2023 IEEE International Conference on Robotics and Automation (ICRA 2023)
May 29 - June 2, 2023. London, UK

Online Visual SLAM Adaptation against Catastrophic Forgetting with Cycle-Consistent Contrastive Learning

Sangni Xu[1,5], Hao Xiong[2,*], Qiuxia Wu[1], Tingting Yao[3], Zhihui Wang[4], Zhiyong Wang[5]

Abstract— **Visual SLAM (Simultaneous Localisation and Mapping) aims to simultaneously estimate camera poses and depth maps from navigation videos captured. While recent deep learning based methods have achieved great success on this task, they tend to work well on source domain data and suffer from performance degradation on the unseen data of target domain. Hence, we propose an online adaptation approach to continuously adapt a pre-trained visual SLAM model to changing environments in a self-supervised manner. To preserve pre-learned knowledge against catastrophic forgetting, we perform updating on a novel adapter proposed rather than fine-tuning the whole model for adaptation. The adapter includes a cross-domain feature translation module that translates pre-learned features into translated features suitable for adaptation. Ideally, the translated new features should not only contain pre-learned knowledge but also substantially distinct from pre-learned features since these two features represent different domains. We thus introduce cycle-consistent contrastive learning to maximize the dissimilarity between these two features by enlarging the distance between them in the feature space. Besides, our contrastive learning method exploiting cycle-consistency contraint enables the translated features to be transferred back to the pre-learned ones, which helps the translated features better preserve pre-learned knowledge. Comprehensive experiments on both synthetic and real-world datasets demonstrate superior adaptation performance of our proposed method over several state-of-the-art baselines.**

I. Introduction

Visual SLAM plays a key role in many real-world applications, such as self-driving car, robot navigation, and AR/VR. Traditional visual SLAM [1] rely on hand-craft features, potentially combined with image dehazing [2], [3] to deal with extreme weather conditions. Recently, deep learning based SLAM methods [4], [5], [6], [7] have demonstrated great success with superior performance in challenging situations. In most real-world scenarios, the ground-truth depth maps and camera poses are generally expensive to obtain, therefore several self-supervised learning based methods [8], [9], [10], [11] have been proposed as a promising alternative, which utilizes an image reconstruction loss to minimize reconstruction inconsistency between a reference image and the image reconstructed by estimated pose and depth.

However, these methods trained using data from a specific domain or a specific type of scenes often do not perform well on a different unseen scene. As a result, the SLAM performance generally degrades significantly due to domain shift. Various domain adaptation (DA) methods [12], [13], [14], [15], [16] have been proposed to generalize a pre-trained model to an unknown domain. In general, they are categorized into two types: 1) appearance alignment, and 2) feature alignment. For the first type, domain shift is eliminated by aligning images from two domains via techniques like image-to-image translation [17]. The second one intends to learn domain invariant features [18] across different domains. However, these domain adaptation approaches are generic and not specific to visual SLAM problem. Meanwhile, these methods tend to learn adaptation offline, which means they cannot start to learn and adapt until all the target domain data are collected. Consequently, they are not ideal for real-world applications since the open-world environments change continuously and thus require fast adaptation.

Accordingly, online adaptation is considered as an effective solution to adapt fast and continuously to changing environments. To the best of our knowledge, [19], [20] are the most relevant online domain adaptation methods to visual SLAM. They both performed online adaptation using the meta-learning framework that incorporates training and adaptation into a single phase to achieve fast adaptation. Specifically, they kept optimizing the pre-trained model on current data from target domain so that the optimized model well adapts to the current data without waiting for all data of target domain available. However, [19] only evaluated camera poses, while [20] focused on depth estimation. Therefore, they partially addressed the domain drift issue in SLAM and did not demonstrate their effectiveness on all the aspects of SLAM. Moreover, catastrophic forgetting of past experiences was still not addressed explicitly in [19]. Though [20] performed domain adaptation against forgetting, it heavily focused on preserving pre-learned knowledge and thus overlooked learning sufficiently discriminative features on target domain, which could compromise the performance of domain adaptation.

To address these limitations, we propose a novel visual SLAM adaptation approach to perform online adaptation in a self-supervised manner. We also further devise a novel adapter that is added to the pre-trained network for adapting feature representations. To preserve pre-learned knowledge, we only perform updating on the proposed adapters using the meta-learning framework [21] rather than optimizing the whole pre-trained network. In our proposed adapter, a cross-domain feature translation module (denoted as $\mathbf{F}$) translates pre-learned feature x_s to new feature representation $\bar{x}$ suitable for being adaptated to target domain via $\bar{x} = \mathbf{F}(x_s)$. Such module involves two components: 1) Domain Specific Attention

[1]South China University of Technology, [2]Macquarie University, [3]Dalian Maritime University, [4]Dalian University of Technology, [5]The University of Sydney
*Corresponding author, email: hao.xiong@mq.edu.au

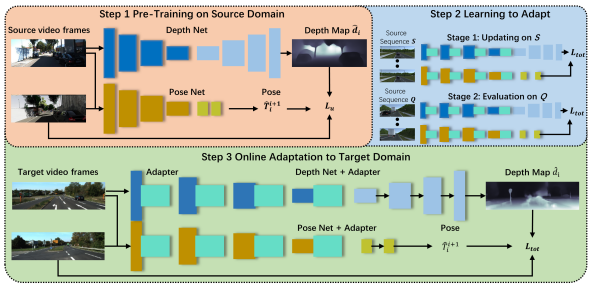

Fig. 1: Overview of the proposed methods which mainly consists of 3 steps: Step 1) pre-training the pose and depth networks on source domain dataset, Step 2) learning to adapt by training our forgetting prevention adapters on source domain dataset, and Step 3) adapting the trained model to target domain dataset.

(DSA) that gives more attentions to target domain-specific information, and 2) Feature Fusion (FF) which integrates the target domain information discovered with x_s to generate $\bar{x}$. To encourage $\bar{x}$ being more discriminative from x_s, we propose cycle-consistent contrastive learning defining $\bar{x}$ and x_s as a negative sample pair to maximize their dissimilarity in feature space. To better preserve pre-learned knowledge in $\bar{x}$, our contrastive learning method exploits cycle-consistent constraint to enforce $\bar{x}$ translated from pre-learned feature x_s should be transferred back to x_s via $\mathbf{G}(\bar{\mathbf{x}}) \approx \mathbf{x_s}$, where $\mathbf{G}$ is another cross-domain feature translation module that transfers $\bar{x}$ back to x_s. Hence, our adapter can learn discriminative target domain feature effectively, while better preserving pre-learned knowledge, and then exploit both features to achieve more robust adaptation.

The key contributions of our work are as follows:

- We propose a novel online adaptation mechanism for visual SLAM to continuously adapt to the changing environments without any supervision.
- We propose a novel adapter to adjust the feature representation pre-learned from source domain to be suitable for being adaptated to target domain. Through cycle consistent contrastive learning devised, the adapter is able to adjust feature representation by incorporating discriminative target domain-specific features, while preserving pre-learned knowledge.
- Comprehensive experiments on both synthetic and real-world datasets, including Virtual KITTI, KITTI and CityScapes, demonstrate the effectiveness and efficiency of our proposed method with superior adaptation performance over several state-of-the-art methods.

II. Related Work

A. Supervised and Self-Supervised Visual SLAM

Supervised SLAM aims to learn depth and pose estimations using ground truth information. Early pose estimation [4], [6] combined CNN and LSTM to extract both spatial and temporal features. In [22], [23], [24], attention was further utilised to extract non-local temporal features for more accurate estimation. For depth estimation, while some works [25], [26] solely predicted depths, other works [27], [28] predicted poses and depths concurrently.

Self-supervised visual SLAM aims to explore various clues for network training, without requiring ground truth poses and depths. Conventional methods [5], [10] utilised two networks to estimate poses and depths, respectively and trained the networks with appearance consistency. Some methods [29], [30], [31] were further proposed to exploit temporal information by using recurrent neural networks such as LSTM. To mitigate the scale inconsistency issue, self-supervised loss with optical flow was investigated [32], [33]. In addition to optical flow, geometric consistency constraint was exploited in [10], [7] to predict scale consistent depths and poses. However, these methods often performed well only on the training dataset, not on an unknown dataset due to domain discrepancy between different datasets. In this work, we aim to enhance the generalizability of our SLAM model on unknown datasets.

B. Domain Adaptation

Most existing domain adaptation (DA) methods rely on Generative Adversarial Network (GAN) to achieve adaptation. For instance, in [34], [35] images between source and target domains were aligned with GAN for similar styles and appearances. GAN was also used to learn invariant representations across domains [36], [37]. Similarly, Gradient-Reversal-Layer (GRL) [38] was utilised to learn a domain invariant encoder shared by different domains [39]. Unlike [39], [12] learned domain specific features and aligned features in task-specific layers. In fact, one of the key issues in DA is catastrophic forgetting of pre-learned knowledge, which severely degrades adaptation performance. Hence, in [40], [41] meta-learning was utilised to alleviate such an issue. Some other works [42], [43] utilised a memory buffer to replay learned knowledge during adaptation.

In addition, a few DA methods focused on either depth prediction [20], [39] or pose estimation [19]. For instance, [19] proposed an online meta-learning framework that allowed the model, which was pre-trained on a source data distribution, to be adapted to an unseen target data distribution. Besides, [20] utilised the meta-learning framework as well, but with additional efforts on forgetting prevention during adaptation. Though the above-mentioned DA methods were based on online adaptation for fast adaptation, they were not designed for the SLAM task. Meanwhile, they paid more attention to preserve pre-learned knowledge and thus overlooked learning discriminative features from target domain. By contrast, we propose an approach for online SLAM adaptation by learning more discriminative knowledge from the unknown domain and fusing them with pre-learned knowledge.

III. Proposed Method

As illustrated in Fig. 1, our proposed method mainly consists of three steps: 1) Pre-training the depth network and pose network on source domain dataset, 2) Learning to adpat

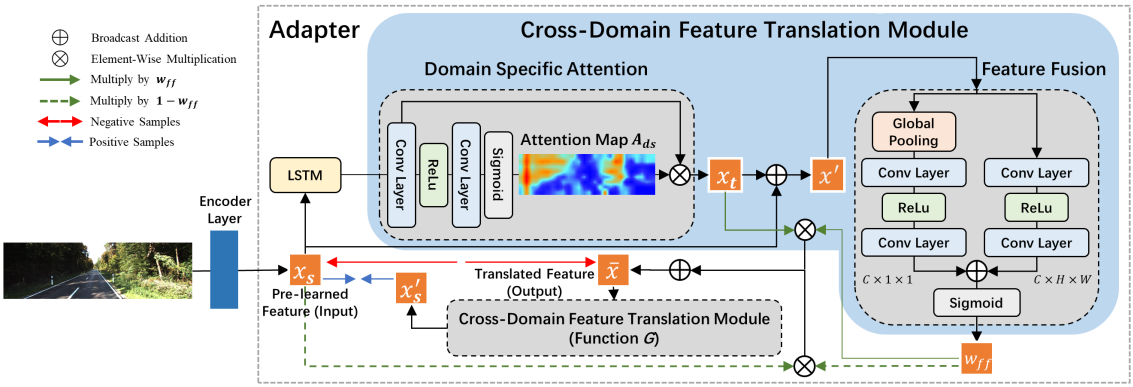

Fig. 2: Illustration of the proposed adapter. During adaptation, it first utilises **Cross-Domain Feature Translation Module** to translate pre-learned features x_s to new features $\bar{x}$ for adaptation. To achieve that, it extracts target domain-specific features x_t with **Domain Specific Attention**, and then fuses x_s and x_t into $\bar{x}$ using **Feature Fusion**. Finally, another cross-domain feature translation module (**function *G***) is applied to transfer $\bar{x}$ to x'_s such that in our **Cycle-Consistent Contrastive Learning** we define x_s as anchor, x'_s and $\bar{x}$ are its positive and negative samples, respectively.

which integrates the networks with the proposed adaptaters shown as green boxes and initialises these adapters using the meta-training scheme [21] with self-supervised loss, and 3) Adapting the networks and adapters to target domain dataset in a self-supervised manner.

In the following sections, we first illustrate the proposed adapter of which detailed architecture is shown in Fig. 2. Then, we introduce the self-supervised loss used in training those models.

A. Proposed Adapter

In this section, we introduce the proposed online adapter that is utilised to adjust the feature representations suitable for adaptation without the need to update the whole network. Briefly, it exploits a cross-domain feature translation module to translate the pre-learned features to suitable feature representations for adaptation. Besides, it further incorporates cycle-consistent contrastive learning to learn more discriminative target domain features while better preserving pre-learned knowledge to avoid catastrophic forgetting.

1) Cross-Domain Feature Translation Module: The cross-domain feature translation module aims to translate features from the source domain to suitable feature representations for adaptation. It consists of two components: **Domain Specific Attention (DSA)** and **Feature Fusion (FF)**. Here, the DSA block extracts target domain-specific feature $x_t \in \mathbb{R}^{L\times C\times H\times W}$ (L, C, H and W are sequence length, channel numbers, height and width) based on the pre-learned source domain feature $x_s \in \mathbb{R}^{L\times C\times H\times W}$. In essence, x_s is generated using network parameters learned from source domain, and thus considered as pre-learned feature. Afterwards, the FF block fuses both x_s and x_t into fused feature $\bar{x} \in \mathbb{R}^{L\times C\times H\times W}$ that involves target domain information without sacrificing past experience.

Domain Specific Attention (DSA): Rather than feeding x_s directly into DSA, we first apply the LSTM layer to exploit temporal information, such that $\hat{x}_s = convLSTM(x_s)$. Based on $\hat{x}_s$, we exploit pixel-wise attention to discover target domain features by paying more attentions to the domain-specific information, such as the tree regions that have been hardly seen in source domain training samples as shown in Fig. 3.

The pixel-wise attention includes two colutional layers with ReLU and sigmoid function as follows:

$$A_{ds} = \sigma(conv(ReLu(conv(\hat{x}_s)))), \tag{1}$$

where A_{ds} refers to the attention map with a shape of $L\times1\times H\times W$. That is, A_{ds} will give higher weights to the pixels in the feature map that have more domain-specific information and vice versa. Finally, the target domain-specific feature x_t can be obtained via:

$$x_t = x_s \odot A_{ds}, \tag{2}$$

where $\odot$ is the element-wise multiplication.

Feature Fusion (FF): Given the pre-learned feature x_s and target domain-specific feature x_t, we next combine them into a fused feature $\bar{x}$ using our FF block. In principal, the fused feature $\bar{x}$ should well balance the pre-learned and newly-learned information for robust adaptation. The final fused feature $\bar{x}$ for adaptation can be obtained via:

$$\bar{x} = w_{ff} \odot x_t + (1 - w_{ff}) \odot x_s, \tag{3}$$

where the weight w_{ff} acts as a balance between x_s and x_t. To determine the value of w_{ff}, we first concatenate x_s and x_t into a combined feature x' using a simple summation. An attention operation is then applied to the combined feature x', for which the attention operation aims to extract both global and local information. The process could be described as follows:

$$g_c = \frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W} x'_c(i,j), \tag{4}$$

$$A_g = conv(ReLu(conv(g_c))), \tag{5}$$

$$A_l = conv(ReLu(conv(x'))). \tag{6}$$

(a) Source domain image (b) Target domain image (c) Pixel attention map A_{ds} of (b)

Fig. 3: Visualization of generated attention map A_{ds}. The difference between source domain image (a) and target domain image (b) is significant. That is, the source domain image has many city-style scenarios with crowded buildings, while the target domain image contains more countryside-style objects such as trees. Therefore, in this case, the attention map (c) of image (b) provides more attention clues (denoted in red in (c)) to extract target domain-specific features such as trees.

Here, Eqn. (4) refers to the global pooling function that changes the shape of x' from $\mathbb{R}^{L\times C\times H\times W}$ to $\mathbb{R}^{L\times C\times 1\times 1}$, and $x'_c(i,j)$ stands for the value of c-th channel at position (i,j). Then, the output g_c feeds into channel-wise attention (Eqn. (5)) for global information extraction. Meanwhile, Eqn. (6) is a pixel-wise attention that extracts local information.

In Eqns. (5) and (6), $A_g \in \mathbb{R}^{L\times C\times 1\times 1}$ and $A_l \in \mathbb{R}^{L\times C\times H\times W}$ represent global and local contexts of concatenated feature x', respectively. Finally, they are combined to create attentional weights w_{ff}:

$$w_{ff} = \sigma(A_g \oplus A_l). \tag{7}$$

It is worth noting that the value of w_{ff} is between 0 and 1 due to the sigmoid function, and $\oplus$ here denotes the broadcast summation.

2) Cycle-Consistent Contrastive Learning : We propose cycle-consistent contrastive learning to further enforce that the feature $\bar{x}$ generated by Eqn. (3) for adaptation well preserve pre-learned knowledge and also contain more discriminative target domain-specific information.

To better preserve pre-learned knowledge, it exploits the cycle-consistent constraint to ensure that the features $\bar{x}$ transfered from pre-learned features x_s can be transferred back to x_s via:

$$x'_s = G(\bar{x}). \tag{8}$$

Here, the function G is another **cross-domain feature translation module** as introduced in section III-A.1, which first translates from $\bar{x}$ to x'_s. Then, our contrastive learning mechanism minimizes the similarity between x_s and x'_s to satisfy the cycle-consistent constraint.

By contrast, we maximise the similarity between x_s and $\bar{x}$ to discrminate them better and thus learn more target domain-specific information for $\bar{x}$. Then, our contrastive loss is:

$$L_{con} = \|x_s - x'_s\|_2^2 + max\{0, m - \|x_s - \bar{x}\|_2^2\}, \tag{9}$$

where $\|\|_2^2$ denotes the $l2$ distance, and we set m as 1 in our case.

B. Self-Supervised Training

The self-supervised loss contains a reconstruction loss, a geometry constraint, and a depth smoothness loss. The reconstruction loss is defined as:

$$L_{rec} = \sum_{i=1}^{L-1} \alpha||I_i - \hat{I}_i||_1^2 + (1-\alpha)\frac{1 - SSIM(I_i, \hat{I}_i)}{2}, \tag{10}$$

where α= 0.15, and $\hat{I}_i$, I_i are the reconstructed image and the input image, respectively. $SSIM()$ is the SSIM loss [44] commonly used in previous works like [10], [19] to measure the structural similarity between two images. $\hat{I}_i$ can be derived from I_{i+1} with predicted pose $\hat{T}_i^{i+1}$ between I_i and I_{i+1}.

To make the predicted depth map smooth, we also utilise the depth smoothness loss:

$$L_{smooth} = \sum_{i=1}^{L}\sum_{x,y} ||\nabla_{x,y}\hat{d}_i(x,y)||e^{-||\nabla_{x,y}I_i(x,y)||}, \tag{11}$$

where $\hat{d}_i$ is the predicted depth map of image I_i, and (x,y) denotes the pixel coordinates. This edge-aware smoothness loss can preserve the sharpness of the edges in the predicted depth map while smoothing out the noise.

Then, the overall unsupervised loss L_u is defined as:

$$L_u = ML_{rec} + \lambda_s L_{smooth} + M\lambda_g L_{gc}, \tag{12}$$

where L_{gc} denotes geometric consistency [10] to enforce the scale consistency of depth map. Similar to photometric consistency, the scale consistency between depth maps $\hat{d}_i$ and $\hat{d}_{i+1}$ can be evaluated by measuring the difference between $\hat{d}_i$ and the reconstructed depth map $\hat{d}'_i$ which is inferred from depth map $\hat{d}_{i+1}$ of image I_{i+1} with predicted pose $\hat{T}_i^{i+1}$. To mask out invalid pixels (e.g. pixels from moving objects), we apply a mask M to the reconstruction loss L_{rec} to remove the influence of moving objects. The mask M is generated from a small network f_M that consists of two convolutional layers and a sigmoid function. The network f_M outputs a mask based on the difference between the predicted depth map $\hat{d}_i$ and the reconstructed depth map $\hat{d}'_i$.

During adaptation, the total loss L_{tot} also includes the cycle-consistent contrastive loss L_{con}:

$$L_{tot} = ML_{rec} + \lambda_s L_{smooth} + M\lambda_g L_{gc} + \lambda_c L_{con}, \tag{13}$$

where λ_s, λ_g, λ_{cr} are set as 0.1, 0.5, 0.5.

IV. Experiments

A. Dataset and Evaluation Metrics

Virtual KITTI: Virtual KITTI [49] is a synthetic dataset containing videos from 6 different virtual scenes of an urban setting under different weather conditions. It has 50 monocular videos (21,260 frames in total) with ground truth camera poses and dense depth maps. We use all the sequences as source domain.

Methods	Error (lower is better)				Accuracy (higher is better)		
	AbsRel ↓	SqRel ↓	RMSE ↓	RMSLog ↓	$< 1.25 \uparrow$	$< 1.25^2 \uparrow$	$< 1.25^3 \uparrow$
vKITTI to KITTI							
[20]	0.153	-	5.508	-	0.776	0.923	-
Ours	**0.151**	1.221	**5.507**	0.224	**0.801**	**0.925**	0.965
Cityscapes to KITTI							
[20]	0.138	-	5.348	-	**0.819**	0.931	-
Ours	**0.136**	1.170	**5.301**	0.217	0.818	**0.935**	0.967
KITTI to KITTI (No Domain Adaptation)							
[18]	0.105	0.753	4.389	0.179	0.890	0.965	0.983
[45]	0.113	0.704	4.581	0.184	0.871	0.961	0.984
[46]	0.114	0.813	4.706	0.191	0.873	0.960	0.982
[47]	0.099	0.708	4.372	0.175	0.900	0.967	0.984

TABLE I: Quantitative comparison of depth prediction on two adaptation scenarios.

Methods	Seq 09		Seq 10	
	$t_{rel} \downarrow$	$r_{rel} \downarrow$	$t_{rel} \downarrow$	$r_{rel} \downarrow$
vKITTI to KITTI				
[19]	17.94	5.36	26.27	**7.01**
Ours	**17.51**	**4.98**	**24.97**	7.97
CityScapes to KITTI				
[19]	18.25	**5.97**	28.16	11.08
Ours	**18.07**	6.15	**26.51**	**9.01**
KITTI to KITTI (No Domain Adaptation)				
[30]	3.49	0.010	5.81	0.018
[19]	5.89	3.34	4.79	0.83
[46]	5.08	1.05	4.32	2.34
[48]	2.36	1.06	3.00	1.28

TABLE II: Quantitative comparison of pose prediction on two adaptation scenarios. t_{rel} [%]: average translational RMSE, r_{rel} [deg/m]: average rotational RMSE.

Methods	Error (lower is better)				Accuracy (higher is better)		
	AbsRel ↓	SqRel ↓	RMSE ↓	RMSLog ↓	$< 1.25 \uparrow$	$< 1.25^2 \uparrow$	$< 1.25^3 \uparrow$
vKITTI to KITTI							
PF	0.304	1.917	8.109	0.393	0.607	0.813	0.946
DF	0.315	2.095	9.973	0.437	0.583	0.783	0.861
PF+DF	0.155	1.242	5.603	0.230	0.795	0.918	0.961
PF+DF+CS	0.153	1.239	5.535	0.228	0.799	0.922	0.963
Ours	**0.151**	**1.221**	**5.507**	**0.224**	**0.801**	**0.925**	**0.965**
CityScapes to KITTI							
PF	0.310	1.911	8.086	0.388	0.606	0.810	0.945
DF	0.324	2.101	9.981	0.435	0.590	0.789	0.864
PF+DF	0.144	1.352	5.448	0.224	0.811	0.930	0.965
PF+DF+CS	0.139	1.184	5.372	**0.213**	**0.820**	0.932	0.965
Ours	**0.136**	**1.170**	**5.301**	0.217	0.818	**0.935**	**0.967**

TABLE III: Ablation study on depth prediction.

Methods	Seq 09		Seq 10	
	$t_{rel} \downarrow$	$r_{rel} \downarrow$	$t_{rel} \downarrow$	$r_{rel} \downarrow$
vKITTI to KITTI				
PF	20.41	43.79	40.82	16.22
DF	23.57	50.13	51.90	22.87
PF+DF	18.33	6.07	25.98	9.51
PF+DF+CS	17.96	5.13	26.01	8.62
Ours	**17.51**	**4.98**	**24.97**	**7.97**
CityScapes to KITTI				
PF	19.94	45.18	38.24	17.46
DF	25.72	52.81	48.04	20.59
PF+DF	18.15	6.36	28.03	9.24
PF+DF+CS	18.46	5.27	26.95	**8.01**
Ours	**18.07**	**5.24**	**26.51**	**8.01**

TABLE IV: Ablation study on pose prediction.

CityScapes: CityScapes [50] is a real-world dataset including video sequences captured by cameras mounted on cars moving in various cities. We follow the official train/test split to generate our training set, for which it eventually generates videos from 18 different cities as source domain.

KITTI: KITTI raw [51] and KITTI odometry [52] are both real-world datasets containing video sequences recorded from a moving vehicle, in which KITTI odometry includes 11 video sequences with ground truth poses. For depth prediction evaluation, we use Eigen's train/test split [53] as target domain to obtain the test set of KITTI raw. For pose estimation evaluation, we follow the same train/test split as target domain in [5] which used sequences 09 and 10 for performance evaluation.

Depth Evaluation Metrics: To evaluate depth estimation, we adopt the same metrics used in existing methods [5], [54], including the mean absolute relative error, the average squared relative error, the root mean squared error, the root mean squared log error, and the accuracy under threshold $\delta \in \{1.25^1, 1.25^2, 1.25^3\}$. As self-supervised methods are not able to predict depth in absolute scale, the predicted depth maps are multiplied by a scale factor to align the medians of the prediction with those of the ground truths [5].

Pose Evaluation Metrics: For pose prediction evaluation, we measure the averaged root mean square error (RMSE) [52] of translation and rotation. Similar to depth estimation, the predicted poses are not in absolute scale and are multiplied by a scale factor to align with the ground truth camera poses.

B. Implementation Details

Our model is implemented using PyTorch. The depth network and pose network have similar structures as [10] other than using separate prediction heads to predict translations and rotations, and the resolution of the input images is 256×832. In the pre-training stage, we initialise our depth network and pose network with ResNet-50 pre-trained on ImageNet. The model is trained for 10,000 iterations with ground truth depth maps, and for another 10,000 iterations in a self-supervised manner. The learning rate in the pre-training stage is set to $1e^{-4}$. After that, the adapters and decoders are trained using the meta-learning framework MAML [21]. In the meta-learning stage, we set the learning rate for the inner optimisation to $1e^{-4}$, and for the outer optimisation to $1e^{-5}$. The adapters and decoders are meta-trained for 20,000 iterations. Meanwhile, the parameters of encoders are fixed during the meta-learning stage. Stochastic Gradient Descent (SGD) is used for the inner optimisation of meta-learning, while Adam optimiser is used for pre-training stage, the outer optimisation of meta-learning and the adaptation. We set the batch size to 4, 2 and 1 for pre-training, meta-training and adaptation, respectively. The length of the input sequence is set to 5.

C. Performance Evaluation

1) Depth Prediction Evaluation: The proposed model is trained on two source domains - virtual KITTI and CityScapes, and then evaluated on KITTI. As shown Table I, our proposed method considering both pre-learned and target domain-specific features outperforms the baseline method [20] in both adaptation scenarios (i.e., from virtual KITTI to KITTI and from CityScapes to KITTI) with respect to most of the evaluation metrics. We also show the results of the methods not performing domain adaptation (i.e., KITTI to KITTI) to reflect the challenge of cross-domain SLAM. The results suggest that the methods [18], [45], [46], [47] pre-trained and evaluated on the same KITTI dataset are able to perform much better than those performing cross-domain SLAM. In Fig. 4, we can further observe that depth maps estimated with adapters capture more fine-grained details and are closer to the ground truths than those without adapters.

(a) Image (b) Ground Truths (c) Without Adapters (d) Ours

Fig. 4: Visualization of depth estimation. The first two rows in (d) show the depths estimated by our model pretrained on vKITTI, while the last two rows in (d) show the results of our model pretrained on CityScapes. Results with adaptation (d) are consistently better than those without adaptation (c) with reference to the ground truths (b).

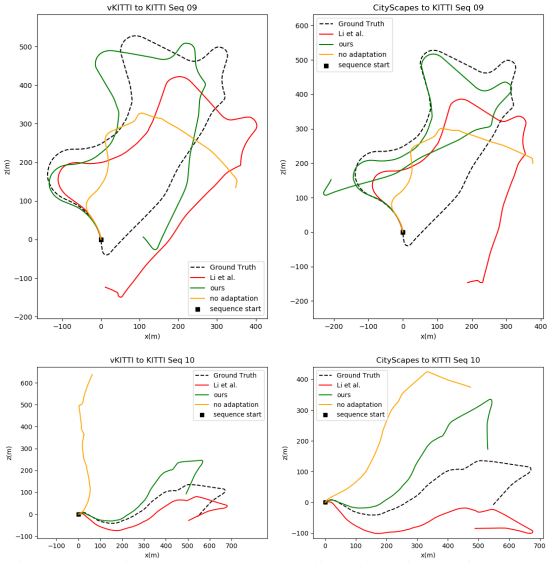

Fig. 5: Comparison between predicted trajectories on KITTI.

2) Pose Prediction Evaluation: Table II shows a comparison between our method and several baselines including the state-of-the-art online adaptation pose estimation method [19], also together with the non-adaptation based methods (i.e., KITTI to KITTI). While those without domain adaptation perform best since they were pre-trained and evaluated on the same KITTI dataset, similarly, our method is superior to the other online adaptation baseline [19] with respect to most of the metrics. This is more obvious as shown in Fig. 5 where the beginning part of our estimated trajectory (green line) is significantly closer than that of [19] (red line) to the ground truth (dotted line). That means our method drifted from ground truth less, and achieved more robust and accurate adaptation. Besides, the trajectory generated by our model without adapter (yellow line) performed poorly when no adaptation is applied.

D. Ablation Studies

First, we denote **PF** as a model without adapters, which performs adaptation using pre-learned features only. Meanwhile, **DF** represents a model only utilising target domain-specific features generated by the **domain specific attention** as introduced in III-A.1. From Tables III and IV, we observe that online adaptation with either pre-learned features (**PF**) or target domain-specific features (**DF**) is insufficient to ensure accurate depth and pose estimation. Therefore, the combination of pre-learned and target domain-specific features (**PF+DF**) substantially improves accuracy.

Next, we evaluate the effectiveness of the proposed cycle-consistent contrastive loss. Our contrastive loss involves two terms - one exploits cycle-consistent constraint to better preserve pre-learned knowledge, and the other one aims to learn more discriminative target domain features. Here, we first study integrating fused features with cycle-consistent constraint only (**PF+DF+CS**) before adding entire contrastive loss to generate our full model (**Ours**). As shown in Tables III and IV, adding cycle-consistent constraint can further improve the accuracy, but our method with complete contrastive loss achieves the best depth and pose predictions in most cases.

V. Conclusions

In this paper, we present a novel self-supervised SLAM method to achieve online adaptation in the open world. To our best knowledge, it is one of the first studies on monocular visual SLAM with online domain adaptation. We devise a novel adapter to adjust feature representations suitable for adaptation, but fixing the pre-trained model to prevent catastrophic forgetting on past experiences. Besides, we propose the cycle-consistent contrastive learning to further learn discriminative features from target domain, while better preserving pre-learned knowledge to achieve more robust adaptation. Comprehensive experiments on both synthetic and real-world datasets, including Virtual KITTI, KITTI and CityScapes, demonstrate that our proposed method outperform several state-of-the-art baselines.

REFERENCES

[1] C. Campos, R. Elvira, J. J. Gómez, J. M. M. Montiel, and J. D. Tardós, "ORB-SLAM3: An accurate open-source library for visual, visual-inertial and multi-map SLAM," *arXiv preprint arXiv:2007.11898*, 2020.

[2] R. He, Z. Wang, Y. Fan, and D. Dagan Feng, "Multiple scattering model based single image dehazing," in *2013 IEEE 8th Conference on Industrial Electronics and Applications (ICIEA)*, 2013.

[3] R. He, Z. Wang, H. Xiong, and D. D. Feng, "Single image dehazing with white balance correction and image decomposition," in *2012 International Conference on Digital Image Computing Techniques and Applications (DICTA)*, 2012.

[4] S. Wang, R. Clark, H. Wen, and N. Trigoni, "Deepvo: Towards end-to-end visual odometry with deep recurrent convolutional neural networks," in *ICRA*, 2017.

[5] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe, "Unsupervised learning of depth and ego-motion from video," in *CVPR*, 2017.

[6] R. Clark, S. Wang, H. Wen, A. Markham, and N. Trigoni, "Vinet: Visual-inertial odometry as a sequence-to-sequence learning problem," in *AAAI*, 2017.

[7] M. Xiong, Z. Zhang, W. Zhong, J. Ji, J. Liu, and H. Xiong, "Self-supervised monocular depth and visual odometry learning with scale-consistent geometric constraints," in *IJCAI*, 2020.

[8] R. Li, S. Wang, Z. Long, and D. Gu, "Undeepvo: Monocular visual odometry through unsupervised deep learning," in *ICRA*, 2018.

[9] Y. Almalioglu, M. R. U. Saputra, P. P. de Gusmao, A. Markham, and N. Trigoni, "Ganvo: Unsupervised deep monocular visual odometry and depth estimation with generative adversarial networks," *ICRA*, 2019.

[10] J.-W. Bian, Z. Li, N. Wang, H. Zhan, C. Shen, M.-M. Cheng, and I. Reid, "Unsupervised scale-consistent depth and ego-motion learning from monocular video," in *NeurIPS*, 2019.

[11] U.-H. Kim, S.-H. Kim, and J.-H. Kim, "Simvodis++: Neural semantic visual odometry in dynamic environments," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 4244–4251, 2022.

[12] S. Li, C. Liu, Q. Lin, B. Xie, Z. Ding, G. Huang, and J. Tang, "Domain conditioned adaptation network," in *AAAI*, 2020.

[13] H. Akada, S. F. Bhat, I. Alhashim, and P. Wonka, "Self-supervised learning of domain invariant features for depth estimation," in *WACV*, 2022.

[14] S. Lee, J. Hyun, H. Seong, and E. Kim, "Unsupervised domain adaptation for semantic segmentation by content transfer," in *AAAI*, 2021, pp. 8306–8315.

[15] Y. Zhang, Z. Qiu, T. Yao, D. Liu, and T. Mei, "Fully convolutional adaptation networks for semantic segmentation," in *CVPR*, 2018.

[16] S. Zhang, J. Zhang, and D. Tao, "Towards scale consistent monocular visual odometry by learning from the virtual world," in *2022 International Conference on Robotics and Automation (ICRA)*, 2022.

[17] S. Sankaranarayanan, Y. Balaji, C. D. Castillo, and R. Chellappa, "Generate to adapt: Aligning domains using generative adversarial networks," in *CVPR*, 2018.

[18] A. Gurram, A. F. Tuna, F. Shen, O. Urfalioglu, and A. M. López, "Monocular depth estimation through virtual-world supervision and real-world sfm self-supervision," *IEEE Transactions on Intelligent Transportation Systems*, 2021.

[19] S. Li, X. Wang, Y. Cao, F. Xue, Z. Yan, and H. Zha, "Self-supervised deep visual odometry with online adaptation," in *CVPR*, 2020.

[20] Z. Zhang, S. Lathuiliere, E. Ricci, N. Sebe, Y. Yan, and J. Yang, "Online depth learning against forgetting in monocular videos," in *CVPR*, 2020.

[21] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *ICML*, 2017.

[22] F. Xue, X. Wang, S. Li, Q. Wang, J. Wang, and H. Zha, "Beyond tracking: Selecting memory and refining poses for deep visual odometry," in *CVPR*, 2019.

[23] E. Parisotto, D. S. Chaplot, J. Zhang, and R. Salakhutdinov, "Global pose estimation with an attention-based recurrent network," in *CVPRW*, 2018.

[24] S. Xu, H. Xiong, Q. Wu, and Z. Wang, "Attention-based long-term modeling for deep visual odometry," in *2021 Digital Image Computing: Techniques and Applications (DICTA)*, 2021, pp. 1–8.

[25] D. Xu, W. Wang, H. Tang, H. Liu, N. Sebe, and E. Ricci, "Structured attention guided convolutional neural fields for monocular depth estimation," in *CVPR*, 2018.

[26] X. Chen, X. Chen, and Z.-J. Zha, "Structure-aware residual pyramid network for monocular depth estimation," in *IJCAI*, 2019.

[27] H. Zhou, B. Ummenhofer, and T. Brox, "Deeptam: Deep tracking and mapping," in *ECCV*, 2018.

[28] M. Bloesch, J. Czarnowski, R. Clark, S. Leutenegger, and A. J. Davison, "Codeslam — learning a compact, optimisable representation for dense visual slam," in *CVPR*, 2018.

[29] S. Li, F. Xue, X. Wang, Z. Yan, and H. Zha, "Sequential adversarial learning for self-supervised deep visual odometry," in *ICCV*, October 2019.

[30] Y. Zou, P. Ji, Q.-H. Tran, J.-B. Huang, and M. Chandraker, "Learning monocular visual odometry via self-supervised long-term modeling," in *ECCV*, 2020.

[31] C. Wang, Y.-P. Wang, and D. Manocha, "Motionhint: Self-supervised monocular visual odometry with motion constraints," in *2022 International Conference on Robotics and Automation (ICRA)*, p. 1265–1272.

[32] R. Wang, S. M. Pizer, and J.-M. Frahm, "Recurrent neural network for (un-)supervised learning of monocular video visual odometry and depth," in *CVPR*, 2019.

[33] X. Li, Y. Hou, P. Wang, Z. Gao, M. Xu, and W. Li, "Transformer guided geometry model for flow-based unsupervised visual odometry," *arXiv preprint arXiv:2101.02143*, 2021.

[34] K. Bousmalis, N. Silberman, D. Dohan, D. Erhan, and D. Krishnan, "Unsupervised pixel-level domain adaptation with generative adversarial networks," in *CVPR*, 2017, pp. 95–104.

[35] C. Zheng, T.-J. Cham, and J. Cai, "T2net: Synthetic-to-realistic translation for solving single-image depth estimation tasks," in *ECCV*, 2018.

[36] J. N. Kundu, P. K. Uppala, A. Pahuja, and R. V. Babu, "Adadepth: Unsupervised content congruent adaptation for depth estimation," in *CVPR*, 2018.

[37] K. PNVR, H. Zhou, and D. Jacobs, "Sharingan: Combining synthetic and real data for unsupervised geometry estimation," in *CVPR*, 2020.

[38] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *ICML*, 2015.

[39] A. Gurram, A. F. Tuna, F. Shen, O. Urfalioglu, and A. M. López, "Monocular depth estimation through virtual-world supervision and real-world sfm self-supervision," *arXiv preprint arXiv:2103.12209*, 2021.

[40] H. Liu, M. Long, J. Wang, and Y. Wang, "Learning to adapt to evolving domains," in *NeurIPS*, 2020.

[41] K. Javed and M. White, "Meta-learning representations for continual learning," in *NeurIPS*, 2019.

[42] G. Gupta, K. Yadav, and L. Paull, "Look-ahead meta learning for continual learning," in *NeurIPS*, 2020.

[43] J. Dong, Y. Cong, G. Sun, B. Ma, and L. Wang, "I3dol: Incremental 3d object learning without catastrophic forgetting," in *AAAI*, 2021.

[44] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[45] W. Zhao, S. Liu, Y. Shu, and Y.-J. Liu, "Towards better generalization: Joint depth-pose learning without posenet," in *CVPR*, 2020.

[46] J.-W. Bian, H. Zhan, N. Wang, Z. Li, L. Zhang, C. Shen, M.-M. Cheng, and I. Reid, "Unsupervised scale-consistent depth learning from video," *International Journal of Computer Vision (IJCV)*, 2021.

[47] C. Zhao, Y. Zhang, M. Poggi, F. Tosi, X. Guo, Z. Zhu, G. Huang, Y. Tang, and S. Mattoccia, "Monovit: Self-supervised monocular depth estimation with a vision transformer," in *International Conference on 3D Vision*, 2022.

[48] X. Li, Y. Hou, P. Wang, Z. Gao, M. Xu, and W. Li, "Transformer guided geometry model for flow-based unsupervised visual odometry," *Neural Computing and Applications*, vol. 33, no. 13, pp. 8031–8042, 2021.

[49] A. Gaidon, Q. Wang, Y. Cabon, and E. Vig, "Virtual worlds as proxy for multi-object tracking analysis," in *CVPR*, 2016.

[50] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *CVPR*, 2016.

[51] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *IJRR*, 2013.

[52] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *CVPR*, 2012.

[53] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," in *NIPS*, 2014.

[54] C. Godard, O. Mac Aodha, M. Firman, and G. J. Brostow, "Digging into self-supervised monocular depth estimation," in *ICCV*, 2019.