EchoDiffusion: Waveform Conditioned Diffusion Models for Echo-Based Depth Estimation

Wenjie Zhang[1,2,3], Jun Yin[1], Long Ma[1], Peng Yu[1], Xiaoheng Jiang[1,2,3], Zhen Tian[1,2,3*], Mingliang Xu[1,2,3*]

[1]School of Computer Science and Artificial Intelligence, Zhengzhou University, Zhengzhou, 450001, China
[2]Engineering Research Center of Intelligent Swarm Systems, Ministry of Education, Zhengzhou, 450001, China
[3] National Supercomputing Center in Zhengzhou, Zhengzhou, 450001, China
{wjzhang, jiangxiaoheng, ieztian, iexumingliang}@zzu.edu.cn, {yinjun, malong, pengyuu}@gs.zzu.edu.cn

Abstract

To extract spatial information, depth estimation using conventional echo-based methods typically employs models with encoder-decoder architectures, such as UNet. However, these methods may face challenges in extracting fine details from echo waveforms and handling multi-scale feature extraction with high precision. To address these challenges, we introduce EchoDiffusion, a framework that incorporates diffusion models conditioned on waveform embeddings for echo-based depth estimation. This framework employs the Multi-Scale Adaptive Latent Feature Network (MALF-Net) to extract multi-scale spatial features and perform adaptive fusion, encoding the echo spectrograms into the latent space. Additionally, we propose the Echo Waveform Detail Embedder (EWDE), which leverages a pre-trained Wav2Vec model to extract detailed spatial information from echo waveforms, using these details as conditional inputs to guide the reverse diffusion process in the latent space. By embedding the echo waveforms into the reverse diffusion process, we can more accurately guide the generation of depth maps. Our extensive evaluations on the Replica and Matterport3D datasets demonstrate that EchoDiffusion establishes new benchmarks for state-of-the-art performance in echo-based depth estimation.

Code — https://github.com/wjzhang-ai/EchoDiffusion

Introduction

Depth estimation has become a critical task in computer vision, enabling applications such as autonomous driving (Hu et al. 2024; Zheng et al. 2024), attitude estimation (Ren et al. 2023; Kawai et al. 2023), and 3D object detection (Li et al. 2023). However, depth estimation faces challenges across various environments, prompting research into multiple modalities. Conventional methods, including image-based techniques (Zhang et al. 2023; Wang et al. 2024), infrared sensors (Shimada et al. 2022), and LiDAR (Shao et al. 2023; Singh et al. 2023), each offer distinct advantages. Image-based methods excel at capturing detailed information in well-lit conditions. LiDAR provides high-resolution 3D data, while infrared sensors are effective for detecting depth information at close range. Echo-based approaches, inspired by echolocation (Senocak et al. 2019), offer an alternative by emitting sound waves and analyzing their reflections to infer depth. Echo-based methods receive the transmitted sound signals through the air medium to perceive spatial information, are not affected by light, and provide a relatively wide field of view. Given these advantages, many studies focus on using echoes for depth estimation.

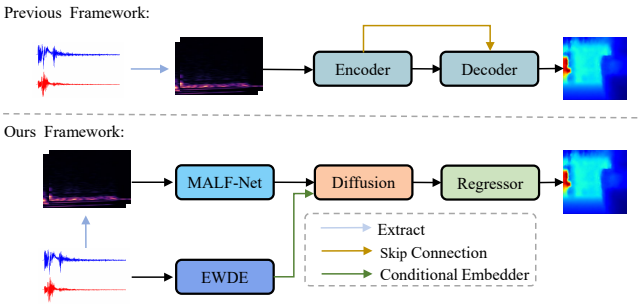

Figure 1: Previous Framework v.s. Our Framework. The conventional training model uses echo spectrograms as input, extracts features through an encoder, and then directly predicts depth information via a decoder. In contrast, our approach incorporates the echo waveforms as an additional input to guide the diffusion process of the echo spectrograms, with the depth map being generated by a regressor.

To capture spatial features from echo spectrograms, research on echo-based depth estimation often employs UNet-like architectures. The UNet architecture effectively leverages its encoder-decoder structure to extract relevant contextual features from the echo spectrograms and accurately reconstruct depth maps. In Christensen et al. (2020), the authors used the UNet network to perform depth estimation from echo spectrograms and initially explored the possibility of direct depth estimation using echo waveforms. Vasudevan et al. (2020) utilized a modified UNet architecture without skip connections, applying Atrous Spatial Pyramid Pooling (ASPP) (Chen et al. 2017) during the encoder phase to extract and fuse multi-scale information from echo spectrograms. Irie, Shibata et al. (2022) introduced a co-attentive model that uses echo spectrograms and angular spectra as in-

*Corresponding Authors

puts, with fusion occurring at the encoding stage to capture more detail for depth map prediction. Parida et al. (2021) adapted the encoder structure in their Echo-Net to better accommodate the time-frequency distribution of echo spectrograms. Additionally, Brunetto et al. (2023) utilized a deeper UNet architecture to capture richer multi-scale information. Although previous studies have made significant progress, depth estimation based on echo still has considerable room for improvement due to inherent limitations in resolution.

To further improve the performance of echo-based depth estimation, we draw upon the strengths of diffusion models, which have demonstrated remarkable capabilities in modeling complex data distributions. As researched by Ho, Jain et al. (2020), diffusion models excel at generating high-fidelity outputs with rich detail preservation. They have been successfully applied to monocular depth estimation (Duan, Guo et al. 2023; Ke et al. 2024; Patni, Agarwal et al. 2024), utilizing sophisticated feature extraction mechanisms to capture intricate spatial details and iteratively refine depth information through denoising processes. In traditional depth estimation, an image is a two-dimensional projection where each pixel directly corresponds to a spatial position in the scene. This allows for straight forward application of diffusion models. However, echo spectrograms represent echo waveforms in the time-frequency domain and lack direct spatial correspondence, which complicates direct diffusion. Therefore, developing specialized diffusion models tailored for echo-based depth estimation is essential.

In this study, we propose a novel diffusion model framework for echo-based depth estimation, as illustrated in Figure 1, which utilizes both echo spectrograms and echo waveforms as inputs. To effectively extract and encode spatial information from echo spectrograms into a latent space suitable for diffusion processes, we propose the Multi-Scale Adaptive Latent Feature Network (MALF-Net). This network combines Atrous Spatial Pyramid Pooling (ASPP) and Adaptive Spatial Feature Fusion (ASFF) (Liu et al. 2019) to capture multi-scale spatial features within the echo spectrograms. By leveraging ASPP, MALF-Net is able to extract features at various scales, while ASFF enables the adaptive fusion of these features to create more comprehensive latent space. This final latent space is subsequently employed as the initial condition for the forward diffusion process in the diffusion model.

While MALF-Net improves the forward diffusion process by extracting multi-scale spatial features from echo spectrograms, spectrograms alone may miss fine-grained temporal details crucial for accurate depth estimation. To address this, we introduce the Echo Waveform Detail Embedder (EWDE) to process echo waveforms. EWDE utilizes the Wav2Vec model (Baevski et al. 2020) to extract rich, high-resolution features from echoes, capturing the complex characteristics of the original echo signals. These features are utilized to generate conditional embeddings through Probability Embedder and an Embed Adapter. These embeddings retain fine-grained information that traditional preprocessing methods may overlook and are subsequently integrated with the latent features in the spectrogram. During the reverse diffusion process, the conditional embedding obtained from the EWDE module is utilized to guide the denoising steps.

Our main contributions can be summarized as:

- This study introduces EchoDiffusion, a novel framework utilizing diffusion models for echo-based depth estimation. EchoDiffusion employs MALF-Net to generate the latent space that serves as the initial condition for the forward diffusion process and uses EWDE to capture waveform details that guide noise removal during reverse diffusion. This approach enables the generation of a more accurate depth map. Our method achieves state-of-the-art performance, surpassing existing approaches on benchmark datasets Replica and Matterport3D.
- We propose MALF-Net, a novel architecture for latent feature extraction from echo spectrograms. MALF-Net captures multi-scale spatial information and adaptively fuses features across multiple scales, thereby enhancing the quality of the latent space and improving the forward diffusion process in depth estimation.
- To leverage the fine-grained temporal details of the echo waveforms and guide noise removal process of reverse diffusion, we propose EWDE. This module extracts conditional embeddings from echo waveforms, providing critical temporal information that refines the latent spectrogram features and enhances depth map generation.

Related Work

Diffusion for Monocular Depth Estimation

Recent research has explored various methodologies that leverage diffusion models for monocular depth estimation (Ji et al. 2023; Duan, Guo et al. 2023; Patni, Agarwal et al. 2024; Ke et al. 2024). DiffusionDepth (Duan, Guo et al. 2023) reinterprets monocular depth estimation as a denoising diffusion process, where an initial random depth distribution is progressively refined into an accurate depth map based on a single image input. ECoDepth (Patni, Agarwal et al. 2024) enhances the precision and detail of depth estimates by integrating diffusion models with sophisticated conditioning strategies, utilizing depth priors and image features as conditions within the diffusion process. Ke et al. (2024) present an innovative approach that repurposes pre-trained diffusion-based image generators for monocular depth estimation, adapting the generative process to produce depth maps by capitalizing on the models' inherent spatial understanding. In summary, advances in diffusion models demonstrate their effectiveness in depth estimation. These models show the ability to transform noisy data into accurate representations by utilizing additional contextual information.

Echo-based Depth Estimation

There has been substantial work on echo depth estimation, as evidenced by studies such as (Christensen et al. 2020; Vasudevan et al. 2020; Irie, Shibata et al. 2022). Early research in this area can be traced back to (Christensen et al. 2020), which employed an encoder-decoder architecture. This study utilized echo spectrograms as input to a UNet

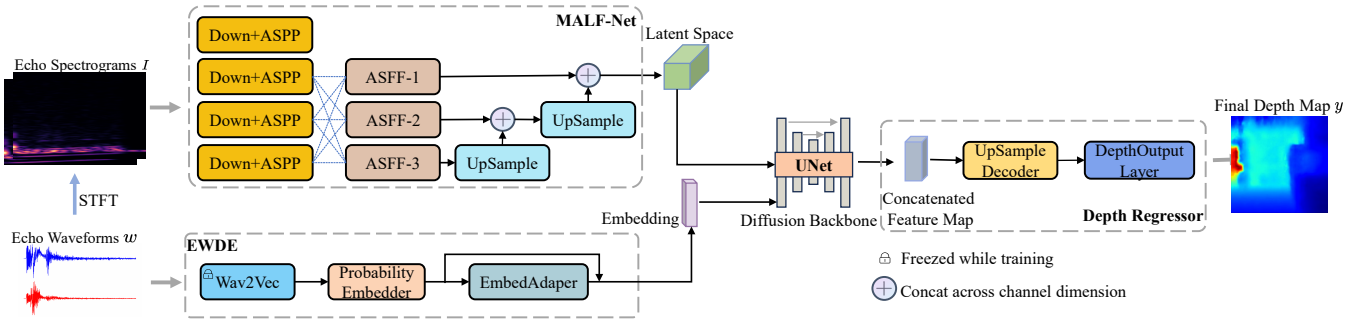

Figure 2: An overview of our proposed model. The MALF-Net generates a latent space of the echo spectrograms I, which is derived from the echo waveforms w using the STFT. Simultaneously, the EWDE module extracts conditional embeddings directly from the waveforms. These conditional embeddings guide the reverse diffusion of the latent space through the diffusion backbone, achieving conditional diffusion. The resulting feature maps are then combined into a concatenated feature map, which is subsequently processed by the depth regressor to produce the final depth map.

network for depth estimation and explored the potential of directly predicting depth maps from echo waveforms. Vasudevan et al. (2020) integrated ASPP into the final layer of the encoder to facilitate multi-scale feature extraction. Their experiments demonstrated that ASPP significantly enhances the capture of relevant echo features, leading to an improvement in the model's depth estimation performance. Additionally, Irie, Shibata et al. (2022) showcased the effectiveness of multiple input fusion for depth estimation by combining the angle spectrum and spectrogram within a co-attention-guided model. However, despite these advancements, prior research has not fully exploited the fine-grained temporal details inherent in echo waveforms.

Method

Overview

The main framework of the proposed EchoDiffusion architecture is illustrated in Figure 2. Our model comprises four main components: MALF-Net generates a latent space representation of the echo spectrograms for the forward diffusion process. Concurrently, the EWDE module extracts conditional embeddings from the echo waveforms. During the reverse diffusion process, the diffusion backbone utilizes these conditional embeddings to guide the denoising of the noisy latent space. Finally, a depth regressor processes the refined feature map to produce the final depth map.

Within MALF-Net, during the downsampling phase, we utilize ASPP to capture features at various spatial scales from the echo spectrograms. ASPP employs atrous convolutions of various sizes to extract features across different scales. Following this features extracted at different levels of the network are adaptively fused using multiple ASFF modules, maximizing the utilization of multi-scale feature information. The features derived from the three ASFF modules are then integrated to achieve encoding into the latent space.

To generate conditional embeddings, we utilize the EWDE module. These embeddings guide the reverse diffusion process within the latent space. Echo waveform features are first extracted using the Wav2Vec model, with their dimensionality reduced from 768 to 100 through downsampling and a linear activation function. The resulting features are then multiplied by the conditional embedding matrix to produce conditional probabilities that direct the reverse diffusion process.

We adopt the latent diffusion framework and utilize the UNet diffusion backbone, as implemented in Stable Diffusion (Rombach et al. 2022), to execute the diffusion process. The UNet backbone generates four feature maps, which are aggregated using a depth regressor to generate a feature map of dimensions $64 \times \frac{H}{4} \times \frac{W}{4}$. Subsequently, the predicted depth map y is generated using the UpSample Decoder and Depth Output Layer.

Multi-Scale Adaptive Latent Feature Network

Encoding the input into a latent space is a crucial step in diffusion models, as it reduces the dimensionality of the input and facilitates subsequent diffusion and reconstruction processes. After encoding into the latent space, the integration of waveform embeddings can guide the model toward generating a more accurate and comprehensive depth map. However, unlike images, the pixels in echo spectrograms do not directly correspond to depth information. Instead, spatial information is inferred indirectly through the analysis of frequency components over time. To effectively utilize these characteristics, we propose the MALF-Net, a network specifically designed for multi-scale feature extraction and adaptive fusion. This design enables the MALF-Net to fully leverage spatial information at different scales, resulting in a higher-quality latent space representation.

As shown in Figure 3, the network begins with the input spectrograms I, which are extracted from the echo waveforms using the Short-Time Fourier Transform (STFT). The spectrograms are then processed by the Downsample and ASPP (DA) Modules, which performs downsampling to reduce spatial dimensions and increase the receptive field. The DA Module first applies max pooling to efficiently downsample the input, preserving essential features while reducing computational complexity. Following the downsampling, the module utilizes two DoubleConv layers to further

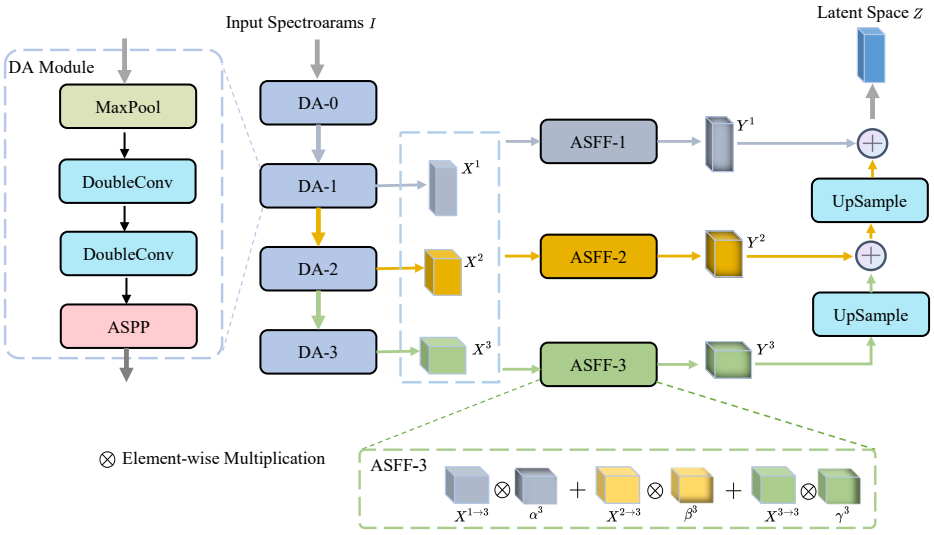

Figure 3: The detailed structure of the MALF-Net. The echo spectrograms are passed as input through a series of ASPP and ASFF modules to perform multi-scale feature extraction and adaptive fusion, respectively. We employ DA modules for downsampling and to enhance multi-scale feature extraction by appending an ASPP module to the end of each DA module. The last three feature maps X^1, X^2, X^3 generated by the DA modules are fed into three ASFF modules. Each ASFF module operates on feature maps of corresponding scales, performing adaptive feature fusion. The resulting fused feature maps Y^1, Y^2, Y^3 from the three ASFF modules are then upsampled and concatenated to generate the final latent space Z.

extract spatial information from the spectrogram. These layers apply convolution followed by the ReLU activation function, introducing non-linearity and enhancing the model's ability to capture complex patterns in the data. Moreover, the final layer of the DA Module incorporates an ASPP module, employing four dilated convolutions with varying dilation rates to extract multi-scale features from the feature maps. We use four DA Modules to generate the corresponding four feature maps $X^i, i \in \{0, 1, 2, 3\}$, and use the last three feature maps as input for the subsequent ASFF modules.

To effectively combine the multi-scale spatial features extracted by the ASPP, we utilize the ASFF module for adaptive fusion. The ASPP first extracts feature maps at various scales, capturing different levels of spatial information. The ASFF module then takes these multi-scale feature maps and applies an adaptive attention mechanism to weigh and fuse them. Specifically, ASFF achieves this by adaptively learning fusion spatial weights for each scale feature map through identity scaling and adaptive fusion. For example, in ASFF-3, the process involves max pooling and a 3×3 convolution on the X^1 feature map to obtain $X^{1\to3}$, as well as a 3×3 convolution on the X^2 feature map to obtain $X^{2\to3}$, ensuring consistency in size across the three feature maps by,

$$Y^3 = X^{1\to3} \cdot \alpha^3 + X^{2\to3} \cdot \beta^3 + X^{3\to3} \cdot \gamma^3 \quad (1)$$

where α, β, and γ represent the weight parameters obtained from X^0, X^1, and X^2 through a 1×1 convolution.

Subsequently, adaptive fusion is applied, where softmax is utilized to weigh, sum, and normalize the adjusted feature maps, resulting in the final fused feature map Y^3. This normalization ensures that the weighted parameters fall within the range of $[0, 1]$ and sum to 1, thereby enabling dynamic adjustment of feature importance across spatial locations.

By utilizing three ASFF models, we obtain fused feature maps Y^1, Y^2, and Y^3 sourced from three different sizes of the initial feature maps. These fused feature maps are then upsampled, concatenated, and combined to yield the final encoded feature map Z in the latent space by,

$$Z = ((Y^3)^{\uparrow} \oplus Y^2)^{\uparrow} \oplus Y^1 \quad (2)$$

where, $\uparrow$ represents UpSample and $\oplus$ represents concatenation. This final step in MALF-Net effectively combines the multi-scale features into a single, comprehensive latent space Z.

Echo Waveform Detail Embedder

To obtain the fine-grained temporal details from echo waveforms and enhance the guidance provided to the conditional diffusion model during reverse diffusion, we utilize the pre-trained Wav2Vec model (Baevski et al. 2020) to extract features from echo signals. Originally designed for speech recognition tasks, Wav2Vec is adept at capturing subtle echo features. By leveraging this model, we aim to harness its feature extraction capabilities for the specific context of echo-based depth estimation. We take the pre-trained Wav2Vec model and freeze the parameters of its feature extractor during model training to acquire the output of its last hidden layer, resulting in a vector of 768 dimensions.

In the EWDE module, as shown in Figure 4, the probabilistic embedder processes the output of Wav2Vec, reducing the dimensionality and conditioning the features to align with the embedding space requirements. This process includes average pooling and two linear transformations, interspersed with GELU (Gaussian Error Linear Unit) activations

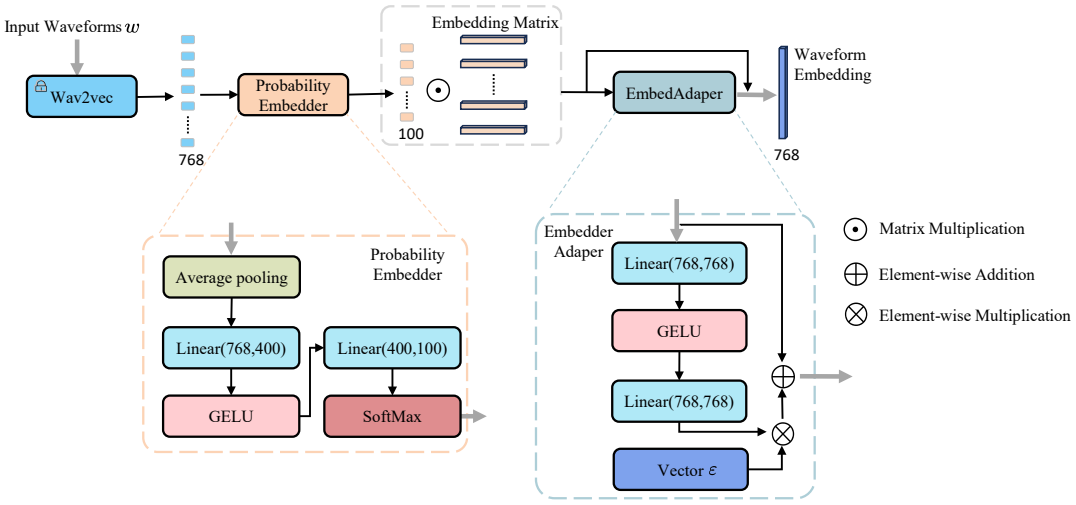

Figure 4: The detailed structure of the EWDE module. The EWDE module uses echo waveforms to create conditional embeddings that guide the reverse diffusion process. It starts with the Wav2Vec model to extract detailed echo features from the input waveforms, resulting in a 768-dimensional vector. These features are then reduced to 100 dimensions through a linear transformation. A probabilistic embedder is used to create probabilistic embeddings, which are then combined with the embedding matrix. The final output is refined by the embedding adapter to produce the ultimate conditional embeddings.

and a final softmax layer. The Probabilistic Embedder compresses the Wav2Vec output from 768 to 100 dimensions, creating a learnable embedding matrix that seamlessly integrates with the conditional embedding matrix.

The Embed Adapter is responsible for refining the Wav2Vec-extracted features to make them suitable for guiding the diffusion process. The core of the Embed Adapter consists of two sequential linear layers, each with an input and output dimension of 768. The first layer applies a linear transformation to the input features, followed by a GELU activation function, which introduces non-linearity and enhances the expressive capacity of the model. The output is then passed through another linear layer, maintaining the same dimensionality. Upon obtaining the transformed features from the fully connected layers, the module refines these features using a learnable parameter Vector ε, which is initialized to a small value (1e-4) and shares the same dimensionality as the feature vector. This parameter allows the model to fine-tune the scaling of the features during training. This parameter vector modulates the importance of individual elements in the transformed features. The transformed features are scaled by Vector ε and then added back to the original input features, effectively enhancing the feature space with relevant details captured during the transformation process.

Depth Regressor

We employ the UNet diffusion backbone to extract four feature maps at different hierarchical levels. These feature maps are subsequently concatenated to facilitate feature fusion, resulting in a combined feature map with dimensions of $352 \times 8 \times 8$. The upsampling decoder, consisting of a series of transposed convolutional layers, is then utilized to progressively restore the spatial resolution of the feature map. Finally, the depth output layer, composed of two convolutional layers with 3×3 kernels, is applied to further refine the output and generate the final depth map.

Experiments

Datasets

We conduct experiments using the Replica (Straub et al. 2019) and Matterport3D (Chang et al. 2017) datasets. These datasets are widely utilized in echo depth estimation research due to their rich variety of scenes and diverse data. Studies such as (Parida et al. 2021; Irie, Shibata et al. 2022; Brunetto et al. 2023) have employed these datasets to evaluate model performance.

The Replica dataset comprises 18 scenes, including hotels, offices, and various room types. For training, 15 scenes with 5,496 instances are utilized, while 3 scenes containing 1,464 instances are reserved for testing. The Matterport3D dataset contains more scenes, consisting of a total of 90 scenes, including apartments, meeting rooms, shops, and more. Of these, 59 scenes (comprising 40,176 instances) are used for training, 10 scenes (comprising 13,592 instances) for validation, and 8 scenes (comprising 13,602 instances) for testing.

Parameter Setting

The EchoDiffusion model was implemented using PyTorch and trained on an NVIDIA GeForce RTX 4090 D. Training times per epoch varied depending on the dataset: approximately 80 seconds for the Replica dataset and around 400 seconds for the Matterport3D dataset. The training process spanned 150 epochs for each dataset. A learning rate of 0.0001 was employed, coupled with an L2 regularization

Dataset	Method	RMSE↓	REL↓	log10↓	$\delta < 1.25$↑	$\delta < 1.25^2$↑	$\delta < 1.25^3$↑
Replica	Echo-Net	0.995	0.638	0.208	0.338	0.599	0.742
	Co-attention	0.921	**0.560**	0.203	0.419	0.636	0.763
	Bat-Net	0.956	0.622	0.206	0.448	0.636	0.750
	EchoDiffusion (Ours)	**0.913**	0.604	**0.194**	**0.515**	**0.668**	**0.764**
Matterport3D	Echo-Net	1.778	0.569	0.192	0.464	0.642	0.759
	Bat-Net	1.752	0.583	0.201	0.422	0.633	0.755
	EchoDiffusion (Ours)	**1.702**	**0.512**	**0.187**	**0.481**	**0.659**	**0.770**

Table 1: Experimental results on the Replica and Matterport3D datasets.

coefficient of 0.0005 (Paszke et al. 2019). The AdamW optimizer (Paszke et al. 2019) was utilized for optimization, with a batch size set to 32.

Evaluation Metrics

Following earlier works in echo-based depth estimation (Parida et al. 2021; Irie, Shibata et al. 2022; Brunetto et al. 2023), we evaluate the performance of the proposed method based on root mean squared error (RMSE), mean relative error (REL), mean log10 error, and the threshold accuracy ($\delta < 1.25$, $\delta < 1.25^2$, $\delta < 1.25^3$).

Experimental Results

To accurately assess the predictive capabilities of the EchoDiffusion model, we evaluated its performance using both Replica and Matterport3D datasets and compared it with leading echo depth prediction models Bat-Net (Brunetto et al. 2023) and Echo-Net (Parida et al. 2021). Since the co-attention model (Irie, Shibata et al. 2022) only published test results on the Replica and did not publish the code, we limited our comparison to that dataset. Eco-Net, a multimodal depth estimation model, was tested for single echo depth estimation by setting the image input to zero, following the approach used in (Irie, Shibata et al. 2022; Brunetto et al. 2023). Bat-Net and the co-attention model are single-mode echo depth estimation models. The experimental results Table 1 shows that the EchoDiffusion model outperforms Echo-Net, Co-attention, and Bat-Net on the Replica and Matterport3D datasets, demonstrating its superior robustness and accuracy across most metrics.

On the Replica dataset, as shown in Table 1, the EchoDiffusion model shows slightly lower performance in the REL metric but outperforms other models across all other metrics. Notably, our model shows significant improvement in the three threshold error indicators, with an approximate 8% enhancement over the previously best-performing Co-attention model. The performance on the Matterport3D dataset is similarly commendable. Despite a slight decrease in the REL metric, the overall results remain largely consistent with those observed on the Replica dataset, with even more pronounced improvements in threshold accuracy. This underscores the EchoDiffusion model's superior ability to efficiently capture fine details.

Figure 5 demonstrates how our model effectively captures fine-grained object details and preserves overall depth cues. In the first row, EchoDiffusion accurately outlines potted plants and refrigerators, clearly depicting their spatial

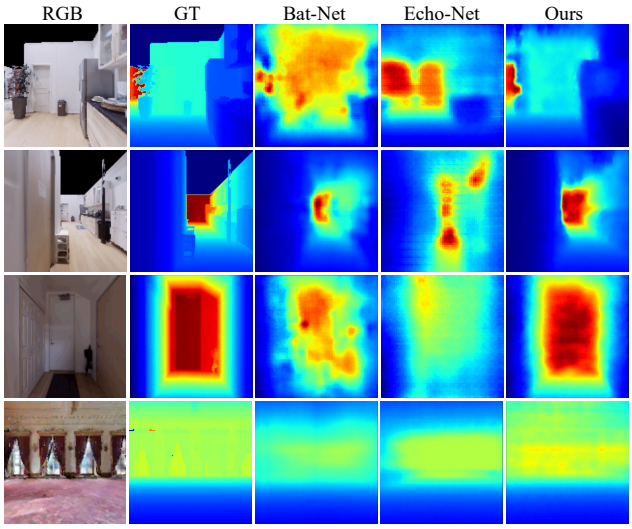

Figure 5: Qualitative results on the Replica and Matterport3D datasets. Among these images, the first three are from the Replica dataset. The first and second depict living room scenes from long and close views, respectively. The third image shows a bedroom scene under low-light conditions. The fourth image represents a hotel scene from the Matterport3D dataset.

arrangement. In the second row, we show the corner information from the refrigerator to the stove and accurately maintain the size information of the back wall. Moreover, our model distinctly represents the depth information across various scenes. In the third row, EchoDiffusion reconstructs depth details inside the doorway, even in low light, revealing a clear spatial hierarchy. The fourth row highlights the model's accuracy in identifying the boundary between the floor and window, as well as the distance to the balcony.

Ablation Study

In our ablation study, as shown in Table 2, we conducted a comprehensive analysis to emphasize the critical role and efficacy of the MALF network. The MALF-Net integrates ASFF and ASPP modules, as depicted in the figure above. Each component uniquely contributes to the overall performance, yet their combination in the MALF network yields superior results compared to individual components and

Dataset	Latent Space Generator	RMSE↓	REL↓	log10↓	$\delta < 1.25$↑	$\delta < 1.25^2$↑	$\delta < 1.25^3$↑
Replica	FPN	0.929	0.608	0.198	0.501	0.662	0.758
	ASFF	0.938	0.630	0.198	0.498	0.660	0.757
	ASPP	0.925	0.627	0.200	0.500	0.662	0.761
	MALF-Net	**0.913**	**0.604**	**0.194**	**0.515**	**0.668**	**0.764**
Matterport3D	FPN	1.741	0.574	0.193	**0.487**	0.656	0.761
	ASFF	1.722	0.573	0.188	0.470	0.649	0.768
	ASPP	1.742	0.597	0.197	0.484	0.655	0.759
	MALF-Net	**1.702**	**0.512**	**0.187**	0.481	**0.659**	**0.770**

Table 2: Ablation experiments on the Replica and Matterport3D datasets.

other baseline models.

To demonstrate the contributions of ASFF and ASPP within the MALF network, we compared the following configurations:

- FPN: A basic Feature Pyramid Network without advanced fusion and pyramid pooling techniques.
- ASFF: Utilizing only the Adaptive Structure Feature Fusion modules without the ASPP component.
- ASPP: Incorporating only the Atrous Spatial Pyramid Pooling without ASFF.

Here we only replace the MALF network, leaving the other model components unchanged. As shown in Table 2, the results on the Replica dataset show that the MALF network excels across all metrics. Specifically, the MALF network achieves an RMSE of 0.913, reducing the error by 1.30% compared to 0.925 for ASPP and by 2.66% compared to 0.938 for ASFF. On the Matterport3D dataset, the MALF performance index has also basically reached the best, especially REL up to 0.512. In order to demonstrate the validity of the framework that echo waveforms guides spectrogram diffusion, visualization of ablation experiments is also presented. As shown in Figure 6, our model still predicts better results than Echo-Net and Bat-Net, even when encoding into latent space using traditional FPN modules. This effectively illustrates the validity of our framework, which uses echo waveforms to guide the reverse diffusion, ultimately resulting in a more comprehensive and accurate depth map.

ASPP utilizes atrous convolutions with varying dilation rates to effectively capture features at multiple scales. However, while ASPP is adept at capturing multi-scale contexts, it does not inherently provide mechanisms to adaptively weigh and fuse these features based on their relevance. In contrast, the ASFF module excels at adaptively combining features from different scales and levels. By leveraging an attention mechanism, it effectively learns spatial weights, allowing the network to emphasize the most pertinent features for the specific task at hand. Nevertheless, ASFF in isolation may not comprehensively address the full spectrum of scale variations as effectively as ASPP.

By integrating ASPP and ASFF within the MALF network, our model capitalizes on the strengths of both components: ASPP's extensive multi-scale feature extraction and ASFF's adaptive, attention-driven feature fusion. This integration not only enhances the model's ability to extract a broad spectrum of feature scales but also allows the network

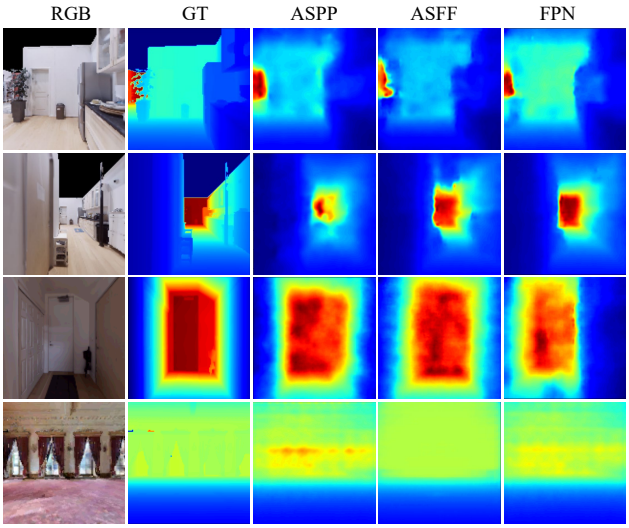

Figure 6: Qualitative ablation results on the Replica and Matterport3D datasets.

to dynamically prioritize the most critical features. Consequently, the model's capacity to capture multi-scale spatial information is significantly improved. Moreover, when compared to traditional FPN networks, the MALF network consistently demonstrates superior performance, further validating its effectiveness.

Conclusion

In this paper, we present EchoDiffusion, a novel approach where the echo spectrograms are encoded into a latent space for diffusion, while the echo waveforms guide the reverse diffusion process to estimate depth. We introduce MALF-Net, which effectively extracts and merges multi-scale spatial features, optimizing the encoding of spectrograms into the latent space. Additionally, we propose the EWDE module, which generates conditional embeddings from echo waveforms, thereby improving the accuracy of depth map reconstruction during reverse diffusion. Experiments on the Replica and Matterport3D datasets demonstrate that EchoDiffusion outperforms existing methods across key metrics. Ablation studies further validate the effectiveness of MALF-Net and the superiority of the waveform-guided diffusion approach.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grants 62325602 and U21B2037, and in part by the Natural Science Foundation of Henan Province under Grant 232300421093.

References

Baevski, A.; Zhou, Y.; Mohamed, A.; and Auli, M. 2020. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. *Advances in Neural Information Processing Systems*, 33: 12449–12460.

Brunetto, A.; et al. 2023. The Audio-Visual BatVision Dataset for Research on Sight and Sound. In *International Conference on Intelligent Robots and Systems*, 1–8. IEEE.

Chang, A.; Dai, A.; Funkhouser, T.; Halber, M.; Niessner, M.; Savva, M.; Song, S.; Zeng, A.; and Zhang, Y. 2017. Matterport3D: Learning from RGB-D Data in Indoor Environments. *International Conference on 3D Vision*, 667–676.

Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; and othres. 2017. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40: 834–848.

Christensen, J. H.; et al. 2020. BatVision: Learning to See 3D Spatial Layout with Two Ears. In *International Conference on Robotics and Automation*, 1581–1587.

Duan, Y.; Guo, X.; et al. 2023. DiffusionDepth: Diffusion Denoising Approach for Monocular Depth Estimation. *arXiv preprint arXiv:2303.05021*.

Ho, J.; Jain, A.; et al. 2020. Denoising Diffusion Probabilistic Models. *Advances in Neural Information Processing Systems*, 33: 6840–6851.

Hu, K.; Cao, T.; Li, Y.; Chen, S.; and Kang, Y. 2024. DALDet: Depth-Aware Learning Based Object Detection for Autonomous Driving. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 2229–2237. AAAI.

Irie, G.; Shibata, T.; et al. 2022. Co-Attention-Guided Bilinear Model for Echo-Based Depth Estimation. In *International Conference on Acoustics, Speech and Signal Processing*, 4648–4652. IEEE.

Ji, Y.; Chen, Z.; Xie, E.; Hong, L.; Liu, X.; Liu, Z.; Lu, T.; Li, Z.; et al. 2023. DDP: Diffusion Model for Dense Visual Prediction. In *Proceedings of the IEEE International Conference on Computer Vision*, 21741–21752. IEEE.

Kawai, H.; et al. 2023. DNN Based Camera Attitude Estimation Using Aggregated Information from Camera and Depth Images. In *Proceedings of the IEEE International Symposium on System Integration*, 1–6. IEEE.

Ke, B.; Obukhov, A.; Huang, S.; Metzger, N.; Daudt, R. C.; et al. 2024. Repurposing Diffusion-Based Image Generators for Monocular Depth Estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 9492–9502. IEEE.

Li, Y.; Bao, H.; Ge, Z.; Yang, J.; Sun, J.; et al. 2023. BEVStereo: Enhancing Depth Estimation in Multi-View 3D Object Detection with Temporal Stereo. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 1486–1494. AAAI.

Liu, S.; et al. 2019. Learning Spatial Fusion for Single-Shot Object Detection. *arXiv preprint arXiv:1911.09516*.

Parida, K. K.; et al. 2021. Beyond Image to Depth: Improving Depth Prediction using Echoes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 8268–8277. IEEE.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32.

Patni, S.; Agarwal, A.; et al. 2024. ECoDepth: Effective Conditioning of Diffusion Models for Monocular Depth Estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 28285–28295. IEEE.

Ren, P.; Chen, Y.; Hao, J.; Sun, H.; Qi, Q.; Wang, J.; and Liao, J. 2023. Two Heads Are Better than One: Image-Point Cloud Network for Depth-Based 3D Hand Pose Estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 2163–2171. AAAI.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 10684–10695. IEEE.

Senocak, A.; Oh, T.-H.; Kim, J.; Yang, M.-H.; and Kweon, I. S. 2019. Learning to Localize Sound Sources in Visual Scenes: Analysis and Applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(5): 1605–1619.

Shao, S.; Pei, Z.; Chen, W.; Liu, Q.; Yue, H.; and Li, Z. 2023. Sparse Pseudo-LiDAR Depth Assisted Monocular Depth Estimation. *IEEE Transactions on Intelligent Vehicles*, 9(1): 917–929.

Shimada, T.; Nishikawa, H.; Kong, X.; and Tomiyama, H. 2022. Depth Estimation from Monocular Infrared Images for Autonomous Flight of Drones. In *Proceedings of the International Conference on Electronics, Information, and Communication*, 1–6. IEEE.

Singh, A. D.; Ba, Y.; Sarker, A.; Zhang, H.; Kadambi, A.; Soatto, S.; Srivastava, M.; and Wong, A. 2023. Depth Estimation from Camera Image and mmWave Radar Point Cloud. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 9275–9285. IEEE.

Straub, J.; Whelan, T.; Ma, L.; Chen, Y.; Wijmans, E.; Green, S.; Engel, J. J.; Mur-Artal, R.; Ren, C.; Verma, S.; et al. 2019. The Replica Dataset: A Digital Replica of Indoor Spaces. *arXiv preprint arXiv:1906.05797*, 1–10.

Vasudevan, A. B.; et al. 2020. Semantic Object Prediction and Spatial Sound Super-Resolution with Binaural Sounds. In *European Conference on Computer Vision*, 638–655. Cham: Springer.

Wang, Y.; Liang, Y.; Xu, H.; Jiao, S.; and Yu, H. 2024. SQLdepth: Generalizable Self-Supervised Fine-Structured Monocular Depth Estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 5713–5721. AAAI.

Zhang, N.; Nex, F.; Vosselman, G.; and Kerle, N. 2023. Lite-Mono: A Lightweight CNN and Transformer Architecture for Self-Supervised Monocular Depth Estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 18537–18546. IEEE.

Zheng, J.; Lin, C.; Sun, J.; Zhao, Z.; Li, Q.; and Shen, C. 2024. Physical 3D Adversarial Attacks against Monocular Depth Estimation in Autonomous Driving. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 24452–24461. IEEE.