from 50k tries and 700 robot hours. In: ICRA (2016)
63. Purushwalkam, S., Gupta, A., Kaufman, D.M., Russell, B.: Bounce and learn: Modeling scene dynamics with real-world bounces. In: ICLR (2019)
64. Quattoni, A., Torralba, A.: Recognizing indoor scenes. In: CVPR (2009)
65. Ranjan, A., Jampani, V., Balles, L., Kim, K., Sun, D., Wulff, J., Black, M.J.: Competitive collaboration: Joint unsupervised learning of depth, camera motion, optical flow and motion segmentation. In: CVPR (2019)
66. Ren, Z., Jae Lee, Y.: Cross-domain self-supervised multi-task feature learning using synthetic imagery. In: CVPR (2018)
67. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: MICCAI (2015)
68. Rosenblum, L.D., Gordon, M.S., Jarquin, L.: Echolocating distance by moving and stationary listeners. Ecological Psychology (2000)
69. de Sa, V.R.: Learning classification with unlabeled data. In: NeurIPS (1994)
70. 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)
71. Senocak, A., Oh, T.H., Kim, J., Yang, M.H., So Kweon, I.: Learning to localize sound source in visual scenes. In: CVPR (2018)
72. Silberman, N., Hoiem, D., Kohli, P., Fergus, R.: Indoor segmentation and support inference from rgbd images. In: ECCV (2012)
73. Straub, J., Whelan, T., Ma, L., Chen, Y., Wijmans, E., Green, S., Engel, J.J., Mur-Artal, R., Ren, C., Verma, S., et al.: The replica dataset: A digital replica of indoor spaces. arXiv preprint arXiv:1906.05797 (2019)

74. Stroffregen, T.A., Pittenger, J.B.: Human echolocation as a basic form of perception and action. Ecological psychology (1995)
75. Tian, Y., Shi, J., Li, B., Duan, Z., Xu, C.: Audio-visual event localization in unconstrained videos. In: ECCV (2018)
76. Ummenhofer, B., Zhou, H., Uhrig, J., Mayer, N., Ilg, E., Dosovitskiy, A., Brox, T.: Demon: Depth and motion network for learning monocular stereo. In: CVPR (2017)
77. Vanderelst, D., Holderied, M.W., Peremans, H.: Sensorimotor model of obstacle avoidance in echolocating bats. PLoS computational biology (2015)
78. Vasiljevic, I., Kolkin, N., Zhang, S., Luo, R., Wang, H., Dai, F.Z., Daniele, A.F., Mostajabi, M., Basart, S., Walter, M.R., Shakhnarovich, G.: DIODE: A Dense Indoor and Outdoor DEpth Dataset. arXiv preprint arXiv:1908.00463 (2019)
79. Veach, E., Guibas, L.: Bidirectional estimators for light transport. In: Photorealistic Rendering Techniques (1995)
80. Vijayanarasimhan, S., Ricco, S., Schmid, C., Sukthankar, R., Fragkiadaki, K.: Sfm-net: Learning of structure and motion from video. arXiv preprint arXiv:1704.07804 (2017)
81. Wang, P., Shen, X., Lin, Z., Cohen, S., Price, B., Yuille, A.L.: Towards unified depth and semantic prediction from a single image. In: CVPR (2015)
82. Wang, X., Gupta, A.: Unsupervised learning of visual representations using videos. In: ICCV (2015)
83. Xia, F., Zamir, A.R., He, Z., Sax, A., Malik, J., Savarese, S.: Gibson env: Real-world perception for embodied agents. In: CVPR (2018)
84. Xu, D., Ricci, E., Ouyang, W., Wang, X., Sebe, N.: Multi-scale continuous crfs as sequential deep networks for monocular depth estimation. In: CVPR (2017)
85. Yang, Z., Wang, P., Xu, W., Zhao, L., Nevatia, R.: Unsupervised learning of geometry with edge-aware depth-normal consistency. In: AAAI (2018)
86. Ye, M., Zhang, Y., Yang, R., Manocha, D.: 3d reconstruction in the presence of glasses by acoustic and stereo fusion. In: ICCV (2015)
87. Zamir, A.R., Sax, A., Shen, W., Guibas, L.J., Malik, J., Savarese, S.: Taskonomy: Disentangling task transfer learning. In: CVPR (2018)
88. Zhang, R., Isola, P., Efros, A.A.: Colorful image colorization. In: ECCV (2016)
89. Zhao, H., Gan, C., Rouditchenko, A., Vondrick, C., McDermott, J., Torralba, A.: The sound of pixels. In: ECCV (2018)
90. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: CVPR (2017)
91. Zhou, T., Brown, M., Snavely, N., Lowe, D.G.: Unsupervised learning of depth and ego-motion from video. In: CVPR (2017)
92. Zhou, Y., Wang, Z., Fang, C., Bui, T., Berg, T.L.: Visual to sound: Generating natural sound for videos in the wild. In: CVPR (2018)