PANO-ECHO: PANOramic depth prediction enhancement with ECHO features

Xiaohu Liu*, Amandine Brunetto†, Sascha Hornauer†, Fabien Moutarde† and Jialiang Lu*
*SJTU Paris Elite Institute of Technology, Shanghai Jiao Tong University, Shanghai, China
{liuxiaohu, jialiang.lu}@sjtu.edu.cn
†Center for Robotics, MINES Paris, PSL University, Paris, France
{amandine.brunetto, sascha.hornauer, fabien.moutarde}@minesparis.psl.eu

***Abstract*—Panoramic depth estimation gains importance with more 360° images being widely available. However, traditional mono-to-depth approaches, optimized for a limited field of view, show subpar performance when naively adapted. Methods tailored to process panoramic input improve predictions but can not overcome ambiguous visual information and scale-uncertainty inherent to the task.**

In this paper we show the benefits of leveraging sound for improved panoramic depth estimation. Specifically, we harness audible echoes from emitted *chirps* as they contain rich geometric and material cues about the surrounding environment. We show that these auditory cues can enhance a state-of-the-art panoramic depth prediction framework. By integrating sound information, we improve this vision-only baseline by $\approx 12\%$.

Our approach requires minimal modifications to the underlying architecture, making it easily applicable to other baseline models. We validate its efficacy on the Matterport3D and Replica datasets, demonstrating remarkable improvements in depth estimation accuracy. Our code is available here: https://github.com/peter12398/PANO-ECHO

***Index Terms*—Audio-Visual learning, panoramic depth estimation, multi-modal fusion**

I. Introduction

Our understanding of the world is mediated through multiple senses, each contributing to our overall perception. In particular, sound is more than background noise, it intricately shapes our perception of our environment. From everyday sounds like conversations and microwave *beeps* to the immersive experience of virtual reality, it carries information not just about its source but also about the physical space around us. By listening to a concert recorded in a cathedral we are able to sense the geometric features of the environment without being physically there. Spatial acoustic cues enable us to perceive the distance, direction and semantic of sounds without visual aid, offering crucial insights into our environment.

These cues arise from how sound waves propagate through space: sound spreads in all directions and is transformed by various acoustic phenomena, affected by the layout of rooms, materials and objects.

Recently, sound has proven to be a promising modality on its own but also to enhance traditional vision tasks. Echoes from short linear frequency modulated signals (*chirps*) can be used to predict depth in front of a robot [11], [25], [29], [31]. Monocular to depth methods are enhanced by the addition of sound [12], [13], [32]. The conversation between multiple people can be used in addition to limited visual information to infer a scene occupancy map [33]. Sound can add semantic information to floor plans outside of the field-of-view [34]. When visual sensors fail because of adversarial condition, sound can provide missing information [35].

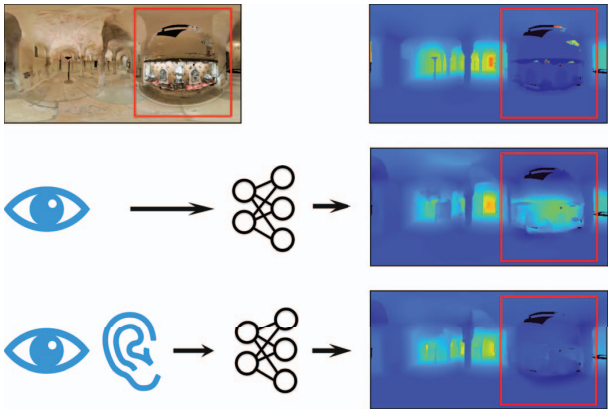

Fig. 1. Monocular depth prediction for indoor panoramas (ground truth in top row) can struggle with ambiguous areas. Visual illusions of objectiveness or openings can lead to systematic prediction errors (middle row). In this example, an illuminated showcase can be misinterpreted as window to a space behind it. Our proposed PANO-ECHO uses sound for improved depth estimation (bottom row) leading to overall better scale and reduced illusions for challenging distorted panoramas.

Depth information is crucial in various fields such as computer vision and robotics. It enhances spatial understanding, providing a three-dimensional perspective. Depth can be obtained using sensors such as an RGB-D camera and LiDAR. However, reflective and transparent surfaces degrade their precision.

Monocular depth estimation methods are low-cost and widely accessible alternatives that are easier to integrate compared to depth sensors, making them advantageous in various applications. Still, extracting depth information from monocular RGB images is challenging due to the inherent lack of direct depth cues in a single 2D image. Unlike stereo vision, which relies on the disparity between two images recorded from offset viewpoints, a monocular image has only one viewpoint, leading to an ill-posed problem due to scale

979-8-3503-5409-6/24/$31.00 ©2024 IEEE
DOI 10.1109/CAI59869.2024.00193

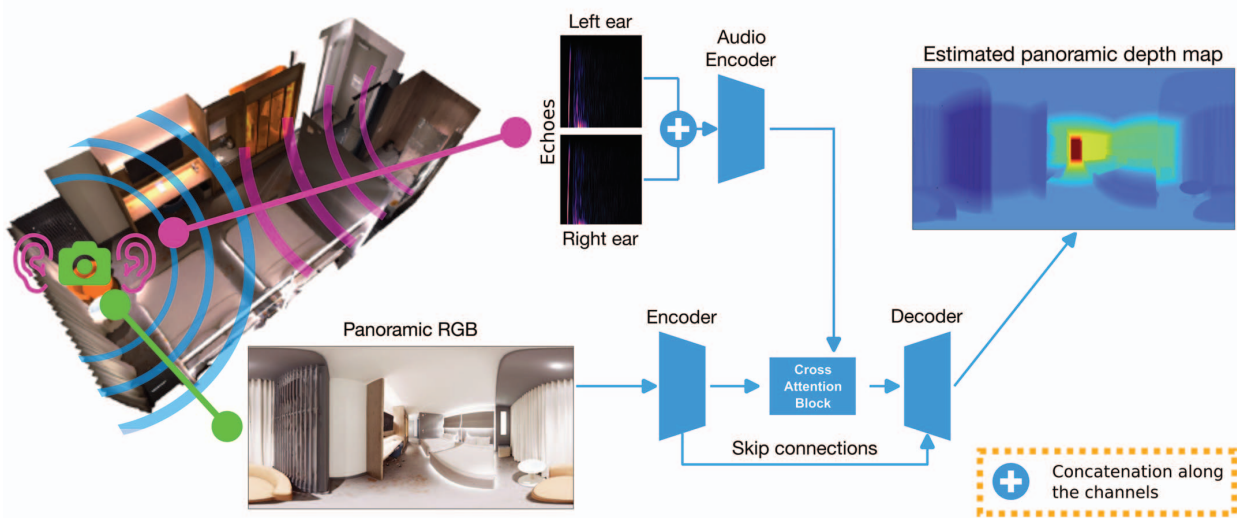

Fig. 2. Our proposed PANO-ECHO framework. Pairs of panoramic equirectangular RGB images and echoes are recorded in simulation at a specific location and orientation. Images enter the Panoformer and are combined with echo features after the *PST Block* in the bottleneck using our cross-attention mechanism. STFT's of Echoes are concatenated along the channel dimension and processed by a Resnet18 to echo features. Finally the Panoformer decoder creates an estimated depth map of the panorama.

ambiguity.

Furthermore, occlusions, shadows and certain scene configurations can create illusions of shapes from context and mislead depth estimation algorithms.

Sound encapsulates depth-like information and is conveniently recordable, given the widespread availability of microphones in most electronic devices. In a video, sound is usually recorded at the same time as the RGB images. Thus, improving monocular to depth algorithms with sound depth cues is a realistic option in many use cases and yields promising results.

We propose to use echoes of a frequency modulated signal, called *chirp*, to enhance a state-of-the-art panoramic depth prediction method: PanoFormer [2]. We use the SoundSpaces [1] simulator to generate 360° equirectangular RGB images and corresponding binaural Room Impulse Responses (RIR) at the ego-location. We convolve them with our generated *chirp* to obtain binaural echoes.

Many vision-based panoramic depth prediction methods [2], [3], [6], [7], [20] adopt an encoder-decoder based architecture. We show that it is possible to further improve their depth estimation and eliminate visual artefacts by fusing sound knowledge in the latent space.

Our contribution can be summarized as follows:

1) We demonstrate that, as sound propagation is 360° and is modified by the room architecture, material and objects, it can provide valuable depth cues, which are not available from a single RGB image.
2) We show how to improve 360° monocular depth prediction by incorporating echoes features in the latent space. We improve the current SOTA baseline "PanoFormer" [2] by $\approx 12\%$ MRE and MAE. Qualitatively, we show this is partly achieved by correctly interpreting ambiguous areas in the image when using audio-visual input.
3) Our method is low cost and easy to set up in real-world systems. Thus, it has the potential to provide accurate depth information without the need of complex and expensive traditional depth sensors. Especially, it solves issues inherent to visual sensors such as transparent surfaces leading to wrong depth information.
4) The proposed method can be easily transferred to any panoramic monocular to depth encoder-decoder architecture. We show its effectiveness in two other methods, Bifuse [6] and Unifuse [3].

II. Related Work

Panoramic Depth Estimation. Estimating 360° depth poses challenges due to panoramic distortions, as highlighted by Zioulis et al. [20]. They found that regular depth models trained on standard images don't perform well on 360° datasets, emphasizing the need for specialized panoramic methods.

One possible solution to tackle this challenge involves designing specialized network structures. Tateno et al. [22] introduced distortion-aware convolution filters, enabling a network trained on regular images to predict panoramic depth without requiring an additional 360° dataset. Similarly, Chen et al. [23] suggested using deformable convolution, strip pooling modules, and a spherical-aware weight matrix to adapt the network to various panoramic distortions. SliceNet [7] segments the scene into vertical slices, leveraging a multi-layer LSTM to capture high-frequency features through long and short-term dependencies between slices. In contrast, PanoFormer [2] addresses equirectangular projection (ERP) distortions by

dividing patches in the spherical tangent domain into tokens, incorporating learnable token flow and a proposed ERP-specific coherence loss. A second solution involves mitigating distortion by combining ERP and cube map projection. While ERP images offer a complete Field of View (FoV) with simultaneous distortion, especially in polar areas, cube map projection provides a narrow FoV without distortion but suffers from boundary discontinuity. To solve different projection drawbacks, Bifuse [6] adopts a two-branch network to leverage both projections, introducing learnable masks for fusion in the encoder and decoder, though at the expense of additional computational overhead. Shen et al. [21] address the dual-cubemap method, rotating the ERP image by 45° and employing two encoder branches with a Boundary Revision module to mitigate boundary discontinuity. Bifuse++ goes a step further by enhancing scalability and computation cost-effectiveness through self-training and Contrast-Aware Photometric Loss [19]. Similarly, Unifuse introduces a novel fusion framework at the decoding stage to combine two projections and addresses the computational complexity of Bifuse [3].

Audio-Visual learning. Sound arrives omnidirectional and provides rich geometric and semantic information about the space it traverses. Certain species, are mastering echolocation, meaning they are capable of navigating and localizing prey using spatial acoustic cues. Drawing inspiration from this phenomenon, various machine learning methods exploit echoes generated by audible short linear frequency signals to get insights about the surrounding space. [1], [13] demonstrate the efficacy of leveraging echoes for the navigation of embodied agents within simulated environments. Echoes emerge as a valuable asset not only for FoV depth estimation as a standalone modality [11], [26], [31] but also as a complementary source alongside vision. This synergy extends to enhancing monocular [12], [13], [27], [32] and stereo [36] depth estimation methods. The versatility of echoes further allows 2D floor plan estimation outside the field-of-view [34], [37] and facilitates sound event detection and localization [24].

In the broader context, the 360° propagation of sound has been leveraged for sound-to-panoramic depth estimation [25], [29], [30]. However, sound-only methods have limited performances compared to their vision-based counterparts.

Motivated by this, we propose a novel approach that integrates echoes with vision to enhance the state-of-the-art in 360° vision-based depth prediction.

III. Approach

We are interested in enhancing 360° depth prediction from monocular RGB images by incorporating geometric-aware echoes features. The objective is to improve depth accuracy and reduce visual artefacts. We introduce a novel approach called PANO-ECHO that leverage complementary information from audio-visual signals while being easy to transfer to other panoramic depth estimation models.

Our method has two main components: 1) a vision-based monocular to panoramic depth encoder-decoder, and 2) a binaural echoes feature extractor. Visual and sound features are fused in the encoder-decoder latent space (fig. 2). Specifically, we choose the current state-of-the-art PanoFormer [2] as the vision-based model, but we show in section IV-C that this audio-visual fusion can be easily extended to other panoramic depth prediction methods such as Unifuse [3] and Bifuse [6].

Next, we describe these two components, the multimodal fusion and training objective.

A. Audio-Visual Features Extractor

We first generate observations in the SoundSpaces simulator [1] where an agent is set at various location of a scene. For each position we generate a binaural RIR and a corresponding 360° equirectangular RGB image at the ego-location.

Acoustic Embedding. To obtain echoes, we convolve generated RIR with a 3ms-long audible *chirp* i.e. a linear frequency sweep signal between 20 Hz and 20 kHz. Thus, we obtain echoes of a *chirp* emitted and recorded at the agent location. We then use the short-time Fourier transform (STFT) to represent all echoes as magnitude STFT $2 \times F \times T$, where F is the number of frequency bins, T is the number of overlapping time windows, and each STFT has 2 channels. We extract the spatial acoustic cues contained in the echoes STFT with a Resnet18 [4] encoder.

Visual Embedding. The corresponding 360° RGB images are resized to resolution 512×256 using bilinear interpolation, standardized and encoded by PanoFormer encoder and bottleneck blocks [2].

B. Multimodal Fusion Module

Most methods fuse visual and acoustic latent vectors along the channel dimension [1], [13], [15]–[17], [27], [28]. Others explore individual bilinear transformation before concatenation followed by attention masks to decide on the contribution of each modality [12] or use a complete cross-modal transformer [14].

[43] develop a cross-modality encoder to identify agreement between textual and visual features. This concept most closely resembles our need to see if visual and sound information agree on depth estimation. Inspired by their work, we propose a cross-modal fusion method to pay attention to suitable features.

Our cross-modal fusion queries distance aware audio features to select propagation of visual features and queries in parallel semantic rich visual features about which audio features to propagate. Finally both are summed to one latent vector, from which the network can not only learn the feature content but also the agreement (see fig. 3).

This can be used by the network to be aware of hard instances. For example, while a transparent window is invisible, a *chirp* will be reflected and provide depth information resulting in audio-visual information mismatch. The network can thereby learn to put its attention to the more reliable modality for each instance conditioned on this mismatch.

The attention matrix, $\mathcal{A}$, can be formulated as follow:

$$\mathcal{A}\left(f^{\mathcal{V}}, f^{\mathcal{S}}\right) = \mathrm{softmax}\left(\frac{Q^{\mathcal{V}}(K^{\mathcal{S}})^{T}}{\sqrt{s}}\right) V^{\mathcal{S}} \qquad (1)$$

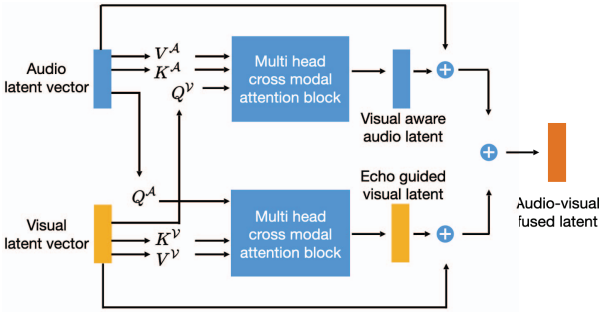

Fig. 3. Illustration of our proposed two-branch cross-modal attention module for audio-visual fusion: In residual connections latent vectors of one modality are added to their respective cross-modal aware vector and finally summed for the final feature.

With Q^m, K^m and V^m respectively the query, the key and the value for the latent sequence of one modality m. The modality can be $\mathcal{S}$ for sound or $\mathcal{V}$ for visual. f^m represents the feature sequence of one modality. The query can be expressed as $Q^{\mathcal{V}} = W_Q^{\mathcal{V}} f^{\mathcal{V}}, K^{\mathcal{S}} = W_K^{\mathcal{S}} f^{\mathcal{S}}, V^{\mathcal{S}} = W_V^{\mathcal{S}} f^{\mathcal{S}}$. s is a scaling factor and set to 512 which equals to dimension of cross-modal attention module.

Symmetrically, we also calculated the echo-guided visual attention map $\mathcal{A}\left(f^{\mathcal{S}}, f^{\mathcal{V}}\right)$, by using audio information as attention key. The goal is to guide the network to pay attention to different part of visual input according to different echoes components. Thus the final audio-visual fusion latent can be presented as:

$$L^{\mathcal{S},\mathcal{V}} = \mathcal{A}\left(f^{\mathcal{V}}, f^{\mathcal{S}}\right) + \mathcal{A}\left(f^{\mathcal{S}}, f^{\mathcal{V}}\right) \tag{2}$$

Following [14], [18], [42] we use pre-normed residual units with dropout for better regularization.

C. Model Training

Our model learns to predict a panoramic depth map using a corresponding pair of echo and 360° equirectangular RGB image during training in a supervised manner.

Following related work in SliceNet [7] and Panoformer [2], we adopt the combination of BerHu loss [8] on depth maps, horizontal and vertical gradients to preserve high frequency details. The BerHu loss between predicted depth map $\hat{D}$ and the ground truth one D can be written as follows:

$$B_c(\hat{D}, D) := \begin{cases} |\hat{D} - D| & |\hat{D} - D| \leq c \\ \frac{(\hat{D}-D)^2 + c^2}{2c} & |\hat{D} - D| > c \end{cases} \tag{3}$$

where c is the parameter controlling switch between $L1$ and $L2$ loss.

Similarly to [6], the loss function exclusively considers pixels with valid depth values in the ground truth map, achieved by applying a mask to both the ground truth and predicted maps.

By denoting respectively the horizontal and vertical gradient operator ∇_x and ∇_y, our final training objective can be formulated as follows:

$$\begin{aligned} L(\hat{D}, D) := B_c(\hat{D}, D) + \\ B_c(\nabla_x(\hat{D}), \nabla_x(D)) + \\ B_c(\nabla_y(\hat{D}), \nabla_y(D)) \end{aligned} \tag{4}$$

We train our model on two RTX3090 GPUs using $c = 0.2$, a learning rate of 10^{-4} and Adam optimizer [5]. Similar to PanoFormer, we use a batch size of 2. Ground truth depth maps are also clipped to 16 m and normalized.

IV. Experiments

A. Dataset

SoundSpaces [1], [38], built upon the Habitat simulator [39], [40], introduces a realistic sound propagation simulation, enhancing its capabilities for generating sound observations at various agent locations across scenes sourced from diverse vision datasets, such as Replica [10] and Matterport3D (MP3D) [9]. In the initial version of SoundSpaces [1], pre-computed binaural Room Impulse Responses (RIRs) and 90° FoV RGB and depth images are provided for these datasets. We generate 360° equirectangular images by using the corresponding simulated sensor in Habitat. The corresponding RIRs are generated with SoundSpaces 1 provided code. We use only the RIRs at the camera positions, where emitter and listener are at the same position. This choice is inspired by echolocation and the desire for our method to seamlessly integrate with easily accessible data, as found in common devices like smartphones with co-located microphones and speakers.

Considering the orientation dependence of binaural microphone shape, we render binaural RIRs for different orientations, specifically $\theta \in [0°, 90°, 180°, 270°]$. Panoramic RGB images are created with the center aligning to the direction faced by the binaural microphone.

Each generated RIR undergoes convolution with a 3 ms-long linear frequency modulated signal spanning 0 Hz to 22.05 kHz for Replica and 0 Hz to 8 kHz for MP3D. The magnitude Short-Time Fourier Transform (STFT) of the resulting echoes is then computed. For Replica and MP3D data, we compute the STFT with an Hann window and distinct parameters— FFT size of 512, window length of 1.4ms, and hop length of 0.3 ms for Replica, and FFT size of 512, window length of 2 ms, and hop length of 0.5 ms for Matterport3D. This results in two-channel STFTs where each channel has 257 frequency bins and respectively 312 and 226 overlapping temporal windows. The parameter differences stem from the distinct sample rates of Replica (44.1 kHz) and MP3D (16 kHz).

In total, we obtain 6,928 audio-visual samples for the Replica dataset and 8,281 for Matterport3D [1]. For Matter-

[1]The total number of rendered RGB-D observation samples for MP3D in SoundSpace 1 is 82810, we decided to randomly sample 10% of the position-orientation pairs with a constant random seed for each scene to be more consistent with previous work in terms of data number. In addition, our experiments with Unifuse showed that training with 10% or 100% of rendered samples had little effect on the final results.

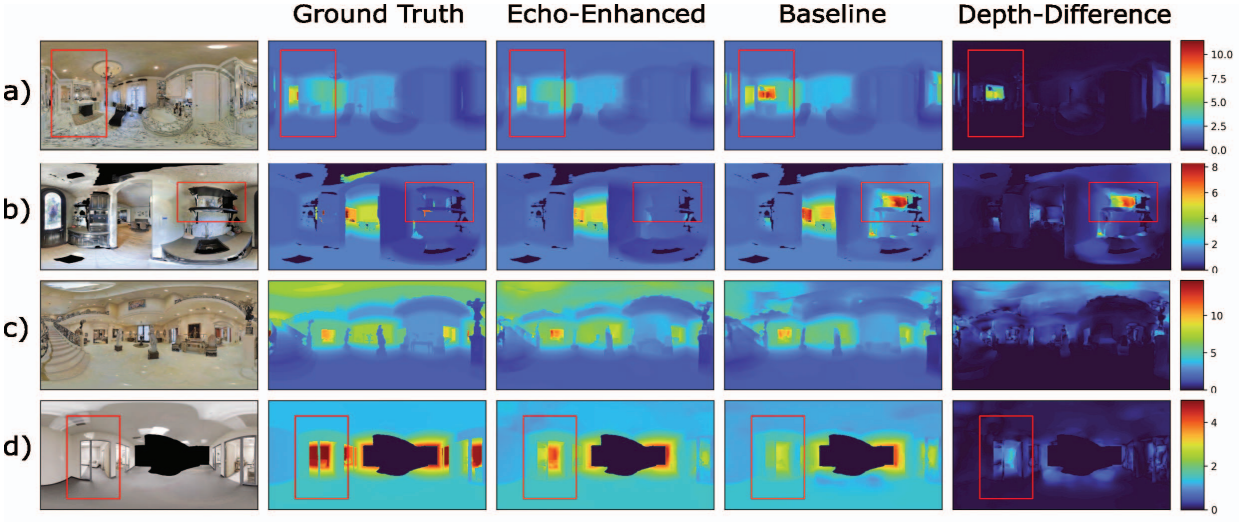

Fig. 4. Issues improved when adding echo features to training of PanoFormer a), Unifuse b), c) and Bifuse d) with the proposed cross-modal fusion strategy. The visual only PanoFormer baseline may misinterpret the mirror in a) as open window while the Unifuse baseline b) may predict a space between two shelves as hollow. Using echoes can disambiguate perception in these local areas. Estimation of the global dimensions of the room is also improved c) which can be based on cues in the reverberations of the echoes. Bifuse is not able to predict if an invisible glass door is open or closed in case d), while echoes reflect distinctively different. The colour bar unit is in metres.

TABLE I

QUANTITATIVE RESULTS USING PANOFORMER, UNIFUSE AND BIFUSE AS BASELINES ON MATTERPORT3D AND REPLICA DATASETS. ECHO *Architecture* DENOTES BASELINE MODEL ENHANCED WITH ECHOES WITH OUR PROPOSED TWO-BRANCH CROSS ATTENTION MODULE. CAT AND SUM MEANS USE CONCATENATION AND SUMMATION AS LATENT FUSION METHOD RESPECTIVELY. METRICS ARE COMPUTED ON UN-NORMALIZED DEPTH. RMSE IS IN METERS.

Dataset	Method	MRE ↓	MAE ↓	**RMSE** ↓	**RMSE(log)** ↓	$\boldsymbol{\delta_1}$ ↑	$\boldsymbol{\delta_2}$ ↑	$\boldsymbol{\delta_3}$ ↑
Matterport3D	PanoFormer	0.2725	0.1193	1.1097	0.0934	0.8287	0.9104	0.9408
	Echo PF (Cross-Att.)	0.2399	0.1057	1.0562	**0.0863**	**0.8549**	**0.9187**	**0.9518**
	Echo PF (cat)	0.2457	0.1106	1.0537	0.0892	0.8450	0.9154	0.9486
	Echo PF (sum)	**0.2398**	**0.1106**	**1.0317**	0.0865	0.8478	0.9178	0.9509
	Unifuse	0.3135	0.1412	1.2114	0.1115	0.7941	0.8945	0.9354
	Echo Unifuse (Cross-Att.)	**0.2728**	**0.1250**	**1.1398**	**0.0989**	**0.8251**	**0.9080**	**0.9505**
	Echo Unifuse (cat)	0.2927	0.1338	1.1804	0.1053	0.8090	0.8986	0.9400
	Echo Unifuse (sum)	0.2888	0.1326	1.1679	0.1045	0.8092	0.9012	0.9406
	Bifuse	0.3330	0.1492	1.2573	0.1147	0.7829	0.8882	0.9303
	Echo Bifuse (Cross-Att.)	**0.2806**	**0.1257**	**1.1585**	**0.0999**	**0.8271**	**0.9093**	**0.9438**
	Echo Bifuse (cat)	0.3061	0.1328	1.2048	0.1063	0.8077	0.8988	0.9351
	Echo Bifuse (sum)	0.2929	0.1364	1.1806	0.1041	0.8068	0.9005	0.9396
Replica	PanoFormer	0.0238	0.0654	0.2063	0.0497	0.9398	0.9838	0.9924
	Echo PF (Cross-Att.)	0.0209	0.0643	0.1969	0.0455	0.9409	0.9876	0.9956
	Echo PF (cat)	0.0211	0.0644	0.1944	0.0458	0.9392	0.9862	0.9949
	Echo PF (sum)	**0.0201**	**0.0630**	**0.1874**	**0.0438**	**0.9416**	**0.9883**	**0.9961**
	Unifuse	0.0295	0.0702	0.2343	0.0549	0.9236	0.9782	0.9915
	Echo Unifuse (Cross-Att.)	0.0279	**0.0660**	**0.2201**	**0.0507**	**0.9299**	0.9813	**0.9926**
	Echo Unifuse (cat)	**0.0273**	0.0664	0.2262	0.0522	0.9295	**0.9815**	0.9921
	Echo Unifuse (sum)	0.0290	0.0694	0.2355	0.0536	0.9236	0.9786	0.9918
	Bifuse	0.0308	0.0833	0.2526	0.0568	0.9142	0.9824	0.9938
	Echo Bifuse (Cross-Att.)	0.0255	0.0677	0.2306	0.0506	0.9339	0.9842	0.9939
	Echo Bifuse (cat)	0.0258	0.0693	0.2304	0.0509	0.9362	0.9850	0.9937
	Echo Bifuse (sum)	**0.0244**	**0.0652**	**0.2196**	**0.0482**	**0.9379**	**0.9857**	**0.9947**

port3D, we use the official split of the SoundSpaces 1 dataset, resulting in 57 scenes for training, 11 for validation and 15 for testing. For Replica, we split the dataset into training, validation, and test sets based on different scenes, resulting in 11 scenes for training, 4 for validation, and 3 for testing. [2]

[2]We split the Replica dataset to balance each room type in the train, validation and test sets as follows: Train: {apartment_0, frl_apartment_{0,1,2,3}, hotel_0, office_{0,1,2}, room_0} Val: {apartment_1, frl_apartment_4, office_3, room_1} Test: {apartment_2, frl_apartment_5, office_4, room_2}

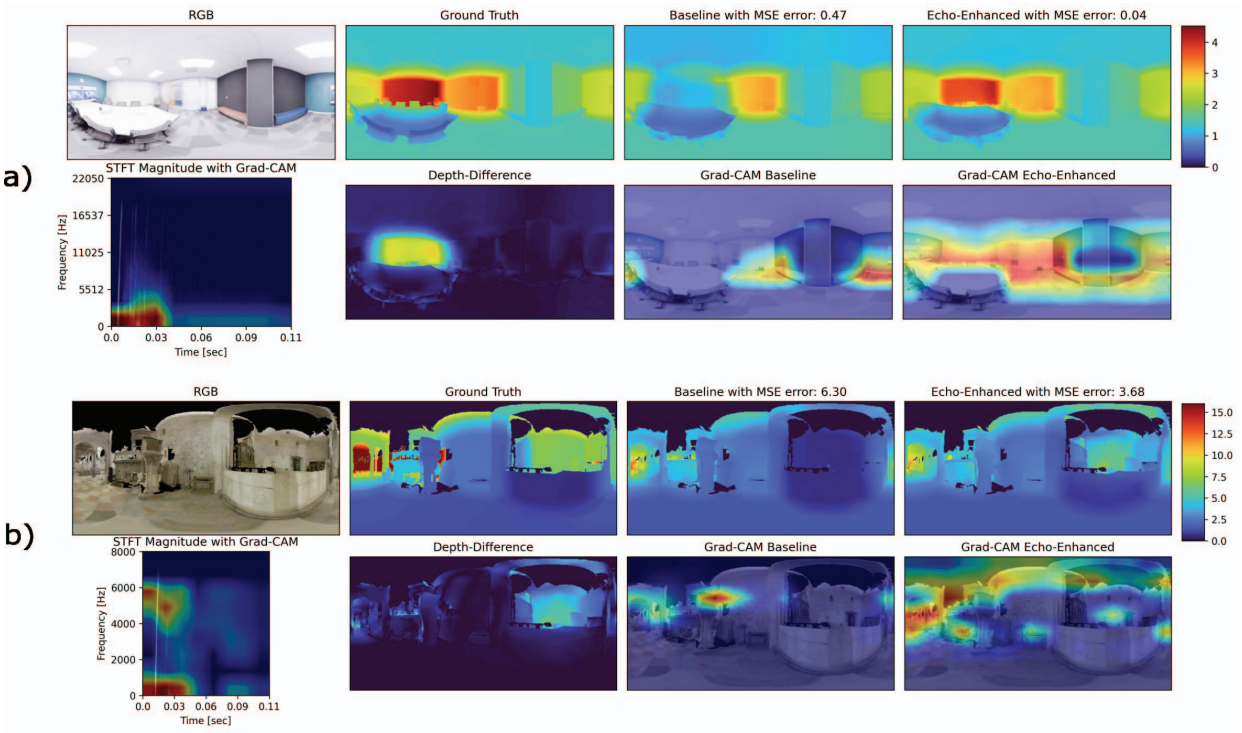

Fig. 5. Grad-CAM visualizations using Unifuse. We investigate network activations in the RGB image and STFT for two cases having large performance gap between the vision-only baseline and the echo-enhanced approach. Shown are the RGB image, the two depth predictions, Gradcam activations on the STFT, depth-prediction difference between predictions and finally Gradcam activations on each RGB image. It shows how our approach estimates distances to walls correctly and how it puts attention on the first echo. The baseline fails with an entire wall, potentially due to low spatial cues. The last two rows investigate a scene with a low, broken wall. The background and the wall are of similar colour and texture but at very different distances. The Grad-CAM [44] activations are harder to interpret but the echo-enhanced approach seems to pay attention on the RGB image and STFT at several locations for this complex scene. It pays attention to the early reflections of the signal and to the reverberant tails, which contain useful cues to distinguish fore- and background. The colour bar unit is in metres.

This framework enables us to assess our task across a multitude of environments, facilitating testing across diverse scene types, method comparison under uniform settings, and the reporting of reproducible results. Notably, there is currently no publicly available real-world dataset that includes panoramic RGB and depth images coupled with binaural RIRs or echoes.

B. Evaluation Metrics

Similar to related works on depth estimation [2], [3], [6], [41] we consider 5 metrics: 1) Mean Absolute Error (MAE), 2) Mean Relative Error (MRE), 3) Root Mean Squared Error (RMSE), 4) RMSE in logarithmic space, and 5) threshold accuracies denoted as δ_1, δ_2, δ_3.

C. Results

With our experimental design, we aim to highlight specific achievements:

1) **Depth Estimation Enhancement**: Our PANO-ECHO method, enriched with echoes, outperforms vision-only state-of-the-art approaches, exemplified by the comparison with its vision-only counterpart, PanoFormer.
2) **Adaptability to Various Architectures**: We establish the versatility of our method by seamlessly integrating it into encoder-decoder and CNN-based panoramic mono-to-depth methods, as shown by its augmentation of PanoFormer, Unifuse and Bifuse.

For the first point we compare PANO-ECHO and PanoFormer to showcase the improved depth estimation achieved by our echo-augmented method. Then, we demonstrate the adaptability of our approach by incorporating echoes into two further CNN-based 360° monocular to depth methods. Our benchmarking involves a comparison with vision-based state-of-the-art methods, specifically Panoformer, Unifuse and Bifuse, providing a comprehensive evaluation of our echo-enhanced methods.

Depth Estimation Enhancement. In Section III, we extend the capabilities of PanoFormer by incorporating echo-features and integrating them within the latent space. Our novel echo-guided method demonstrates a substantial performance improvement compared to the original PanoFormer baseline when trained and evaluated on the same datasets. Notably, we observe more than 11% reduction in both MRE and MAE on

Matterport3D and more than 12% improvement in MRE on Replica. Please find all details in table table I.

Based on a quantitative and qualitative analysis we have several hypothesis what leads to these performance gains.

Including echoes seems to mitigate the generation of artifacts, presumably resulting from visual illusions. In vision-only methods there are unresolvable ambiguities about scale and shapes as with a showcase in case b) and a glass door in case d) of fig. 4. Mirrors as in case a) in fig. 4 can be impossible to distinguish from the scene without additional sensors. Confusion about fore- and background of a scene, as in case b) of fig. 5, due to textures and viewpoint are another source of depth estimation errors. Echoes carry information of distance to obstacles and can therefore disambiguate perception in all of these cases. Moreover, even in the case where the semantic layouts are correctly predicted, the absolute value of distances and spatial dimensions can still be a challenge for vision-only methods, as shown in case c) in fig. 4 and case a) in fig. 5. However, echoes, which inherently contain the absolute distance values and spatial dimensions in the first echoes and reverberations, can help to resolve such scale uncertainties.

Consequently, PANO-ECHO improves prediction of depth values in several ways, even in low-texture areas, such as walls, as depicted in fig. 5.

Adaptibility to Various Architectures. Our proposed echo-guided framework is not limited to improve the performance of PanoFormer, but can be extended to other models with an encoder-decoder structure, commonly used in SOTA models [3], [6], [7], [20].

We showcase and evaluate integration into two other panoramic depth prediction methods, Bifuse and Unifuse [3], [6]. While relative performance of these methods trained on our datasets is consistent with the vision-only case (see table I), integration of our method improves performance for MRE and MAE by an average of $\approx 14\%$ for Unifuse and Bifuse on Matterport3D and $\approx 12\%$ on the Replica dataset.

We also studied alternative fusion methods, shown in table I. Summation and concatenation of visual and audio features performs in some conditions well on PanoFormer and Bifuse. On Unifuse, cross-modal attention is almost always best. While future work could explore the reason and improvements e.g. by combining methods, any fusion method beats the baseline, showing the large and robust contribution of sound.

Limitations. Our approach was tested in simulation with a specific simulated hardware configuration: one binaural microphone directed in one direction, consistent with the camera image. Alternatives such as 4 microphones were not explored. Binaural microphones can be oriented in different ways in a panorama but we have chosen to create data samples from 4 discrete directions. That means the method is orientation dependent, we need to choose an orientation for each panorama to emit the *chirp*.

Furthermore, SoundSpaces 1 provides RIR samples for the MP3D dataset at the sample rate of 16 kHz which is low compared to the *chirp* maximum frequency (20 kHz). Thus it limits the maximum frequency which we can use without aliasing and keeps us from using the same chirp for both datasets. This can be improved in future work by creating one large custom dataset sampled with a constant, and higher rate for every scene.

We use STFT's as input whose maximum length limits the distance at which we can measure depth due to the speed of sound. However, in this work we follow the baseline in clipping the maximum depth therefore the length of the STFT is more than sufficient.

Our results were achieved in simulation and performance in the real-world may differ. We are aware of issues such as that meshes of the simulator contain holes and un-annotated surfaces. Furthermore, material properties are estimated with a high uncertainty. All this impacts the sound simulation. Finally, potential use of ambient sound instead of echoes could make the approach even wider applicable.

V. Conclusions

In this paper, we propose PANO-ECHO, an echo-enhanced method improving 360° mono to panoramic depth estimation models. By taking the current state-of-the-art method, PanoFormer, as a vision-only baseline, experimental results show that the proposed PANO-ECHO improves depth prediction performance on all metrics on two simulated datasets: Matterport3D and Replica. Qualitative results show the effectiveness of echoes in resolving visual illusions and improving overall depth estimation even under heavy distortions. Finally, we demonstrate the easy extensibility of our method to other frameworks by adapting two recent panoramic depth estimation approaches, Unifuse and Bifuse. Thus we show that using echoes is an easy to use and implement addition to mono to panoramic depth estimation pipelines leading to significant positive impacts on performance.

Acknowledgment

This work is supported by the National Natural Science Foundation of China under Grants 62132013. The authors acknowledge the support of the China Scholarship Council at NO.202206230045, and the support of French Agence Nationale de la Recherche (ANR), under grant ANR22-CE94-0003 (projet Omni-BatVision).

References

[1] C. Chen, U. Jain, C. Schissler, S. V. A. Gari, Z. Al-Halah, V. K. Ithapu, P. Robinson, and K. Grauman, "Soundspaces: Audio-visual navigation in 3d environments," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*. Springer, 2020, pp. 17–36.

[2] Z. Shen, C. Lin, K. Liao, L. Nie, Z. Zheng, and Y. Zhao, "Panoformer: Panorama transformer for indoor 360° depth estimation," in *European Conference on Computer Vision*. Springer, 2022, pp. 195–211.

[3] H. Jiang, Z. Sheng, S. Zhu, Z. Dong, and R. Huang, "Unifuse: Unidirectional fusion for 360 panorama depth estimation," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 1519–1526, 2021.

[4] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[5] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[6] F.-E. Wang, Y.-H. Yeh, M. Sun, W.-C. Chiu, and Y.-H. Tsai, "Bifuse: Monocular 360 depth estimation via bi-projection fusion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 462–471.

[7] G. Pintore, M. Agus, E. Almansa, J. Schneider, and E. Gobbetti, "Slicenet: deep dense depth estimation from a single indoor panorama using a slice-based representation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 11 536–11 545.

[8] S. Lambert-Lacroix and L. Zwald, "The adaptive berhu penalty in robust regression," *Journal of Nonparametric Statistics*, vol. 28, no. 3, pp. 487–514, 2016.

[9] A. Chang, A. Dai, T. Funkhouser, M. Halber, M. Niessner, M. Savva, S. Song, A. Zeng, and Y. Zhang, "Matterport3d: Learning from rgb-d data in indoor environments," *arXiv preprint arXiv:1709.06158*, 2017.

[10] J. Straub, T. Whelan, L. Ma, Y. Chen, E. Wijmans, S. Green, J. J. Engel, R. Mur-Artal, C. Ren, S. Verma *et al.*, "The replica dataset: A digital replica of indoor spaces," *arXiv preprint arXiv:1906.05797*, 2019.

[11] J. H. Christensen, S. Hornauer, and X. Y. Stella, "Batvision: Learning to see 3d spatial layout with two ears," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 1581–1587.

[12] K. K. Parida, S. Srivastava, and G. Sharma, "Beyond image to depth: Improving depth prediction using echoes," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 8268–8277.

[13] R. Gao, C. Chen, Z. Al-Halah, C. Schissler, and K. Grauman, "Visualechoes: Spatial image representation learning through echolocation," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*. Springer, 2020

[14] C. Chen, R. Gao, P. Calamia, and K. Grauman, "Visual acoustic matching," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18 858–18 868.

[15] C. Chen, W. Sun, D. Harwath, and K. Grauman, "Learning audio-visual dereverberation," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023

[16] R. Gao and K. Grauman, "Visualvoice: Audio-visual speech separation with cross-modal consistency," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2021

[17] R. Gao and K. Grauman, "2.5 d visual sound," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 324–333.

[18] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu *et al.*, "Conformer: Convolution-augmented transformer for speech recognition," *arXiv preprint arXiv:2005.08100*, 2020.

[19] F.-E. Wang, Y.-H. Yeh, Y.-H. Tsai, W.-C. Chiu, and M. Sun, "Bifuse++: Self-supervised and efficient bi-projection fusion for 360 depth estimation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 5, pp. 5448–5460, 2022.

[20] N. Zioulis, A. Karakottas, D. Zarpalas, and P. Daras, "Omnidepth: Dense depth estimation for indoors spherical panoramas," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 448–465.

[21] Z. Shen, C. Lin, L. Nie, K. Liao, and Y. Zhao, "Distortion-tolerant monocular depth estimation on omnidirectional images using dual-cubemap," in *2021 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2021

[22] K. Tateno, N. Navab, and F. Tombari, "Distortion-aware convolutional filters for dense prediction in panoramic images," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 707–722.

[23] H.-X. Chen, K. Li, Z. Fu, M. Liu, Z. Chen, and Y. Guo, "Distortion-aware monocular depth estimation for omnidirectional images," *IEEE Signal Processing Letters*, vol. 28, pp. 334–338, 2021.

[24] M. Yasuda, Y. Ohishi, and S. Saito, "Echo-aware adaptation of sound event localization and detection in unknown environments," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 226–230.

[25] E. Tracy and N. Kottege, "Catchatter: Acoustic perception for mobile robots," *IEEE Robotics and Automation Letters*, vol. 6, no. 4, pp. 7209–7216, 2021.

[26] G. Irie, T. Shibata, and A. Kimura, "Co-attention-guided bilinear model for echo-based depth estimation," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 4648–4652.

[27] A. Wang, Z. Fang, X. Jiang, Y. Gao, G. Cao, and S. Ma, "Depth estimation of multi-modal scene based on multi-scale modulation," in *2023 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2023

[28] A. Owens and A. A. Efros, "Audio-visual scene analysis with self-supervised multisensory features," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 631–648.

[29] A. B. Vasudevan, D. Dai, and L. Van Gool, "Semantic object prediction and spatial sound super-resolution with binaural sounds," in *European conference on computer vision*. Springer, 2020, pp. 638–655.

[30] D. Dai, A. B. Vasudevan, J. Matas, and L. Van Gool, "Binaural soundnet: predicting semantics, depth and motion with binaural sounds," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 1, pp. 123–136, 2022.

[31] A. Brunetto, S. Hornauer, S. X. Yu, and F. Moutarde, "The audio-visual batvision dataset for research on sight and sound," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 1-8, 2023.

[32] L. Zhu, E. Rahtu, and H. Zhao, "Beyond visual field of view: Perceiving 3d environment with echoes and vision," *arXiv preprint arXiv:2207.01136*, 2022.

[33] S. Majumder, H. Jiang, P. Moulon, E. Henderson, P. Calamia, K. Grauman, and V. K. Ithapu, "Chat2map: Efficient scene mapping from multi-ego conversations," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 10 554–10 564.

[34] S. Purushwalkam, S. V. A. Gari, V. K. Ithapu, C. Schissler, P. Robinson, A. Gupta, and K. Grauman, "Audio-visual floorplan reconstruction," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 1183–1192.

[35] F. R. Valverde, J. V. Hurtado, and A. Valada, "There is more than meets the eye: Self-supervised multi-object detection and tracking with sound by distilling multimodal knowledge," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021

[36] C. Zhang, K. Tian, B. Ni, G. Meng, B. Fan, Z. Zhang, and C. Pan, "Stereo depth estimation with echoes," in *European Conference on Computer Vision*. Springer, 2022, pp. 496–513.

[37] B. Zhou, M. Elbadry, R. Gao, and F. Ye, "Batmapper: Acoustic sensing based indoor floor plan construction using smartphones," in *Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services*, 2017, pp. 42–55.

[38] C. Chen, C. Schissler, S. Garg, P. Kobernik, A. Clegg, P. Calamia, D. Batra, P. W. Robinson, and K. Grauman, "Soundspaces 2.0: A simulation platform for visual-acoustic learning," in *NeurIPS 2022 Datasets and Benchmarks Track*, 2022.

[39] A. Szot, A. Clegg, E. Undersander, E. Wijmans, Y. Zhao, J. Turner, N. Maestre, M. Mukadam, D. Chaplot, O. Maksymets, A. Gokaslan, V. Vondrus, S. Dharur, F. Meier, W. Galuba, A. Chang, Z. Kira, V. Koltun, J. Malik, M. Savva, and D. Batra, "Habitat 2.0: Training home assistants to rearrange their habitat," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[40] Manolis Savva*, Abhishek Kadian*, Oleksandr Maksymets*, Y. Zhao, E. Wijmans, B. Jain, J. Straub, J. Liu, V. Koltun, J. Malik, D. Parikh, and D. Batra, "Habitat: A Platform for Embodied AI Research," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[41] J. Hu, M. Ozay, Y. Zhang, and T. Okatani, "Revisiting single image depth estimation: Toward higher resolution maps with accurate object boundaries," in *2019 IEEE winter conference on applications of computer vision (WACV)*. IEEE, 2019, pp. 1043–1051.

[42] Q. Wang, B. Li, T. Xiao, J. Zhu, C. Li, D. F. Wong, and L. S. Chao, "Learning deep transformer models for machine translation," *arXiv preprint arXiv:1906.01787*, 2019.

[43] P. Li, J. Gu, J. Kuen, V. I. Morariu, H. Zhao, R. Jain, V. Manjunatha, and H. Liu, "Selfdoc: Self-supervised document representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 5652–5660.

[44] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," *International Journal of Computer Vision*, vol. 128, no. 2, p. 336–359, Oct. 2019.

[45] J. Tan, W. Lin, A. X. Chang, and M. Savva, "Mirror3d: Depth refinement for mirror surfaces," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 15 990–15 999