CO-ATTENTION-GUIDED BILINEAR MODEL FOR ECHO-BASED DEPTH ESTIMATION

Go Irie *Takashi Shibata* *Akisato Kimura*
NTT Corporation

ABSTRACT

Echoes reflect a geometric structure of a scene surrounding a sound source. In this paper, we address the problem of estimating depth maps of indoor scenes based on echoes. First, we experimentally show that fusing multiple acoustic features, especially spectrogram and angular spectrum, can improve estimation accuracy. We then propose a novel bilinear model that incorporates dense co-attention for effective feature fusion. Our model is able to obtain a compact fused feature while capturing the second-order correlations of intra- and inter-features. Thorough evaluations on two datasets demonstrate the superiority of the proposed method over the state-of-the-art echo-based depth estimation and feature fusion methods.

***Index Terms*—** bilinear model, co-attention, echo-based depth estimation, feature fusion

1. INTRODUCTION

3D structure of a scene is an essential piece of information in various applications such as scene understanding, localization, navigation, path planning, and smart home. To date, a variety of cutting-edge depth sensors including LiDARs and infrared stereo cameras have been developed. Around the same time, monocular depth estimation, the task of estimating depth maps from RGB images, has been actively studied in the fields of image processing and computer vision [1, 2, 3, 4]. However, even these state-of-the-art techniques are not a panacea; RGB images cannot be acquired in a room with inadequate illumination, and infrared light is difficult to acquire stably in outdoor scenes. LiDAR is vulnerable to airborne particles such as dust and fog, and advanced sensors tend to be large and expensive.

Inspired by recent work examining audio-visual modeling for various geometric prediction tasks [5, 6, 7, 8], in this paper, we explore an alternative direction for depth estimation of indoor scenes – *an echo-based approach*. Suppose we have a known sound source and a microphone array. The sound emitted from the source bounces off surrounding walls and objects, and the time of arrival at each microphone varies depending on their locations and shapes. Hence, the time difference of arrival (TDoA) or direction of arrival (DoA) of echoes are strongly correlated with the depth map of the scene.

In general, the problem of reconstructing 3D shape of a scene from echoes is ill-posed and cannot be solved analytically without any assumptions, e.g., the room is a convex polyhedron [9]. Since such an assumption does not always hold in natural indoor scenes where many furniture or fixtures are placed, researchers have investigated solutions based on deep learning. Turpin et al. [10] empirically showed that depth maps can be recovered from a single-channel time-of-arrival histogram of echoes. Christensen et al. [11] proposed to use U-Net to estimate depth maps from spectrograms of echoes acquired from binaural recordings. Vasudevan et al. [12] instead used a convolutional neural network with atrous spatial pyramid pooling module (ASPP). A few recent methods [13, 14] have shown that depth estimation accuracy can be significantly improved by integrating both audio and visual features.

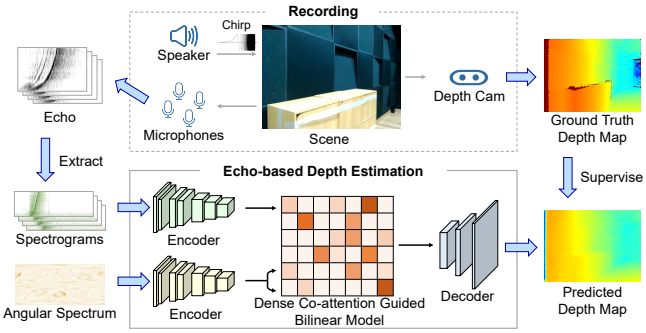


Fig. 1: **Overview of Our Task and Framework.** A speaker emits a chirp signal into the scene, and a microphone array collects the echoes. At the same time, a depth camera measures the ground truth depth map. Our task is to predict the depth map from the echoes. We propose a bilinear feature fusion model with dense co-attention for improving depth estimation quality, which is to capture the second-order correlation between the spectrogram and the angular spectrum.

All these existing methods use a single acoustic feature, typically spectrogram, and little has been paid attention to other features, or combinations of them. In fact, phase features closely related to TDoA/DoA, such as generalized cross-correlation with phase transformation (GCC-PHAT) [15] and angular spectrum, are often used in geometric prediction tasks such as source localization and sound event localization tasks [16, 17]. Similar to these tasks, depth estimation is also expected to gain useful information from phase features.

Motivated by these observations, we first evaluate phase features and their combinations with spectrogram, and demonstrate that fusing spectrogram and angular spectrum can improve depth estimation accuracy. We then propose a new bilinear model to effectively fuse these features. Unlike existing models, ours is based on a unary quadratic form and dense co-attention that efficiently captures the second-order correlations in and across the features. Exhaustive evaluations with two datasets show that our model outperforms state-of-the-art echo-based depth estimation and feature fusion methods.

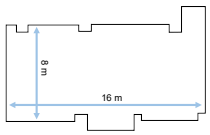
2. FEATURE STUDY

An overview of our problem and framework is shown in Fig. 1. We extract features from the echoes and feed them into our depth estimation network to output the depth map. The network is trained to minimize the difference between the predicted and ground truth depth maps. Our aim is to improve the depth estimation accuracy with this framework. The first point to consider is the features. In this section, we conduct a feature study to evaluate two representative phase features, GCC-PHAT and angular spectrum.

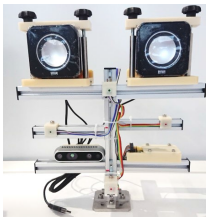
2.1. Studio Dataset

In order to investigate the practical performance, we built a real sound dataset using the recording environment and recording device shown in Fig. 2. Hereafter, this dataset will be referred to as **Studio**.

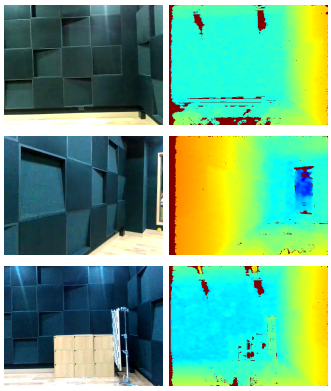
The recording environment is a near-rectangular reverberant studio of 16m width $\times$ 8m depth $\times$ 3m height. Our recording device is



(a) Floor plan



(b) Recording device



(c) Examples of scenes and depth maps

Fig. 2: **Studio Dataset.** (a) Floor plan of the sound collection environment. (b) Our recording device. (c) Examples of scenes and recorded raw depth maps acquired by the depth camera. Depth maps are color coded for visualization (red: nearest - blue: farthest).

structured by a steel rig, equipped with an RGB-Depth camera (Intel RealSense D435), stereo speakers (only one of which is used), and four omnidirectional microphones mounted in a tripod configuration. We emitted a chirp (time-stretched pulse up to 16kHz) from the speaker at various locations and orientations and recorded echoes and a raw depth map of 480×360 in temporal synchronization.

The number of collected samples was $1,478$. We split them into training/validation and test sets in the ratio of 9:1 ($1,331$ and 147, respectively). The duration of each audio sample was 1 second (16kHz sampling rate), which was sufficiently long to accommodate the chirp echoes. Each sample is acquired at a different location and orientation from each other, and thus originates from different RIRs.

2.2. Network Architecture

Following the monocular depth estimation networks [1, 2, 3, 4], we adopted a encoder-decoder type architecture. The encoder part receives the features (detailed in the next subsection) and outputs an intermediate feature map. We adapted the ResNet-50 architecture to our task by discarding the last fully-connected and global average pooling layers. For the decoder part, we used the standard decoder architecture used in [2]. It consists of a stack of four upsampling blocks that each doubles the spatial resolution of the input feature map while reducing the number of channels by half.

2.3. Features

We evaluate spectrogram (Spec), GCC-PHAT (GCC), and angular spectrum (AS), and two fused features of GCC+Spec and AS+Spec.

Spectrogram (Spec). We used the short-term Fourier transform (STFT) with a window length of 512 samples and a window shift of 64 samples. A Hann window function was used.

GCC-PHAT (GCC). We obtained the cross-power spectrum of GCC-PHAT [15] with the same STFT setup as for Spec. The feature has exactly the same time-frequency dimensions as Spec.

Angular Spectrum (AS). Following [16], we estimated the 2D angular spectrum of azimuth and elevation from the GCC feature.

GCC+Spec (concat). We concatenated GCC and Spec along their channel directions.

AS+Spec (linear). Unlike GCC+Spec, the sizes of AS and Spec are different, making it impossible to simply concatenate them. So we put a simple feature fusion block between the encoder and decoder. Specifically, after obtaining the intermediate feature maps of AS and Spec, we apply a linear fully-connected layer to each of them to align their dimensions to $1,024$ and then concatenate them. To be fair, we also evaluated the case where the exact same feature fusion block is applied to GCC+Spec and call this case GCC+Spec (linear).

Table 1: **Feature Study Results.**

Feature	higher is better			lower is better		
	δ_1	δ_2	δ_3	AbsRel	RMSE	$\log_{10}$
Median	0.432	0.777	0.937	0.322	0.927	0.344
Spec	0.880	0.988	**0.999**	0.110	0.371	0.047
GCC	0.683	0.898	0.972	0.213	0.607	0.082
AS	0.771	0.955	0.994	0.156	0.489	0.065
GCC+Spec (concat)	0.729	0.924	0.983	0.182	0.541	0.073
GCC+Spec (linear)	0.865	0.983	0.997	0.118	0.376	0.049
AS+Spec (linear)	**0.885**	**0.989**	**0.999**	**0.107**	**0.348**	**0.044**

2.4. Results

We evaluate the features in the following common evaluation metrics [2]. The predicted and ground truth depth values at i-th pixel $(1 \leq i \leq n)$ are denoted by $\hat{t}_i$ and t_i^*, respectively.

- δ_α: ratio of pixels whose relative error is within 1.25^α ($\alpha \in \{1, 2, 3\}$). Higher is better.
- AbsRel: mean absolute relative error, i.e., $\frac{1}{n}\sum_i^n |t_i^* - \hat{t}_i|/t_i^*$. Lower is better.
- RMSE: root mean square error, i.e., $\sqrt{\frac{1}{n}\sum_i^n (t_i^* - \hat{t}_i)^2}$. Lower is better.
- $\log_{10}$: mean $\log_{10}$ error, i.e., $\frac{1}{n}\sum_i^n |(\log_{10} t_i^* - \log_{10} \hat{t}_i)|$. Lower is better.

All the network was implemented in PyTorch and trained with Adam for 300 epochs. The learning rate was set at 1.0×10^{-4}, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and weight decay of 1.0×10^{-3}. The loss function used was scale-invariant mean square error in log space [1].

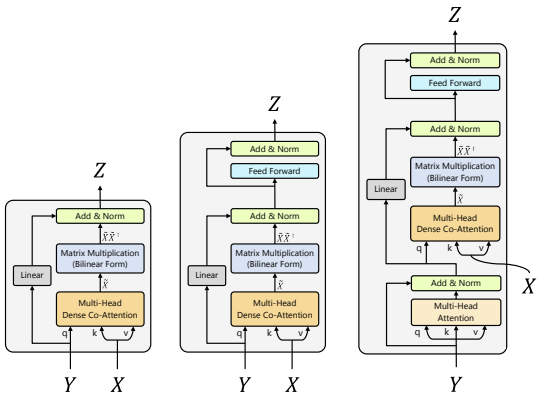
The results are listed in Table 1. We can draw the following three key observations. First, all of the features reflect information effective for depth estimation. This can be seen from the fact that the accuracy by each feature is predominantly better than that by Median, i.e., the case where the depth value of each pixel is predicted by the median of the training samples. Second, when we compare single features (Spec, GCC, and AS) with each other, Spec is clearly better than GCC and AS. Finally, AS+Spec (linear) outperforms Spec. Meanwhile, we found that GCC+Spec, even if it is GCC+Spec (linear), could not improve Spec. The reason may be that AS directly represents spatial information in azimuth and elevation angles, and thus is more suitable for the convolution operations.

3. DENSE CO-ATTENTION GUIDED BILINEAR MODEL

So far, we confirmed that AS+Spec provides better performance than Spec or AS alone, even when using a simple feature fusion model with a linear fully-connected layer. One interesting question here would be whether we can further improve the estimation accuracy by enhancing the feature fusion model. To achieve this goal, in this section we propose a new bilinear feature fusion model.

3.1. Model Formulation

A desired property of feature fusion in general is to capture intra- and inter-feature interactions while reducing the size of the fused feature. This is indeed crucial in our problem. Since AS and Spec represent different information of echos (spatial cues vs. time-frequency cues), higher-order correlations between them are necessary to accurately recover the depth map of the scene. Meanwhile, the computational cost of the decoder is dominant in depth estimation networks [4], so the fused feature fed to the decoder should be compact enough. To



(a) Base (b) Base+FFN (c) Base+FFN+SA

Fig. 3: **Variations of Our Feature Fusion Block.** (a) The base architecture. Two extended versions combined with (b) a feed-forward network (Base+FFN) and (c) a feed-forward network and self-attention module (Base+FFN+SA).

satisfy these properties, we build a new bilinear feature fusion model to efficiently capture the second-order correlation of features.

Suppose we have two feature maps (three-way tensors) to fuse, $\mathcal{X} \in \mathbb{R}^{c_x \times h_x \times w_x}$ and $\mathcal{Y} \in \mathbb{R}^{c_y \times h_y \times w_y}$, where c_x (c_y), h_x (h_y), and w_x (w_y) are the number of channels, height, and width of $\mathcal{X}$ ($\mathcal{Y}$), respectively. Let $s_x = h_x w_x$ and $s_y = h_y w_y$. Without loss of generality, we reshape the tensors to matrices as $\mathcal{X} \to X \in \mathbb{R}^{c_x \times s_x}$ and $\mathcal{Y} \to Y \in \mathbb{R}^{c_y \times s_y}$ and consider to fuse X and Y.

Bilinear Model for Single Feature Case. Let us first consider a single feature case and start from a simple bilinear model [18]:

$$Z = XX^\top \in \mathbb{R}^{c_x \times c_x}, \tag{1}$$

Drawbacks of this simple bilinear model are that the size of Z becomes huge and the interactions in the spatial bins along s_x are squashed. To avoid these problems, we consider the following form:

$$Z = U^\top XSX^\top U \in \mathbb{R}^{k \times k}. \tag{2}$$

$U \in \mathbb{R}^{c_x \times k}$ is a projection matrix to reduce the size of Z to $k \times k$ ($k \leq c_x$) while taking into account the correlation of X in its channel direction. $S \in \mathbb{R}^{s_x \times s_x}$ is a positive semi-definite (PSD) matrix to model the interactions between all the spatial bins of X.

Feature Fusion with Dense Co-attention. Now we extend Eq. 2 to fuse X and Y. A straightforward way is to use a binary quadratic form of X and Y to evenly fuse them, i.e., $Z = U^\top XSY^\top V$. $V \in \mathbb{R}^{c_y \times k}$ is a projection matrix for Y, the counterpart of U. In fact, many existing bilinear feature fusion models are built upon this idea [19, 20, 21, 22]. However, as in our case where there is a significant gap in performance between the features to fuse (see Sec. 2.4), such a model that evenly fuses the features does not always provide significant improvement (as we will show later).

Instead, we propose to use a unary quadratic form with S guided by dense co-attention [23, 20, 24] between the two features. S is PSD so can always be decomposed as $S = A^\top A$. Eq. 2 turns into:

$$Z = U^\top XA^\top AX^\top U. \tag{3}$$

Let $\tilde{X} = U^\top XA^\top$. Eq. 3 is rewritten as:

$$Z = \tilde{X}\tilde{X}^\top, \tag{4}$$

which is still a unary quadratic form. Here, $\tilde{X}$ can be naturally modeled as a dense co-attention, i.e., self-attention [25] using $V^\top Y$, $U^\top X$, and $U^\top X$ for query, key, and value, respectively:

$$\tilde{X} = U^\top XA^\top = U^\top X \underbrace{\left[\text{softmax}\left(\frac{X^\top UV^\top Y}{\sqrt{k}}\right)\right]}_{A^\top}. \tag{5}$$

Plugging this into Eq. (4), our final model is given as:

$$Z = U^\top X \left[\text{softmax}\left(\frac{X^\top UV^\top Y}{\sqrt{k}}\right)\right] \left[\text{softmax}\left(\frac{Y^\top VU^\top X}{\sqrt{k}}\right)\right] X^\top U. \tag{6}$$

Table 2: **Results on Studio.**

Method	higher is better			lower is better		
	δ_1	δ_2	δ_3	AbsRel	RMSE	$\log_{10}$
Median	0.432	0.777	0.937	0.322	0.927	0.344
SELDnet [17]	0.750	0.951	0.993	0.161	0.520	0.192
Echo-Net [14]	0.904	0.991	**0.999**	0.103	0.355	0.044
Linear	0.885	0.989	**0.999**	0.107	0.348	0.044
Bilinear [22]	0.897	0.988	**0.999**	0.102	0.336	0.042
Dense Co-attention [24]	0.911	0.990	**0.999**	0.098	0.320	0.041
Ours	**0.921**	**0.992**	**0.999**	**0.093**	**0.308**	**0.039**

Table 3: **Results on Replica.**

Method	higher is better			lower is better		
	δ_1	δ_2	δ_3	AbsRel	RMSE	$\log_{10}$
Median	0.310	0.514	0.675	0.779	1.156	0.652
SELDnet [17]	0.403	0.598	0.727	1.078	1.086	0.226
Echo-Net [14]	0.338	0.599	0.742	0.638	0.995	0.208
Linear	0.390	0.613	0.749	0.567	0.956	0.211
Bilinear [22]	0.399	0.620	0.757	0.567	0.950	0.206
Dense Co-attention [24]	0.416	0.631	0.760	0.581	0.931	0.207
Ours	**0.419**	**0.636**	**0.763**	**0.560**	**0.921**	**0.203**

3.2. Feature Fusion Block Implementation

Our model is formulated as self-attention of Transformer [25]. Hence, following the implementation of encoder/decoder of Transformer, we consider the following three implementations of our feature fusion block as illustrated in Fig. 3.

Base Architecture. We follow the base architecture of self-attention module [25]. Specifically, we use multi-head dense co-attention with eight attention heads for computing A in Eq. 5 and involve residual connection (Add) followed by layer normalization (Norm). Note that we apply a simple linear projection to Y before Add, in order to fit the size of the fused feature Z.

Base+FFN. Following the encoder of Transformer, which has the basic self-attention module followed by a feed-forward network (Feed Forward: FFN), we build a variation consisting of the Base Architecture and Feed Forward.

Base+FFN+SA. Following the decoder of the Transformer, which has yet another self-attention module inside, we apply a self-attention module to Y, followed by feature fusion with Base+FFN. Unless otherwise noted, we use Base+FFN+SA in our experiments.

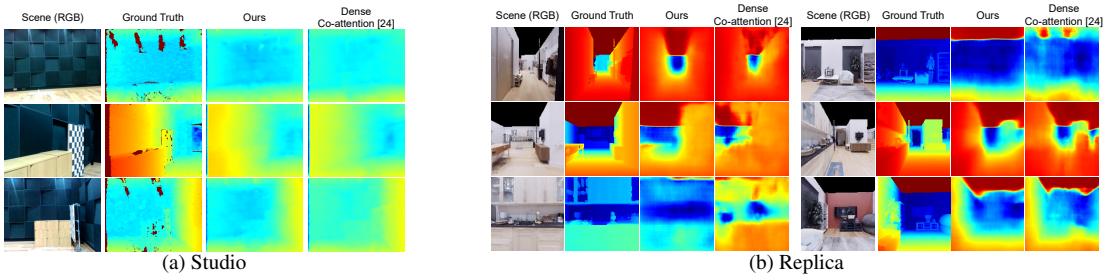
4. EXPERIMENTS

We evaluate our feature fusion model. Besides Studio, we use **Replica** [26], a synthetic indoor scene dataset, to facilitate the evaluation of the generalization ability across different scenes. We consistently use AS (X) and Spec (Y) as the features to be fused.

4.1. Baselines

We evaluate the following three existing feature fusion models relevant to our model under the same encoder-decoder configuration:

Linear: Feature fusion with linear fully-connected layer. This corresponds to AS+Spec (linear) tested in Sec. 2.



(a) Studio (b) Replica

Fig. 4: **Qualitative Results on (a) Studio and (b) Replica.**

Table 4: **Ablation Study on Studio.**

Method	higher is better			lower is better		
	δ_1	δ_2	δ_3	AbsRel	RMSE	$\log_{10}$
Base	0.919	**0.994**	**0.999**	0.096	0.318	0.040
Base+FFN	0.912	0.992	**0.999**	0.095	0.312	0.040
Base+FFN+SA	**0.921**	0.992	**0.999**	**0.093**	**0.308**	**0.039**

Table 5: **Ablation Study on Replica.**

Method	higher is better			lower is better		
	δ_1	δ_2	δ_3	AbsRel	RMSE	$\log_{10}$
Base	0.405	0.626	0.758	0.609	0.929	0.204
Base+FFN	**0.426**	**0.640**	0.761	**0.560**	**0.921**	0.204
Base+FFN+SA	0.419	0.636	**0.763**	**0.560**	**0.921**	**0.203**

Bilinear: Bilinaer feature fusion based on binary quadratic form [22].

Dense Co-attention: Feature fusion with dense co-attention [24].

In addition to these, we also evaluate the following two different versions of echo-based depth estimation network:

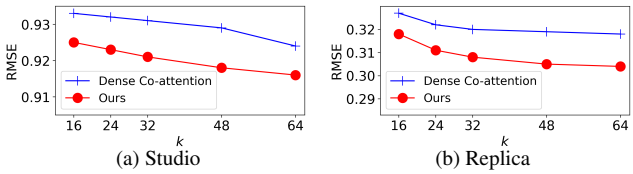
SELDnet: As a representative network that takes both phase and magnitude features as input for geometric prediction, we evaluate SELDnet [17] for the sound event localization and detection task. Specifically, we built a depth estimation network by using SELDnet (without the top fully-connected layers) as its encoder and connecting it to the same decoder as our network.

Echo-Net: We also evaluate **Echo-Net** [14], the state-of-the-art network for echo-based depth estimation.

4.2. Results

Results on Studio. We first show the results on Studio. The protocol is the same as that used in Sec. 2.4. For fairness, the size of the fused feature is set to $1,024$ (i.e., $k = 32$) for all the methods.

The results are shown in Table 2. First, Ours achieved the best accuracy in all the metrics. The improvement in RMSE from Linear to Ours reaches 11.4%, which highlights the effectiveness of our model. Second, the superiority of Ours to Dense Co-attention and Bilinear justifies our fusion model. Bilinear provides limited performance improvement. This may be because the binary quadratic form assumed in Bilinear, which fuses two features evenly, may hinder performance improvement when there is a significant gap in performance between the features to be fused (see Table 1). The fact that Dense Co-attention is more promising than Bilinear suggests the effectiveness of dense co-attention, and the superiority of Ours over Dense Co-attention emphasizes the validity of our formulation that integrates the dense co-attention mechanism with unary quadratic form. Moreover, we found that the results degraded by about 30% when the two features were swapped, supporting our assumption that there is an imbalance in the effectiveness of the features and that they should not be fused evenly. Third, although SELDnet and Echo-Net did not perform as well as our network, the differences from Median show that their predictions are reasonable. Finally, we found that Ours and Dense Co-attention are faster than Linear and Bilinear; the running times for a single forward pass on a single Tesla P100 GPU were 21.9 ms, 22.3 ms, 24.1 ms, and 26.9 ms, in that order.



(a) Studio (b) Replica

Fig. 5: **RMSE vs. Feature Dimension k.**

Results on Replica. Replica [26] is a synthetic dataset containing RGB images, depth maps, and echoes rendered/generated using Habitat-Sim [27] and SoundSpaces [28] from 3D scans of 18 scenes including apartments, offices, and hotels. We followed the protocol used in [14]; we used 15 scenes for training and the remaining 3 scenes for testing. The echo at each location was obtained by convolving a pre-computed binaural RIR with a 3ms chirp up to 20kHz. The first 60ms of the echo is used to extract the features.

The results are shown in Table 3. The overall trend is similar to the case of Studio; Dense Co-attention performs better than Bilinear, and Ours outperforms both. Ours is also better than Echo-Net [14]. These results again support the validity of our model.

Qualitative Results. Fig. 4 shows some examples of the predictions. Ours better recovers the ground truth depth maps than Dense Co-attention [24]. In Studio, Dense Co-attention failed to capture the changes in depth and tended to output flatter maps than the ground truth. In contrast, Ours was able to capture small changes, such as the shelf placed on the left side of the scene in the second row. The similar tendency is observed in Replica; Ours could predict the depth maps of various dynamic ranges more accurately than Dense Co-attention. These examples highlight the effectiveness of our model.

Ablation Study. Tables 4 and 5 show the difference in performance of the three variants of our feature fusion block (see Fig. 3). Overall, the performance tends to improve as we add more components for both datasets (Base+FFN+SA $\geq$ Base+FFN $\geq$ Base). In Replica, the difference between Base+FFN and Base+FFN+SA is not very significant, which suggests that depending on the data set, a lighter architecture can be chosen without sacrificing performance.

Sensitivity to Feature Dimension k. We analyzed the impact of the size of the fused feature k on RMSE in Fig. 5. As k increases, the accuracy tends to improve. Ours is consistently better than Dense Co-attention, which shows the strong effectiveness of our method.

5. CONCLUSION

We addressed the problem of echo-based depth estimation. We showed that fusing angular spectrum and spectrogram improves the depth estimation accuracy and proposed a new bilinear feature fusion model that incorporates dense co-attention. Our experiments proved that our model outperforms the state-of-the-art methods.

6. REFERENCES

[1] David Eigen, Christian Puhrsch, and Rob Fergus, "Depth map prediction from a single image using a multi-scale deep network," in *Proc. NeurIPS*, 2014.

[2] Iro Laina, Christian Rupprecht, and Vasileios Belagiannis, "Deeper depth prediction with fully convolutional residual networks," in *Proc. 3DV*, 2016.

[3] Fangchang Ma and Sertac Karaman, "Sparse-to-Dense: Depth prediction from sparse depth samples and a single image," in *Proc. ICRA*, 2018.

[4] Go Irie, Daiki Ikami, Takahito Kawanishi, and Kunio Kashino, "Cascaded transposed long-range convolutions for monocular depth estimation," in *Proc. ACCV*, 2020.

[5] Ariel Ephrat, Inbar Mosseri, Oran Lang, Tali Dekel, Kevin Wilson, Avinatan Hassidim, William T. Freeman, and Michael Rubinstein, "Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation," in *Proc. SIGGRAPH*, 2018.

[6] Chuang Gan, Hang Zhao, Peihao Chen, David Cox, and Antonio Torralba, "Self-supervised moving vehicle tracking with stereo sound," in *Proc. ICCV*, 2019.

[7] Go Irie, Mirela Ostrek, Haochen Wang, Hirokazu Kameoka, Akisato Kimura, Takahito Kawanishi, and Kunio Kashino, "Seeing through sounds: Predicting visual semantic segmenation results from multichannel audio signals," in *Proc. ICASSP*, 2019.

[8] Arda Senocak, Tae-Hyun Oh, Junsik Kim, Ming-Hsuan Yang, and In So Kweon, "Learning to localize sound sources in visual scenes: Analysis and applications," *PAMI*, vol. 43, no. 5, pp. 1605–1619, 2021.

[9] Ivan Dokmanić, Reza Parhizkar, Andreas Walther, Yue M. Lu, and Martin Vetterli, "Acoustic echoes reveal room shape," *PNAS*, vol. 110, pp. 12186–12191, 2013.

[10] Alex Turpin, Valentin Kapitany, Jack Radford, Davide Rovelli, Kevin Mitchell, Ashley Lyons, Ilya Starshynov, and Daniele Faccio, "3D imaging from multipath temporal echoes," *Physical Review Letters*, vol. 126, 2021.

[11] Jesper Haahr Christensen, Sascha Hornauer, and Stella X. Yu, "BatVision: Learning to see 3D spatial layout with two ears," in *Proc. ICRA*, 2020.

[12] Arun Balajee Vasudevan, Dengxin Dai, and Luc Van Gool, "Semantic object prediction and spatial sound super-resolution with binaural sounds," in *Proc. ECCV*, 2020.

[13] Ruohan Gao, Changan Chen, Ziad Al-Halah, Carl Schissler, and Kristen Grauman, "VisualEchoes: Spatial image representation learning through echolocation," in *Proc. ECCV*, 2020.

[14] Kranti Kumar Parida, Siddharth Srivastava, and Gaurav Sharma, "Beyond image to depth: Improving depth prediction using echoes," in *Proc. CVPR*, 2021.

[15] C. Knapp and G. Carter, "The generalized cross-correlation method for estimation of time delay," *IEEE TASSP*, vol. 24, pp. 320–327, 1976.

[16] Charles Blandin, Alexey Ozerov, and Emmanuel Vincent, "Multi-source TDOA estimation in reverberant audio using angular spectra and clustering," *Signal Processing*, vol. 92, pp. 1950–1960, 2012.

[17] Sharath Adavanne, Archontis Politis, Joonas Nikunen, and Tuomas Virtanen, "Sound event localization and detection of overlapping sources using convolutional recurrent neural networks," *IEEE Journal of Selected Topics in Signal Processing (JSTSP)*, vol. 13, pp. 34–48, 2018.

[18] Tsung-Yu Lin, Aruni RoyChowdhury, and Subhransu Maji, "Bilinear convolutional neural networks for fine-grained visual recognition," *PAMI*, vol. 40, no. 6, pp. 1309–1322, 2018.

[19] J.-H. Kim, K.-W. On, J. Kim, J.-W. Ha, and B.-T. Zhang, "Hadamard product for low-rank bilinear pooling," in *Proc. ICLR*, 2017.

[20] Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang, "Bilinear attention networks," in *Proc. NeurIPS*, 2018.

[21] Zhou Yu, Jun Yu, Jianping Fan, and Dacheng Tao, "Multi-modal factorized bilinear pooling with co-attention learning for visual question answering," in *Proc. ICCV*, 2017.

[22] H. Ben-younes, R. Cadene, M. Cord, and N. Thome, "MUTAN: Multimodal Tucker fusion for visual question answering," in *Proc. ICCV*, 2017.

[23] Duy-Kien Nguyen and Takayuki Okatani, "Improved fusion of visual and language representations by dense symmetric co-attention for visual question answering," in *Proc. CVPR*, 2018.

[24] Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian, "Deep modular co-attention networks for visual question answering," in *Proc. CVPR*, 2019.

[25] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *Proc. NeurIPS*, 2017.

[26] Julian Straub et al., "The Replica dataset: A digital replica of indoor spaces," in *arXiv Preprint, 1906.05797*, 2019.

[27] Manolis Savva et al., "Habitat: A platform for embodied AI research," in *Proc. ICCV*, 2019.

[28] Changan Chen, Unnat Jain, Carl Schissler, Sebastia V. Amengual Gari, Ziad Al-Halah, Vamsi Krishna Ithapu, Philip Robinson, and Kristen Grauman, "SoundSpaces: Audio-visual navigation in 3D environments," in *Proc. ECCV*, 2020.