2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

Visual Timing For Sound Source Depth Estimation in the Wild

Wei Sun[1] and Lili Qiu[1]

***Abstract*— Depth estimation enables a wide variety of 3D applications, such as robotics and autonomous driving. Despite significant work on various depth sensors, it is challenging to develop an all-in-one method to meet multiple basic criteria. In this paper, we propose a novel audio-visual learning scheme by integrating semantic features with physical spatial cues to boost monocular depth with only one microphone. Inspired by the flash-to-bang theory, we develop FBDepth, the first passive audio-visual depth estimation framework. It is based on the difference between the time-of-flight (ToF) of the light and the sound. We formulate sound source depth estimation as an audio-visual event localization task for collision events. To approach decimeter-level depth accuracy, we design a coarse-to-fine pipeline to push the temporary localization accuracy from event-level to millisecond-level by aligning audio-visual correspondence and manipulating optical flow. FBDepth feeds the estimated visual timestamp together with the audio clip and objects visual features to regress the source depth. We use a mobile phone to collect 3.6K+ video clips with 24 different objects at up to $65m$. FBDepth shows superior performance especially at a long range compared to monocular and stereo methods.**

I. Introduction

Depth estimation is crucial for 3D perception and manipulation. Despite advancements, current methods lack a balance in accuracy, range, angular resolution, cost, and power consumption as shown in Table I. Active depth estimation methods actively emit sensing signals. They can achieve high accuracy because of the physical fundamental and well-designed modulated sensing signals. Lidar is the most attractive one due to its large sensing range and dense point clouds. However, the density is not sufficient enough to recognize small objects at a long distance. Besides, the prohibitive cost and power consumption limit the availability of most devices. Passive depth estimation takes signals from the environment. RGB camera can achieve pixel-wise density and consume much less energy. However, the effective range and accuracy of stereo depth is limited by the baseline and monocular depth is ill-posed without any physical formulation. Besides, it requires domain adaption and camera calibration for various camera intrinsics [1].

In this paper, we introduce a groundbreaking method to enhance single RGB camera performance by incorporating just one microphone, enabling explicit physical depth measurement. Our approach, named Flash-to-Bang Depth (FBDepth), draws inspiration from the natural phenomenon used to calculate the distance of lightning strikes by measuring the time difference between the visibility of lightning (flash) and the audibility of thunder (bang). This principle capitalizes on the vast speed differential between light and sound, allowing for the perceptible delay in sound to be a reliable depth marker.

Sensor	Device/Method	Acc	Range	Angular Resolution	Power	Cost
LiDAR	Velodyne HDL-32E [2]	2cm	100m	1.33°(V) 0.1°-0.4°(H)	10W	$5K
structured light	Realsense D455	2%	6m	pixel-level	3.5W	$400
ToF camera	Azure Kinect [3]	1cm	6m	pixel-level	5.9W	$600
mmWave	Navtech CTS350-X [2]	4.38cm	163m	1.8°	20w	$500
Ultrasound	Rtrack [4]	2cm	5m	object-level	0.5W	$10
WiFi	Chronos [5]	65 cm	$15m$	object-level	10W	$< $50
RGB camera	NeWCRFs[6]	9.52%[7] 5.20%[8]	10m 80m	pixel-level	1W	$30
stereo camera	ZED 2i[9]	2% 7%	10m 30m	pixel-level	2W	$450
camera + mic	FBDepth	2.98%	60m	obj-level	1W	$30

TABLE I
OVERVIEW OF EXISTING ACTIVE DEPTH AND PASSIVE DEPTH SENSORS.

Applying this concept, FBDepth innovatively estimates the depth of audio-visual events triggered by collisions. The collision event has been explored for navigation and physical search in [10], but our work is the first that uses the collision for depth estimation. While collisions—ranging from a bouncing ball to a musician striking a drum—serve as common triggers, our research is pioneering in utilizing these occurrences for depth estimation. We delve into the unique characteristics of collision events, noting their brief duration and rarity, leading to minimal overlap. Significantly, objects are almost stationary at the moment of collision, despite their subsequent dynamic movement. Additionally, the collision's sound is sufficiently loud to be detected over long distances, providing a robust basis for our depth estimation scheme. This novel approach leverages the inherent properties of collisions to offer a precise and efficient solution for depth estimation challenges.

Our FBDepth method advances beyond the 1-second-segment localization common in prior work[11], [12], [13], achieving precise event-level localization by syncing audio and video data. Unlike previous methods that solely rely on audio-visual semantic features, we utilize optical flow to differentiate between moving and static objects, using changes in optical flow to pinpoint collision moments with frame-level accuracy. We then refine our approach to millisecond-level precision by treating it as an optimization challenge, where we interpolate the most accurate collision moment by analyzing the flow of objects before and after

[1]Department of Computer Science, The University of Texas at Austin

collision. Leveraging visual collision timestamps, FBDepth learns depth from audio clips and visual cues without needing the exact time of the audio collision. This method accounts for the nuanced differences in audio-visual alignment among various objects, such as the distinct sounds produced by rigid versus elastic bodies upon collision. We incorporate semantic features to inform the network about the materials, sizes, and other characteristics of objects, enhancing depth estimation accuracy. The main contributions of this work are: (1) introducing FBDepth, the first passive audio-visual depth estimation method (2) proposing a millisecond-level audio-visual localization technique and (3) demonstrating the effectiveness through comprehensive testing on over 3.6K audio-visual samples from 24 distinct objects.

II. Related Work

Multi-modality Depth Estimation: Integrating cameras with active sensors has proven advantageous in recent depth estimation efforts. Works like [14], [15] have merged sparse Lidar data with imagery to reconstruct dense depth maps, while [16], [17], [18] have enhanced accuracy and range by pairing camera data with Lidar or Radar. However, these approaches often face cost and power consumption challenges. In contrast, [19], [20] utilize audio chirps for depth mapping, though their applicability is mostly confined to echo-rich environments like rooms. FBDepth innovates by employing a single microphone to capture natural sounds, maintaining a passive setup while leveraging physical sound propagation for extended range depth estimation.

Sound Source Localization: Traditional systems localize sound sources using microphone arrays [21] or combining a single microphone with a camera [22], focusing on direction of arrival (DOA) or distance estimation through arrival time differences [4], [23] or visual-semantic matching [13], [24]. Others estimate distance using triangulation from DOAs and room geometry [25], [26], with some studies examining room acoustics and reverberation cues for distance [27], [28]. FBDepth distinguishes itself by directly calculating distance through time-of-flight (ToF), showing enhanced accuracy over indirect and reverberation-based methods.

Audio-visual Event Localization: This field focuses on identifying and pinpointing events within videos. Initiatives like [13] introduced audio-visual event (AVE) datasets, employing audio-guided visual attention to isolate sounding objects or actions. Subsequent research has developed frameworks and mechanisms to exploit dual-modality and attention features [29], [11], though they often deal with coarse temporal event boundaries. FBDepth addresses instant collision events, overcoming boundary issues through a unique collision-focused approach and a coarse-to-fine strategy, achieving millisecond-level temporal resolution without needing exact timestamps.

Comparative works, such as [10], which explores embodied agent navigation towards dropped objects in virtual environments, showcase the integration of asynchronous audio-visual data for navigation. Despite the realism of these simulations, they lack the precision required for real-world, millisecond-level collision analysis. Similarly, datasets focused on falling objects and rapid movements [30], [31], [32] contribute to understanding motion but fall short in audio and depth integration, underscoring FBDepth's novel approach to depth estimation through natural sound analysis.

III. Physical Formulation

We formulate the depth estimation by the physical law of wave propagation, the fundamental equation $\frac{d}{v} - \frac{d}{c} = T$ where the depth of the sound source is d and the difference between the ToF of sound and light is T. c and v denote the propagation speeds of light and sound, respectively. We can estimate d based on $d = \frac{cvT}{c-v} \approx vT$ since $c \gg v$. We observe $T = T_{audio} - T_{video} + T_{hardware}$, where T_{audio} and T_{video} denote the event time in the audio and video recordings, respectively, and $T_{hardware}$ denotes the start time difference in the audio and video recordings. It can be small as well as have a small variance with a well-designed media system such as the Apple AVFoundation framework. We regard it as a constant unknown bias to learn.

It is not feasible to label the precise T_{video} and T_{audio} manually. T_{video} can be tagged at frame-level. Even though many commercial cameras can support up to 240 FPS, it results in a 4-ms segment and 1.43m depth variation. Moreover, it is tough to determine the exact frame that is nearest to the collision in high FPS mode by a human being due to the constrained view of the camera. T_{audio} is challenging to recognize in the wild as well. Although the audio sampling rate is high enough, we can only recognize early peaks instead of the first sample triggered by the collision. The best effort of segmentation is 10-ms level based on real data.

We cannot learn the timestamp with supervision. We propose a 2-stage estimation framework. The goal of the first stage is to estimate the numerical T_{video}. As figure 1 shows, we localize the audio-visual event in the stream and then take advantage of the unique optical flow of the collision to estimate T_{video} at ms-level. In the second stage, we place the T_{video} as an anchor into the audio clip and regress the depth with depth supervision. We make the network optimize T_{audio} automatically with knowledge of the T_{video}, the audio waveform and visual features.

IV. Approach

We demonstrate a novel coarse-to-fine pipeline to localize the collision with a super temporal resolution in the video shown in Figure 1. This method does not require annotations on ms-level, which is at least two orders of magnitude finer than previous approaches. They rely on the supervision of segment annotations, such as AVE dataset with 1-second segments [13], Lip Reading Sentences 2 dataset with word-level segments [33], BOBSL with sentence-level alignments [34].

A. Event-Level Localization

Our work focuses on enhancing audio-visual modeling for collision events, aiming for precise localization with high

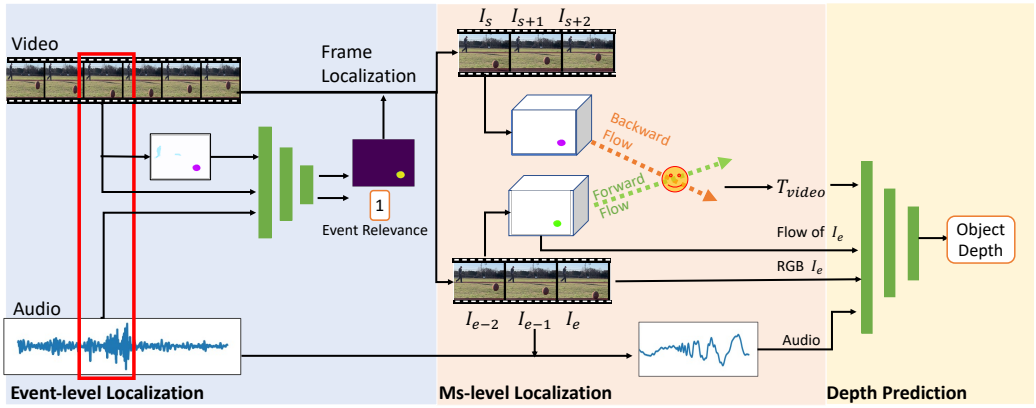


Fig. 1. Model architecture. Our audio-visual depth estimation uses the video, audio, and optical flow to perform the event-level localization to retrieve the collision event. It analyzes the collision flow and estimates the collision timestamp in the video. It uses multiple modalities including RGB, flow, audio, and the timestamp to estimate the depth.

granularity. Unlike previous studies [10], the distinct characteristics of collisions—such as their significant motion and the highly informative nature of their impact sounds relating to object properties like shape and material—facilitate a simpler yet effective cross-matching learning strategy. Our observations indicate that collisions are sparse, with minimal overlap observed in natural settings, exemplified by a study on a basketball court showing rare instances of concurrent collisions within a vast dataset.

We introduce MAVNet, a motion-guided audio-visual correspondence network that enhances cross-matching between audio features and RGB-F channels for precise audio-visual segmentation. This network captures the entire target object's pixels, facilitating a detailed scene understanding. It comprises dual backbones for processing RGB-F visual data and audio signals: a U-Net style encoder extracts frame features conditioned by optical flow, while audio branch employs a 1D convolution layer for initial feature extraction, followed by 2D convolutions for semantic audio feature learning. Features from both modalities are combined, and the network outputs include a binary segmentation mask and a classification of event relevance.

MAVNet's training utilizes a Binary Cross Entropy (BCE) loss, optimizing both segmentation accuracy and event relevance with a weighted sum to balance the contributions of each component. The total loss equation is $_{total} = BCE(M, \hat{M}) + \lambda * BCE(y, \hat{y})$ where λ is the hypermeter to set. During inference, MAVNet processes video and audio streams with low FPS to reduce computational load, activating the segmentation head only when a high match between audio and visual data is detected. This enables the network to efficiently identify and analyze collision events within the audio-visual streams.

B. Frame-level Localization

Given a sequence of video frames, our objective is to categorize them into two groups: frames before the collision $\mathbf{V_0}$ and frames after the collision $\mathbf{V_1}$. This involves identifying the last frame I_e in $\mathbf{V_0}$ immediately before the collision and the first frame I_s in $\mathbf{V_1}$ right after the collision, thus locating the collision between frames I_e and I_s.

Through motion analysis, we observe a significant indicator: a sharp acceleration change at the collision moment, attributed to the impact's impulse force. We define the acceleration at frame I_t as $a_t = v_t - v_{t-1}$, and the change in acceleration as $\delta a_t = a_t - a_{t-1}$. A large δa is noted between I_e and I_s, contrasting with smaller δa values in frames preceding or following the collision. If the object ceases movement post-collision, the subsequent static frame, I_{e+1}, is considered I_s. Frames before I_e are selected to form $\mathbf{V_0}$, and those after I_s compose $\mathbf{V_1}$. In the analysis phase, we leverage a retrieved mask to track object positions within frames, calculating velocity, acceleration, and changes in acceleration. Initially determining I_e and I_s at a lower FPS, we then extend our analysis to the interval between I_e and I_s at a higher FPS for accurate collision localization.

C. Ms-level Localization

To refine the localization of collision moments, we face challenges with traditional frame interpolation, as collisions disrupt typical motion continuity, necessitating the approach to capture the abrupt motion changes caused by impacts.

Motion Consistency: Essential for spatio-temporal analysis, *motion first consistency* describes scenarios where an object's motion remains stable over frames, allowing for predictable future positions and frame interpolation. However, collision-induced impulse forces disrupt this stability, challenging motion prediction and introducing *motion second consistency*. This new consistency type indicates that pre- and post-collision motions converge at a common point while maintaining separate *motion first consistencies*.

Utilizing *motion second consistency*, we extrapolate motion from *motion first consistency* to identify collision timestamps, focusing on the timing rather than motion specifics at the collision point. While [31], [32] also explore sub-frame motion recovery, they rely on high-FPS ground truth for training, whereas our approach prioritizes determining the collision's occurrence over detailed motion analysis.

Optical flow extrapolation Optical flow, crucial for frame prediction and interpolation [35], [36], [37], [38], [39], captures pixel movements to understand object dynamics. Traditional optical flow generation from adjacent frames faces limitations in extrapolation accuracy due to pixel drift and accumulative errors.

To pinpoint collision moments, we propose utilizing optical flow for fine-grained event analysis, overcoming the inaccuracies of center-point trajectory methods which lose critical pixel information and pose estimation. Our approach computes optical flow sequences from an anchor frame I_a across a sequence $= I_0, I_1, ...I_n$, defined as $_{a\rightarrow} = f_{a\rightarrow 0}, f_{a\rightarrow 1}, ...f_{a\rightarrow n}$. Here, $f_{a\rightarrow n}(x, y)$ denotes the pixel $I_a(x, y)$'s displacement to I_n, allowing for global motion tracking without iterative warpings.

For collision analysis, we select k frames preceding ($_{pre} = I_{e-k+1}, ..., I_e$) and following ($_{post} = I_{s+k-1}, ..., I_s$) the collision with I_e as the anchor frame. This choice facilitates accurate optical flow estimation due to minimal motion variance near collision. Optical flow sequences $_{e\rightarrow_{pre}}$ and $_{e\rightarrow_{post}}$ are estimated, and a segmentation mask for I_e isolates target object pixels. Finally, we employ regressors for individual pixel motion to predict future positions in any sub-frame, enabling precise collision timing and depth estimation.

Optical flow interpolation We utilize pixel-Level regressors for optical flow sequences $_{e\rightarrow_{pre}}$ and $_{e\rightarrow_{post}}$ to perform flow extrapolation $f_{e\rightarrow e+\delta t_0}$ and $f_{a\rightarrow s+\delta t_1}$, where δt_0 and δt_1 represent the extrapolation steps. The objective is to minimize the difference between extrapolated flows, optimizing under the constraint that $e + \delta t_0 < s + \delta t_1$, ensuring a positive collision duration $(s + \delta t_1 - (e + \delta t_0) > 0)$. The goal is

$$\min_{e-s\leq\delta t_1\leq 0\leq\delta t_0\leq s-e} ||f_{e\rightarrow e+\delta t_0}, f_{a\rightarrow s+\delta t_1}||_2,$$
$$s.t. e + \delta t_0 < s + \delta t_1$$

It aims for millisecond-level localization ($\hat{T}_{video}$) by leveraging the precision of optical flow, which captures comprehensive pixel movements, unlike methods relying on key points or bounding box intersections that lack granularity.

D. Depth Regression

To refine the depth estimation process, we correlate the estimated video timestamp $\hat{T}_{video}$ with the audio timestamp T_{audio} and hardware bias $T_{Hardware}$, guided by ground truth depth. Recognizing the challenge posed by the variability in sound generation across different objects, materials, shapes, and motions, we enhance the model's ability to discern T_{audio} amidst diverse waveforms and background noise. To this end, we incorporate semantic and motion features from the target object's RGB-F frame at I_e into the depth predictor, aiding in the identification of relevant waveform patterns.

We initiate the audio processing by selecting a sequence beginning from I_e, marking anchor points at $\hat{T}_{video}$ to synchronize the audio with the visual collision timestamp. This sequence is then processed through a 1D convolution layer to obtain a 2D representation, which is further refined by two residual blocks for enhanced feature extraction. Concurrently, ResNet-18 extracts the RGB-F features of the target object, which are then tiled and concatenated with the audio features along the channel dimension. This combination is further processed by two additional residual blocks to integrate the features effectively. The fusion of features proceeds through a pooling layer and a fully connected layer to predict the depth, translating into a depth value through 2D projection. The learning objective is optimized using the Mean Square Error (MSE) formula, $\mathcal{L}_{depth} = ||d, \hat{d}||_2$, where d and $\hat{d}$ represent the target and predicted depths, respectively.

V. IMPLEMENTATION

Multi-Sensor Data Collection: Our data collection leverages a multi-sensor platform, as depicted in Figure 2, using an iPhone XR for its 240-fps slow-motion video and 48 kHz audio capabilities. The device's minimal $T_{hardware}$ variance ensures precise timestamping, crucial for our analysis. Despite the iPhone XR's lack of a telephoto lens, we attached an ARPBEST monocular telescope to enhance remote scene capture, albeit without surpassing the image quality of commercial smartphone lenses. Attempts to utilize Android devices revealed significant audio-video synchronization challenges, reaffirming iOS as our preferred platform due to its reliability.

Sensor Configuration: Our setup includes a ZED 2 stereo camera and a Livox Mid-70 lidar alongside the iPhone. The lidar's role was primarily to measure anchor positions due to its limited efficacy in capturing detailed depth data for smaller, distant objects.

Raw Collision Dataset: Figure 3 showcases the 24 objects utilized, spanning various materials (wood, metal, foam, rubber, plastic, paper) and designed to simulate collision sounds across distances of 2 to 65 meters. This resulted in over 3.6K audio-visual sequences and 280K+ frames, each enriched with a corresponding audio clip. Each sequence ranges from 40 to 120 frames, accompanied by a relevant audio clip. To capture intricate details, we employed a stereo camera for static imagery and lidar technology for precise depth mapping. We partitioned the raw sequences into training, validation, and testing sets with ratios of 80%, 10%, and 10%, respectively. Dataset augmentation was subsequently performed within each of these splits.

AVD Dataset: We then expanded our dataset into a comprehensive audio-visual depth (AVD) dataset, incorporating 10K sequences enriched with multi-collision events. This enhancement involved selectively cropping a moving object from one sequence and integrating it into another, with corresponding audio adjustments to mirror these temporal shifts. In the event-level localization phase, we meticulously segmented 66.7ms audio clips capturing the essence of impact sounds, alongside 20 frames from each sequence showcasing visible objects. These were then categorized as positive or negative pairs, based on the presence or absence of impact sounds, resulting in a robust collection of approximately 400K audio-visual pairs.

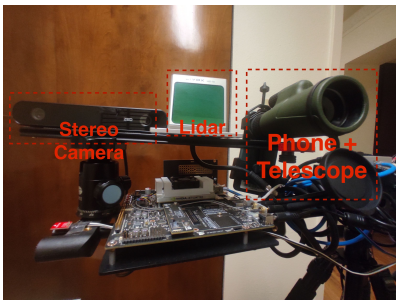


Fig. 2. Data collection platform with multiple sensors

Fig. 3. A set of objects are used for experiments. They cover most common items and materials in daily life

To further refine our dataset, we introduced a variance of up to 3% in the raw depth data to account for the variability in depth perception and aligned the audio samples with the video timestamps, addressing the issue of discrete anchor depths. This modification, crucial for maintaining depth integrity, was complemented by image enhancements (including adjustments for low light conditions, flips, and rotations) and audio diversification with background noise sourced from WHAM! [40], ensuring a rich, versatile dataset suitable for rigorous depth estimation research.

Model Implementation We utilized the RAFT[41] model, pre-trained on both synthetic and real data, to calculate optical flow, demonstrating strong performance on our dataset. For audio feature extraction in MVANet, we adapted a ResNet-18[42] model, pre-trained on our dataset, omitting its final layer to produce 256-dimensional audio embeddings, with weights updated during MVANet training. MVANet employs a U-Net architecture with five down- and up-convolution blocks, altering channel numbers for segmentation mask generation. Additionally, it integrates a convolutional and fully connected layer sequence for event relevance classification. The depth prediction component starts with a 1D convolution layer, transitioning to a 2D representation, followed by Resnet-like modules for both audio and RGB-F inputs, each yielding 1024-dimensional embeddings. These embeddings are then merged and processed through fully connected layers for depth regression.

MVANet and the depth regression network were developed in Python using PyTorch, trained on a system equipped with an Intel® Xeon® Gold 6230 CPU and an NVIDIA Quadro RTX 6000 GPU. Training parameters were set to a learning rate of 0.001 and a batch size of 32.

Method	Input	FPS	AbsErr(m)	AbsRel(%)	RMSE(m)
NeWCRFs	V	-	1.68	9.48	3.49
ZED SDK	S	-	2.03	7.28	3.69
LEAStereo	S	-	0.88	4.09	2.95
FBDepth	A+V	30	0.95	4.26	1.51
FBDepth	A+V	60	0.72	3.34	1.27
FBDepth	A+V	120	0.67	3.11	1.09
FBDepth	A+V	240	**0.64**	**2.98**	**1.03**

TABLE II

V, S, A REPRESENT VISUAL, STEREO, AUDIO RESPECTIVELY.

VI. EVALUATION

Baselines We include three types of baseline for comparison. We compare to a monocular depth estimation method NeWCRFs [6], a state-of-the-art(SOTA) on multiple benchmarks. We also compare to stereo matching methods including the ZED built-in ultra depth estimation SDK and a SOTA method LEAStereo [43]. We use dense depth maps collected by the lidar to finetune NeWCRFs. Stereo methods do not need finetune. Despite optical flow-based interpolation, we compare to interpolation using key points such as the trajectories of center or bounding boxes.

Metrics We use the mean absolute depth errors as $AbsErr = \frac{1}{n}\sum_{i=1}^{n}|d-\hat{d}|$, root mean square absolute relative errors $RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(d-\hat{d})^2}$, and absolute relative error $AbsRel = \frac{1}{n}\sum_{i=1}^{n}\frac{|d-\hat{d}|}{d}$ as the end-to-end performance metrics.

FFBDepth focuses on estimating depth for colliding objects sparsely, whereas monocular and stereo methods provide dense depth estimations for all object pixels. To ensure fairness in comparison, we use the median depth from the dense depth maps estimated by these methods and compare it to the median depth from ground-truth dense maps collected via lidar across different distances: close ($\leq$ $10m$), mid ($10m$-$30m$), and far ($\geq$ $30m$). Due to limitations such as hardware timestamp variance and restricted FPS, there's an inherent limit to temporal resolution, leading to a minimum achievable $AbsErr$. This limitation impacts $AbsRel$ more at closer distances, but its effect diminishes with increasing distance. Our goal is to identify the optimal range.

A. Results

In the depth estimation results (Table III), FBDepth showcases superior accuracy across varying distances and FPS rates when compared to both NeWCRFs and stereo-based methods like ZED SDK and LEAStereo. Notably, stereo matching with LEAStereo achieves the lowest absolute error (AbsErr) at close range (0.07m), but FBDepth at 240 FPS outperforms in mid (0.58m) and far distances (0.95m), highlighting its effectiveness over greater distances with AbsRel percentages improving to 7.25% (close), 3.55% (mid), and 2.41% (far). The RMSE metric further underscores FBDepth's robust performance, with values improving

Method	Input	FPS	AbsErr(m) close/mid/far	AbsRel(%) close/mid/far	RMSE(m) close/mid/far
NeWCRFs	V	-	0.553/1.09/3.27	11.1/6.74/8.64	0.895/1.51/5.82
ZED SDK	S	-	0.083/0.96/5.10	1.78/6.05/12.7	0.108/1.07/6.30
LEAStereo	S	-	**0.067**/0.66/2.47	1.48/4.24/5.98	**0.083**/0.76/5.08
FBDepth	A+V	30	0.485/0.83/1.33	10.9/5.20/3.32	0.731/1.01/2.29
FBDepth	A+V	60	0.418/0.70/1.11	8.94/4.33/2.79	0.597/0.83/1.86
FBDepth	A+V	120	0.392/0.61/0.98	8.42/3.79/2.49	0.534/0.75/1.68
FBDepth	A+V	240	0.337/**0.58**/**0.95**	7.25/**3.55**/**2.41**	0.476/**0.69**/**1.61**

TABLE III

A DETAILED COMPARISON OF HOW DIFFERENT DEPTH ESTIMATION APPROACHES PERFORM AT VARIOUS DISTANCES

significantly at 240 FPS to 0.48m (close), 0.69m (mid), and 1.61m (far).

In contrast, NeWCRFs and ZED SDK, despite their strong performance in certain areas, demonstrate limitations with increased distance, as evidenced by their higher AbsErr and RMSE values, particularly in the far range with NeWCRFs reaching up to 3.27m AbsErr and 5.82m RMSE, and ZED SDK at 5.10m AbsErr and 6.30m RMSE.

FBDepth's consistent improvement across metrics with higher FPS—specifically, a notable decrease in AbsErr and RMSE from 30 FPS to 240 FPS—emphasizes the importance of temporal resolution in depth estimation. At 240 FPS, FBDepth not only reduces AbsErr and RMSE but also lowers AbsRel, showcasing its capability to accurately estimate depth across all distances. This performance is particularly significant given the method's application of both audio and visual inputs (A+V), which distinguishes it from the solely visual (V) or stereo (S) input methods of the baselines, proving its effectiveness and versatility in depth estimation tasks.

B. Ablation Study

Event-level localization Our study assesses optical flow's utility in collision detection and object contouring (Table IV). We use recall and precision, defined by correctly recognized events with an Intersection over Union (IoU) above 0.5, to gauge performance. Optical flow acts as a preliminary mask, enhancing both metrics. Main recall challenges arise from subdued collision sounds or concurrent collision occurrences, while misrecognition typically stems from frame-confined similar objects or indistinct events.

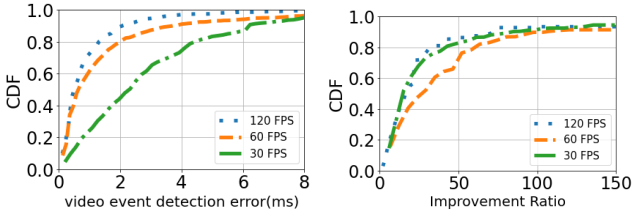
Frame-level localization Frame rate significantly influences error metrics, with a notable decrease observed when increasing from 30 FPS to 60 FPS, as detailed in Table II. This improvement diminishes at higher rates, with 30 FPS proving inadequate for capturing rapid movements, and 60 FPS marking an optimal threshold for discerning collision events. It is consistent with the trend to set 60 FPS as the default video recording and playing. The motion in 120/240 FPS is even slower so it is more difficult to distinguish the frame I_e. The frame error is no more than the one in the low FPS mode. Thus, 120/240 FPS brings less improvement.

Ms-level localization Our investigation into the specialized interpolation technique is conducted from two angles:

Method	AbsErr(m)	AbsRel(%)	RMSE(m)	Recall	Precision
event loc w/o flow	-	-	-	87.3	93.7
FBDepth w/o interp	1.95	9.16	4.07	-	-
FBDepth w/ center	1.39	6.57	2.31	-	-
FBDepth w/ bbox	1.23	5.65	2.06	-	-
FBDepth w/o RGB-F	0.92	4.25	1.42	-	-
FBDepth	**0.64**	**2.98**	**1.03**	94.5	98.7

TABLE IV

ABLATION STUDY FOR FBDEPTH USING DIFFERENT SETUPS AT EACH STAGE. THE INPUT IS 240 FPS.



(a) Temporal error of low FPS estimation compared to 240 FPS (b) Improvement ratio of temporal resolution

Fig. 4. Effectiveness of the video event detection in the second stage

Verification of Interpolation Accuracy: Given the absence of ground truth timestamps, we establish 240 FPS estimations as our benchmark and compare lower FPS estimates against this standard to evaluate the interpolation's effectiveness. High FPS settings are presumed to more accurately capture collision moments within shorter frame durations, reducing temporal error. Our analysis, as shown in Figure 4, records median temporal errors of 2.3ms, 0.65ms, and 0.5ms for 30, 60, and 120 FPS, respectively. Notably, the 60 FPS setting achieves a 25-fold improvement in temporal resolution, underscoring the method's reliability and robustness for millisecond-level localization.

Depth Estimation Performance with Various Interpolation Approaches: Table IV examines how interpolation methods influence depth estimation accuracy. Optical flow-based interpolation emerges as superior, enabling more precise timestamp estimation essential for depth calculation. Traditional interpolation methods, such as those relying on object centers or bounding boxes, fall short due to their reliance on limited keypoints, failing to capture the object's detailed dynamics. While center-based interpolation tracks an object's geometric center, it overlooks significant 3D motion changes. Similarly, bounding box approaches provide a broader motion context but still cannot accurately represent complex movements like rotations, offering only a rudimentary approximation of 3D motion based on 2D translations.

Depth regression Excluding RGB-F channel data in depth regression increases error due to ambient and background noises (Table IV). Depth estimation accuracy varies across materials, with wood exhibiting the lowest median AbsErr (0.47m) due to its loud collision sound and minimal spin, contrasting with foam's higher error (0.81m) due to its softer impact sounds and flexibility.

Impact of object materials Material properties significantly

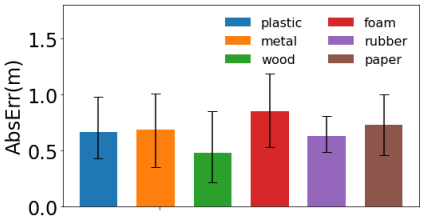
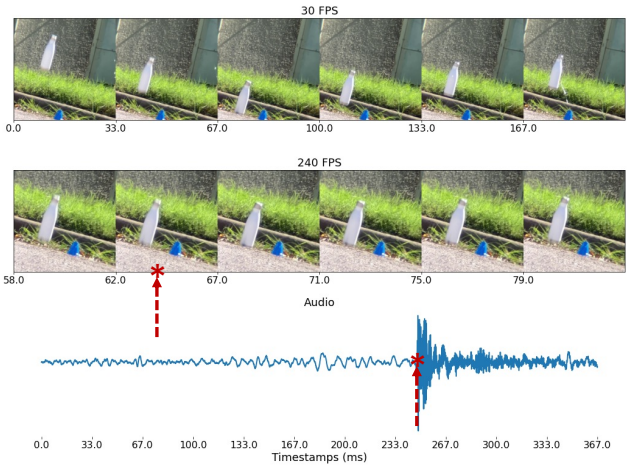
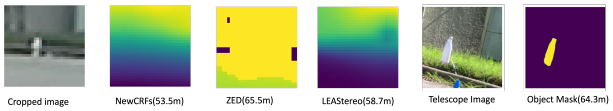


Fig. 5. Depth estimation error across different materials

affect collision dynamics, influencing both visual and auditory signatures. Our dataset covers six materials, with performance metrics per material shown in Figure 5. Metal and rubber exhibit distinct behaviors due to their physical properties, yet FBDepth adeptly manages these variations.



(a) Slow motion and the corresponding waveform



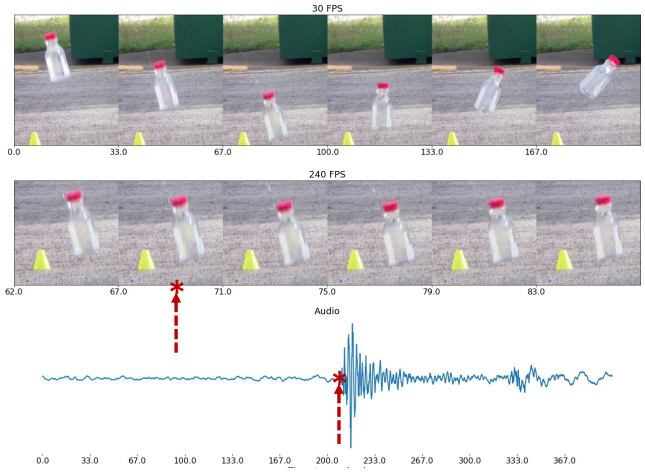
(b) Qualitative results

Fig. 6. A collision event at 63.2m

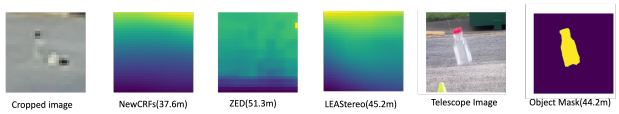
C. Qualitative Results

Our qualitative analysis showcases FBDepth's enhanced performance against traditional depth estimation techniques. In Figure 7, we present comparative depth estimation visuals from various methods. Specifically, Figure 7(b) highlights the region of interest (ROI) for the target event alongside ROIs from the raw RGB image and depth maps predicted by NewCRFs, ZED Ultra mode, and LEAStereo.

Key observations indicate that in the raw image, the target object appears too small and indistinct, making depth interpretation particularly challenging at greater distances. Both monocular and stereo methods struggle to accurately estimate depth from such limited pixel data. Contrastingly, the telescope view and the estimated segmentation mask, depicted in the latter part of Figure 7(b), significantly enhance object visibility. FBDepth, designed for sparse depth estimation, does not produce a dense depth map for straight-



(a) Slow motion and the corresponding waveform



(b) Qualitative results

Fig. 7. A collision event at 43.4m

forward comparison. Instead, Figure 7(a) integrates audio and visual timelines, illustrating object motion at 30 fps and in slow motion at 240 fps for detailed collision analysis, alongside the corresponding audio waveform. The anchor timestamp is set at the first frame of 30 fps, with collision timestamps labeled on the 240 fps video timeline and audio waveform, facilitating direct depth estimation and intuitive depth understanding through observed delays.

In essence, FBDepth leverages physical principles for authentic depth perception, offering a significant advantage over camera-based methods, which often rely heavily on training data and specific camera configurations, limiting their depth access and performance accuracy.

VII. Conclusion

Our research introduces FBDepth, an innovative depth estimation method that utilizes the "Flash-to-Bang" principle. It learns to synchronize video and audio data to identify events, facilitating depth estimation in uncalibrated, natural environments. Through extensive testing, we've shown that FBDepth consistently delivers accurate depth measurements across various distances, outperforming traditional methods that falter as distance increases.

This method is particularly effective for scenarios that combine visible actions with audible sounds, such as in sports analytics or human motion studies. By gathering sparse depth points over time, FBDepth enhances monocular depth estimation, advancing beyond the limitations of current depth completion studies.

Moving forward, our goals are to further improve FBDepth's precision, broaden its application to a wider range of scenarios, and investigate its capacity for depth estimation of additional objects within scenes. Such advancements promise

to make FBDepth a more comprehensive and accurate tool for depth sensing in diverse environments.

REFERENCES

[1] Z. Li, Z. Chen, A. Li, L. Fang, Q. Jiang, X. Liu, and J. Jiang, "Unsupervised domain adaptation for monocular 3d object detection via self-training," *arXiv preprint arXiv:2204.11590*, 2022.

[2] D. Barnes, M. Gadd, P. Murcutt, P. Newman, and I. Posner, "The oxford radar robotcar dataset: A radar extension to the oxford robotcar dataset," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 6433–6438.

[3] Z. Zhang, "Microsoft kinect sensor and its effect," *IEEE multimedia*, vol. 19, no. 2, pp. 4–10, 2012.

[4] W. Mao, M. Wang, W. Sun, L. Qiu, S. Pradhan, and Y.-C. Chen, "Rnn-based room scale hand motion tracking," in *The 25th Annual International Conference on Mobile Computing and Networking*, 2019, pp. 1–16.

[5] D. Vasisht, S. Kumar, and D. Katabi, "{Decimeter-Level} localization with a single {WiFi} access point," in *13th USENIX symposium on networked systems design and implementation (NSDI 16)*, 2016, pp. 165–178.

[6] W. Yuan, X. Gu, Z. Dai, S. Zhu, and P. Tan, "Newcrfs: Neural window fully-connected crfs for monocular depth estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022.

[7] P. K. Nathan Silberman, Derek Hoiem and R. Fergus, "Indoor segmentation and support inference from rgbd images," in *ECCV*, 2012.

[8] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *The International Journal of Robotics Research*, vol. 32, no. 11, pp. 1231–1237, 2013.

[9] StereoLab, "Zed 2 camera dataset," https://www.stereolabs.com/assets/datasheets/zed2-camera-datasheet.pdf, 2021.

[10] C. Gan, Y. Gu, S. Zhou, J. Schwartz, S. Alter, J. Traer, D. Gutfreund, J. B. Tenenbaum, J. H. McDermott, and A. Torralba, "Finding fallen objects via asynchronous audio-visual integration," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10 523–10 533.

[11] Y. Wu, L. Zhu, Y. Yan, and Y. Yang, "Dual attention matching for audio-visual event localization," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 6292–6300.

[12] Y. Xia and Z. Zhao, "Cross-modal background suppression for audio-visual event localization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 19 989–19 998.

[13] Y. Tian, J. Shi, B. Li, Z. Duan, and C. Xu, "Audio-visual event localization in unconstrained videos," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 247–263.

[14] J. Qiu, Z. Cui, Y. Zhang, X. Zhang, S. Liu, B. Zeng, and M. Pollefeys, "Deeplidar: Deep surface normal guided depth prediction for outdoor scene from sparse lidar data and single color image," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3313–3322.

[15] S. Imran, X. Liu, and D. Morris, "Depth completion with twin surface extrapolation at occlusion boundaries," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 2583–2592.

[16] Y. Long, D. Morris, X. Liu, M. Castro, P. Chakravarty, and P. Narayanan, "Radar-camera pixel depth association for depth completion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 12 507–12 516.

[17] K. Zhang, J. Xie, N. Snavely, and Q. Chen, "Depth sensing beyond lidar range," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 1692–1700.

[18] D. Zhang, A. Prabhakara, S. Munir, A. Sankaranarayanan, and S. Kumar, "A hybrid mmwave and camera system for long-range depth imaging," *arXiv preprint arXiv:2106.07856*, 2021.

[19] R. Gao, C. Chen, Z. Al-Halah, C. Schissler, and K. Grauman, "Visualechoes: Spatial image representation learning through echolocation," in *European Conference on Computer Vision*. Springer, 2020, pp. 658–676.

[20] 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.

[21] J.-M. Valin, F. Michaud, J. Rouat, and D. Létourneau, "Robust sound source localization using a microphone array on a mobile robot," in *Proceedings 2003 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2003)(Cat. No. 03CH37453)*, vol. 2. IEEE, 2003, pp. 1228–1233.

[22] J. Hershey and J. Movellan, "Audio vision: Using audio-visual synchrony to locate sounds," *Advances in neural information processing systems*, vol. 12, 1999.

[23] W. Sun, M. Wang, and L. Qiu, "Spatial aware multi-task learning based speech separation," *arXiv preprint arXiv:2207.10229*, 2022.

[24] R. Arandjelovic and A. Zisserman, "Objects that sound," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 435–451.

[25] M. Wang, W. Sun, and L. Qiu, "{MAVL}: Multiresolution analysis of voice localization," in *18th USENIX Symposium on Networked Systems Design and Implementation (NSDI 21)*, 2021, pp. 845–858.

[26] S. Shen, D. Chen, Y.-L. Wei, Z. Yang, and R. R. Choudhury, "Voice localization using nearby wall reflections," in *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking*, 2020, pp. 1–14.

[27] N. Singh, J. Mentch, J. Ng, M. Beveridge, and I. Drori, "Image2reverb: Cross-modal reverb impulse response synthesis," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 286–295.

[28] C. Chen, W. Sun, D. Harwath, and K. Grauman, "Learning audio-visual dereverberation," *arXiv preprint arXiv:2106.07732*, 2021.

[29] Y.-B. Lin, Y.-J. Li, and Y.-C. F. Wang, "Dual-modality seq2seq network for audio-visual event localization," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 2002–2006.

[30] J. Kotera, J. Matas, and F. Šroubek, "Restoration of fast moving objects," *IEEE Transactions on Image Processing*, vol. 29, pp. 8577–8589, 2020.

[31] J. Kotera, D. Rozumnyi, F. Sroubek, and J. Matas, "Intra-frame object tracking by deblatting," in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2019, pp. 0–0.

[32] D. Rozumnyi, J. Kotera, F. Sroubek, and J. Matas, "Sub-frame appearance and 6d pose estimation of fast moving objects," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 6778–6786.

[33] J. S. Chung and A. Zisserman, "Lip reading in the wild," in *Asian conference on computer vision*. Springer, 2016, pp. 87–103.

[34] H. Bull, T. Afouras, G. Varol, S. Albanie, L. Momeni, and A. Zisserman, "Aligning subtitles in sign language videos," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 11 552–11 561.

[35] Y. Wang, L. Jiang, M.-H. Yang, L.-J. Li, M. Long, and L. Fei-Fei, "Eidetic 3d lstm: A model for video prediction and beyond," in *International conference on learning representations*, 2018.

[36] J. Zhang, Y. Wang, M. Long, W. Jianmin, and S. Y. Philip, "Z-order recurrent neural networks for video prediction," in *2019 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2019, pp. 230–235.

[37] S. Meyer, A. Djelouah, B. McWilliams, A. Sorkine-Hornung, M. Gross, and C. Schroers, "Phasenet for video frame interpolation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 498–507.

[38] S. Niklaus, L. Mai, and F. Liu, "Video frame interpolation via adaptive separable convolution," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 261–270.

[39] X. Xu, L. Siyao, W. Sun, Q. Yin, and M.-H. Yang, "Quadratic video interpolation," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[40] G. Wichern, J. Antognini, M. Flynn, L. R. Zhu, E. McQuinn, D. Crow, E. Manilow, and J. L. Roux, "Wham!: Extending speech separation to noisy environments," *arXiv preprint arXiv:1907.01160*, 2019.

[41] Z. Teed and J. Deng, "Raft: Recurrent all-pairs field transforms for optical flow," in *European conference on computer vision*. Springer, 2020, pp. 402–419.

[42] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *arXiv preprint arXiv:1512.03385*, 2015.

[43] X. Cheng, Y. Zhong, M. Harandi, Y. Dai, X. Chang, H. Li, T. Drummond, and Z. Ge, "Hierarchical neural architecture search for deep stereo matching," *Advances in Neural Information Processing Systems*, vol. 33, 2020.