Lifelong-MonoDepth: Lifelong Learning for Multi-Domain Monocular Metric Depth Estimation

Junjie Hu, *Member, IEEE,* Chenyou Fan, Liguang Zhou, Qing Gao, Honghai Liu, *Fellow, IEEE* and Tin Lun Lam, *Senior Member, IEEE*

***Abstract*—With the rapid advancements in autonomous driving and robot navigation, there is a growing demand for lifelong learning models capable of estimating metric (absolute) depth. Lifelong learning approaches potentially offer significant cost savings in terms of model training, data storage, and collection. However, the quality of RGB images and depth maps is sensor-dependent, and depth maps in the real world exhibit domain-specific characteristics, leading to variations in depth ranges. These challenges limit existing methods to lifelong learning scenarios with small domain gaps and relative depth map estimation. To facilitate lifelong metric depth learning, we identify three crucial technical challenges that require attention: i) developing a model capable of addressing the depth scale variation through scale-aware depth learning, ii) devising an effective learning strategy to handle significant domain gaps, and iii) creating an automated solution for domain-aware depth inference in practical applications. Based on the aforementioned considerations, in this paper, we present i) a lightweight multi-head framework that effectively tackles the depth scale imbalance, ii) an uncertainty-aware lifelong learning solution that adeptly handles significant domain gaps, and iii) an online domain-specific predictor selection method for real-time inference. Through extensive numerical studies, we show that the proposed method can achieve good efficiency, stability, and plasticity, leading the benchmarks by 8% $\sim$ 15%. The code is available at https://github.com/FreeformRobotics/Lifelong-MonoDepth.**

***Index Terms*—Monocular depth estimation, lifelong learning, cross-domain learning**

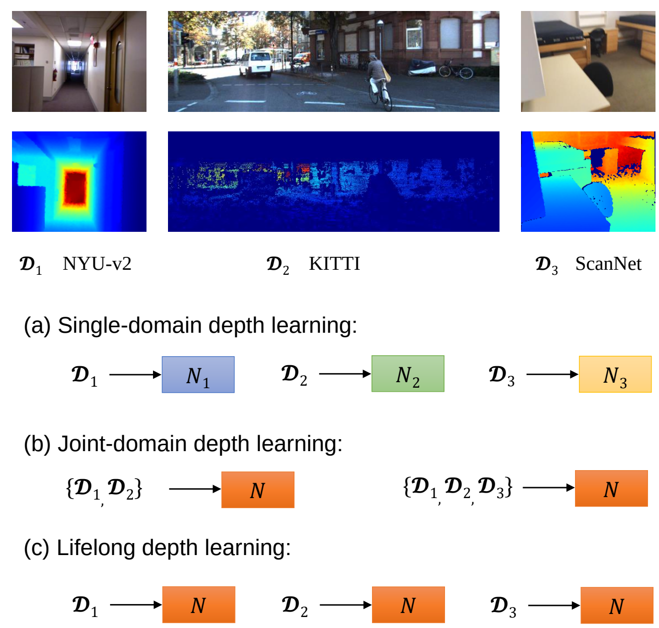


Fig. 1. Depth learning in the real world where the same color and different colors denote the same and different models. Traditional approaches include single-domain learning to train a domain-specific model as (a), and joint-domain learning to obtain a domain-robust model as (b). We aim to learn a model that can infer metric depth maps for multiple domains in a lifelong learning manner as (c).

I. Introduction

ACQUIRING scene depths in real depth scale is an essential requirement for real-world applications, e.g., SLAM [38], self-driving [35], robot navigation [5], 3D reconstruction [9], human-computer interaction [7], augmented reality [6], etc. As a cost-effective solution to depth sensors, monocular depth estimation (MDE) aims to infer depth maps from visual images. MDE has gained great success by learning with deep convolutional neural networks (CNNs) in a data-driven fashion. In the early stage, traditional studies handled MDE by training and testing on a single-domain [18], [8], [26], [2], [16], [42], as shown in Fig. 1. (a).

However, learning-based methods have often been criticized and questioned due to their poor generalizability for out-of-distribution data. Despite the recent trend of tackling poor generalizability by covering possible domains as much as possible [29], [28], [41], [36], as seen in Fig. 1. (b), it is impossible to exhaust all possible patterns of data in the real world. When there are some new patterns of data or target domains, a pre-trained model has to be re-trained from scratch, resulting in a tremendous waste of time and cost. Inspired by human cognition, researchers have attempted to empower CNNs with lifelong learning mechanisms which aim to perform incremental learning on new domains or tasks with the minimum increase over model complexity, training time, and reuse of data on old tasks. This practice has already seen promising results on image recognition [4], [31], [21], [1].

J.Hu, L.Zhou, and Tin Lun Lam are with the Shenzhen Institute of Artificial Intelligence and Robotics for Society (AIRS), The Chinese University of Hong Kong, Shenzhen, China. E-mail: hujunjie@cuhk.edu.cn.

C.Fan is with the School of Artificial Intelligence, South China Normal University, China. E-mail: fanchenyou@scnu.edu.cn.

Q.Gao is with the School of Electronics and Communication Engineering, Sun Yat-sen University, Shenzhen, China. E-mail: gaoqing.ieee@gmail.com.

H.Liu is with the State Key Laboratory of Robotics and systems, Harbin Institute of Technology (Shenzhen), China. E-mail: honghai.liu@hit.edu.cn.

L.Zhou and T.L.Lam are with the School of Science and Engineering, The Chinese University of Hong Kong, Shenzhen, China. E-mail: liguangzhou@link.cuhk.edu.cn, tllam@cuhk.edu.cn.

T.L.Lam is the corresponding author.

This work was supported in part by the National Natural Science Foundation of China under Grants 62073274 and 62106156, in part by Shenzhen Science and Technology Program under Grant JCYJ20220818103000001, and in part by the Shenzhen Institute of Artificial Intelligence and Robotics for Society under Grant AC01202101103.

On the other hand, since there is a significant difference between image recognition and MDE, it is largely unknown how to enable lifelong learning for MDE (Fig. 1. (c)). Most previous approaches [29], [28], [36] of multi-domain learning choose only to infer relative depth maps to tackle domain gaps. Besides, only a few studies [22], [20], [45] have tried empowering MDE with lifelong learning, and none of them can infer scale-aware metric depth maps. In this paper, we extensively study this under-explored problem and provide some valuable insights. We identify two major challenges of scale-ware MDE that cause catastrophic forgetting (forgetting learned knowledge after updating a trained model on a new domain) when performing lifelong learning, including

- Significant domain gap: both visual images and depth images are significantly different across different domains. Thus, a trained model on old domains tends to shift its parameters to accommodate a new domain significantly.
- Depth scale imbalance: scene depth scales are usually domain-dependent and dominated by a specific range such that lifelong learning across two domains of different scales is ineffective.

The above challenges limit existing methods to lifelong learning scenarios with small domain gaps and relative depth map estimation, as demonstrated in [45], [22], [20]. To facilitate lifelong metric depth learning, we propose a general framework, *Lifelong-MonoDepth* for lifelong learning on MDE. We posit that three pivotal technical challenges necessitate attention: i) developing a model capable of addressing the depth scale imbalance issue through scale-aware depth learning, ii) devising an effective learning strategy to handle significant domain gaps, and iii) creating an automated solution for domain-aware depth inference in practice.

We consider MDE under natural circumstances where agents work in complex real-world environments, including both indoor and outdoor scenarios. Fig. 1 shows several samples of images and depth maps from three different domains. As seen, depth maps captured in the real world are significantly different across domains; their quality and scales are domain-dependent. Therefore, the model has to efficiently assemble multiple prediction branches for multi-domain metric depth inference. Otherwise, if the model only has a single branch for outputting depth, its parameters will gradually shift to accommodate the depth range of a new domain when performing lifelong learning. Additionally, we need to also ensure the memory-efficient and computationally frugal computation for MDE task considering their real-world deployment. To this end, we present a lightweight multi-head framework that consists of a domain-shared encoder and domain-specific layers. The framework dynamically grows a domain-specific predictor with only a 0.21M increase in parameters when learning on a new domain. The framework allows robust metric depth learning across multi-domains.

To facilitate effective and practical learning of our proposed model, we adhere to the principles of lifelong learning that minimizes the use of historical data to reduce data storage costs and keeps the increase in parameters for new domains as minimal as possible to improve efficiency in practical applications. To achieve this, we adopt a regularization term that applies a knowledge distillation loss as [24], thereby eliminating the need for extensive storage of historical data as prior works [22], [20] and avoiding excessive parameter increases as [45]. To handle the significant domain gaps, we introduce an uncertainty-aware knowledge preservation solution by incorporating uncertainty estimation and uncertainty consistency into our lifelong learning framework. The former addresses potential performance degradation resulting from outliers present in ground-truth depth maps of certain domains, thereby achieving a better balance between different domains. The latter poses a strong regularization for preserving the knowledge of original domains. To further mitigate the significant domain gap, we also apply a replay loss using only 500 randomly selected image and depth pairs from each original domain. By strategically selecting this limited subset of data, we effectively bridge the domain gap while maintaining computational efficiency and ensuring reliable learning outcomes.

We then consider how to dynamically select the optimal domain-specific predictor given an input image during inference. We assume the input image belongs to one of the target domains, and the key is how to identify that domain. As the replay data is a small subset of each domain, we propose to compare the distance between the image and each domain in the feature space. Then, the closest domain is the one with the minimum distance.

To validate the effectiveness of the proposed method, we perform lifelong depth learning on three real-world datasets with significant domain gaps. We show through experiments that the proposed method can i) enable lifelong learning for scale-aware depth estimation, ii) cope with significant domain shift during lifelong learning, and iii) infer a depth map in real-time.

In summary, our contributions are:

- We present an efficient multi-head framework that enables lifelong, cross-domain, and scale-aware monocular depth learning. To our best knowledge, we are the first to fulfill multi-domain metric depth estimation via lifelong learning.
- We introduce an uncertainty-aware knowledge preservation solution by incorporating uncertainty estimation and uncertainty consistency into our lifelong learning framework. The former strikes a balance between different domains since the quality of depth is domain-dependent, and the latter provides a strong regularization for better preserving the model's knowledge on original domains.
- We propose to automatically select the domain-specific predictor for an image during inference based on the minimum distance to mean features of each domain.
- We perform extensive experiments to demonstrate a promising balance between the stability (remembering old knowledge) and the plasticity (acquiring new knowledge) of the proposed method.

The remainder of this paper is organized as follows. In Sec. II, we discuss the necessary background and related studies. We present the proposed lifelong depth learning

TABLE I
COMPARISONS BETWEEN SEVERAL REPRESENTATIVE EXISTING WORKS AND OUR WORK.

Methods	Lifelong learning	Scale-aware	Cross-domain learning strategy	(Un)supervised learning
Virtual Normal v1 [42]	✗	✓	✗	Supervised
Virtual Normal v2 [43]	✗	✗	Mixed data	Supervised
DABC [23]	✗	✓	Mixed data	Supervised
MiDas [29]	✗	✗	Mixed data	Supervised
CoSelfDepth [20]	✓	✗	Mixed data	Unsupervised
Ours	✓	✓	Sequential learning	Supervised

framework in Sec. III. We then provide extensive numerical evaluations in Sec. IV and finally conclude our work in Sec. V.

II. Related Works

A. Monocular Depth Estimation

In recent years, monocular depth estimation has been formulated in a data-driven fashion either by penalizing pixel-wise loss between predicted depth maps and ground truth depth maps in supervised learning [40], [18], [8], [25], [16], [17] or complying with the geometry consistency of multi-views in unsupervised learning [47], [35], [44], [19], [46], [37]. The advantage of unsupervised approaches is they can learn from videos and thus are easy to implement. However, their greatest drawback is that they only estimate relative depth maps and are highly limited for many applications, *e.g.*, robot navigation.

While there has been significant progress in learning specific data domains, previous methods face severe challenges when deploying in real-world applications due to their poor performance in robustness and generalization. One effective approach is to address these challenges by collecting a large-scale dataset encompassing diverse domains and training a domain-invariant model, as demonstrated in recent research works such as [29], [41], [28], [36], [43]. This solution is rather straightforward and still costly. Furthermore, if some new target domains appear, the model has to be learned from scratch. Therefore, there is an urgent demand to develop lifelong learning models as human vision systems.

In this paper, we study a lifelong learning paradigm that enables extending a single model for MDE to multiple domains sequentially. We list the comparisons between this work and existing works in Table I. Compared to other methods of multi-domain learning using mixed data training strategy, we only reuse a few training data (less than 1%) from each of the old domains. A few works also studied LL for unsupervised MDE. by using replay data [22], [20] or additional parameters for domain adaptation [45]. They perform cross-domain learning with a small domain gap [22], [45] or pre-training with mixed domain data [20]. We differ from them in three major aspects: i) we can perform metric (absolute) depth estimation, which is more challenging than those methods in LL, ii) we can perform cross-domain learning with significant domain gaps, and iii) we can perform long-range lifelong learning across three different domains while previous works only successfully conducted lifelong learning between two similar domains.

B. Lifelong Learning

Lifelong learning (LL), also called incremental learning or continual learning, has been an active topic in machine learning. It aims to enable ifelong learning of a model on new concepts/tasks/domains while preventing forgetting the previously learned knowledge. Most existing works solve LL on image recognition, and various approaches have been proposed. In general, those methods can be categorized into three types [4]: i) replay methods [30], [11] that store some training samples for each of the previous tasks and reuse them while learning a new task, ii) regularization-based methods that prevents forgetting by imposing extra regularization instead of storing training samples, such as LwF [24] using a knowledge distillation loss on previous tasks or EWC [21] enforcing an additional loss term to alleviate changing on the weights important for previous tasks, and iii) parameter isolation methods that fix trained parameters on old tasks and employ extra network branches for training a new task [33], [31]. It is worth mentioning that ExpertGate [1] assumes there are multiple expert models corresponding to multi-tasks. It proposed to learn domain-specific auto-decoder for each domain and use the minimum image reconstruction error to select the corresponding expert model. This method is straightforward and memory inefficient. For parameter isolation methods, the minimum increase of parameters is expected.

However, it is largely unknown how to enable LL for dense regression tasks, such as MDE. This work aims to disclose the difficulties and provide solutions for LL on MDE. To minimize the use of historical data and keep the increase in parameters for new domains as minimal as possible, our method enables lifelong learning through regularization by introducing uncertainty-aware knowledge preservation. Besides, as mentioned before, we also use a few replay data to mitigate significant domain gaps further. The replay data is also utilized for online domain identification during inference.

III. Method

A. Multi-head Depth Prediction Framework

We follow most previous works in a practical setting that assumes training data of previous tasks are unavailable when learning a new task. Since the scene scale of depth is domain-dependent, we design a framework with multi-head depth predictors for domain-specific inference and a shared encoder for feature extraction. Each predictor is learned to estimate depth maps of a specific domain with a fixed depth range. A visualization of our framework is given in Fig. 2 (a), where

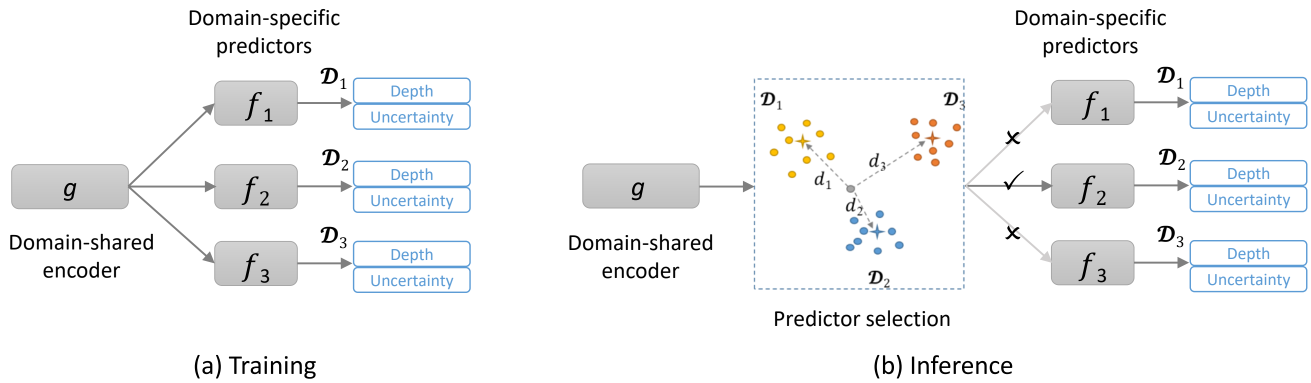


Fig. 2. The diagram of the proposed lifelong learning framework for multi-domain metric depth estimation. $\mathcal{D}_i$ denotes i-th domain/task and f_i denotes the corresponding predictor for scale-aware depth estimation.

TABLE II
NUMBER OF PARAMETERS OF THE FEATURE EXTRACTOR AND DOMAIN-SPECIFIC PREDICTORS.

Module	Parameters (M)
g	21.81
f_1	0.21
f_2	0.21
f_3	0.21

the model is shown for learning three different domains. The model starts from one depth predictor, *i.e.*, f_1, for learning on $\mathcal{D}_1$ and extends its predictors dynamically and sequentially for learning on domains $\mathcal{D}_2$ and $\mathcal{D}_3$.

Then, the problem is that such multi-head architecture will significantly increase model complexity for learning on multi-domains. To alleviate this issue, we introduce an extremely lightweight design using only two convolutional layers for depth prediction in each predictor. We take a compact pyramid feature network [13] in which features extracted at each scale from the encoder are fused and compressed to a small fixed number; then, they are concatenated and inputted to the predictors. In our framework, the shared encoder is built on a ResNet-34 [10] and has 21.81 M parameters; each domain-specific predictor has 0.21 M parameters, respectively. In the case of learning three domains, the framework yields over 97% shared parameters to promote computational efficiency. A more detailed description of the framework is given in the supplementary file.

During inference, the framework needs to identify the correct predictor for an input image. We propose an efficient method for this purpose.

B. Lifelong Cross-Domain Depth Learning

1) Uncertainty-aware Knowledge Acquisition: Given a target domain $\mathcal{D}^t$, where $\mathcal{D}^t = \{x^t, y^t\}$, x^t and y^t denote images and their corresponding depth maps, we can directly let the model learn to estimate depth maps in the target domain.

Depth maps captured by sensors are usually sparse, suffer from outliers, and miss valid information around object boundaries, as seen in Fig. 1. To eliminate the effect of outliers and improve the robustness, we employ an uncertainty-aware loss [12] as follows:

$$\ell_{ud} = \sum (e^{-s^t}(\hat{y}^t - y^t)^2 + s^t), \tag{1}$$

where $\hat{y}^t$ is predicted depth maps from x^t, e denotes the exponential operation, s^t denotes pixel-wise uncertainty maps estimated simultaneously with depth maps, such that:

$$\hat{y}^t, s^t = f_t(g(x^t)). \tag{2}$$

Similar to the depth estimation layers, we also use the two convolutional layers for uncertainty estimation, resulting in a total of 0.21 M parameters for each domain-specific predictor. Besides, the uncertainty estimation layers can be dropped during inference for efficient computation.

2) Uncertainty-aware Knowledge Preservation: For learning on a new domain $\mathcal{D}^{t+1}$, we accordingly add a new domain-specific depth predictor f_{t+1} such that $\hat{y}^{t+1} = f_{t+1}(g(x^{t+1}))$ and learn its parameters with Eq. (1). However, this will shift the parameters of the encoder, thus leading the estimation on $\mathcal{D}^{1,2,...,t}$ to malfunction, *i.e.*, causing catastrophic forgetting.

As studied in [14], applying KD with out-of-distribution data is able to distill the knowledge of a model learned on the original domain for MDE. In our method, the trained model on $\mathcal{D}^1, ..., \mathcal{D}^t$ serves as an expert teacher and provides desired predictions on each domain. Formally, we let g and $f_1, ..., f_{t+1}$ denote the new encoder and domain-specific predictors while performing lifelong learning on $\mathcal{D}^{t+1}$, let g' and $f'_1, ..., f'_t$ be the old model learned on $\mathcal{D}^1, ..., \mathcal{D}^t$. We apply regularization on both depth consistency and uncertainty consistency on $\mathcal{D}^i, i \in \{1, 2, ..., t\}$ as follows:

$$\begin{aligned} &\ell_{cons} = \sum (|\hat{y}^i_n - \hat{y}^i_o| + |s^i_n - s^i_o|), \\ &\text{s.t.} \quad \hat{y}^i_n, s^i_n = f_i(g(x^{t+1})) \\ &\qquad \hat{y}^i_o, s^i_o = f'_i(g'(x^{t+1})), \end{aligned} \tag{3}$$

where $\hat{y}^i_n$ and $\hat{y}^i_o$ denotes predicted depth images of $\mathcal{D}^i$ with the new model and old model, respectively; similarly, s^i_n and s^i_o are predicted uncertainty with new model and old model.

Algorithm 1 Lifelong-MonoDepth: Training

Input: $\mathcal{D}^{t+1}$: new target domain;
$N^t = \{g', f'_1, ..., f'_t\}$: old model;
λ^t: weight coefficients;
$\mathcal{P} = \{\mathcal{P}_1, ..., \mathcal{P}_t\}$: replay sets;
Output: $N^{t+1} = \{g, f_1, ..., f_{t+1}\}$: new model;

```
Freeze N^t;
for j = 1 to iterations do
    ▷ % knowledge acquisition from new domain %
    Set gradients of N^{t+1} to 0;
    Select a batch (x^{t+1}, y^{t+1}) from D^{t+1};
    Get predictions ŷ^{t=1}, s^{t+1} ← f_{t+1}(g(x^{t+1}));
    Compute uncertainty-aware depth loss ℓ_ud by Eq.(1);
    ▷ % knowledge preservation for old domains %
    for i = 1 to t do
        Get consistency loss ℓ_cons by Eq.(3);
        Select a batch (x^i, y^i) from P_i;
        Compute replay loss ℓ_replay by Eq.(4);
    end for
    Get the total loss L = ℓ_ud + λ^i Σ_{i=1}^t (ℓ_cons + ℓ_replay);
    Backpropagate L;
    Update N^{t+1};
end for
```

Algorithm 2 Lifelong-MonoDepth: Inference

Input: $N^t = \{g, f_1, ..., f_t\}$: learned model on $\mathcal{D}^1$ to $\mathcal{D}^t$;
$\mu = \{\mu_1, ..., \mu_t\}$: domain-specific mean features;
x: an image from any domain $\mathcal{D}^i, i \in \{1, ..., t\}$;
Output: $\hat{y}$: a depth map;

```
Compute intermediate features by g(x);
for i = 1 to t do
    Compute the distance d_i between g(x) and μ_i;
end for
Select predictor f_i ← arg min d_i;
Output depth map ŷ ← f_i(g(x));
```

TABLE III
DETAILS OF THE RGBD DATASETS USED IN THE EXPERIMENTS.

Dataset	Depth range (m)	Training scenarios / images	Test scenarios / images
NYU-v2	$0 \sim 10$	249 / 50688	215 / 654
KITTI	$0 \sim 80$	138 / 85898	18 / 1000
ScanNet	$0 \sim 6$	1513 / 50473	100 / 17607

3) Replay for Memory Enhancement: If images from D^{t+1} lie in the same distribution as images from D^i, *i.e.*, $\mathcal{P}(x^i) == \mathcal{P}(x^{t+1})$, Eq.(3) will be fully effective for preserving knowledge on $\mathcal{D}^i$. Otherwise, its performance tends to degrade due to the domain gap. Therefore, there is a risk that the model will significantly deteriorate its performance on previous domains because of a significant domain shift between $\mathcal{D}^{t+1}$ and D^i where $i \in \{1, 2, ..., t\}$.

To handle this issue, we take a replay strategy as many classical lifelong learning methods [30], [11], which is more consistent with human cognition by periodically and repeatedly reviewing historical data. We randomly preserve limited training data (500 images) of each of the previous domains and replay them for learning on new domains. Then, the replay loss is formulated as:

$$\ell_{replay} = \ell_{ud}(\hat{y}^i, y^i). \tag{4}$$

Then, the loss for lifelong learning on $\mathcal{D}^{t+1}$ can be written as:

$$\mathcal{L} = \sum_{i=1}^{t} \lambda^i \left(\ell_{cons}(\hat{y}^i_n, s^i_n, \hat{y}^i_o, s^i_o) + \ell_{replay}(\hat{y}^i, y^i)\right) + \ell_{ud}(\hat{y}^{t+1}, y^{t+1}), \tag{5}$$

where λ is a vector and λ^t denotes the weight coefficient for domain $\mathcal{D}^t$. The first and the second loss term in Eq.(5) alleviate knowledge forgetting on domains $\mathcal{D}^1$ to $\mathcal{D}^t$, the third loss term in Eq.(5) promotes learning knowledge on the new target domain $\mathcal{D}^{t+1}$.

C. Online Cross-Domain Depth Inference

After incremental learning on $\mathcal{D}^1$ to $\mathcal{D}^t$, ideally, the model is able to correctly estimate a depth map $\hat{y}$ from any image x sampled from D^i, $i \in \{1, 2, ..., t\}$. A practical challenge is how to identify the domain of x^i and accordingly select the corresponding predictor f_i automatically during inference.

To address this problem, we propose to identify the minimum distance between a given image and each domain in the feature space. Since we preserve a small subset of each domain, we can obtain the mean features of each domain approximated with these replay data, that is:

$$\mu^i = \sum_{k=1}^{k} g(x^i_k), \tag{6}$$

where x^i_k is k-th image of the replay set of the domain D^i, μ^i is the mean features of D^i calculated by the replay set. Then, identifying the domain to which the input image belongs can be formulated as:

$$\begin{aligned} & i \leftarrow \arg\min_i d_i, \\ & \text{s.t.} \quad d_i = \|g(x) - u^i\|_2, \end{aligned} \tag{7}$$

where $\|\cdot\|_2$ denotes the ℓ_2 norm. Finally, the i-th predictor, i.e., f_i corresponding to the domain $\mathcal{D}^i$ is selected for depth inference.

IV. EXPERIMENTS

A. Experimental Setup

1) Datasets: We evaluate our method on three benchmark datasets, including two indoor and one outdoor dataset. The details are given as follows.

a) NYU-v2 [34]: The NYU-v2 dataset is one of the most commonly used benchmarks for indoor depth estimation. NYU-v2 has 464 indoor scenes captured by Microsoft Kinect with an original resolution of 640×480. Among them, 249 scenes are used for training, and the rest 215 scenes are used for testing. We use the pre-processed data by [16], [15] with about 50,000 RGBD pairs. Following previous studies, we

TABLE IV

QUANTITATIVE COMPARISONS BETWEEN EXISTING METHODS AND THE PROPOSED METHOD IN WHICH & DENOTES DATA MIXING AND → DENOTES SEQUENTIAL ORDER FOR LIFELONG LEARNING. NOTE THAT WE SPECIFY THE CORRECT DOMAIN-SPECIFIC PREDICTOR FOR EACH INPUT IMAGE. ∗ DENOTES RESULTS TAKEN FROM [20].

	NYU-v2			KITTI			Average		
Method	RMSE	REL	δ_1	RMSE	REL	δ_1	RMSE	REL	δ_1
SDT	0.532	0.130	0.836	3.286	0.070	0.939	1.909	0.100	0.888
JDT (NYU-v2 & KITTI)	0.581	0.151	0.803	3.658	0.086	0.911	2.120	0.119	0.857
Comoda* [22] (NYU-v2 & KITTI → KITTI)	0.673	0.191	0.706	6.249	0.158	0.769	3.461	0.175	0.738
CoSelfDepth* [20] (NYU-v2 & KITTI → KITTI)	0.626	0.187	0.728	5.809	0.154	0.784	3.218	0.171	0.756
FT (NYU-v2 → KITTI)	1.133	0.328	0.451	3.655	0.079	0.918	2.394	0.204	0.685
FAL (NYU-v2 → KITTI)	0.532	0.130	0.836	8.946	0.252	0.600	4.739	0.191	0.718
EWC (NYU-v2 → KITTI)	1.007	0.251	0.475	4.550	0.100	0.876	2.779	0.176	0.676
Ours (NYU-v2 → KITTI)	0.622	0.162	0.768	3.829	0.081	0.910	**2.226**	**0.122**	**0.839**
FT (KITTI → NYU-v2)	0.555	0.137	0.820	13.22	0.450	0.179	6.888	0.294	0.500
FAL (KITTI → NYU-v2)	0.991	0.318	0.523	3.286	0.070	0.939	2.139	0.194	0.731
EWC (KITTI → NYU-v2)	0.650	0.173	0.755	7.178	0.243	0.573	3.914	0.208	0.664
Ours (KITTI → NYU-v2)	0.567	0.142	0.812	5.060	0.136	0.813	**2.814**	**0.139**	**0.813**

resize the images to 320×240 pixels and then crop their central parts of 304×228 pixels as inputs. For testing, we use the official small subset of 654 RGBD pairs.

b) KITTI [39]: : This outdoor dataset, collected by car-mounted cameras and a LIDAR sensor, was also widely used as a benchmark in previous studies of MDE. We use the official KITTI depth prediction dataset with the official split of scenes for training and validation. The training and validation set has 138 and 18 driving sequences, respectively. The resolution is about 1216×352 for most images. We randomly crop a patch with 480×320 resolution for training and use the original resolution for testing.

c) ScanNet [3]: ScanNet is a large-scale indoor RGBD dataset that contains 2.5 million RGBD images. We randomly and uniformly select a subset of approximately 50,000 samples from the training splits of 1513 scenes for training and evaluate the models on the test set of another 100 scenes with 17K RGB pairs. The resolution of RGB images is 1296 × 968. We apply image resizing and cropping as utilized on the NYU-v2 dataset.

2) Implementation Details: We train the model for 20 epochs using the Adam optimizer with an initial learning rate of 0.0001 for each dataset and reduce it to 50% for every five epochs. While learning on $\mathcal{D}^{t+1}$, the hyper-parameters λ^i for preventing forgetting on $\mathcal{D}^i$ are set to 10 for the indoor dataset and 100 for the outdoor dataset for all experiments throughout the paper. We trained models with a batch size of 8 in all the experiments using PyTorch [27]. For the sake of fair comparison, we train with the uncertainty-aware loss function for all baseline methods. Notably, as the depth scale is significantly different across domains, as seen in Table III, we apply a scale-invariant operation to depth maps in the loss function by dividing the median depth value of ground truth to exclude potential disturbance.

For evaluation, we use the most popular three measures, including RMSE, REL, and δ_1. The first is a scale-aware measure, and the latter two are scale-invariant.

3) Baselines: Since no previous methods have been proposed for lifelong metric depth learning, we consider different learning strategies as baselines for our method as follows.

Single-Domain Training (SDT): is the standard learning protocol for single domain depth learning, *i.e.*, training and evaluating on the same dataset. The performance of SDT provides an upper bound that we aim to reach.

Joint-Domain Training (JDT): randomly selects a batch from each domain and then mixes data to perform joint learning.

Fine-tuning (FT): is a common baseline for lifelong learning. We also compare it in experiments.

Freezing And Learning (FAL): is a parameter isolation strategy. It freezes old model parameters including $g, f_1, ..., f_t$, and only updates new model parameters, *i.e.*, f_{t+1} while learning on $\mathcal{D}^{t+1}$.

Elastic Weight Consolidation [21] (EWC): is a classical method for lifelong learning. It overcomes catastrophic forgetting by discouraging modifying weights important for old tasks.

Among baseline methods, **SDT** and **JDT** use a single-head network to demonstrate the upper bound of single-domain learning and multi-domain learning. The other methods adopt the same multi-head framework as our method. We also compare two existing methods of lifelong depth learning, including **Comoda [22]** and **CoSelfDepth [20]**. These two methods are proposed for unsupervised depth learning and use data replay to avoid catastrophic forgetting. Note that the original Comoda targets outdoor autonomous driving scenes with small domain gaps and the code of CoSelfDepth is not publicly available. We take results implemented in CoSelfDepth [20] for reference in which experiments on NYU-v2 and KITTI are performed.

B. Results of Stability and Plasticity

1) Results on Two Domains: We first conduct experiments on two domains, including NYU-v2 and KITTI, the domain gap between which is significant. We compare the proposed method against all baseline approaches. Since the learning order has a large impact on the results of each domain, we perform experiments in the order of both NYU-v2 → KITTI

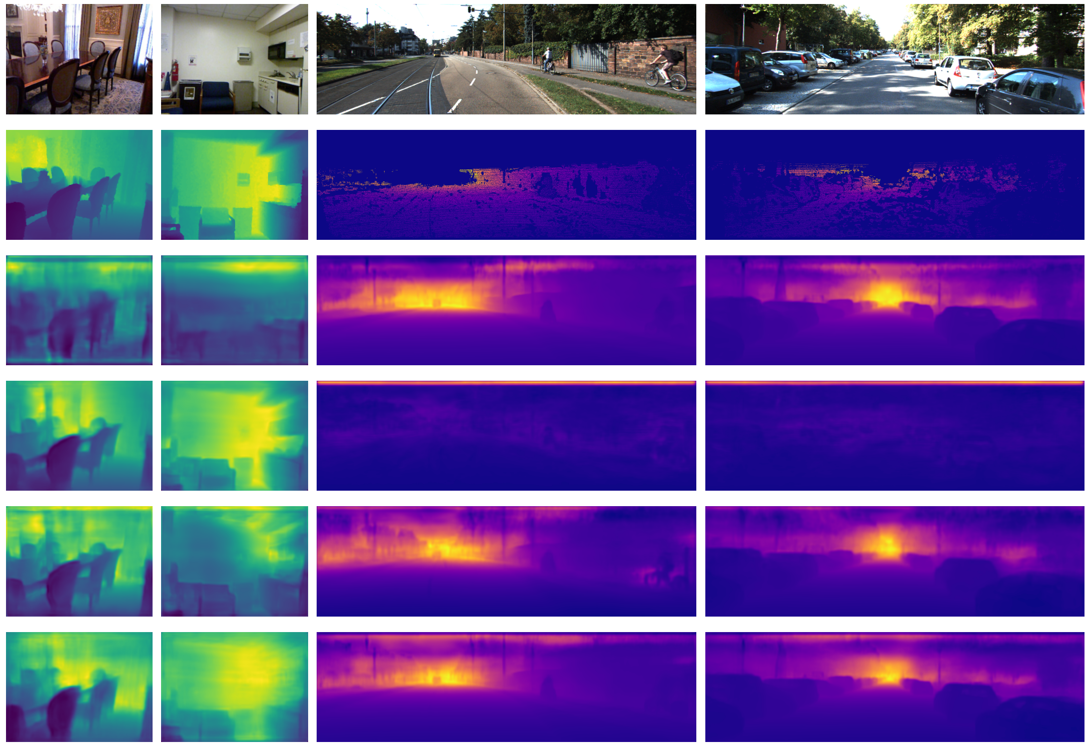


Fig. 3. Qualitative comparison of depth maps predicted by different methods in the learning order of NYU-v2 → KITTI.

and KITTI → NYU-v2. The old domain is NYU-v2 and the new domain is KITTI for the former and reversely for the latter. Notably, FT and FAL inherently yield the best plasticity and stability due to their training strategy. Therefore, we also compute the average accuracy on the two datasets for better quantifying the trade-off between stability and plasticity.

The results are given in Table IV. No doubt, SDT and JDT gained the best and second-best performance, respectively. For results of cross-domain learning, although FT obtained the best accuracy on the new domain, it would yield extremely poor performance on the old domain, showing the worst result of mean accuracy. EWC can be seen as an improved method of FT that tackles this problem by employing an additional regularization term. We observe that EWC demonstrates slightly low performance than FT on the new domain, whereas it gained much better performance on the old domain, thus achieving better average accuracy. In contrast, FAL does not suffer from catastrophic forgetting at the cost of sacrificing the plasticity on a new domain. As a result, our method achieves promising results for both the old and the new domain. Although it slightly underperforms FAL in stability showing the second-best performance on the old domain, we gained the best average accuracy for all three measures, *e.g.*, outperforming FT, FAL, and EWC in δ_1 by 15.4%, 12.1%, 13.6% on NYU-v2 and 31.3%, 8.2%, 14.9% on KITTI, respectively.

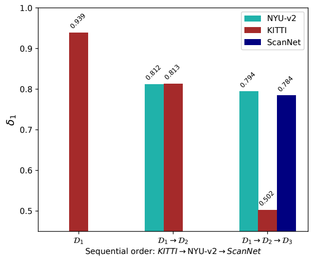
The results of Comoda [22] and CoSelfDepth [20] are taken from [20]. Note that the implementations are different from ours. Thus, we mark them in $*$. These two methods used half of the training data from NYU-v2 and KITTI for pre-training and then performed lifelong learning with the other half of the data. Hence, they suffer marginally from large domain shifts. Nevertheless, our approach demonstrates clearly better performance than the two methods.

TABLE V
THE AVERAGE ACCURACY OF LIFELONG DEPTH LEARNING ON ALL THREE DOMAINS WITH ALL SIX DIFFERENT LEARNING ORDERS.

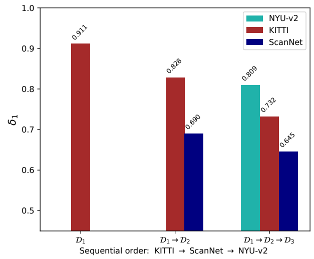
Learning order	RMSE	REL	δ_1
KITTI → NYU-v2 → ScanNet)	3.314	0.181	0.694
KITTI → ScanNet → NYU-v2)	2.871	0.173	0.729
NYU-v2 → KITTI → ScanNet)	2.263	0.145	0.780
ScanNet → KITTI → NYU-v2)	2.075	0.146	0.774
NYU-v2 → ScanNet → KITTI)	1.716	0.145	0.779
ScanNet → NYU-v2 → KITTI)	1.644	0.146	0.782
Average	2.314	0.156	0.756

Fig. 3 shows qualitative comparisons between our method and baseline approaches. It is seen that FT predicted good depth maps on KTTI; however, it failed on NYU-v2. FAL inferred the best depth maps on NYU-v2, on the other hand, failed on KITTI. Both EWC and our method could produce perceptually correct depth maps and our method yield more accurate predictions on NYU-v2.

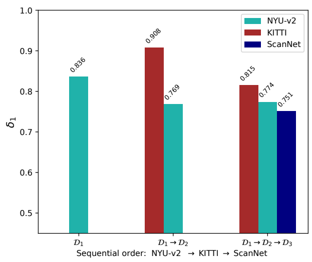
2) Results on Three Domains: We then perform experiments on all three domains. In this case, the first learned domain will suffer from more long-term forgetting. As both NYU-v2 and ScanNet are composed of indoor scenes, the domain gap between them is small; on the other hand, they have a tremendous domain gap from KITTI. Thus, the domain order performing lifelong learning also affects the final performance. For a fair evaluation, we conduct experiments



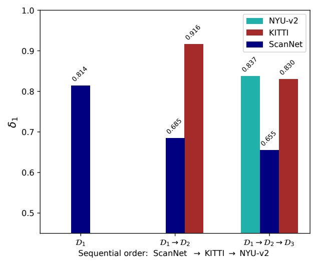
(a) KITTI → NYU-v2 → ScanNet



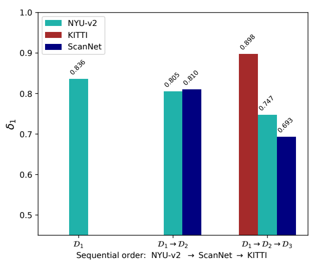
(b) KITTI → ScanNet → NYU-v2



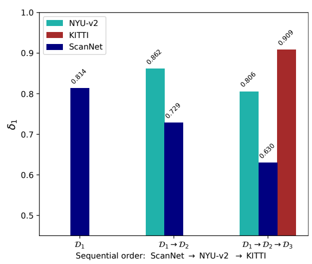
(c) NYU-v2 → KITTI → ScanNet



(d) ScanNet → KITTI → NYU-v2



(e) NYU-v2 → ScanNet → KITTI



(f) ScanNet → NYU-v2 → KITTI

Fig. 4. The δ_1 accuracy on three domains for different learning orders.

for all possible combinations. Totally there are six different combinations regards to learning order. We take a standard evaluation protocol for lifelong learning that computes the mean accuracy over multi-domains.

We report the mean δ_1 accuracy in Table V. As seen, ScanNet → NYU-v2 → KITTI demonstrates the best accuracy, whereas KITTI → NYU-v2 → ScanNet shows the worst performance. Also, KITTI → ScanNet → NYU-v2 leads to penultimate accuracy. It is not surprising since learning on KITTI would cause the model to forget the domain for a longer time.

Fig. 4 gives a more detailed visualization of performance evolution. It is observed in Fig. 4 (a) that the performance on KITTI in KITTI → NYU-v2 → ScanNet degraded significantly compared to other methods. The results agree well with our expectations that we should lastly learn KITTI at best. Overall, Fig. 4 (c) demonstrates the best trade-off between stability and plasticity as the variance of δ_1 among the three domains is small.

We also provide qualitative results in Fig. 5. As seen, our approach can accurately infer uncertainty maps for the images of the corresponding domain. Generally, high uncertainty is around missing points, far regions, and object boundaries. We will quantify the effects of uncertainty estimation in ablation studies.

C. Results of Online Predictor Selection

Online predictor selection is a relatively practical requirement. Since there are multiple depth predictors, given an input image, the model must automatically identify its domain and select the correct branch to infer a depth map. Therefore, we conduct experiments to validate the effectiveness of the proposed predictor selection method for the trained framework varying the learning order. The results are given in Table VI in which Domain Prior denotes the results of pre-specifying the corresponding predictor for input images. It provides an upper bound to our predictor selection method. As seen, for results of lifelong learning on two domains, *i.e.*, NYU-v2 → KITTI and KITTI → NYU-v2, our predictor selection method demonstrates a 100% success rate. For results on three domains, our method still attained 100% success rate for KITTI while yielding a slight accuracy drop (within 4%) for NYU-v2 or ScanNet. We consider the miss between NYU-v2 and ScanNet reasonable as they contain some similar indoor images. It can be better observed in Fig. 6, which shows the number of categorized images on test sets of the three domains. Fig. 6 (a) and (b) show results of Ours (NYU-v2 → KITTI → ScanNet) and Ours (KITTI → NYU-v2 → ScanNet), respectively. It is seen that our predictor selection method could identify data from KITTI without misclassification. Although there are some misclassified images between NYU-v2 and ScanNet, as we discussed above, the depth maps can still be accurately inferred due to the small domain gap. Therefore, the accuracy drop is slight and acceptable. Besides, as shown in Table VI, the accuracy could be improved sometimes.

We also evaluate the computational efficiency of our proposed method, running on a computer with Intel(R) Xeon(R) Gold 6230 CPU and a GeForce RTX 2080 Ti GPU card. We calculate the GPU time by running the model for an input image 10000 times and calculate the mean time. Table VII shows the results for the three domains. As seen, our predictor selection module spends only 2.7 ms more time for NYU-v2 and ScanNet, and 0.5 ms for KITTI, respectively, demonstrating superior efficiency.

D. Ablation Study

We perform several ablation studies to analyze our method better. For simplicity, we conduct experiments on two domains. We take ours (NYU-v2 → KITTI) as the base method and remove some critical operations, including data replay, uncertainty consistency, and scale-invariant operation, for comparison. The results are given in Table VIII.

Without uncertainty estimation: the uncertainty is used in the uncertainty-aware loss Eq.(1) and consistency loss Eq.(3). We remove the uncertainty estimation module to evaluate the performance. As a result, we observe performance degradation both on NYU-v2 and KITTI.

Without data replay: data replay is used to enhance the stability of the model. The results without replay demonstrate 0.3% and 2% accuracy drop on KITTI and NYU-v2, respectively. It indicates that replay is more important in improving stability.

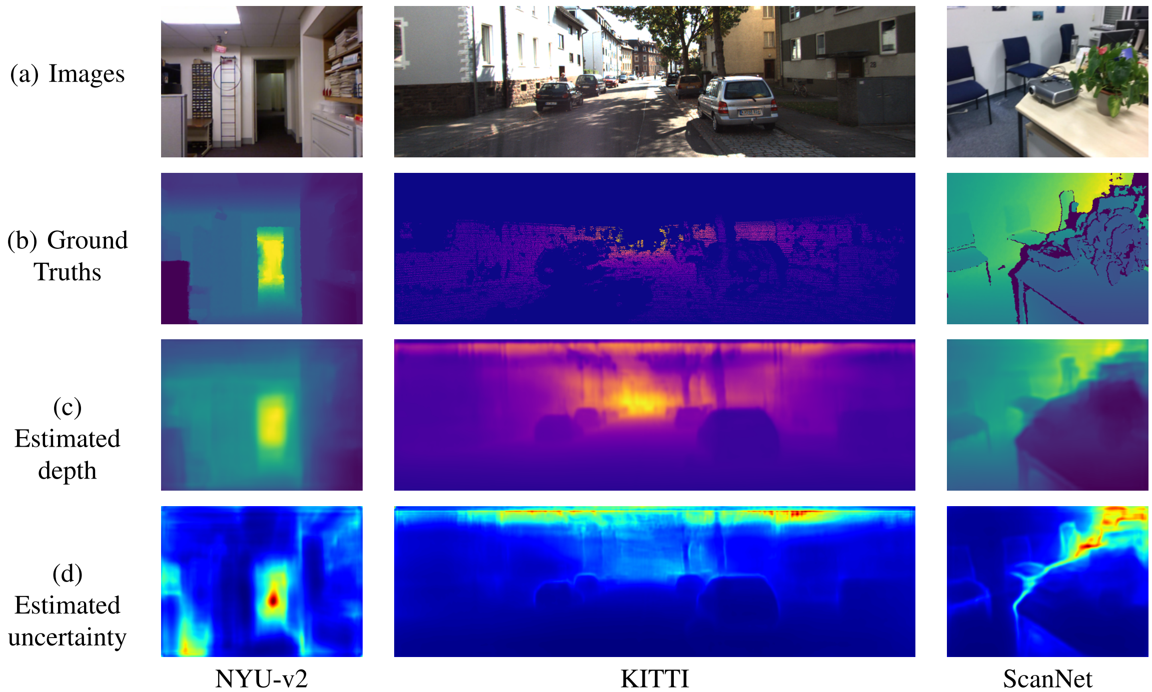
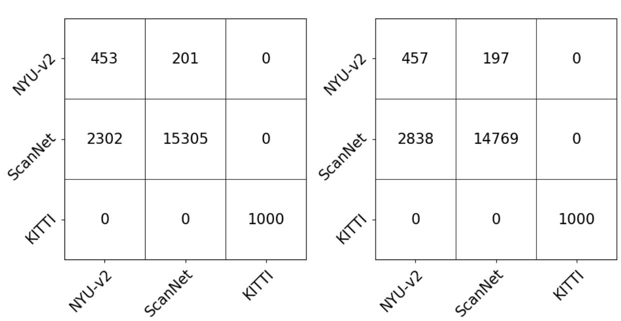


Fig. 5. Qualitative comparison of depth and uncertainty maps predicted by our method in the learning order of NYU-v2 → KITTI → ScanNet.

TABLE VI
THE δ_1 ACCURACY OF LIFELONG DEPTH LEARNING ON ALL THREE DOMAINS WHERE DOMAIN PRIOR DENOTES RESULTS OF MANUALLY SPECIFY DOMAIN-SPECIFIC PREDICTOR FOR INPUT IMAGES. ON THE CONTRARY, OUR PREDICTOR SELECTION AUTOMATICALLY CHOOSES THE PREDICTOR BASED ON THE MINIMUM FEATURE DISTANCE.

	Domain Prior			Our Predictor Selection			Accuracy Drop		
Learning order	NYU-v2	KITTI	ScanNet	NYU-v2	KITTI	ScanNet	NYU-v2	KITTI	ScanNet
Ours (NYU-v2 → KITTI)	0.768	0.910	-	0.768	0.910	-	0%	0%	-
Ours (KITTI → NYU-v2)	0.812	0.813	-	0.812	0.813	-	0%	0%	-
Ours (NYU-v2 → KITTI → ScanNet)	0.774	0.815	0.751	0.769	0.815	0.748	0.5% ↓	0%	0.3% ↓
Ours (ScanNet → KITTI → NYU-v2)	0.837	0.830	0.655	0.805	0.830	0.667	3.2% ↓	0%	0.2% ↑
Ours (KITTI → NYU-v2 → ScanNet)	0.794	0.502	0.784	0.794	0.502	0.764	0%	0%	2% ↓
Ours (KITTI → ScanNet → NYU-v2)	0.809	0.732	0.645	0.770	0.732	0.651	3.9%↓	0%	0.6% ↑
Ours (ScanNet → NYU-v2 → KITTI)	0.806	0.909	0.630	0.780	0.909	0.639	2.6%↓	0%	0.9% ↑
Ours (NYU-v2 → ScanNet → KITTI)	0.747	0.898	0.693	0.751	0.898	0.672	0.4% ↑	0%	2.1%↓



(a) NYU-v2 → KITTI → ScanNet (b) KITTI → NYU-v2 → ScanNet

Fig. 6. Results of the number of categorized images on test sets of the three domains.

Without uncertainty consistency: The uncertainty consistency is applied along with depth consistency in the original method to prevent forgetting. As shown, without uncertainty consistency, the performance further degrades mainly for the old domain, even though we observe a slight improvement for the new domain.

TABLE VII
RESULTS OF COMPUTATIONAL EFFICIENCY.

Datasets	Resolution	GPU [ms] w/o predictor selection	GPU [ms] w predictor selection
NYU-v2	304 × 228	8.2	10.9
ScanNet	304 × 228	8.2	10.9
KITTI	1216 × 352	28.0	28.5

With a different backbone network: We replace the ResNet-34 based encoder with MobileNet-v2 [32]. It gives us a more lightweight network with only 1.99 M parameters. The δ_1 accuracy is 0.733 and 0.901 for NYU-v2 and KITTI, respectively, and the mean accuracy reaches 0.817, which still outperforms other baseline methods built on large networks.

TABLE VIII
RESULTS OF ABLATION STUDIES.

Method	NYU-v2	KITTI	Average
Ours (NYU-v2 → KITTI)	0.768	0.910	0.839
w/o uncertainty estimation	0.740	0.857	0.799
w/o ℓ_{replay}	0.749	0.907	0.828
w/o ℓ_{replay} and uncertainty consistency	0.700	0.913	0.807

E. Summary

- The proposed multi-head lifelong depth learning framework, *i.e.*, *Lifelong-MonoDepth*, can estimate depth maps with the absolute scale from multi-domains even though there exist significant domain gaps.
- *Lifelong-MonoDepth* attains a good balance between stability and plasticity on real-world datasets. It generally outperforms baseline methods by around 8% $\sim$ 15%.
- *Lifelong-MonoDepth* can automatically identify the domain-specific predictor during inference, showing satisfactory accuracy and efficiency.
- The learning order of domains has an essential effect on lifelong depth learning. For example, learning in NYU-v2 → ScanNet → KITTI substantially outperforms KITTI → NYU-v2 → ScanNet in average accuracy over multi-domains. Generally, learning in an indoor → outdoor order contributes to better performance. In practice, the learning order should be decided according to the specific applications.

V. Conclusion

We present a novel lifelong learning framework for multi-domain metric depth estimation, namely *Lifelong-MonoDepth*. We argue that the major challenges are i) large domain gaps and ii) depth scale imbalance, which cause catastrophic forgetting in lifelong learning. We then propose an efficient multi-head network composed of a domain-shared encoder and domain-specific predictors. Such multi-head predictors enable estimate depth maps with different scales and mitigate domain shift. To alleviate catastrophic forgetting, we propose a novel strategy that applies both depth and uncertainty consistency to avoid knowledge forgetting and uses replay regularization to improve stability further.

We conduct extensive numerical studies to demonstrate the effectiveness of our method. We show that our approach outperforms all baseline methods by a good margin. We also provide the effects of varying the learning order of multiple domains. During inference, we propose to calculate the distance between an image and each domain; then, the minimum distance corresponds to the domain-specific predictor to infer a depth map.

For the first time, we are able to enable scale-aware depth prediction across multi-domains with significant domain gaps in lifelong learning. Potential applications of our method include visual navigation, obstacle avoidance, 3D perception. We hope our method can inspire more future explorations on lifelong depth learning.

References

[1] R. Aljundi, P. Chakravarty, and T. Tuytelaars, "Expert gate: Lifelong learning with a network of experts," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3366–3375.

[2] X. Chen, X. Chen, and Z.-J. Zha, "Structure-aware residual pyramid network for monocular depth estimation," in *International Joint Conferences on Artificial Intelligence*, 2019, pp. 694–700.

[3] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner, "Scannet: Richly-annotated 3d reconstructions of indoor scenes," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 5828–5839.

[4] M. De Lange, R. Aljundi, M. Masana, S. Parisot, X. Jia, A. Leonardis, G. Slabaugh, and T. Tuytelaars, "A continual learning survey: Defying forgetting in classification tasks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 7, pp. 3366–3385, 2021.

[5] R. de Queiroz Mendes, E. G. Ribeiro, N. dos Santos Rosa, and V. Grassi, "On deep learning techniques to boost monocular depth estimation for autonomous navigation," *Robotics Auton. Syst.*, vol. 136, p. 103701, 2021.

[6] R. Du, E. Turner, M. Dzitsiuk, L. Prasso, I. Duarte, J. Dourgarian, J. Afonso, J. Pascoal, J. Gladstone, N. Cruces, S. Izadi, A. Kowdle, K. Tsotsos, and D. Kim, "DepthLab: Real-Time 3D Interaction With Depth Maps for Mobile Augmented Reality," in *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*, 2020, pp. 829–843.

[7] Y. Duan, H. Deng, and F. Wang, "Depth camera in human-computer interaction: An overview," in *2012 Fifth International Conference on Intelligent Networks and Intelligent Systems*. IEEE, 2012, pp. 25–28.

[8] H. Fu, M. Gong, C. Wang, K. Batmanghelich, and D. Tao, "Deep ordinal regression network for monocular depth estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2002–2011.

[9] V. Guizilini, R. Ambrus, S. Pillai, A. Raventos, and A. Gaidon, "3d packing for self-supervised monocular depth estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 2485–2494.

[10] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[11] S. Hou, X. Pan, C. C. Loy, Z. Wang, and D. Lin, "Learning a unified classifier incrementally via rebalancing," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2019, pp. 831–839.

[12] J. Hu, C. Bao, M. Ozay, C. Fan, Q. Gao, H. Liu, and T. L. Lam, "Deep depth completion from extremely sparse data: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 7, pp. 8244–8264, 2023.

[13] J. Hu, C. Fan, H. Jiang, X. Guo, Y. Gao, X. Lu, and T. L. Lam, "Boosting light-weight depth estimation via knowledge distillation," in *International Conference on Knowledge Science, Engineering and Management*, 2023, pp. 27–39.

[14] J. Hu, C. Fan, M. Ozay, H. Jiang, and T. L. Lam, "Data-free dense depth distillation," *arXiv preprint arXiv:2208.12464*, 2022.

[15] J. Hu and T. Okatani, "Analysis of deep networks for monocular depth estimation through adversarial attacks with proposal of a defense method," *arXiv preprint arXiv:1911.08790*, 2019.

[16] J. Hu, M. Ozay, Y. Zhang, and T. Okatani, "Revisiting single image depth estimation: Toward higher resolution maps with accurate object boundaries," in *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, 2019, pp. 1043–1051.

[17] L. Huynh, P. Nguyen-Ha, J. Matas, E. Rahtu, and J. Heikkilä, "Guiding monocular depth estimation using depth-attention volume," in *European Conference on Computer Vision*, 2020, pp. 581–597.

[18] L. Iro, R. Christian, B. Vasileios, T. Federico, and N. Nassir, "Deeper depth prediction with fully convolutional residual networks," in *International Conference on 3D Vision*, 2016, pp. 239–248.

[19] H. Jiang, L. Ding, J. Hu, and R. Huang, "Plnet: Plane and line priors for unsupervised indoor depth estimation," in *International Conference on 3D Vision*, 2021, pp. 741–750.

[20] M. U. K. Khan, "Towards continual, online, self-supervised depth," *arXiv preprint arXiv:2103.00369*, 2021.

[21] J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska *et al.*, "Overcoming catastrophic forgetting in neural networks," *Pro-*

ceedings of the national academy of sciences, vol. 114, no. 13, pp. 3521–3526, 2017.

[22] Y. Kuznietsov, M. Proesmans, and L. Van Gool, "Comoda: Continuous monocular depth adaptation using past experiences," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 2907–2917.

[23] R. Li, K. Xian, C. Shen, Z. Cao, H. Lu, and L. Hang, "Deep attention-based classification network for robust depth prediction," in *Asian Conference on Computer Vision*. Springer, 2018, pp. 663–678.

[24] Z. Li and D. Hoiem, "Learning without forgetting," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 12, pp. 2935–2947, 2017.

[25] J. Liu, Y. Wang, Y. Li, J. Fu, J. Li, and H. Lu, "Collaborative deconvolutional neural networks for joint depth estimation and semantic segmentation," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, pp. 5655–5666, 2018.

[26] F. Ma and S. Karaman, "Sparse-to-dense: Depth prediction from sparse depth samples and a single image," in *IEEE International Conference on Robotics and Automation*, 2018, pp. 1–8.

[27] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., 2019, pp. 8024–8035.

[28] R. Ranftl, A. Bochkovskiy, and V. Koltun, "Vision transformers for dense prediction," in *Proceedings of the IEEE International Conference on Computer Vision*, 2021, pp. 12 179–12 188.

[29] R. Ranftl, K. Lasinger, D. Hafner, K. Schindler, and V. Koltun, "Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 3, pp. 1623–1637, 2022.

[30] S.-A. Rebuffi, A. Kolesnikov, G. Sperl, and C. H. Lampert, "icarl: Incremental classifier and representation learning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2001–2010.

[31] A. A. Rusu, N. C. Rabinowitz, G. Desjardins, H. Soyer, J. Kirkpatrick, K. Kavukcuoglu, R. Pascanu, and R. Hadsell, "Progressive neural networks," *arXiv preprint arXiv:1606.04671*, 2016.

[32] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4510–4520.

[33] J. Serra, D. Suris, M. Miron, and A. Karatzoglou, "Overcoming catastrophic forgetting with hard attention to the task," in *International Conference on Machine Learning*. PMLR, 2018, pp. 4548–4557.

[34] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from rgbd images," in *European Conference on Computer Vision*, vol. 7576, 2012, pp. 746–760.

[35] Z. Song, J. Lu, Y. Yao, and J. Zhang, "Self-supervised depth completion from direct visual-lidar odometry in autonomous driving," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 8, pp. 11 654–11 665, 2022.

[36] L. Sun, W. Yin, E. Xie, Z. Li, C. Sun, and C. Shen, "Improving monocular visual odometry using learned depth," *IEEE Transactions on Robotics*, vol. 38, no. 5, pp. 3173–3186, 2022.

[37] Q. Sun, Y. Tang, C. Zhang, C. Zhao, F. Qian, and J. Kurths, "Unsupervised estimation of monocular depth and vo in dynamic environments via hybrid masks," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, pp. 2023–2033, 2022.

[38] K. Tateno, F. Tombari, I. Laina, and N. Navab, "Cnn-slam: Real-time dense monocular slam with learned depth prediction," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6565–6574.

[39] J. Uhrig, N. Schneider, L. Schneider, U. Franke, T. Brox, and A. Geiger, "Sparsity invariant cnns," in *International Conference on 3D Vision*, 2017, pp. 11–20.

[40] D. Wofk, F. Ma, T.-J. Yang, S. Karaman, and V. Sze, "Fastdepth: Fast monocular depth estimation on embedded systems," in *IEEE International Conference on Robotics and Automation*, 2019, pp. 6101–6108.

[41] K. Xian, C. Shen, Z. Cao, H. Lu, Y. Xiao, R. Li, and Z. Luo, "Monocular relative depth perception with web stereo data supervision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 311–320.

[42] W. Yin, Y. Liu, C. Shen, and Y. Yan, "Enforcing geometric constraints of virtual normal for depth prediction," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 5684–5693.

[43] W. Yin, J. Zhang, O. Wang, S. Niklaus, S. Chen, Y. Liu, and C. Shen, "Towards accurate reconstruction of 3d scene shape from a single monocular image," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 5, pp. 6480–6494, 2023.

[44] H. Zhan, R. Garg, C. S. Weerasekera, K. Li, H. Agarwal, and I. D. Reid, "Unsupervised learning of monocular depth estimation and visual odometry with deep feature reconstruction," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 340–349.

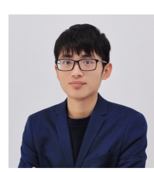
[45] Z. Zhang, S. Lathuiliere, E. Ricci, N. Sebe, Y. Yan, and J. Yang, "Online depth learning against forgetting in monocular videos," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2020, pp. 4494–4503.

[46] C. Zhao, G. G. Yen, Q. Sun, C. Zhang, and Y. Tang, "Masked gan for unsupervised depth and pose prediction with scale consistency," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, pp. 5392–5403, 2021.

[47] T. Zhou, M. R. Brown, N. Snavely, and D. G. Lowe, "Unsupervised learning of depth and ego-motion from video," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 6612–6619.

Junjie Hu (Member, IEEE) received the M.S. and Ph.D. degrees from the Graduate School of Information Science, Tohoku University, Sendai, Japan, in 2017 and 2020, respectively. He is currently a Research Scientist with the Shenzhen Institute of Artificial Intelligence and Robotics for Society. His research interests include machine learning, computer vision, and robotics.

Chenyou Fan is an Associate Professor with the School of Artificial Intelligence, South China Normal University, China. He received the B.S. degree in computer science from the Nanjing University, China, in 2011, and the M.S. and Ph.D. degrees from Indiana University, USA, in 2014 and 2019, respectively. His research interests include machine learning and computer vision. He served in the program committee of CVPR, NeurIPS, ACM MM and top AI journals for more than 20 times.



Liguang Zhou received the B.Eng. degree in Electrical Engineering from China Jiliang University, Hangzhou, China, in 2016, and he is pursing the Ph.D. degree in Computer Information Engineering at The Chinese of Hong Kong, Shenzhen, China. His research interests include robotic vision, scene understanding, and computational photography.

Qing Gao received his Ph.D. degree from the Chinese Academy of Sciences (CAS), Shenyang, China, in 2020. He is currently an Associate Professor at the School of Electronics and Communication Engineering, Sun Yat-sen University, Shenzhen, China. His research interests include robotics, artificial intelligence, machine vision, and human–robot interaction.

Honghai Liu (Fellow, IEEE) received the Ph.D. degree in intelligent robotics from King's College London, London, U.K., in 2003. He is a Professor with the Harbin Institute of Technology (Shenzhen), Shenzhen, China. He is also a Chair Professor of Human–Machine Systems with the University of Portsmouth, Portsmouth, U.K. His research interests include multi-sensory data fusion, pattern recognition, intelligent video analytics, intelligent robotics, and their practical applications.

Tin Lun Lam (Senior Member, IEEE) received the Ph.D. degrees from the Chinese University of Hong Kong, Hong Kong, in 2010. He is an Assistant Professor with the Chinese University of Hong Kong, Shenzhen, China, and the Director of Center for the Intelligent Robots, Shenzhen Institute of Artificial Intelligence and Robotics for Society. His research interests include multi-robot systems, field robotics, and collaborative robotics.