SUPPLEMENTARY MATERIALS for ProtoDepth

Patrick Rim Hyoungseob Park S. Gangopadhyay Ziyao Zeng Younjoon Chung Alex Wong

Yale Vision Lab

1. Domain Descriptor Analysis

To better understand the performance of ProtoDepth in the agnostic setting, we analyze the relationship between sample descriptors and learned domain descriptors using the t-SNE visualization shown in Fig. 1. This analysis is based on the KBNet model trained on the indoor dataset sequence, and it reveals insights into how ProtoDepth selects prototype sets during inference.

Each sample descriptor is computed deterministically using global average pooling (GAP) over the bottleneck features of the frozen model. Since the encoder layers are always frozen during training, the sample descriptors of a certain dataset are a lifelong deterministic function of the features present in that dataset. The domain descriptors, on the other hand, are learned during training to align with the sample descriptors of their respective datasets, enabling effective prototype set selection.

The visualization demonstrates that the majority of sample descriptors for each dataset cluster closely around their respective domain descriptors. This alignment confirms that the training process successfully associates each dataset with its corresponding descriptor at test-time, ensuring accurate prototype selection in the agnostic setting. However, it is noteworthy that some sample descriptors are closer to domain descriptors of other datasets. For example, non-negligible subsets of VOID sample descriptors appear to have higher affinity with the NYUv2 and ScanNet domain descriptors. This overlap introduces a degree of generalization, allowing the model to select prototypes from a different domain if they better align with the input sample's features.

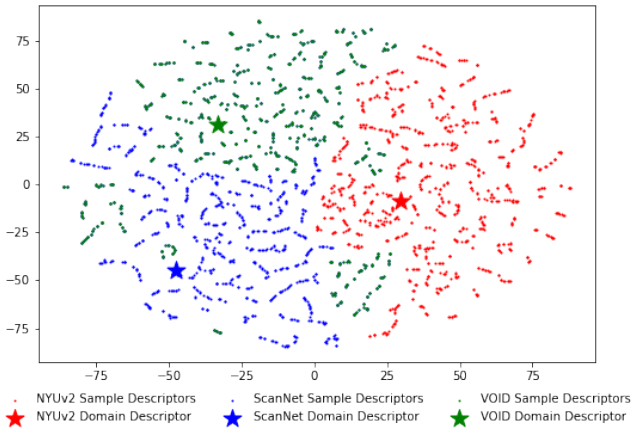

Figure 1. **t-SNE plot** of KBNet sample descriptors for indoor validation datasets (NYUv2, ScanNet, VOID) and their respective domain descriptors learned during training in the agnostic setting. While most sample descriptors align closely with their respective domain descriptors, some overlap enables cross-domain generalization, improving performance in challenging scenarios.

This ability to adaptively select domain descriptors explains why ProtoDepth achieves superior performance in the agnostic setting than in the incremental setting for certain metrics. By relaxing the constraint of fixed domain identity during inference, the agnostic setting enables the model to exploit cross-domain generalization in cases where overlapping features exist between datasets. While this occurs in only a minority of scenarios, it underscores the utility of allowing the model to flexibly choose prototypes, particularly in instances where the distributional characteristics of one domain may overlap with those of another.

Most importantly, the t-SNE plot clearly illustrates that, despite the presence of some overlap, the domain descriptors remain sufficiently distinct to avoid significant performance degradation due to incorrect prototype selection. Instead, this overlap even facilitates generalization (see Tab. 4), enabling the model to leverage features from neighboring domains to improve depth completion on difficult samples. This balance between dataset alignment and cross-domain generalization is central to ProtoDepth's ability to adapt to the challenging domain-agnostic setting.

2. Transformer Experiments

In the main paper, we stated that ProtoDepth is applicable to any model with a latent space, including both CNNs and transformers. To explore the applicability of ProtoDepth to transformer-based architectures, we adapted Uformer [24], a simple encoder-decoder model consisting entirely of transformer blocks, for depth completion. The model takes as input patchified versions of the image and sparse depth, where inputs from each modality are split into 14×14 patches and

Setting	Method	Average Forgetting (%) MAE	RMSE	iMAE	iRMSE	Average Performance MAE	RMSE	iMAE	iRMSE	SPTO MAE	RMSE	iMAE	iRMSE
(1) KBNet	ANCL [11]	9.73	10.75	5.58	16.38	56.89	120.30	13.77	31.85	47.32	103.42	13.88	32.76
	CMP [10]	5.39	5.11	8.25	7.90	55.92	117.83	13.74	31.43	46.03	102.36	13.55	32.03
	Ours	**3.20**	**1.30**	**4.91**	**2.94**	**54.25**	**115.55**	**13.20**	**30.50**	**45.26**	**101.10**	**13.28**	**31.72**
(2) Uformer	Finetuned	87.94	73.61	110.98	852.79	183.24	302.99	51.07	297.92	137.20	238.95	49.54	142.33
	L2P [25]	57.07	43.84	50.82	58.24	171.74	273.75	46.90	121.30	139.08	231.88	51.98	156.41
	Ours	**37.15**	**25.50**	**31.86**	**17.04**	**161.62**	**255.54**	**42.38**	**79.34**	**133.36**	**220.68**	**44.74**	**84.31**
(3) KBNet	ANCL [11]	20.49	8.94	23.11	27.73	438.05	1795.76	1.21	3.56	503.53	2203.44	1.18	3.53
	CMP [10]	15.95	15.47	6.90	7.39	447.09	1887.14	1.09	3.19	507.90	2262.46	1.06	3.21
	Ours	**4.51**	**3.10**	**2.96**	**1.88**	**409.90**	**1730.72**	**1.04**	**3.04**	**478.79**	**2138.35**	**1.01**	**3.07**
(4) KBNet	ANCL [11]	35.10	35.31	18.13	10.04	313.71	1067.35	18.89	30.39	343.06	1129.85	18.66	30.20
	CMP [10]	31.60	36.04	12.63	9.90	307.87	1117.91	16.71	30.41	336.08	1142.94	16.66	30.23
	Ours	**20.61**	**18.75**	**9.79**	**6.25**	**277.04**	**985.58**	**15.07**	**28.42**	**309.57**	**1035.55**	**15.05**	**28.24**
(5) Uformer	L2P [25]	69.28	23.25	81.95	48.78	519.72	1458.78	25.65	36.21	470.84	1407.23	25.38	35.45
	Ours	**45.42**	**7.67**	**46.18**	**22.05**	**451.08**	**1252.88**	**22.34**	**32.00**	**401.95**	**1220.67**	**21.97**	**31.63**

Table 1. **Additional quantitative results** comparing to recent baselines on indoor, outdoor, and mixed sequences with backbone as denoted: (1,2) Indoor: NYUv2 → ScanNet → VOID (3) Outdoor: KITTI → Waymo → VKITTI (4,5) Mixed: KITTI → NYUv2 → Waymo

embedded as $N \times C$ tokens. We adapted Uformer for depth completion by implementing a dual-encoder structure, with one encoder processing image tokens and the other processing sparse depth tokens. Each encoder contains four transformer blocks. After being processed by the encoders, the tokens from both modalities are concatenated and fed into a shared decoder with four additional transformer blocks. Consistent with the CNN-based models used in the main paper, skip connections are included between each encoder block and its corresponding decoder block, allowing multi-scale features to flow between the encoders and decoder.

For ProtoDepth-A and ProtoDepth, we implemented our method in the exact same way as we do for CNN-based models, applying prototype sets to the latent space layers, i.e., the bottleneck and skip connections. The prototype sets learn global (multiplicative) and local (additive) biases for each layer, adapting the frozen transformer layers to each new dataset while mitigating forgetting. This demonstrates that ProtoDepth is fully architecture-agnostic and can be seamlessly applied to both CNNs and transformers.

A notable inclusion in this section is the prompt-based method L2P [25] (Learning to Prompt), which serves as a representative baseline for prompt-based methods. Prompt-based continual learning methods were not included in the main experiments because all existing unsupervised depth completion models are CNN-based, and prompt-based approaches, which operate by prepending prompts to tokenized inputs, are not applicable to CNNs, which operate directly on images without tokenization, which prevents the straightforward insertion of prompts into the input space. However, with the implementation of Uformer, a transformer-based model, we are now able to evaluate L2P, which is a foundational method for prompt-based continual learning.

For L2P, we implement the method as described in the original paper. Specifically, we use a prompt pool of size $M = 20$ and select $N = 5$ prompts for each input during training and inference. To adapt L2P for depth completion, we implement their loss term, which pulls selected keys closer to their corresponding queries, and incorporate it into our overall loss function (Eq. (1) in the main paper) with a weight of 0.5, as suggested in [25]. To evaluate in the domain-agnostic setting, where dataset identity is withheld at test time, we train $M = 20$ new prompts for each new dataset during continual training. At test-time, the model queries all existing learned prompts.

3. Additional Experiments

In Tab. 1-(2), we compare to L2P [Wang et al., CVPR '22] [25], a prompt-based method, where we adapt Uformer for unsupervised depth completion as no transformer-based model currently exists for this task. We have added comparisons to ANCL [Kim et al., CVPR '23] [11], an architecture-based method, and CMP [Kang et al., CVPR '24] [10], a rehearsal-based method, on the indoor Tab. 1-(1) and outdoor Tab. 1-(3) sequences using the KBNet backbone. ProtoDepth-A (*Ours*) outperforms all of these recent methods, reaffirming our findings.

In Tab. 1-(4,5), we add experiments in a mixed setting, where the dataset sequence transitions from outdoor to indoor and back to outdoor. We compare to ANCL, CMP, and L2P in this mixed setting and show that ProtoDepth-A outperforms all of these recent methods.

	MAE	RMSE	iMAE	iRMSE
Depth Anything [35]	49.22	88.74	21.22	51.22
Depth Pro [2]	43.06	93.36	20.80	52.24
Ours	**33.66**	**86.99**	**17.48**	**43.02**

Table 2. Comparison against depth estimation foundation models.

	MAE	RMSE	iMAE	iRMSE
Ours	686.86	**2024.42**	1.58	**3.52**
Upper Bound	**671.95**	2231.97	**1.34**	**3.52**

Table 3. Comparison against joint training (upper bound).

	MAE	RMSE	iMAE	iRMSE
Joint Training	2800.27	6284.63	6.06	11.23
ANCL [11]	2753.07	6195.09	5.69	10.86
CMP [10]	2885.82	6234.33	7.12	13.57
Ours	**2697.47**	**5966.57**	**5.40**	**10.58**

Table 4. Zero-shot generalization to nuScenes.

Tab. 2 shows that recent depth estimation unified/foundation models, Depth Pro [Bochkovskii et al., 2024] [2] and Depth Anything [Yang et al., CVPR '24] [35] (fit to metric scale via median scaling) do *not* outperform ProtoDepth-A (NYU $\rightarrow$ VOID) when evaluated on **VOID**. This validates the advantage of our method over direct depth estimation. Also of note, Depth Pro and Depth Anything are supervised and semi-supervised, while we are unsupervised.

In continual learning, joint training a larger model (e.g., transformer) on all datasets simultaneously serves as a performance upper bound. Tab. 3 shows that ProtoDepth-A achieves comparable mean performance to this upper bound on **{KITTI, Waymo, VKITTI}** using the adapted Uformer. Importantly, we address the scientific question of learning in a sequential manner, where one does not have access to all data at once or must learn a new dataset without breaking backwards-compatibility – a common real-world scenario.

Improved generalization to unseen datasets in the intersection of observed domains helps to motivate our method. Tab. 4 shows generalization to **nuScenes** (outdoor) after training on KITTI $\rightarrow$ Waymo $\rightarrow$ VKITTI. ProtoDepth-A outperforms joint training, ANCL, and CMP, demonstrating its ability to leverage domain-specific prototypes to enhance zero-shot generalization.

4. Dataset Details

Indoor datasets: The **NYU Depth V2** [19] ("NYUv2") dataset comprises 464 diverse indoor scenes from residential, office, and commercial environments captured using a Microsoft Kinect. It contains approximately 400,000 aligned RGB and depth image pairs with a resolution of 640×480. About 1,500 points are sampled for each sparse depth map using the Harris corner detector [9]. This dataset serves as a standard benchmark for indoor depth estimation tasks. For our indoor dataset sequence, we utilize NYUv2 as the initial dataset $\mathcal{D}_1$ for pretraining our depth completion models that are subsequently applied to indoor continual learning scenarios. The **VOID** [29] dataset presents sparse depth maps with $\approx 0.5\%$ density ($\approx$1,500 points), alongside RGB frames from various indoor settings such as laboratories, classrooms, and gardens, totaling approximately 58,000 frames (640×480) captured via XIVO [6]. VOID is designed to address challenges in areas with minimal texture and significant camera motion, key factors for assessing robustness in indoor depth completion tasks. **ScanNet** [5], a comprehensive indoor dataset, encompasses over 2.5 million frames paired with RGB-D data. Depth frames in ScanNet are captured at a resolution of 640×480 pixels, whereas the color frames have a higher resolution of 1296×968 pixels. Again, we use the Harris corner detector [9] to subsample $\approx$ 1,500 points for the sparse depth maps. We use a subset of the dataset with approximately 250,000 frames across 706 scenes. For all indoor datasets, we use a training crop size of 416×576. For evaluation, depth values across all of these indoor datasets are constrained between 0.2 and 5 meters.

Outdoor datasets: The **KITTI** [22] dataset is an established benchmark in autonomous driving that comprises over 93,000 stereo image pairs with a resolution of 1240×376 and sparse LiDAR depth maps ($\approx 5\%$ density), all synchronized and captured across diverse urban and rural landscapes using a Velodyne LiDAR sensor. KITTI is the initial dataset $\mathcal{D}_1$ for pretraining our depth completion models for the outdoor dataset sequence. The **Waymo Open Dataset** [21] ("Waymo") provides roughly 230,000 high-resolution frames (1920×1280 and 1920×1040) along with LiDAR point clouds, captured from scenes that encompass a broad spectrum of driving scenarios and conditions. For Waymo, the depth values during evaluation are capped between 0.001 and 80 meters and during training, a crop size of 800×640 is employed. The **Virtual KITTI** [7] ("VKITTI") dataset offers synthetic, altered re-creations of KITTI scenes captured from virtual worlds created in Unity, with over 21,000 frames at 1242×375 resolution and dense ground truth depth, facilitating the study of domain adaptation. We apply synthetic weather conditions and view rotations to simulate domain shifts that lead to forgetting. For KITTI and VKITTI, we restrict the depth values during evaluation to between 0.001 and 100 meters and utilize a depth cropping of 240×1216. During training, we use a crop size of 320×768.

Given the differences in image resolutions, crop sizes, and evaluation depths, in addition to the different types of scenes captured and sensors used to collect the datasets, we observe large domain gaps between datasets within each sequence, motivating the need for continual learning. We will release code for reproducibility.

5. Depth Completion Metrics

When we reference depth completion metrics in the main paper, we specifically mean the *error* metrics outlined below and formulated in Tab. 5. The metrics include Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Inverse Mean Absolute Error (iMAE), and Inverse Root Mean Squared Error (iRMSE). MAE measures the average $L1$ difference between predicted and ground-truth depths, providing a straightforward indication of prediction accuracy. RMSE measures $L2$ difference which gives higher weight to larger errors, making it sensitive to outliers and thus a robust measure for practical applications. iMAE and iRMSE, on the other hand, are particularly useful for scenarios where errors in smaller depth values are more critical, as they focus on the relative error in inverse depth. Collectively, these metrics allow for a comprehensive evaluation of a model's capability to predict depth from input data under varied environmental settings, e.g., indoor and outdoor. We note that lower values indicate better performance for all four error metrics. All results are reported in 'mm' (millimeters) unless otherwise specified, providing a clear metric standardization.

The results of our experiments are shown in Tab. 1, which compares ProtoDepth, ProtoDepth-A (agnostic setting), L2P, and full finetuning ("Finetuned") on the indoor dataset sequence. ProtoDepth achieves superior performance across all metrics, with zero forgetting in the incremental setting, with one exception: ProtoDepth-A outperforms ProtoDepth in one measure, SPTO for iRMSE, highlighting the benefits of its generalization capability. This result is consistent with our earlier observations: by allowing the model to select domain descriptors and prototype sets dynamically at test time, ProtoDepth-A can leverage features from overlapping domains to improve performance on ambiguous samples. This flexibility enables better generalization, which, in certain scenarios, can lead to improved outcomes compared to the fixed domain identity approach used in ProtoDepth.

Notably, ProtoDepth-A outperforms L2P in the agnostic setting, demonstrating the strength of prototype-based adaptation compared to prompt-based approaches. While L2P shows improvements over finetuning, it performs less well than ProtoDepth, which can be attributed to a fundamental limitation of prompt-based methods. These methods rely on learnable prompts or tokens to adapt frozen vision transformer models for continual learning, but there is no natural scale at which to discretize images or choose an appropriate prompt size, unlike the discrete text tokens used in natural language processing. In contrast, ProtoDepth's prototype-based approach eliminates the need for tokenized inputs, enabling it to operate directly in the latent feature space. This flexibility not only enhances its adaptability across diverse datasets but also allows it to be applied seamlessly to both transformers and convolutional neural networks, which are prevalent in unsupervised depth completion.

Metric	Definition
MAE ↓	$\frac{1}{\lvert\Omega\rvert}\sum_{x\in\Omega}\lvert\hat{d}(x)-d(x)\rvert$
RMSE ↓	$\left(\frac{1}{\lvert\Omega\rvert}\sum_{x\in\Omega}\lvert\hat{d}(x)-d(x)\rvert^2\right)^{1/2}$
iMAE ↓	$\frac{1}{\lvert\Omega\rvert}\sum_{x\in\Omega}\lvert 1/\hat{d}(x)-1/d(x)\rvert$
iRMSE ↓	$\left(\frac{1}{\lvert\Omega\rvert}\sum_{x\in\Omega}\lvert 1/\hat{d}(x)-1/d(x)\rvert^2\right)^{1/2}$

Table 5. **Error metrics for depth completion.** These metrics evaluate the accuracy of predicted depth values $\hat{d}(x)$ compared to ground truth depth values $d(x)$ over the set of pixels Ω.

6. Outdoor Prototype Set Sizes

We extend our investigation of prototype set sizes (i.e., number of prototypes) for the image and sparse depth layers (denoted as $N^{(I)}$ and $N^{(z)}$, respectively) to the outdoor dataset sequence. The results of these experiments are presented in Tab. 6. Based on the findings, we select $N^{(I)} = 25$ and $N^{(z)} = 10$ for the main experiments on the outdoor dataset sequence. Smaller set sizes demonstrate suboptimal performance, as they lack the capacity to adequately capture the diversity of features across datasets. Larger set sizes also result in performance degradation, likely due to the additional parameters learning noise and overfitting to the training data. The best performance is achieved when $N^{(I)} > N^{(z)}$, aligning with our observations in the indoor experiments. This can be attributed to the larger distributional shift between scenes in the image modality compared to the sparse depth modality [14]. For the bottleneck layer, which fuses features from both modalities, we again use $N^{(I)}$ as the prototype set size. As a baseline, we also report the performance of the frozen base model pretrained on KITTI ("Pretrained"), which has no additional parameters or further training. The poor results highlight the necessity of continual learning to adapt to non-stationary data distributions. For both indoor and outdoor settings, the prototype set size analysis is conducted using the KBNet model; we adopt the same prototype set sizes for all other models, as they all have a similar number of parameters.

7. Additional Qualitative Analysis

To illustrate the reduced forgetting achieved by ProtoDepth, we provide a qualitative comparison of depth predictions and error maps for all baseline methods on input samples from NYUv2 after continual training on ScanNet (Fig. 2 and Fig. 3). These figures demonstrate how ProtoDepth and ProtoDepth-A consistently outperform the baselines, specifically in reconstructing crowded indoor scenes with sparse depth measurements and challenging lighting conditions.

In Fig. 2, baseline methods such as Finetuned and EWC

				Waymo				VKITTI			
Method	$N^{(I)}$	$N^{(z)}$	# Params	MAE	RMSE	iMAE	iRMSE	MAE	RMSE	iMAE	iRMSE
Pretrained	-	-	0M (0%)	3930.68	6405.75	9.55	14.34	10527.70	18086.22	17.45	31.50
ProtoDepth	1	1	0.24M (3.5%)	587.92 ±61.20	1900.96 ±145.34	1.41 ±0.12	2.96 ±0.17	937.18 ±60.31	4027.53 ±47.08	1.92 ±0.38	5.82 ±0.42
	10	10	0.25M (3.7%)	524.76 ±37.18	1667.74 ±27.98	1.28 ±0.06	2.74 ±0.03	686.22 ±3.42	3638.20 ±12.29	0.90 ±0.04	3.50 ±0.07
	25	10	0.27M (3.9%)	**483.92** ±27.59	**1656.33** ±16.34	**1.19** ±0.04	**2.68** ±0.02	**676.28** ±4.64	**3608.42** ±16.61	**0.80** ±0.07	**3.25** ±0.24
	25	25	0.28M (4.0%)	508.60 ±20.36	1688.09 ±10.88	1.23 ±0.04	2.72 ±0.03	680.65 ±3.40	3614.61 ±14.82	0.87 ±0.05	3.51 ±0.19
	100	100	0.38M (5.5%)	522.39 ±50.06	1711.44 ±72.41	1.27 ±0.10	2.76 ±0.09	686.89 ±5.45	3635.01 ±27.57	0.93 ±0.09	3.53 ±0.08

Table 6. **Sensitivity study** of prototype set sizes ($N^{(I)}$ and $N^{(z)}$) on ProtoDepth using KBNet for outdoor datasets (Waymo and VKITTI). KBNet is pretrained on the initial dataset (KITTI). Parameter overhead is reported as a percentage of the full KBNet model's parameters. Smaller set sizes show suboptimal performance due to insufficient capacity to capture feature diversity, while larger set sizes also degrade performance, likely from overfitting and learning noise.

exhibit substantial forgetting, resulting in high error concentrations. Finetuned, in particular, struggles to retain photometric priors learned from NYUv2, evident in the poor reconstruction of furniture edges and flat areas with depth gradients. Replay performs marginally better but still fails to recover fine details, as its rehearsal mechanisms are insufficient to address the large distributional shift between NYUv2 and ScanNet. LwF shows improved performance, with fewer errors compared to Finetuned, EWC, and Replay. However, it fails to accurately reconstruct regions with sparse depth measurements (see Sparse Depth), such as the curtain.

ProtoDepth and ProtoDepth-A, on the other hand, produce high-fidelity depth predictions. ProtoDepth benefits from its prototype-based adaptation, effectively preserving features from NYUv2 while adapting to ScanNet. Notably, ProtoDepth-A exhibits comparable performance and even outperforms ProtoDepth in reconstructing certain regions, such as the smooth surface of the curtain. This improvement is due to ProtoDepth-A's generalization capability, which allows it to dynamically select prototype sets from overlapping domains based on the affinity of domain descriptors, thereby enhancing its ability to handle ambiguous inputs.

Fig. 3 reinforces these observations with a second example. Once again, baseline methods exhibit significant forgetting, with Finetuned, EWC, and LwF producing poor depth predictions. In contrast, ProtoDepth and ProtoDepth-A produce high-fidelity reconstructions. The well-defined edges between the furniture, floor, and walls in their predictions highlight their ability to preserve learned features while adapting to new domains. ProtoDepth-A, in particular, demonstrates its generalization strength by leveraging overlapping domain features to improve predictions in certain areas, such as the bedpost edges.

Overall, these qualitative results underscore the ability of ProtoDepth to mitigate catastrophic forgetting and produce high-fidelity depth predictions. By effectively combining domain-specific adaptation and cross-domain generalization, ProtoDepth-A outperforms baseline methods, even under significant domain shifts between NYUv2 and ScanNet.

8. Training Time Comparison

Tab. 7 presents the training time per epoch for each continual learning method on both indoor (ScanNet and VOID) and outdoor (Waymo and VKITTI) datasets using KBNet. These experiments were conducted with a fixed batch size of 12 for indoor datasets and 8 for outdoor datasets, on a single NVIDIA GeForce RTX 3090 GPU. This standardized setup ensures a fair comparison across all methods. The training times vary across datasets because they are measured per epoch, and each training set contains a different number of frames, as detailed in Sec. 4.

ProtoDepth and ProtoDepth-A demonstrate significant improvements in computational efficiency, with training times roughly half those of the baseline methods. This efficiency can be attributed to ProtoDepth's approach of freezing the

	Training Time per Epoch (mins)			
Method	ScanNet	VOID	Waymo	VKITTI
Finetuned	165.8	35.4	84.7	17.3
EWC	168.2	35.9	85.0	18.5
LwF	170.7	38.1	85.4	20.3
Replay	182.9	40.4	88.8	23.0
ProtoDepth-A	92.5	17.9	40.3	10.7
ProtoDepth	85.3	15.7	37.9	9.6

Table 7. **Training times** (minutes per epoch) with KBNet for each continual learning method on both indoor and outdoor datasets.

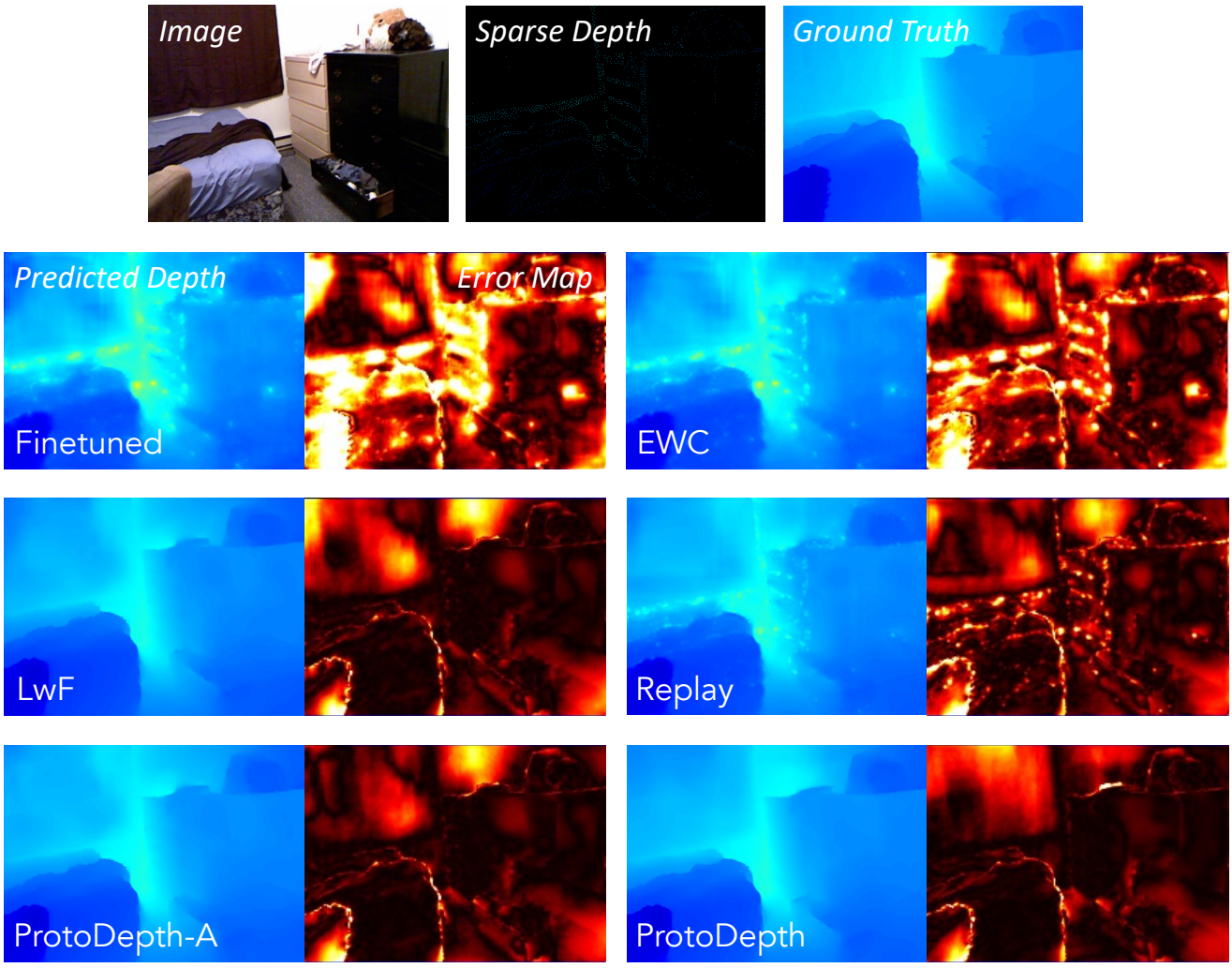

Figure 2. **Qualitative comparison** (1 of 2) of ProtoDepth and baseline methods using FusionNet on **NYUv2** after continual training on **ScanNet**. *Top row:* Input sample from NYUv2. *Following rows:* Output depth and error maps (relative to ground-truth) of same sample from NYUv2 after continual training on ScanNet using each continual learning method.

backbone model and training only the prototype sets, which are applied to the latent space layers (i.e., bottleneck and skip connections). Thus, backpropagation computations are restricted to parameters from the output layer back only to the latent space layers. Since the parameters involved are approximately half of the total parameters, ProtoDepth requires fewer gradient computations compared to methods like EWC, LwF, and Replay that calculate gradients and update parameters across the entire model.

ProtoDepth achieves slightly faster training times than ProtoDepth-A. This difference arises because ProtoDepth-A requires additional computations to train the domain descriptors, which involves calculating and optimizing cosine similarity between sample descriptors and domain descriptors during training. ProtoDepth avoids this step, resulting in a small yet consistent reduction in training time.

Among the baseline methods, Finetuned is the fastest, training slightly faster than EWC, LwF, and Replay. This is because finetuning does not involve the additional regularization or distillation used by EWC and LwF, nor does it use a memory buffer like Replay. However, the simplicity of full finetuning comes at the cost of increased catastrophic forgetting, as evidenced by its consistently poor performance in the main experiments.

The reduced training times of ProtoDepth and ProtoDepth-A are particularly important for real-world applications, where computational efficiency is crucial. By restricting updates to the latent space, ProtoDepth not only reduces computational overhead but also does so while achieving state-of-the-art performance. This efficiency is critical for resource-constrained environments, or scenarios requiring fast adaptation to new datasets. These results highlight

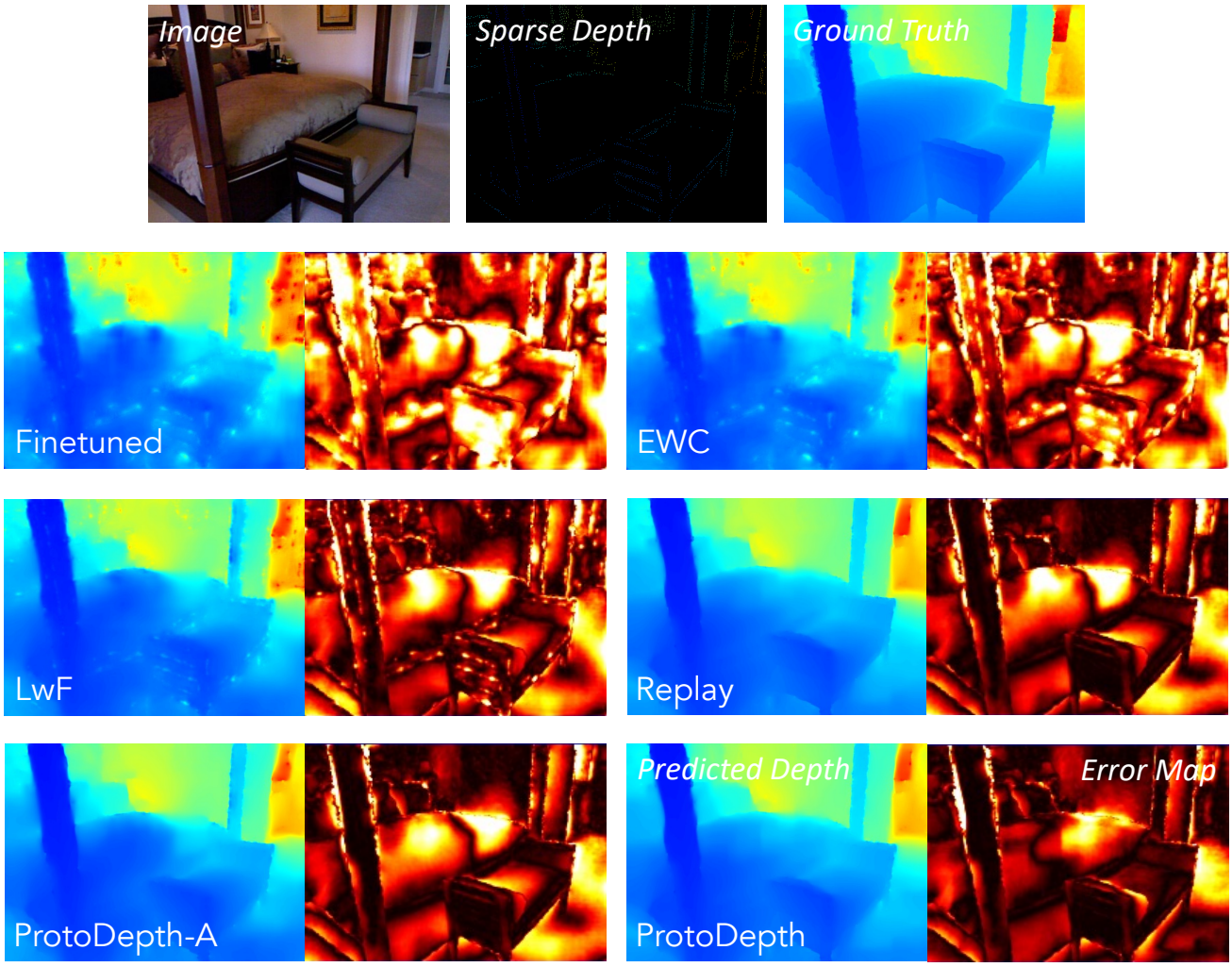

Figure 3. **Qualitative comparison** (2 of 2) of ProtoDepth and baseline methods using FusionNet on **NYUv2** after continual training on **ScanNet**. *Top row:* Input sample from NYUv2. *Following rows:* Output depth and error maps (relative to ground-truth) of same sample from NYUv2 after continual training on ScanNet using each continual learning method.

ProtoDepth's ability to deliver both high performance and practical advantages in training time, underscoring its suitability for continual learning tasks.

9. More Ablation Studies

To further evaluate the importance of prototype sets in ProtoDepth, we conduct additional ablation studies to assess the impact of removing prototype sets from different modalities and latent space layers. Specifically, we analyze the role of prototype sets applied to the image features, sparse depth features, and the bottleneck features. The results, shown in Tab. 8, are evaluated on ScanNet (indoor dataset) and Waymo (outdoor dataset) using KBNet.

The results highlight that removing prototype sets from any of these components significantly degrades performance. When image prototype sets are ablated, we observe a sharp increase in both MAE and RMSE, particularly for ScanNet, where MAE rises from 14.59 to 35.06. This degradation demonstrates the importance of capturing domain-specific biases in image features, as images undergo larger distributional shifts between domains compared to sparse depth, such as changes in lighting, textures, and color distributions.

Similarly, removing the sparse depth prototype sets also results in noticeable performance drops, with MAE increasing from 14.59 to 32.07 for ScanNet. While sparse depth features may exhibit smaller distributional shifts compared to image features, these features are crucial for anchoring the model to the metric scale of the depth predictions. Without the sparse depth prototypes, the model struggles to adapt effectively to the unique distribution of sparse point clouds in each new dataset.

The bottleneck prototype sets play a critical role as well,

	ScanNet		Waymo	
Ablated Component	MAE	RMSE	MAE	RMSE
image prototype sets	35.06	88.23	542.16	1703.01
sparse depth prototype sets	32.07	84.39	537.37	1762.31
bottleneck prototype sets	19.03	60.32	502.21	1680.87
no ablations	14.59	42.20	486.95	1664.18

Table 8. **Ablation studies** on prototype sets for different modalities using KBNet for indoor (ScanNet) and outdoor (Waymo).

as they adapt the fused representations of both image and sparse depth modalities. Ablating the bottleneck prototypes leads to performance degradation, although the impact is less severe than removing the image or sparse depth prototypes. For instance, MAE increases from 14.59 to 19.03 for ScanNet when bottleneck prototypes are removed. This suggests that while the bottleneck prototypes contribute to the overall performance, much of the adaptation occurs in the modality-specific layers.

Notably, when all prototype sets are included (no ablations), ProtoDepth achieves the best performance across both datasets, with significantly lower error metrics compared to any ablated configuration. These results validate the design choice of applying prototype sets to both modality-specific features (image and sparse depth) and their fused representations (bottleneck).

10. Discussion

Accurate 3D reconstruction [23, 33, 34] is crucially important for applications that rely on precise perception of surrounding environments, such as humanoid robotics [16, 18]. One key challenge in this domain is monocular depth estimation (MDE) [2, 12, 26, 28, 35], which aims to recover metric depth from a single image. However, MDE is fundamentally challenging due to scale ambiguity, making it an inherently ill-posed problem. To overcome this challenge, synchronized complementary modalities—such as LiDAR [3, 4, 15, 27, 29], radar [17, 20], inertial sensors [6], additional cameras [1, 8, 32], and even language [37]—can provide additional cues to resolve scale ambiguity. In particular, LiDAR offers high-precision depth measurements that are relatively dense compared to other time-of-flight sensors such as radar, making it a valuable modality for resolving scale ambiguity and enhancing metric depth estimation accuracy. This task of LiDAR-Camera depth estimation, specifically, is commonly referred to as depth completion [13, 30, 31, 36]. In our work, ProtoDepth, we introduce an unsupervised continual depth completion [3] framework that leverages prototypes to continuously learn in challenging and dynamic environments. Unlike traditional approaches that rely on fully supervised training on stationary datasets, ProtoDepth adapts continuously across domains, demonstrating improved generalization without the need for expensive, inaccurate ground truth. Our comprehensive results demonstrate that ProtoDepth effectively mitigates catastrophic forgetting for depth completion, making it a promising solution for real-world applications in autonomous driving, augmented/virtual reality, robotics, and general scene understanding.

References

[1] Zachary Berger, Parth Agrawal, Tian Yu Liu, Stefano Soatto, and Alex Wong. Stereoscopic universal perturbations across different architectures and datasets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15180–15190, 2022. 8

[2] Aleksei Bochkovskii, AmaÃĢl Delaunoy, Hugo Germain, Marcel Santos, Yichao Zhou, Stephan R Richter, and Vladlen Koltun. Depth pro: Sharp monocular metric depth in less than a second. *arXiv preprint arXiv:2410.02073*, 2024. 3, 8

[3] Xien Chen, Rit Gangopadhyay, Michael Chu, Patrick Rim, Hyoungseob Park, and Alex Wong. Uncle: Benchmarking unsupervised continual learning for depth completion. *arXiv preprint arXiv:2410.18074*, 2024. 8

[4] Younjoon Chung, Hyoungseob Park, Patrick Rim, Xiaoran Zhang, Jihe He, Ziyao Zeng, Safa Cicek, Byung-Woo Hong, James S Duncan, and Alex Wong. Eta: Energy-based test-time adaptation for depth completion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6001–6012, 2025. 8

[5] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017. 3

[6] Xiaohan Fei, Alex Wong, and Stefano Soatto. Geo-supervised visual depth prediction. *IEEE Robotics and Automation Letters*, 4(2):1661–1668, 2019. 3, 8

[7] Adrien Gaidon, Qiao Wang, Yohann Cabon, and Eleonora Vig. Virtual worlds as proxy for multi-object tracking analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4340–4349, 2016. 3

[8] Suchisrit Gangopadhyay, Jung-Hee Kim, Xien Chen, Patrick Rim, Hyoungseob Park, and Alex Wong. Extending foundational monocular depth estimators to fisheye cameras with calibration tokens. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5198–5209, 2025. 8

[9] Christopher G. Harris and M. J. Stephens. A combined corner and edge detector. In *Alvey Vision Conference*, 1988. 3

[10] DaeJun Kang, Dongsuk Kum, and Sanmin Kim. Continual learning for motion prediction model via meta-representation learning and optimal memory buffer retention strategy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15438–15448, 2024. 2, 3

[11] Sanghwan Kim, Lorenzo Noci, Antonio Orvieto, and Thomas Hofmann. Achieving a better stability-plasticity trade-off via

auxiliary networks in continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11930–11939, 2023. 2, 3

[12] Dong Lao, Fengyu Yang, Daniel Wang, Hyoungseob Park, Samuel Lu, Alex Wong, and Stefano Soatto. On the viability of monocular depth pre-training for semantic segmentation. In *European Conference on Computer Vision*. Springer, 2024. 8

[13] Tian Yu Liu, Parth Agrawal, Allison Chen, Byung-Woo Hong, and Alex Wong. Monitored distillation for positive congruent depth completion. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part II*, pages 35–53. Springer, 2022. 8

[14] Hyoungseob Park, Anjali Gupta, and Alex Wong. Test-time adaptation for depth completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20519–20529, 2024. 4

[15] Hyoungseob Park, Runjian Chen, Patrick Rim, Dong Lao, and Alex Wong. Orcas: Unsupervised depth completion via occluded region completion as supervision. *The Fourteenth International Conference on Learning Representations*, 2026. 8

[16] Patrick Rim, Kun He, Kevin Harris, Braden Copple, Shangchen Han, Sizhe An, Ivan Shugurov, Tomas Hodan, He Wen, and Xu Xie. Ego-exo 3d hand tracking in the wild with a mobile multi-camera rig. *arXiv preprint arXiv:2510.02601*, 2025. 8

[17] Patrick Rim, Hyoungseob Park, Vadim Ezhov, Jeffrey Moon, and Alex Wong. Radar-guided polynomial fitting for metric depth estimation. *arXiv preprint arXiv:2503.17182*, 2025. 8

[18] Patrick Rim, Kevin Harris, Braden Copple, Shangchen Han, Xu Xie, Ivan Shugurov, Sizhe An, He Wen, Alex Wong, Tomas Hodan, et al. Show3d: Capturing scenes of 3d hands and objects in the wild. *arXiv preprint arXiv:2603.28760*, 2026. 8

[19] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *European Conference on Computer Vision*, 2012. 3

[20] Akash Deep Singh, Yunhao Ba, Ankur Sarker, Howard Zhang, Achuta Kadambi, Stefano Soatto, Mani Srivastava, and Alex Wong. Depth estimation from camera image and mmwave radar point cloud. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9275–9285, 2023. 8

[21] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2446–2454, 2020. 3

[22] Jonas Uhrig, Nick Schneider, Lukas Schneider, Uwe Franke, Thomas Brox, and Andreas Geiger. Sparsity invariant cnns. In *2017 international conference on 3D Vision (3DV)*, pages 11–20. IEEE, 2017. 3

[23] Daniel Wang, Patrick Rim, Tian Tian, Dong Lao, Alex Wong, and Ganesh Sundaramoorthi. Ode-gs: Latent odes for dynamic scene extrapolation with 3d gaussian splatting. *arXiv preprint arXiv:2506.05480*, 2025. 8

[24] Zhendong Wang, Xiaodong Cun, Jianmin Bao, Wengang Zhou, Jianzhuang Liu, and Houqiang Li. Uformer: A general u-shaped transformer for image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17683–17693, 2022. 1

[25] Zifeng Wang, Zizhao Zhang, Chen-Yu Lee, Han Zhang, Ruoxi Sun, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, and Tomas Pfister. Learning to prompt for continual learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 139–149, 2022. 2

[26] Alex Wong and Stefano Soatto. Bilateral cyclic constraint and adaptive regularization for unsupervised monocular depth prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5644–5653, 2019. 8

[27] Alex Wong and Stefano Soatto. Unsupervised depth completion with calibrated backprojection layers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12747–12756, 2021. 8

[28] Alex Wong, Safa Cicek, and Stefano Soatto. Targeted adversarial perturbations for monocular depth prediction. *Advances in neural information processing systems*, 33:8486–8497, 2020. 8

[29] Alex Wong, Xiaohan Fei, Stephanie Tsuei, and Stefano Soatto. Unsupervised depth completion from visual inertial odometry. *IEEE Robotics and Automation Letters*, 5(2):1899–1906, 2020. 3, 8

[30] Alex Wong, Safa Cicek, and Stefano Soatto. Learning topology from synthetic data for unsupervised depth completion. *IEEE Robotics and Automation Letters*, 6(2):1495–1502, 2021. 8

[31] Alex Wong, Xiaohan Fei, Byung-Woo Hong, and Stefano Soatto. An adaptive framework for learning unsupervised depth completion. *IEEE Robotics and Automation Letters*, 6 (2):3120–3127, 2021. 8

[32] Alex Wong, Mukund Mundhra, and Stefano Soatto. Stereopagnosia: Fooling stereo networks with adversarial perturbations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2879–2888, 2021. 8

[33] Chao Xia, Chenfeng Xu, Patrick Rim, Mingyu Ding, Nanning Zheng, Kurt Keutzer, Masayoshi Tomizuka, and Wei Zhan. Quadric representations for lidar odometry, mapping and localization. *IEEE Robotics and Automation Letters*, 8 (8):5023–5030, 2023. 8

[34] Yichen Xie, Chenfeng Xu, Marie-Julie Rakotosaona, Patrick Rim, Federico Tombari, Kurt Keutzer, Masayoshi Tomizuka, and Wei Zhan. Sparsefusion: Fusing multi-modal sparse representations for multi-sensor 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17591–17602, 2023. 8

[35] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10371–10381, 2024. 3, 8

[36] Yanchao Yang, Alex Wong, and Stefano Soatto. Dense depth posterior (ddp) from single image and sparse range. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3353–3362, 2019. 8

[37] Ziyao Zeng, Jingcheng Ni, Daniel Wang, Patrick Rim, Younjoon Chung, Fengyu Yang, Byung-Woo Hong, and Alex Wong. Iris: Integrating language into diffusion-based monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026. 8