Evaluating Visual-to-Echo Distillation for Binaural Depth Prediction beyond Simulations

Nazrul Ismail [*1,2] **Owais Ahmed Malik** [*3] **Ong Wee Hong** [*1,2]

Abstract

Echo reflections encode physical cues about object distance, geometry, and surface material that are useful for spatial reasoning. Prior works proposed to incorporate echo reflections as a modality into depth prediction through direct fusion or cross-modal knowledge distillation from vision to audio, but evaluation has been confined to simulated environments such as Replica and Matterport3D, leaving real-world viability untested. In this short paper, we evaluate Visual2Echo Compositional Contrastive Learning (V2E-CCL), a knowledge distillation framework that predicts depth using binaural echoes by aligning cross-modal representations in a shared latent space, on real binaural recordings from the BatVision dataset. To our knowledge this is the first evaluation of vision-to-echo distillation on real binaural recordings, indicating that the benefit of cross-modal distillation previously observed only in simulation also holds on real-world echoes. We further analyse failure modes specific to real echo capture.

1. Introduction

Acoustic signals propagate independently of visual conditions, capturing geometric information through echo timing, amplitude, and spectral characteristics (Christensen et al., 2020a) for depth estimation. Although monocular depth estimation from vision has advanced rapidly with methods such as MoGE-V1 &V2 (Wang et al., 2025a;b), the audio-based depth estimation domain remains underexplored. Recent work addresses audio-only depth estimation through cross-modal knowledge distillation from vision to audio. In particular, Visual2Echo Compositional Contrastive Learning (V2E-CCL) (Ismail et al., 2026) transfers depth and material knowledge from vision teachers to a binaural-echo student via a Compositional Embedding (CE) module that refines teacher features with audio cues and a Compositional Contrastive Learning (CCL) objective that aligns cross-modal representations in a shared latent space. However, existing evaluations of V2E-CCL are limited to simulated environments (Replica, Matterport3D). Simulation captures unrealistic impulse responses but omits microphone characteristics, hardware noise, and the long-tailed reverberation often present in physical spaces. Hence, it remains an open question whether the gains attributed to vision-to-echo distillation transfer on real recordings. In this extended abstract, we evaluate V2E-CCL on the BatVision dataset (Christensen et al., 2020a), which contains real binaural recordings collected indoors, without architectural or hyperparameter changes.

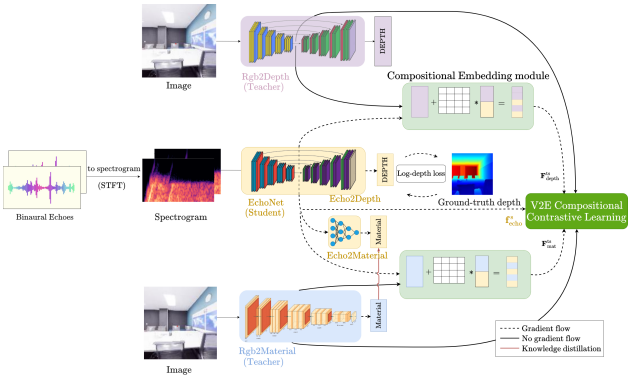

Figure 1. Overview of the V2E-CCL pipeline. Given a binaural chirp, latent embeddings are encoded by teacher and student networks; compositional embeddings and contrastive alignment transfer vision-domain depth and material knowledge to the audio student. Adapted from V2E-CCL paper

2. Related Work

2.1. Audio-visual and Audio-only depth estimation

Knowledge distillation was introduced by (Hinton et al., 2015) for transferring information from a teacher to a stu-

[1]School of Digital Science, Universiti Brunei Darussalam [2]Robotic and Intelligent Systems Laboratory (RoboLab) [3]Atlantic Technological University. Correspondence to: Nazrul Ismail <23h1701@ubd.edu.bn>, Owais Ahmed Malik <owais.malik@atu.ie>, Ong Wee Hong <weehong.ong@ubd.edu.bn>.

Proceedings of the 43^{rd} International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026.

dent through output distributions. (Gupta et al., 2016) later extended this across modalities. Building on cross-modal distillation, several methods have been applied to audio-visual depth estimation (Yun et al., 2023; Gao et al., 2020; Parida et al., 2021; Zhu et al., 2022), but performance degrades when visual input is unavailable at inference.

Audio-only methods exploit the Interaural Time Difference (Wightman & Kistler, 1992) for geometric extraction. (Christensen et al., 2020b) translated binaural spectrograms into depth via conditional adversarial networks, and further iteration (Christensen et al., 2020a) improved this with GCC-PHAT features. (Dai et al., 2022) used a vision network to produce pseudo-ground-truth supervision for outdoor binaural depth. More recently, Zhang et al. (Zhang et al., 2025) proposed EchoDiffusion, which conditions a latent diffusion model on Wav2Vec(Baevski et al., 2020) waveform embeddings to refine depth predictions. Liu et al. (Liu et al., 2025) introduced SAGENet, which encodes 2D geometric cues extracted via GCC-PHAT with PointNet and uses learnable queries initialised from angular spectrum peaks to focus on early-reflection features. However, none of these methods leverage cross-modal distillation from vision; they learn entirely from audio supervision. (Yun et al., 2023) proposed Spatial Alignment Mappping (SAM) to distill spatial knowledge from vision to audio via feature-level alignment, but on panoramic (360°) inputs for semantic localization.

2.2. Real-world evaluation and sim-to-real transfer

Existing audio-based depth estimators (Dai et al., 2022; Zhang et al., 2025; Yun et al., 2023; Ismail et al., 2026) are predominantly trained and evaluated on simulated environments such as Replica and Matterport3D using SoundSpaces (Chen et al., 2023). Simulation yields clean, idealised signals that neglects microphone properties, and thus do not reflect real-world conditions. The BatVision dataset (Christensen et al., 2020a) provides real binaural recordingscollected indoors and has been used to benchmark early echo-only baselines (Christensen et al., 2020a; Dai et al., 2022), but no prior work has evaluated vision-to-echo distillation methods on real binaural recordings.

3. Approach

We briefly recap the V2E-CCL framework (Ismail et al., 2026). We refer readers to the original paper for full architectural details and derivations.

3.1. Framework Overview

Given a dataset $X = \{I_i, A_i, D_i\}_{i=1}^{N}$ of RGB images I_i, binaural echoes A_i, and ground-truth depth maps D_i, V2E-CCL trains two parallel network streams (Figure 2): *teacher* networks (Rgb2Depth, Rgb2Material) producing visual features $\mathbf{f}_x^t$, and *student* networks (Echo2Depth, Echo2Material) producing an audio embedding $\mathbf{f}_{\text{echo}}^s$, where $x \in \{\text{depth}, \text{mat}\}$. The student follows a U-Net architecture taking concatenated left/right STFT spectrograms as input. Teachers are frozen pretrained models (Parida et al., 2021; Bell et al., 2015).

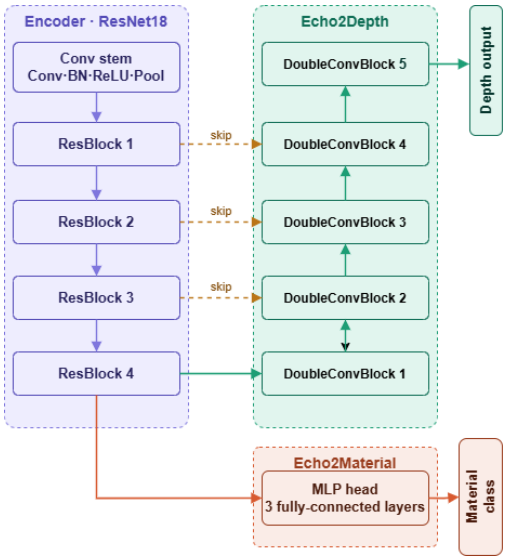

Figure 2. Network achitecture of proposed Echo2Depth and Echo2Material student network

3.2. Compositional Embedding and Contrastive Learning

To bridge the vision-audio domain gap, the framework introduces two complementary components. The **Compositional Embedding (CE)** refines teacher features with audio cues:

$$\mathbf{F}_x^{ts} = \mathbf{f}_x^t + \mathcal{F}_{CE_x}(\mathbf{f}_x^t, \mathbf{f}_{\text{echo}}^s), \tag{1}$$

where $\mathcal{F}_{CE_x}$ learns a residual from teacher–student feature interactions via normalization, concatenation, and linear projection. **Compositional Contrastive Learning (CCL)** projects student, teacher, and composed features into a shared latent space using a set of MLPs $\mathcal{F}_{CCL}$, yielding embeddings $\mathbf{z}_{\text{echo}}^s, \mathbf{z}_x^t, \mathbf{z}_x^{ts}$. These are aligned with the student echo embedding as anchor:

$$\mathcal{L}_{CCL} = \sum_{x \in \{\text{depth,mat}\}} \Big[(1 - \text{sim}(\mathbf{z}_{\text{echo}}^s, \mathbf{z}_x^t)) + (1 - \text{sim}(\mathbf{z}_{\text{echo}}^s, \mathbf{z}_x^{ts})) \Big]. \tag{2}$$

3.3. Training Objective

The full objective combines a base depth + material loss $\mathcal{L}_{\text{base}}$, a compositional depth loss $\mathcal{L}_{\text{comp}}$ supervising the composed embedding, and the CCL alignment:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{base}} + \lambda_{\text{comp}} \mathcal{L}_{\text{comp}} + \lambda_{CCL} \mathcal{L}_{CCL}. \tag{3}$$

Table 1. Depth estimation on the BatVision dataset (BV1). All methods use audio only. Best results in bold, second best underlined.

Method	RMSE↓	REL↓	log10↓	$\delta < 1.25$↑	$\delta < 1.25^2$↑	$\delta < 1.25^3$↑
EchoNet	0.132	0.416	<u>0.155</u>	0.478	<u>0.717</u>	<u>0.844</u>
EchoDiffusion	0.903	0.398	0.170	0.373	0.654	0.828
BatNet	0.131	0.407	<u>0.155</u>	<u>0.480</u>	0.713	0.841
SAGENet	<u>0.123</u>	<u>0.361</u>	0.206	0.448	0.636	0.750
V2E-CCL	**0.120**	**0.304**	**0.125**	**0.569**	**0.821**	**0.924**

The vision teachers remain frozen throughout. We follow the architecture and hyperparameters of (Ismail et al., 2026) without modification.

4. Experimental Setup

Dataset. We evaluate on the BV1 subset of the BatVision dataset (Brunetto et al., 2023), collected at UC Berkeley with a wheeled robot emitting audible sine-swept chirps (20Hz-20kHz, 3 ms) through a forward-facing JBL speaker. Binaural echoes (72.5 ms, 44.1 kHz) are captured by ear-shaped microphones spaced 23.5 cm apart, synchronised with RGB-D images from a ZED stereo camera covering hallways, open areas, conference rooms, and offices. BV1 contains 52,220 instances (39,564 train, 7,618 val, 5,038 test).

Preprocessing. We convert raw binaural audio to spectrograms via Short-Time Fourier Transform (STFT) with 512 frequency bins, 64 window, and a hop length of 16, following the preprocessing of (Brunetto et al., 2023). Unlike the original BatNet configuration (Christensen et al., 2020a), we do not resize the spectrogram to 256×256 instead, we retain the native resolution and make the network agnostic to input size. Depth maps are clipped to 12m following the dataset convention.

Metrics. Following (Eigen et al., 2014), we report RMSE, mean relative error (REL), and threshold accuracy $\delta_{1.25^i}$ for $i \in \{1, 2, 3\}$, where $\delta_{1.25^i}$ is the fraction of pixels with $\max\left(\frac{\hat{\mathbf{D}}}{\mathbf{D}}, \frac{\mathbf{D}}{\hat{\mathbf{D}}}\right) < 1.25^i$, where $i \in 1, 2, 3$

5. Results and Discussion

We compare against prior arts in audio-only depth estimation method such as BatNet (Christensen et al., 2020a), EchoNet (Parida et al., 2021), EchoDiffusion (Zhang et al., 2025) and SAGENet (Liu et al., 2025). We have omit SAM (Yun et al., 2023), as the source code is not publicaly available. To ensure fairness, all methods are fine-tuned on the training set with the hyperparameter settings as (Ismail et al., 2026). Table 1 reports depth estimation results on the BatVision (BV1) dataset and observed the following:

BatNet (Christensen et al., 2020a) and EchoNet (Christensen et al., 2020b) obtain similar RMSE (0.131 and 0.132) despite different designs: BatNet pairs a U-Net decoder with an adversarial PatchGAN discriminator, while EchoNet uses a separate echo encoder–decoder stripped of its material-aware attention module (which requires RGB input unavailable at inference). Both learn an implicit spectrogram-to-depth mapping without explicit geometric modelling, which limits their ability to resolve fine spatial structure. EchoDiffusion (Zhang et al., 2025), which conditions a latent diffusion process on Wav2Vec (Baevski et al., 2020) waveform embeddings, yields a competitive REL (0.398) yet an RMSE an order of magnitude higher than the other methods (0.903). This discrepancy suggests the generative prior preserves relative depth ordering but fails to recover absolute scale on real recordings. We attribute this to two factors: the Wav2Vec encoder is pretrained on speech rather than echo signals, limiting the usefulness of its waveform embeddings for spatial reasoning; and the diffusion process learns the distribution of simulated depth maps, which does not transfer to the acoustic characteristics of real hardware.

V2E-CCL reduces RMSE to 0.120 and REL to 0.304, a relative improvement of 15.8% in REL over the next-best method (SAGENet, 0.361). The largest gains appear in threshold accuracy: $\delta_{1.25}$ reaches 0.569, surpassing BatNet by 18.5%, and $\delta_{1.25^3}$ reaches 0.924. These improvements indicate that the visual representations distilled via Compositional Contrastive Learning encode structural priors, object boundaries, surface layout that pure audio training does not capture. A full per-component breakdown performance is in Appendix A.

Figure 3 shows audio-only predictions on unseen indoor scenes: RGB and ground-truth (GT) depth are shown for reference only. Despite receiving only binaural spectrograms, the distilled student recovers dominant scene geometry in metric scale, correctly localizing corridor depth axes (rows 1-3) and near-to-far floor transitions. Where the GT depth sensor fails on specular surfaces, thin structures, or out-of-range areas (navy regions), the model produces smooth, plausible depth, since acoustic propagation depends on scene geometry rather than surface optics. We observe several limitations in the model. Firstly, the predictions are band-limited: sharp boundaries and small or acoustically transparent objects (stacked chairs in row 6, table legs) are

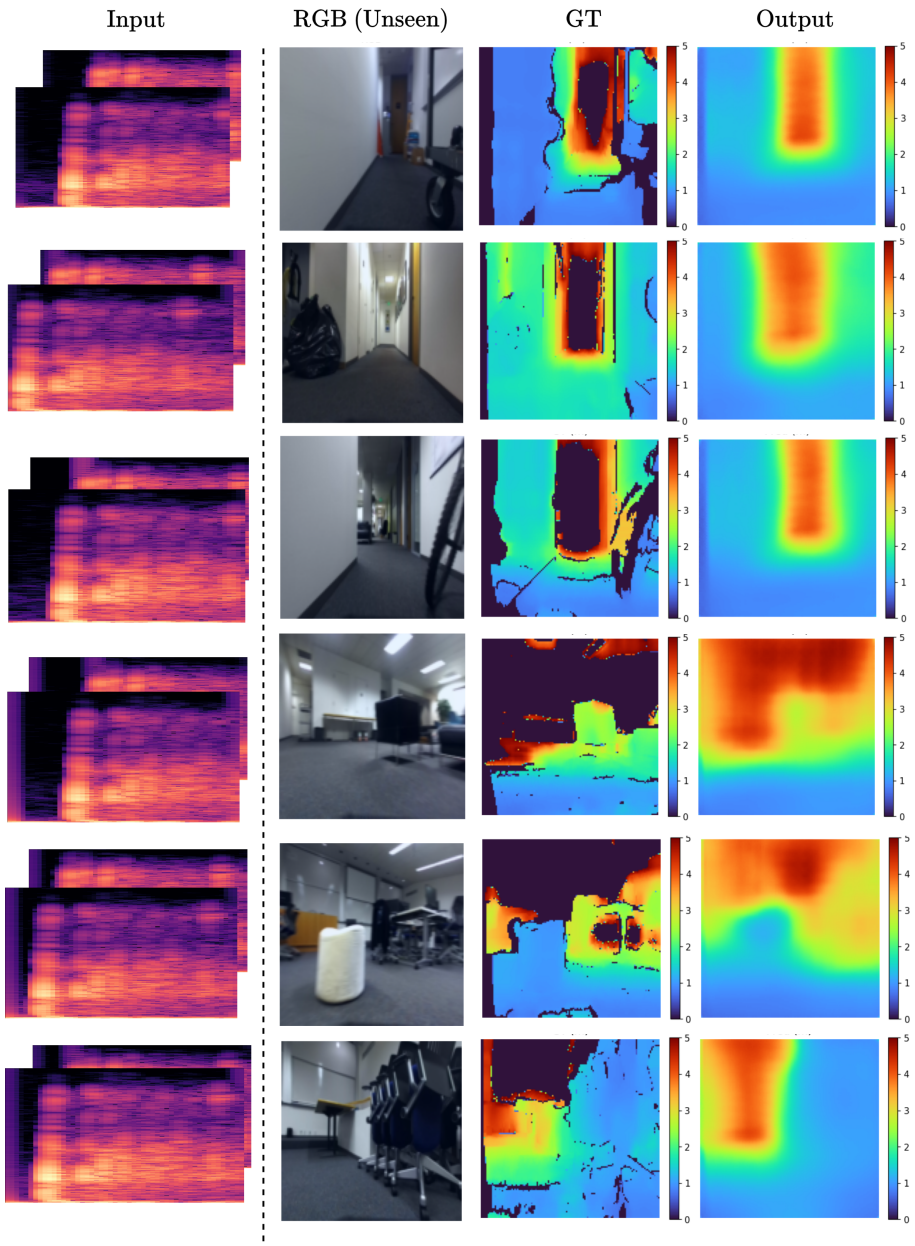

Figure 3. Qualitative results of V2E-CCL. Each row shows (left to right): RGB input, ground-truth depth and predicted depth. Black regions indicate missing depth returns.

smoothed into the surrounding depth gradient, as such structures produce no distinguishable binaural signature. We also observe mild far-field over-extension in cluttered rooms (rows 4-6), consistent with an over-smoothing bias.

6. Conclusion

We evaluated V2E-CCL, a vision-to-echo distillation framework, on real binaural recordings from the BatVision (BV1) dataset without architectural or hyperparameter changes. The distilled student outperforms all audio-only baselines, reducing REL by 15.8% over the next-best method and achieving the highest threshold accuracy across all three δ levels. The main limitation is spatial resolution: fine object boundaries remain unresolved due to the low spatial bandwidth of binaural echoes. Future work includes collecting biosonar recordings, evaluating on the outdoor environments, and combining distilled audio features with lightweight visual encoders for robust audio-visual depth estimation.

References

Baevski, A., Zhou, H., Mohamed, A., and Auli, M. wav2vec 2.0: A framework for self-supervised learning of speech representations, 2020. URL https://arxiv.org/abs/2006.11477.

Bell, S., Upchurch, P., Snavely, N., and Bala, K. Material recognition in the wild with the materials in context database. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3479–3487, 2015.

Brunetto, A., Hornauer, S., Yu, S. X., and Moutarde, F. The audio-visual batvision dataset for research on sight and sound. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 1–8, 2023. doi: 10.1109/IROS55552.2023.10341715.

Chen, C., Schissler, C., Garg, S., Kobernik, P., Clegg, A., Calamia, P., Batra, D., Robinson, P. W., and Grauman, K. Soundspaces 2.0: A simulation platform for visual-acoustic learning, 2023. URL https://arxiv.org/abs/2206.08312.

Christensen, J. H., Hornauer, S., and Yu, S. Batvision with gcc-phat features for better sound to vision predictions. *arXiv preprint arXiv:2006.07995*, 2020a.

Christensen, J. H., Hornauer, S., and Yu, S. X. Batvision: Learning to see 3d spatial layout with two ears. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1581–1587, 2020b. doi: 10.1109/ICRA40945.2020.9196934.

Dai, D., Vasudevan, A. B., Matas, J., and Van Gool, L. Binaural soundnet: predicting semantics, depth and motion with binaural sounds. *IEEE transactions on pattern analysis and machine intelligence*, 45(1):123–136, 2022.

Eigen, D., Puhrsch, C., and Fergus, R. Depth map prediction from a single image using a multi-scale deep network. volume 27, 2014.

Gao, R., Chen, C., Al-Halab, Z., Schissler, C., and Grauman, K. Visualechoes: Spatial image representation learning through echolocation. In *ECCV*, 2020.

Gupta, S., Hoffman, J., and Malik, J. Cross modal distillation for supervision transfer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2827–2836, 2016.

Hinton, G., Vinyals, O., and Dean, J. Distilling the knowledge in a neural network, 2015. URL https://arxiv.org/abs/1503.02531.

Ismail, N., Malik, O. A., and Hong, O. W. Visual2echo compositional contrastive learning (v2e-ccl): Binaural knowledge distilled network for depth prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings*, pp. 6019–6028, June 2026.

Liu, G., Cui, W., Xi, Y., Yang, L., Hu, P., Kong, H., and Wang, Z. Sagenet: Binaural echo-based 3d depth estimation with sparse angular queries and refined geometric cues. In *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 6113–6120, 2025. doi: 10.1109/IROS60139.2025.11245854.

Parida, K., Srivastava, S., and Sharma, G. Beyond image to depth: Improving depth prediction using echoes. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021.

Wang, R., Xu, S., Dai, C., Xiang, J., Deng, Y., Tong, X., and Yang, J. Moge: Unlocking accurate monocular geometry estimation for open-domain images with optimal training supervision. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 5261–5271, 2025a.

Wang, R., Xu, S., Dong, Y., Deng, Y., Xiang, J., Lv, Z., Sun, G., Tong, X., and Yang, J. Moge-2: Accurate monocular geometry with metric scale and sharp details, 2025b. URL https://arxiv.org/abs/2507.02546.

Wightman, F. L. and Kistler, D. J. The dominant role of low-frequency interaural time differences in sound localization. *The Journal of the Acoustical Society of America*, 91(3):1648–1661, 1992.

Yun, H., Na, J., and Kim, G. Dense 2d-3d indoor prediction with sound via aligned cross-modal distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7863–7872, 2023.

Zhang, W., Yin, J., Ma, L., Yu, P., Jiang, X., Tian, Z., and Xu, M. Echodiffusion: Waveform conditioned diffusion models for echo-based depth estimation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(21):22578–22586, Apr. 2025. doi: 10.1609/aaai.v39i21.34416. URL https://ojs.aaai.org/index.php/AAAI/article/view/34416.

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

A. Component Ablation

To further evaluate on V2E-CCL improvement stems from cross-modal distillation rather than from architectural backbone differences or the training protocol, we ablate the two distillation components on the BV1 test split. All variants share the same student backbone, input preprocessing, optimizer, and training budget, and follow the same seed protocol as Section 5; the rows differ only in which distillation component is active. "No distillation" trains the student from depth/material supervision alone, with the vision teachers and both bridging modules removed.

Table 2. Ablation of the V2E-CCL distillation components on BatVision (BV1). All variants use the same student backbone and training budget. Best results in **bold**.

Variant	RMSE↓	REL↓	log10↓	$\delta < 1.25 \uparrow$
No distillation	0.182	0.450	0.325	0.342
+ CE only	0.152	0.423	0.225	0.358
+ CCL only	0.132	0.455	0.289	0.484
Full (CE + CCL)	**0.120**	**0.304**	**0.125**	**0.569**

The "No distillation" student is the weakest variant on every metric (REL 0.450, RMSE 0.182), confirming that the gain reported in Table 1 is driven by the distilled vision priors rather than by model capacity or optimisation, since the backbone and training budget are held fixed across all rows. The two complementary modules: adding only the Compositional Embedding (CE) most improves the scale-sensitive errors (log10 $0.325 \rightarrow 0.225$, REL $0.450 \rightarrow 0.423$), whereas adding only the Compositional Contrastive Learning (CCL) improves the structural metrics (RMSE $0.182 \rightarrow 0.132$, $\delta < 1.25$ $0.342 \rightarrow 0.484$). Notably, CCL alone does not reduce REL (0.455, on par with no distillation), indicating that feature alignment sharpens scene structure but, on its own, does not recover accurate relative depth where CE provides the complementary signal. Combining both yields the best result on all four metrics and drops REL to 0.304, well below either component alone (≥ 0.42) with the two mechanisms are complementary rather than redundant.