Self-Improving Semantic Perception on a Construction Robot

Hermann Blum, Francesco Milano, René Zurbrügg, Roland Siegward, Cesar Cadena, Abel Gawel
{blumh,fmilano,zrene,rsiegwart,cesarc,gawela}@ethz.ch
Autonomous Systems Lab, ETH Zürich

***Abstract*—We propose a novel robotic system that can improve its semantic perception during deployment. Contrary to the established approach of learning semantics from large datasets and deploying fixed models, we propose a framework in which semantic models are continuously updated on the robot to adapt to the deployment environments. Our system therefore tightly couples multi-sensor perception and localisation to continuously learn from self-supervised pseudo labels. We study this system in the context of a construction robot registering LiDAR scans of cluttered environments against building models. Our experiments show how the robot's semantic perception improves during deployment and how this translates into improved 3D localisation by filtering the clutter out of the LiDAR scan, even across drastically different environments. We further study the risk of catastrophic forgetting that such a continuous learning setting poses. We find memory replay an effective measure to reduce forgetting and show how the robotic system can improve even when switching between different environments. On average, our system improves by 60% in segmentation and 10% in localisation compared to deployment of a fixed model, and it keeps this improvement up while adapting to further environments.**

I. Introduction

Mobile robots are expected to be deployed in increasingly unstructured environments. While they will have access to information about the environment, such as basic maps or models, advances in learning-based systems enable robots to partially understand the environment through, e.g., object detection or semantic classification [37, 47]. Such understanding is a key requirement to enable many complex, dynamic robotic applications such as autonomous driving or mobile manipulation [30, 62].

Anticipation of a wide variety of environmental conditions is required for safe operation, however difficult if not impossible, and robotic actors with a high degree of autonomy are required to adapt to unexpected and changing conditions for robust operation. Yet, deployment of learning-based systems typically means pre-training a model on a variety of data and then using this static model during deployment. In this work, we explore how learning can be used as a means to self-improve semantic perception during exploration of the environment. Enabling robots to adapt during the job poses three main challenges on robotic systems:

1) Models need to be efficiently (re)trained to incorporate new data (*incremental / continual learning*).
2) Acquired knowledge should be kept while adapting to new tasks and environments (avert *forgetting*).
3) Training signals of the environment are required during deployment, i.e. without manual human supervision in the loop (*self-supervision*).

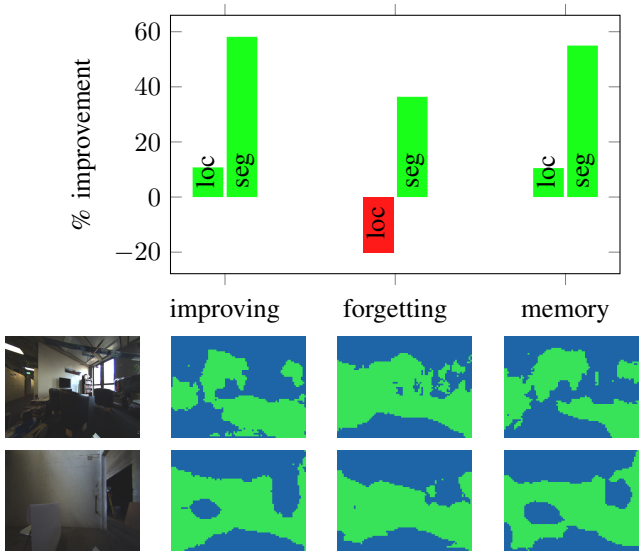


Fig. 1: Our system on average improves semantic segmentation by 60% and median localisation error by 10% during deployment. However, when moving to a different environment type, these improvements can be lost due to *catastrophic forgetting*. We find that memory replay is an effective method against forgetting that can retain large improvements on past and current environments. The graph shows the average relative improvement compared to a fixed pretrained segmentation model. Below are qualitative examples of the segmentation corresponding to the plot above, where blue is *foreground* and green is *background*.

The first two points fit particularly with the field of *continual learning* [42, 32], also referred to as *incremental learning* [8, 52], and *lifelong learning* [14, 57, 54]. This learning paradigm deals with the problem of training neural networks in settings where tasks or classes are presented incrementally, or in which the domain – and more generally the data distribution – changes over time [42]. Furthermore, robotic systems generally rely on constrained resources and can only store a limited amount of information, a limitation that continual learning can tackle by reducing the need for big models and datasets fitting all possible deployment observations [32]. In order to continually learn during robot deployment, robots require to generate streams of

training signals without humans in the loop. We refer to this as *self-supervised pseudolabeling*, where first approaches in this domain leverage multi-sensor [56] or multi-task systems [13] to transfer labels between modalities and tasks respectively.

We particularly study the use-case of deploying robots in environments with existing 3D floorplan maps in which the robot is required to localize. This application case is prevalent in building construction and service robotics as increasingly digital 3D floorplan maps are available in these environments. In contrast to the map data that usually represents the static building structure, most relevant for robotic localization (we will refer to this as *background*), actual environments contain large amounts of un-modelled objects or clutter (we will refer to this as *foreground*), potentially obstructing localization performance [7]. Such robotic applications are particularly interesting for self-supervised systems because there is often not much domain-specific data available [30].

Key to our proposed system is the combination of continual learning and self-supervision. A continually learned segmentation model serves as an input filter to the localisation, segmenting the scene into *foreground* and *background*. Based on the localisation in the 3D floorplan, we then harvest self-supervised pseudo-labels for the continual learning of the segmentation. This coupling generates a feedback loop. The robot localises to generate training data for background-foreground segmentation and segments to localise, resulting in improvement of both segmentation and localisation during deployment (see Figure 1). We thus enable online life-long self-supervised learning of semantic scene understanding. To realize and validate this system, the following sections describe in particular:

- Pseudolabel generation for self-supervision based exclusively on multi-modal calibration and an available floorplan
- Integration of continual learning methods with respect to domain adaptation in semantic segmentation
- Evaluation of segmentation, localisation, and forgetting when deploying the robotic system into three different environments: a construction site, a parking garage, and an office environment.

II. Related Work

A. Self-Improving Robotic Systems

The idea of self-improving, learning robotic agents has been explored before. One framework in which agents are self-improving is reinforcement learning (RL). With RL, robots have been learning to walk [23], grasp objects [41], or fly [26]. All these systems indeed learn by self-improving over time, often failing in the beginning of the learning process. Usually these learned models are fixed once they acquired the necessary skills, instead of life-long learning. This is because they require supervision signals, e.g. from simulators that are not available during deployment. However, online adaptation of model-based RL has been shown for example in [18].

Self-improving robotic systems have also been described outside of RL. For example, [36] and [28] describe online parameter optimisations for model predictive control. The adaptive stereo vision of Tonioni et al. [58] has been a particular inspiration for this work. Very related is also Sofman et al. [55], who learn a probabilistic model for terrain traversability in an online and self-supervised fashion. Interestingly, the mentioned previous works often do not explicitly address the problem of forgetting, which is a more prevalent problem in our semantic domain adaptation.

B. Self-Supervision and Pseudolabels

A large range of works explored different variants of self-supervision to learn useful image features in convolutional neural networks (CNNs). These techniques include learning to (re)color images [63], to (un)rotate images [21], or to relatively position random crops [15]. However, supervision is always required to relate the learned features to any meaning. In mobile agents, egomotion was found to be a promising, cheaply available self-supervision signal for a range of tasks [2]. Photometric consistency between video frames is used to jointly learn camera calibration, visual odometry, depth estimation, and optical flow [13].

A different line of works produces pseudolabels for segmentation by leveraging models trained on more available data, such as image classification. Class activation maps (CAM) of image classifiers [65] can be used to generate sparse regional annotation for an image. However, since CAMs tend to focus on small discriminative regions, directly using them for training a semantic segmentation network can be problematic. Recent methods therefore aim to expand the regions detected by CAMs by either working on the image [11], on the features [31], or by employing region growing [25, 3, 4]. Apart from leveraging large image classification datasets with CAMs, other methods were proposed to translate and refine knowledge from complex trained models. Reza et al. [46] refine semantic annotations by optimizing over Mask R-CNN predictions, room layout estimation, and superpixels relating multiple video frames. Porzi et al. [43] refine a segmentation model by using tracking and optical flow to improve predictions into automatic annotations. Sun et al. [56] map RGB-D images into a point cloud, autonomously labeling these point clouds in 3D space by leveraging bounding box annotations and projecting them back onto the image, in order to obtain pseudo labels that can be used for training.

The above described methods can be used to generate pseudolabels for a range of different target applications. While there is no direct prior work for background-foreground-segmentation, our proposed method builds up on similar ideas to use observable characteristics of the environment to produce a learning signal for the target task, which in our case is image segmentation. At the same time, we leverage existing annotated data from related task as prior knowledge.

C. Continual Learning for Robotics

In recent years, an increasing number of works [29, 34, 45, 50, 42] in the deep learning literature investigated the problem of continual learning, in which neural network models are

trained from non-stationary data distributions over a series of different tasks, across which the domain or the classes in the training data vary [32]. The main objective of continual learning is to optimize for the performance on each task or domain with which the network is presented at any given time, while achieving positive knowledge transfer between the tasks, and preventing performance on the previous tasks from decreasing. This drop in performance on previous tasks is a phenomenon often observed when the drifts in the task and data distribution are not taken into account, and is commonly referred to as *catastrophic* forgetting [17, 42].

One possibility to counteract this negative effect would be to simply store all the data from the previous tasks and retrain the network from scratch at each task. However, this is often not feasible in practice, in particular in scenarios in which the amount of available memory is limited or in which the model needs to be updated and deployed at the same time, and offline training is not possible [32]. For this reason, multiple alternative techniques have been proposed to tackle catastrophic forgetting, ranging from architectural modifications [48, 50] to regularization techniques [29, 34] and methods based on memories [45, 35] or generative models [51, 59]. While each of these techniques can be effective in mitigating catastrophic forgetting to a certain degree depending on the application and on the scenario specifications [42], in all cases a trade-off has to be accounted for between the amount of memory- and computational resources, and extent of the negative backward knowledge transfer. Evaluations of these techniques on complex robotic perception tasks often focus on class-incremental learning [16, 61].

In our work, we supplement the training data at each new environment through the use of a memory that keeps a limited number of samples from the previous environment, a method referred to as *replay buffers* [24, 64]. However, we also evaluate the use of regularization techniques often employed in the literature [29], and show the advantage of using memory-based approaches in our scenario, under different replay regimes.

A number of works have explored the use of continual-learning techniques in the context of semantic segmentation [39, 40, 10, 16, 61]. Michieli and Zanuttigh [39] propose a distillation approach that acts either at the output logits of the network or at an intermediate feature space. The same authors extend this framework in [40] by introducing alternative distillations schemes and incorporating the uncertainty in the estimations from previous models. Cermelli et al. [10] adopt a distillation approach which explicitly models the semantic shift of the pixels from the background class. Douillard et al. [16] and Yu et al. [61] address a similar problem by using pseudo-labels generated from a previous version of the model. All the above methods focus on a class-incremental setting, in which new classes are introduced across tasks, whereas for our application we are interested in the scenario in which classes (background and foreground) do not vary over time, but the environment in which the agent is deployed changes. In this sense, the problem we aim at tackling is also related to a line of works that focus on domain adaptation [66, 49, 33],

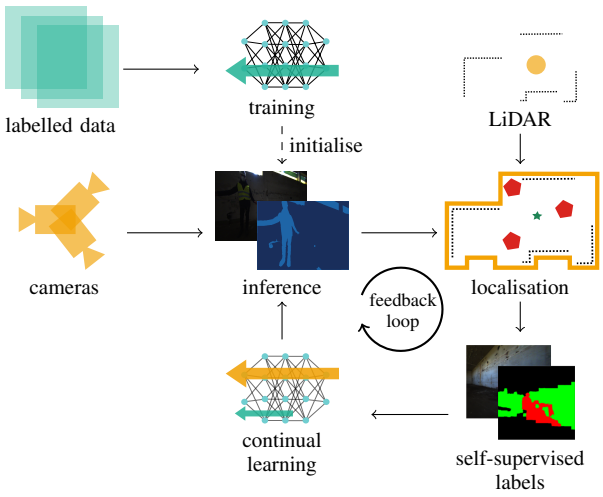


Fig. 2: Overview of the proposed self-improving system. While the perception still incorporates existing training data to form a good prior, it is not fixed during deployment, as would be the established approach. Instead, our semantic segmentation is updated during deployment based on continual learning methods. The necessary training signal is generated from self-supervised pseudolabels, which are therefore available during deployment without manual labelling. We mark signals from the deployment environment in orange and from the pretraining domain in green.

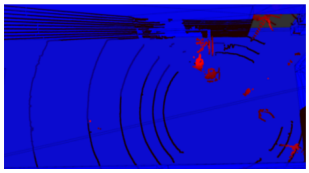
in which the goal is to tune networks models towards better generalization to a new domain (target) which possibly has a large semantic gap from the one of which the models are trained (source). However, these works generally assume that both the source and target domain are known at training time, and the models are not designed to be updated in an online fashion. Conversely, we address the setting in which the deployment domain is not known beforehand, and an agent has to update its semantic knowledge on the current environment without forgetting previously seen environments.

III. Proposed System

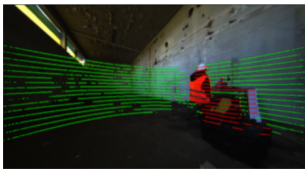
We propose a self-improving perception system that interlinks localisation within a map and semantic segmentation of the scene. Importantly, we define the semantics of the scene not as arbitrary class labels, but by the observable affordance that some parts of the scene are mapped (*background*) and some are not (*foreground*). Therefore, we create pseudolabels based on the localisation in the map to train the semantic segmentation, an we use the segmentation into *foreground* and *background* to inform the localisation. This creates a feedback loop that can yield improvements in both parts, as can be seen in Figure 2.

A. Semantically Informed Localisation

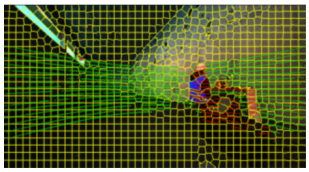
We localize the robot based on aligning 3D LiDAR scans with the given floorplan in the form of a 3D mesh as in [7]. Given the building model mesh M, a pointcloud of the LiDAR scan P, and an initial alignment $T^{(t=0)}_{\text{mesh} \rightarrow \text{lidar}}$, we find



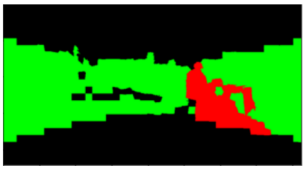
a) Mesh and Pointcloud



b) Projected Pointcloud



c) Oversegmented Image



d) Pseudo Labels

Fig. 3: Overview of the pseudo labeling approach. a) Depicts the pointcloud and the architectural mesh. Points that are closer to the mesh are colored darker. b) Depicts the sparse projected labels after thresholding for a given distance threshold δ. Points colored in red are assigned the foreground class, whereas points colored in green are assigned the background class. c) Shows the oversegmentation of the image using superpixels d) Depicts the final Pseudo labels using superpixels and majority voting.

subsequent robot poses as

$$T^{(t)}_{\text{mesh} \to \text{lidar}} = \text{ICP}(M, P^{(t)}, T^{(t-1)}_{\text{mesh} \to \text{lidar}}).$$

We use point-to-plane ICP [12] and filter out points of large distance and other criteria. Specific parameters for our experiments are reported in Appendix D.

To further divide the scan P into *foreground* and *background* points, we use additional information from a camera system mounted on top of the LiDAR. Once camera images are semantically segmented, we filter $P_{\text{static}} \subseteq P$ as those points $p \in P$ whose reprojected pixel in image frame is segmented as *background* and localise with $\text{ICP}(M, P^{(t)}_{\text{static}}, T^{(t-1)}_{\text{mesh} \to \text{lidar}})$.

B. Pseudolabel Generation

We generate pseudo labels for each camera by labeling the captured LiDAR pointcloud leveraging a current pose estimate as well as an architectural mesh of the building. Our labels contain foreground, background and unknown classes and are created in two steps. First, for each point of the localised LiDAR scan, we calculate the distance to the closest plane of the mesh using fast intersection and distance computation [5]. We then check if the distance surpasses a given threshold δ. If so, the point is assigned the *foreground* class, otherwise the *background* class. In the second stage, we project each point onto the respective camera frames and refine the projection using superpixels created with the SLIC [1] algorithm, which utilizes k-means clustering. In particular, we first oversegment the image into a superpixel set S. A superpixel $s \in S$ is then assigned a class according to a majority voting of the contained projected labels. We further improve the segmentation by discarding superpixels whose depth variance surpasses a given threshold. An overview of the approach is depicted in Figure 3.

This proposed pseudolabel generation does not produce optimal labels[1]. The goal however is not to generate perfect labels, but a useful training signal that can be generated on-the-fly without requiring any external supervision. As our experiments demonstrate, even this noisy learning signal can be useful.

C. Domain Adaptation with Continual Learning

To solve the task of background-foreground segmentation, we incrementally train a neural network architecture on different data sources. There are many large and complex architectures available for semantic segmentation. However, to cater to the goal of online learning, we use a lightweight architecture based on Fast-SCNN [44]. We pre-train the network on the NYU-Depth v2 dataset [53], which contains 1449 images extracted from video sequences of indoor scenes, each with per-pixel semantic annotations. We map the classes *wall*, *ceiling*, and *floor* to background and regard everything else as foreground. We perform this initial (pre-)training step to allow the model to acquire prior knowledge that can then be leveraged as an inductive bias to perform the same segmentation task on subsequent environments that the agent is presented with.

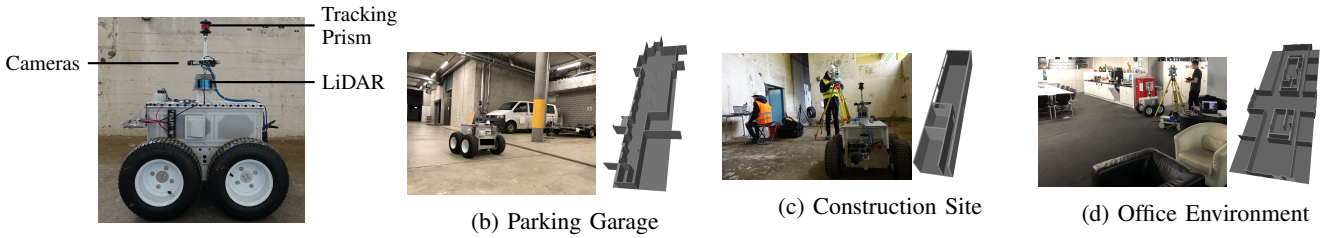
The network is then fine-tuned with self-supervision through the pseudolabels generated on the real-world scene in which the robot is deployed. With reference to the nomenclature often used in continual learning [32], we consider each new environment a *task*, and we assume *task boundaries* to be known. Every time the robot is moved to a new environment, the same scheme as above is applied, i.e., the network trained on the previous environments is provided as new training data the pseudolabels generated from the current environment.

In order to achieve this domain adaptation and allow the network to improve the segmentation accuracy on the environment in which it is currently deployed without forgetting information learned on the previous tasks, we adopt a method based on memory replay buffers. When adapting to a new environment, each training batch is filled with the frames collected in the current environment, along with a small fraction of images collected in the previous environments. Therefore, in each training step, the model has to jointly optimize over current and previous environments. However, storing all observations from past environments in memory would come at huge costs. Instead, a memory buffer for each previous environments only contains a random subset of all image and (pseudo-)label pairs. Training batches are then filled from the memory buffers of previous environments alongside self-supervised labels of the current environment. Furthermore, we also evaluate other continual learning frameworks that aim at mitigating forgetting. An ablation study is reported in Section IV-G.

IV. Experimental Evaluation

We test and verify the applicability of the proposed framework in different steps of increasing complexity. We first

[1]For example, for any clutter object standing on the ground, all points below a height of δ may be labelled as *background* in the pseudolabels, dependent on the boundaries of the superpixel.



(a) Wheeled Robotic Platform (b) Parking Garage (c) Construction Site (d) Office Environment

Fig. 4: Overview of our experimental setup.

validate that our robot can self-improve by deploying it into different unknown environments and measure the gained improvement. We then evaluate the effects of forgetting and knowledge transfer when switching deployment between different environments. Finally, we conduct an experiment in which the robot learns online during the mission.

For each experiment, we measure the localisation error in the x-y plane[2] in mean, median and standard deviation. We also measure the segmentation quality in mean Intersection over Union (mIoU).

A. Robotic System

We conduct all our experiments on an open-source[3] wheeled robotic ground vehicle [20] shown in Figure 4a. To calibrate the sensor system, we calibrate all cameras to an IMU that we attach to the sensor system and then align trajectories from visual-inertial odometry and LiDAR to find the extrinsic calibration between the LiDAR and the cameras. The tracking prism is used to gather ground-truth with a total station for our evaluation and is mounted on the sensor system to be aligned with the optical center of the LiDAR. To improve time synchronisation between the different sensors, we use hardware trigger boards to trigger all cameras simultaneously. From images captured at 20 Hz we then take those closest to the timestamp of the LiDAR scans, which are captured at 5Hz. All sensor drivers run on the same computer where sensors are synchronised to system time. While we also ran time-synchronisation to the same computer from the external total station tracking which supplies ground-truth poses in our evaluations, we found that this synchronisation did not work sufficiently accurately. We therefore manually correct time-offsets based on trajectory alignment between tracked prism and localised robot.

B. Evaluation Environments

We deploy our proposed system into three different environments: a construction site (`Construction`), a parking garage (`Garage`), and an office floor (`Office`). Architectural researchers provided us 3D meshes, constructed from dense 3D scans (`Construction`) and existing 2D floorplans supplemented with additional measurements (`Garage` and `Office`). Figure 4 shows these meshes together with the experimental setup. In each environment, we steer the robot through multiple independent trajectories of 2-3 min while tracking the robot with a total station. The coordinate system of the total station is always initialised to the origin of the building mesh, which we therefore set at corners visible to the total station. To evaluate the learned segmentation models, we sample images from each environment and manually annotate them with segmentation masks. We labelled 30 images from the garage that have in total 65% background pixels, 26 images from the construction site that have 76% background pixels and 31 images from the office that have 40% background pixels. When not specified differently, we evaluate only the region of the images that can also be reprojected to the LiDAR, applying a static field-of-view (FoV) mask.[4]

C. Learning Setup

We use a lightweight architecture based on Fast-SCNN [44] that we train for background-foreground segmentation. We resize input images from both the pre-training dataset and from the cameras to a common size (480×640 pixels). We first train the model on the NYU dataset. Then, when we deploy the robot into a new environment (cf. Sec. IV-D), we fill a replay buffer with samples from NYU in addition to training on the pseudolabels from the current environment. When we evaluate the transfer from a first environment to a second one (cf. Sec. IV-E), we replay images from both NYU and the first environment. We set the replay fraction to 10%, but we evaluate different replay regimes and strategies in Section IV-G. Before feeding images to the network, we augment them with left-right flipping and random perturbation of brightness and hue. See Appendix A for all training parameters.

In each experiment, we hold out 10% of the training samples for validation and train our model using Adam optimizer; we set the learning rate to 10^{-4} for the pre-training on NYU and to 10^{-5} for the remaining experiments, and adaptively decrease it when the validation loss reaches a plateau. We optimize the cross-entropy loss on the binary foreground-background labels. When training on pseudo-labels produced through the method detailed in Sec. III-B, we only apply the loss to those pixels that contain one of the two classes.

[2]The ground truth from the total station only provides translation measurements and does not enable evaluation of correctly estimated orientation.

[3]https://unlimited.ethz.ch/display/ROBOTX/SuperMegaBot

[4]The FoV mask has two reasons: Because we are primarily interested in good localisation, comparing segmentation quality in the region that can be reprojected to the LiDAR relates better to localisation performance. Additionally, pseudolabels are also only available in image regions where the LiDAR provides information, therefore we expect the segmentation to learn mostly the semantics of the scene visible in these regions.

environment	mean/median/std translation error [mm]			segmentation quality [% mIoU]	
	no segmentation	trained on NYU	self-improving	trained on NYU	self-improving
Garage	50 / 41 / 37	58 / 41 / 73	**43 / 35 / 31**	33.9	**62.8**
Construction	488 / 183 / 999*	126 / 78 / 129	**104 / 68 / 105**	27.6	**48.2**
Office	167 / 168 / 88^{+}	196 / 145 / 202	**150 / 138 / 81**	46.5	**53.9**

TABLE I: Comparison of localisation and segmentation quality when using no segmentation, deploying a fixed segmentation model trained on the NYU dataset, and our self-improving approach (including replay). Localisation errors marked with * contain a major ICP failure. We observe that segmentation in general is advantageous and the continual learning on self-supervised pseudolabels improves in all environments compared to deploying a fixed model. $^{+}$Note that in the office, we had to adjust ICP parameters without segmentation due to high amounts of clutter, as described in Appendix D.

D. Deployment in a new environment

To test the effectiveness of the pseudolabel training and the localisation based on filtered pointclouds, we deploy the robot in a single new environment. There, the robot collects information in the form of pseudolabels. The collected pseudolabels are used to train the segmentation, which we initialise with weights trained on NYU. Afterwards, we deploy the robot over a different trajectory in the same environment and measure the performance of both image segmentation and localisation, comparing to unfiltered ICP and filtering with a network trained solely on the available labelled NYU data.

In Table I we compare the performance obtained from our self-supervised foreground-background segmentation with segmentation obtained from the network pre-trained on NYU, and with a baseline that does not semantically filter the pointcloud. We note that segmentation in general is important for good localisation and can prevent failure, as on the construction site. The results also confirm the effectiveness of the pseudolabels, as training on these yields improvements in both segmentation quality and localisation error in all metrics. We conclude that the self-improving setup is working as expected and that through the feedback loop indeed the localisation improves the segmentation, and the segmentation improves the localisation.

E. Transfer into a second environment

In a second stage of experiments, we evaluate the ability of our system to retain knowledge and still adapt to new domains when moved from a first environment to a second one. In particular, we first train on pseudolabels from a *source* environment and afterwards train the same model on pseudolabels from a *target* environment, testing all possible combinations of our three environments. For each of these source-to-target transfers, we evaluate the extent of forgetting on both the pretraining data (NYU) and the source environment, comparing our adopted method based on a replay buffer with simple fine-tuning. We also report the segmentation quality on the target environment with both the replay-based and the finetuning method.

As shown in Table II and summarised in Figure 1, using replay buffers improves the segmentation performance on the previous tasks w.r.t. the case in which no replay is adopted. This is even more prominent when measured on the pseudolabels and measuring forgetting on the NYU data, which we analyse in further detail in Table V in Appendix B. Table II further shows that the forgetting of the models without replay can cause localisation failure on the source environment, as observed for Construction→Garage and Construction→Office. Using replay prevents such failure successfully. In the continual-learning literature, memory replay is usually studied in settings in which high-quality segmentation masks covering all pixels in the image are available [42], which is very different from replaying our pseudolabels. Yet, from our observations we can conclude that memory replay is also effective when the replayed labels are noisy and imperfect.

We note that finetuning can result in better adaptation to the target environment, especially with regard to localisation. This is expected, since the learning process of finetuning is fully tailored to the new environment. This effect is known as *stability-plasticity dilemma* [38], which refers to how retaining old knowledge can inhibit learning of new knowledge, but increasing plasticity can in turn increase the effects of forgetting. In our experiments, the relative improvements of finetuning over memory replay that we observe on the target environment are marginal, suggesting that the chosen 10% replay finds a good balance.

Table II also highlights cases in which deployment in two consecutive environments in general appears advantageous compared to the experiments in Section IV-D, both with and without memory replay. This indicates a general effect of positive knowledge transfer between our evaluation environments. For a self-improving robotic system this is a promising finding, as it suggests that the robot not only adapts to the target environment, but generally can become better at its task with every new deployment.

Finally, we observe that segmentation quality and localisation error are sometimes inconsistent, where a drop in segmentation quality in the source environment does not necessarily transfer into worse localisation. This indicates potential for advanced methods that could decide in a more fine-grained fashion which object types or observations should be kept in memory and which instead are not important for the task at hand and can be forgotten.

F. Online Learning

The previous experiments have been conducted in a multi-mission fashion, in which a first mission gathers data of the environment and the robot learns from it to improve in subsequent missions. We now evaluate how the system can

environment source → target	method	mean/median/std translation error [mm] source	 target	segmentation [% mIoU] source	 target
Garage→ Construction	replay	41 / 33 / 30	98 / 66 / 98	**60.8**	48.6
	finetuning	40 / 33 / 29	87 / 67 / 77	55.1	49.4
Garage→ Office	replay	39 / 31 / 31	168 / 137 / 109	**62.6**	47.2
	finetuning	37 / 30 / 33	196 / 118 / 267	61.0	47.4
Construction→ Garage	replay	**105 / 68 / 108**	40 / 33 / 29	**49.3**	62.2
	finetuning	549 / 84 / 1500*	40 / 31 / 29	42.3	62.0
Construction→ Office	replay	**125 / 72 / 128**	158 / 137 / 93	**50.3**	47.6
	finetuning	514 / 191 / 914*	146 / 123 / 93	45.4	49.4
Office → Garage	replay	**153 / 131 / 95**	44 / 34 / 36	**47.8**	62.1
	finetuning	182 / 151 / 114	38 / 32 / 27	40.2	61.0
Office → Construction	replay	171 / 161 / **86**	91 / 66 / 85	**47.5**	49.9
	finetuning	**168 / 159** / 88	121 / 70 / 128	33.3	49.1

TABLE II: Evaluation of forgetting and knowledge transfer when switching between deployment environments. The perception system is first trained on pseudolabels of the source environment, then on the target, and then evaluated on both. Bold marks cases where one method reduces forgetting compared to the other method. Underlined metrics are better than single-environment deployment from Table I. Finetuning leads to two cases where forgetting causes ICP failures, which are marked with star.

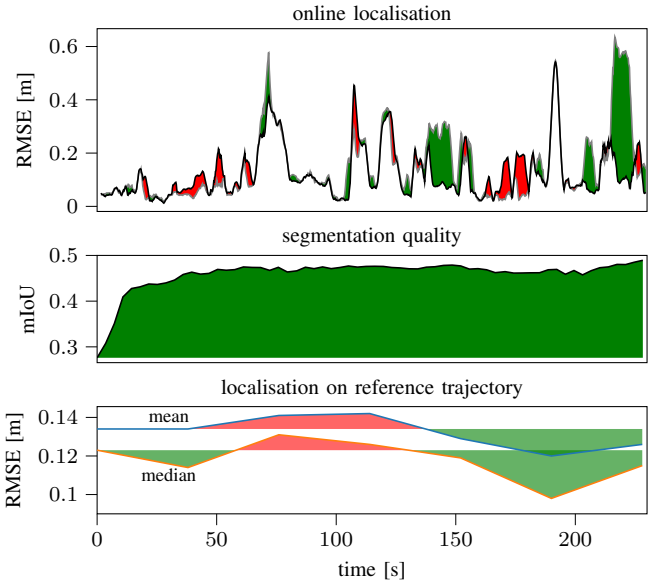


Fig. 5: Online Learning on the construction site. The first row shows the localisation error while the robot drives and learns. The second row shows the evolution of the segmentation quality measured on the ground-truth of the same environment for snapshots of the model at the given time. The third row shows the corresponding localisation error using these snapshots on a different trajectory. Areas in green/red show changes (better/worse) with respect to no online learning. We observe that segmentation quality increases over time and localisation error decreases.

learn online during a mission by learning from the pseudolabels directly as they are generated. For each forward pass, we therefore compile batches of (i) the current camera images that need to be segmented to perform localisation, (ii) 10 images randomly sampled from a buffer A where we already have associated pseudolabels, and (iii) 1 image from a buffer B of NYU images for memory replay. After the forward pass, we extract the prediction of (i) and backpropagate based on the loss from (ii) and (iii). After localising the current scan, we generate pseudolabels for (i) and fill them into buffer A. Due to memory constraints, buffer A holds at maximum 500 image-label pairs. Once it reaches the maximum, we remove a randomly sampled half of the content. In total, the system never stores more than 500 images of the current scene and 144 randomly selected images from NYU.

Our evaluation of online learning is shown in Figure 5. We observe that over time segmentation quality increases and localisation error decreases. Notably, already a few seconds of online learning increase the segmentation quality significantly. However, it takes longer until we measure a notable effect on the localisation. Due to the limited time until the end of the trajectory, we are therefore unable to see if there is a feedback effect where the improved localisation would create better pseudolabels that can in turn increase the segmentation quality further. We conclude that the self-improving framework also works in an online setting, but further investigation on longer time deployments is necessary.

G. Ablation Studies

We investigate two details in ablation studies to assess how the use of different continual learning methods – and in particular different replay schemes – impacts the segmentation accuracy in our domain adaptation setting.

We first evaluate two different strategies for memory replay. In the first strategy, which is the one that we adopt in the main experiments, on each source-to-target experiment (e.g., NYU→`Garage`), we fill the replay buffer with a fraction of samples from the source dataset(s) (NYU in the example), which we select randomly. We then fill training batches from the replay buffer and target dataset according to their relative sizes. In the second strategy, we fill the replay buffer with the full source dataset but fill training batches with a pre-defined target-source ratio. For instance, a ratio `Garage` : NYU $= 4 : 1$ with a batch size of 10 indicates that batches on average contain 8 images from Garage and 2 images from NYU. As shown in Table III, milder replay regimes (larger target-source ratios, or smaller replay fractions) achieve higher performance on the target domain, but cause the amount of information retained

from the source domain to drop. This forgetting phenomenon is particularly evident in the adaptation tasks in which the semantic gap between source and domain is larger. Indeed, for instance, in the NYU→`Garage` experiment we observe a drop of 31.9% in mIoU on the NYU labels between a replay strategy with a fraction of 10% and simple fine-tuning, while the same decrease in performance when the target domain is `Office` – semantically closer to the indoor dataset NYU – is 14.8%. At the same time, the segmentation quality on the pseudo-labels of the target dataset follows an inverse trend, generally increasing for smaller amounts of replay. This highlights the trade-off between the accuracy on the target and on the source domain.

We further compare regularization techniques from the continual-learning literature. In particular, we investigate the use of a distillation approach, in which a weighted regularization term $\mathcal{L}_d$ is added to the cross entropy loss $\mathcal{L}_{ce}$, to encourage the network to retain the knowledge from previous tasks: $\mathcal{L} = \mathcal{L}_{ce} + \lambda\mathcal{L}_d$. Similarly to [39], we experiment this distillation either on the output logits produced by the network (*Output distillation*) or on the intermediate features extracted from the model architecture before the final classification module (*Feature distillation*). Furthermore, we evaluate Elastic Weight Consolidation (EWC) [29]. The loss optimized in EWC is of a similar form: $\mathcal{L} = \mathcal{L}_{ce} + \lambda\mathcal{L}_{ewc}$, where $\mathcal{L}_{ewc}$ penalizes deviations of the network parameters across tasks. We present further details on the methods and parameters in Appendix C.

As shown in Table III, in our experiments replay buffers prove to be the most effective among the examined methods in minimizing the amount of forgetting on the NYU dataset, and generally allow attaining a good trade-off with the segmentation quality on the pseudo-labels from the target domain. We also note that, with limited exceptions, both regularization approaches fail to maintain a good performance on the source dataset NYU. We believe that for distillation methods this can be ascribed to the fact that, as opposed to related works that explored similar techniques in a class-incremental setting [39, 40], we conduct domain adaptation. More importantly, unlike [39, 40] our supervision signal does not consist of accurate ground-truth annotations available at all pixels, but of noisy pseudo-labels that often cover a limited region of the image. Finally, while similar considerations also apply for EWC, we believe that the limited effectiveness of this technique in our setting is also related to the method being designed for classification as opposed to image segmentation.

H. Runtime

We conduct our experiments on 6-year-old hardware with a 8-core i7-6700K CPU and GeForce GTX 980 Ti GPU. While our implementations are not heavily optimised for runtime, we carefully select a fast rather than precise neural network architecture. Accordingly, the segmentation of all three camera images that are projected into the LiDAR scan takes 127 ± 23ms. The following ICP localisation then takes 529 ± 132 ms on our hardware (CPU only). Given the LiDAR frequency of 5 Hz (or 200 ms per scan), the total delay from begin of the scan to the localised pose is approximately 856 ms. This requires a factor 5 optimisation for real-time deployment. After localisation, our pseudolabel generation takes 1.327 ± 0.127 s, most of which is taken by the superpixel method. However, this process is not time-critical as subsequent scans are highly correlated and we therefore always subsample the generated data, taking pseudolabels from at most every third scan.

I. Limitations

Our system required some manual intervention in the sense that a few parameters were changed between environments (ICP and superpixels, see appendix). While coming close, the system also did not run in real-time. We are confident that both points can be tackled with better software implementations.

This study is limited to binary segmentation, and the segmented classes have to be linked to observable affordance. To extend the approach to more classes, affordances could be observed e.g. by manipulating objects [9], or observing contexts and spatial co-occurence [6, 19].

V. Conclusion & Outlook

In this work we propose a framework for self-improving semantic perception by combining continual learning with self-supervision. We study this on a robotic system that localises in 3D floorplan meshes. Our experiments validate the gains of the self-improving systems in diverse environments. In particular, we analyse the effects of knowledge transfer and forgetting when switching between environments. We find that memory replay is an effective solution that can mitigate forgetting, and observe that exposure to multiple environments is sometimes even beneficial for overall performance.

The concept of self-improving, continual, online learning robots opens up exciting questions for future research. The related self-supervision approaches that we describe may facilitate the transfer of our proposed framework to other robotic applications. Moreover, our evaluations of knowledge transfer and forgetting show potential for effective combinations of self-supervision and continual learning. Finally, we identify long-term deployment of online learning systems as an important future research direction for self-improving perception.

Acknowledgments

This work was partially funded by the Hilti Group. Eberhard Unternehmungen kindly allowed us to conduct experiments on their premises.

Furthermore, we thank Selen Ercan for creating the building models, Florian Tschopp and his VersaVIS board for all the multi-sensor calibration, Shen Kaiyue for implementing and testing various regularisation based continual learning methods, and Andrei Cramariuc for GPU cluster support.

References

[1] Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, and Sabine Süsstrunk. SLIC Superpixels Compared to State-of-the-Art Superpixel Methods. *IEEE Trans. Pattern Anal. Machine Intell.*, 34(11):2274–2282, 2012.

		NYU→Garage			NYU→Construction			NYU→Office			
		NYU	Garage		NYU	Construction		NYU	Office		
			Pseudo	GT		Pseudo	GT		Pseudo	GT	
Finetuning		36.4	96.3	61.8	36.6	79.5	48.9	66.2	70.9	51.2	
Replay buffer with ratio target : NYU	1 : 1	**87.2**	86.5	61.5	**82.9**	66.6	46.1	**83.5**	57.8	49.9	memory replay
	3 : 1	81.1	91.7	60.7	79.8	73.1	47.8	81.4	68.1	**52.2**	
	4 : 1	79.8	92.4	**62.3**	78.7	73.1	**48.8**	82.0	65.8	51.7	
	10 : 1	73.4	94.7	61.3	75.3	75.6	47.6	77.7	**71.2**	52.1	
	20 : 1	67.5	95.3	62.0	72.4	76.3	48.3	76.0	69.1	52.0	
	200 : 1	53.9	**96.1**	61.6	53.2	**77.2**	48.7	74.8	68.7	50.9	
Replay buffer with fraction replay	10%	**68.3**	95.4	62.8	**78.6**	**77.0**	48.2	**81.0**	**69.7**	**53.9**	
	5%	65.0	**95.9**	62.0	76.2	76.9	**48.5**	79.9	69.3	51.5	
Feature distillation	$\lambda = 0.5$	35.4	**96.1**	**61.9**	33.2	**77.1**	50.3	**65.3**	71.1	**50.7**	regularisation methods
	$\lambda = 1$	34.8	95.9	61.2	30.6	76.9	**50.8**	58.6	**78.9**	49.1	
	$\lambda = 10$	**37.4**	94.2	61.6	**33.6**	72.1	48.3	48.7	72.2	48.0	
	$\lambda = 50$	33.6	92.7	61.1	32.6	63.1	46.5	44.0	64.5	46.8	
Output distillation	$\lambda = 0.5$	33.1	**94.4**	63.3	34.0	**76.5**	**47.8**	**62.9**	**68.4**	**49.9**	
	$\lambda = 1$	32.4	85.3	**64.4**	**38.2**	59.4	46.6	53.0	60.4	44.3	
	$\lambda = 10$	37.8	40.8	47.5	37.9	32.9	37.1	45.3	35.8	36.1	
	$\lambda = 50$	**39.0**	48.9	53.0	31.7	28.7	31.1	46.3	30.5	35.9	
EWC [29]	$\lambda = 0.5$	36.1	**96.4**	**61.5**	34.4	**76.5**	47.9	65.7	69.2	51.6	
	$\lambda = 1$	36.2	96.3	61.4	**37.0**	76.4	**48.0**	66.2	**74.0**	50.8	
	$\lambda = 10$	**37.9**	96.2	61.1	35.4	76.2	**48.0**	**70.6**	69.1	**51.8**	
	$\lambda = 50$	**37.9**	95.9	**61.5**	35.0	75.6	47.9	65.5	73.2	51.0	

TABLE III: Ablation study over different continual learning methods. After adapting from the source domain (NYU) to the different deployment environments, we measure segmentation quality [% mIoU] on the source data as well as the self-supervised pseudolabels (Pseudo) and our ground-truth annotations (GT). We compare the finetuning baselines with memory replay and regularisation methods. We observe that memory replay is more successful than regularisation at preventing catastrophic forgetting in our application.

[2] Pulkit Agrawal, Joao Carreira, and Jitendra Malik. Learning to See by Moving. In *Intl. Conf. on Computer Vision (ICCV)*, pages 37–45, 2015.

[3] Jiwoon Ahn and Suha Kwak. Learning Pixel-level Semantic Affinity with Image-level Supervision for Weakly Supervised Semantic Segmentation. *CoRR*, abs/1803.10464, 2018.

[4] Jiwoon Ahn, Sunghyun Cho, and Suha Kwak. Weakly Supervised Learning of Instance Segmentation with Inter-pixel Relations. *CoRR*, abs/1904.05044, 2019.

[5] Pierre Alliez, Stéphane Tayeb, and Camille Wormser. 3D fast intersection and distance computation. In *CGAL User and Reference Manual*. CGAL Editorial Board, 5.2 edition, 2020. URL https://doc.cgal.org/5.2/Manual/packages.html#PkgAABBTree.

[6] Joël Bachmann, Kenneth Blomqvist, Julian Förster, and Roland Siegwart. Points2Vec: Unsupervised Object-level Feature Learning from Point Clouds. *CoRR abs/2102.04136*, 2021.

[7] Hermann Blum, Julian Stiefel, Cesar Cadena, Roland Siegwart, and Abel Gawel. Precise robot localization in architectural 3d plans. *arXiv preprint arXiv:2006.05137*, 2020.

[8] Raffaello Camoriano, Giulia Pasquale, Carlo Ciliberto, Lorenzo Natale, Lorenzo Rosasco, and Giorgio Metta. Incremental Robot Learning of New Objects with Fixed Update Time. In *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, 2017.

[9] C. Castellini, T. Tommasi, N. Noceti, F. Odone, and B. Caputo. Using Object Affordances to Improve Object Recognition. *IEEE Trans. Autonomous Mental Development*, 3(3):207–215, 2011.

[10] Fabio Cermelli, Massimiliano Mancini, Samuel Rota Bulò, Elisa Ricci, and Barbara Caputo. Modeling the Background for Incremental Learning in Semantic Segmentation. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[11] Yu-Ting Chang, Qiaosong Wang, Wei-Chih Hung, Robinson Piramuthu, Yi-Hsuan Tsai, and Ming-Hsuan Yang. Mixup-CAM: Weakly-supervised Semantic Segmentation via Uncertainty Regularization. *CoRR*, abs/2008.01201, 2020.

[12] Yang Chen and Gérard Medioni. Object modelling by registration of multiple range images. *Image and Vision Computing*, 10(3):145–155, 1992.

[13] Yuhua Chen, Cordelia Schmid, and Cristian Sminchisescu. Self-Supervised Learning With Geometric Constraints in Monocular Video: Connecting Flow, Depth, and Camera. In *Intl. Conf. on Computer Vision (ICCV)*, pages 7063–7072, 2019.

[14] Zhiyuan Chen and Bing Liu. Lifelong Machine Learning. *Synthesis Lectures on Artificial Intelligence and Machine*

Learning, 12(3):1–207, 2018.

[15] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised Visual Representation Learning by Context Prediction. In *Intl. Conf. on Computer Vision (ICCV)*, pages 1422–1430, 2015.

[16] Arthur Douillard, Yifu Chen, Arnaud Dapogny, and Matthieu Cord. PLOP: Learning without Forgetting for Continual Semantic Segmentation. *CoRR abs/2011.11390*, 2020.

[17] Robert M. French. Catastrophic forgetting in connectionist networks. *Trends in Cognitive Sciences*, 3(4):128–135, 1999.

[18] Justin Fu, Sergey Levine, and Pieter Abbeel. One-shot learning of manipulation skills with online dynamics adaptation and neural network priors. In *IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, pages 4019–4026. IEEE, 2016.

[19] C. Galleguillos, A. Rabinovich, and S. Belongie. Object categorization using co-occurrence, location and appearance. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, 2008.

[20] Abel Gawel, Hermann Blum, Johannes Pankert, Koen Krämer, Luca Bartolomei, Selen Ercan, Farbod Farshidian, Margarita Chli, Fabio Gramazio, Roland Siegwart, et al. A fully-integrated sensing and control system for high-accuracy mobile robotic building construction. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2300–2307. IEEE, 2019.

[21] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. In *Intl. Conf. on Learning Representations (ICLR)*, 2018.

[22] Rémi Giraud, Vinh-Thong Ta, and Nicolas Papadakis. SCALP: superpixels with contour adherence using linear path. *CoRR*, abs/1903.07149, 2019. URL http://arxiv.org/abs/1903.07149.

[23] Tuomas Haarnoja, Sehoon Ha, Aurick Zhou, Jie Tan, George Tucker, and Sergey Levine. Learning to Walk via Deep Reinforcement Learning. *arXiv preprint arXiv:1812.11103*, 2018.

[24] Tyler L. Hayes, Nathan D. Cahill, and Christopher Kanan. Memory Efficient Experience Replay for Streaming Learning. In *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, 2019.

[25] Zilong Huang, Xinggang Wang, Jiasi Wang, Wenyu Liu, and Jingdong Wang. Weakly-Supervised Semantic Segmentation Network With Deep Seeded Region Growing. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[26] J. Hwangbo, I. Sa, R. Siegwart, and M. Hutter. Control of a Quadrotor With Reinforcement Learning. *IEEE Robotics and Automation Letters*, 2(4):2096–2103, 2017.

[27] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Intl. Conf. on Machine Learning (ICML)*, 2015.

[28] J. Kabzan, L. Hewing, A. Liniger, and M. N. Zeilinger. Learning-Based Model Predictive Control for Autonomous Racing. *IEEE Robotics and Automation Letters*, 4(4):3363–3370, 2019. doi: 10.1109/LRA.2019.2926677.

[29] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017.

[30] Lars Kunze, Nick Hawes, Tom Duckett, Marc Hanheide, and Tomáš Krajník. Artificial intelligence for Long-Term Robot Autonomy: A Survey. *IEEE Robotics and Automation Letters*, 3(4):4023–4030, 2018.

[31] Jungbeom Lee, Eunji Kim, Sungmin Lee, Jangho Lee, and Sungroh Yoon. FickleNet: Weakly and Semi-supervised Semantic Image Segmentation using Stochastic Inference. *CoRR*, abs/1902.10421, 2019.

[32] Timothée Lesort, Vincenzo Lomonaco, Andrei Stoian, Davide Maltoni, David Filliat, and Natalia Díaz-Rodríguez. Continual Learning for Robotics: Definition, Framework, Learning Strategies, Opportunities and Challenges. *Information Fusion*, 58:52–68, 2020.

[33] Yunsheng Li, Lu Yuan, and Nuno Vasconcelos. ESL: Entropy-guided Self-supervised Learning for Domain Adaptation in Semantic Segmentation. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[34] Zhizhong Li and Derek Hoiem. Learning without Forgetting. *IEEE Trans. Pattern Anal. Machine Intell.*, 40(12):2935–2947, 2018.

[35] David Lopez-Paz and Marc'Aurelio Ranzato. Gradient Episodic Memory for Continual Learning. In *Conf. on Neural Information Processing Systems (NIPS)*, 2017.

[36] Matthias Lorenzen, Mark Cannon, and Frank Allgöwer. Robust MPC with recursive model update. *Automatica*, 103:461–471, 2019. ISSN 0005-1098.

[37] John McCormac, Ronald Clark, Michael Bloesch, Andrew Davison, and Stefan Leutenegger. Fusion++: Volumetric Object-Level SLAM. In *Intl. Conf. on 3D Vision (3DV)*, pages 32–41. IEEE, 2018.

[38] Martial Mermillod, Aurélia Bugaiska, and Patrick Bonin. The stability-plasticity dilemma: investigating the continuum from catastrophic forgetting to age-limited learning effects. *Frontiers in Psychology*, 4:504, 2013.

[39] Umberto Michieli and Pietro Zanuttigh. Incremental Learning Techniques for Semantic Segmentation. In *Intl. Conf. on Computer Vision (ICCV), Workshop on Transferring and Adapting Source Knowledge in Computer Vision (TASK-CV)*, 2019.

[40] Umberto Michieli and Pietro Zanuttigh. Knowledge distillation for incremental learning in semantic segmentation. *Computer Vision and Image Understanding*, 205:103167, 2021.

[41] M. Q. Mohammed, K. L. Chung, and C. S. Chyi. Review

of Deep Reinforcement Learning-Based Object Grasping: Techniques, Open Challenges, and Recommendations. *IEEE Access*, 8:178450–178481, 2020.

[42] German I. Parisi, Ronald Kemker, Jose L. Part, Christopher Kanan, and Stefan Wermter. Continual lifelong learning with neural networks: A review. *Neural Networks*, 113:54–71, 2020.

[43] Lorenzo Porzi, Markus Hofinger, Idoia Ruiz, Joan Serrat, Samuel Rota Bulo, and Peter Kontschieder. Learning Multi-Object Tracking and Segmentation From Automatic Annotations. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[44] Rudra P K Poudel, Stephan Liwicki, and Roberto Cipolla. Fast-SCNN: Fast Semantic Segmentation Network. In *British Machine Vision Conf. (BMVC)*, 2019.

[45] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H. Lampert. iCaRL: Incremental Classifier and Representation Learning. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[46] Md Alimoor Reza, Akshay U Naik, Kai Chen, and David J Crandall. Automatic Annotation for Semantic Segmentation in Indoor Scenes. In *IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, pages 4970–4976. IEEE, 2019.

[47] Antoni Rosinol, Marcus Abate, Yun Chang, and Luca Carlone. Kimera: an Open-Source Library for Real-Time Metric-Semantic Localization and Mapping. In *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, pages 1689–1696. IEEE, 2020.

[48] Andrei A. Rusu, Neil C. Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, and Raia Hadsell. Progressive Neural Networks. *arXiv preprint arXiv:1606.04671*, 2016.

[49] Antoine Saporta, Tuan-Hung Vu, Matthieu Cord, and Patrick Pérez. ESL: Entropy-guided Self-supervised Learning for Domain Adaptation in Semantic Segmentation. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), Workshop on Scalability in Autonomous Driving*, 2020.

[50] Jonathan Schwarz, Wojciech Czarnecki, Jelena Luketina, Agnieszka Grabska-Barwinska, Yee Whye Teh, Razvan Pascanu, and Raia Hadsell. Progress & Compress: A scalable framework for continual learning. In *Intl. Conf. on Machine Learning (ICML)*, 2018.

[51] Hanul Shin, Jung Kwon Lee, Jaehong Kim, and Jiwon Kim. Continual Learning with Deep Generative Replay. In *Conf. on Neural Information Processing Systems (NIPS)*, 2017.

[52] Konstantin Shmelkov, Cordelia Schmid, and Karteek Alahari. Incremental Learning of Object Detectors without Catastrophic Forgetting. In *Intl. Conf. on Computer Vision (ICCV)*, 2017.

[53] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor Segmentation and Support Inference from RGBD Images. In *European Conf. on Computer Vision (ECCV)*, 2012.

[54] Daniel L. Silver, Qiang Yang, and Lianghao Li. Lifelong Machine Learning Systems: Beyond Learning Algorithms. In *AAAI Conf. on Artificial Intelligence (AAAI)*, 2013.

[55] Boris Sofman, Ellie Lin, J Andrew Bagnell, John Cole, Nicolas Vandapel, and Anthony Stentz. Improving robot navigation through self-supervised online learning. *Journal of Field Robotics*, 23(11-12):1059–1075, 2006.

[56] Weixuan Sun, Jing Zhang, and Nick Barnes. 3D Guided Weakly Supervised Semantic Segmentation. *CoRR*, abs/2012.00242, 2020.

[57] Sebastian Thrun and Tom M. Mitchell. Lifelong robot learning. *Robotics and Autonomous Systems*, 15:25–46, 1995.

[58] Alessio Tonioni, Oscar Rahnama, Thomas Joy, Luigi Di Stefano, Thalaiyasingam Ajanthan, and Philip H S Torr. Learning to Adapt for Stereo. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 9661–9670, 2019.

[59] Chenshen Wu, Luis Herranz, Xialei Liu, Yaxing Wang, Joost van de Weijer, and Bogdan Raducanu. Memory Replay GANs: Learning to Generate Images from New Categories without Forgetting. In *Conf. on Neural Information Processing Systems (NIPS)*, 2018.

[60] Yuxin Wu and Kaiming He. Group Normalization. In *European Conf. on Computer Vision (ECCV)*, 2018.

[61] Lu Yu, Xialei Liu, and Joost van de Weijer. Self-Training for Class-Incremental Semantic Segmentation. *CoRR abs/2012.03362*, 2020.

[62] Ekim Yurtsever, Jacob Lambert, Alexander Carballo, and Kazuya Takeda. A Survey of Autonomous Driving: Common Practices and Emerging Technologies. *IEEE Access*, 8:58443–58469, 2020.

[63] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful Image Colorization. In *European Conf. on Computer Vision (ECCV)*, pages 649–666. Springer, 2016.

[64] Shangtong Zhang and Richard S. Sutton. A Deeper Look at Experience Replay. In *Conf. on Neural Information Processing Systems (NIPS) - Deep Reinforcement Learning Symposium*, 2017.

[65] Bolei Zhou, Aditya Khosla, Àgata Lapedriza, Aude Oliva, and Antonio Torralba. Learning Deep Features for Discriminative Localization. *CoRR*, abs/1512.04150, 2015.

[66] Yang Zou, Zhiding Yu, B. V. K. Vijaya Kumar, and Jinsong Wang. Unsupervised Domain Adaptation for Semantic Segmentation via Class-Balanced Self-training. In *European Conf. on Computer Vision (ECCV)*, 2018.

APPENDIX

A. Details on the Network Architecture

In all our experiments we use a batch size of 10 and train the network for up to 100 epochs, using early stopping with a patience of 20 epochs based on the validation loss. Our network architecture, based on Fast-SCNN [44], has a total of 1,775,110

		NYU		Garage (Pseudo)	
		BN	GN	BN	GN
Ratio target : NYU	1 : 1	84.9	**87.2**	**87.7**	86.5
	3 : 1	77.8	**81.1**	90.6	**91.7**
	4 : 1	76.4	**79.8**	92.0	**92.4**
	10 : 1	70.3	**73.4**	93.6	**94.7**
	20 : 1	66.7	**67.5**	94.5	**95.3**
	200 : 1	**54.6**	53.9	95.3	**96.1**
Fraction replay NYU	10%	67.6	**68.3**	94.0	**95.4**
	5%	63.6	**65.0**	94.9	**95.9**
	0% (fine-tuning)	**37.3**	36.4	95.5	**96.3**

TABLE IV: Comparison of segmentation quality [% mIoU] on NYU→`Garage` between models trained with batch normalization (BN) and models trained with group normalization (GN), under different replay regimes.

trainable parameters. We use group normalization [60] in all layers; we conducted a preliminary ablation study (cf. Table IV) comparing this design choice with the alternative batch normalization [27]. In accordance with [60], we found group normalization to be more indicated for our transfer-learning tasks, in which the statistics of the *source* training data, used by batch normalization to fit per-layer parameters [27], do not match in general those of the *target* domain. This is reflected in the models trained with group normalization performing consistently better or comparably to those trained with batch normalization, as soon as a non-negligible amount of replay is used.

B. Details on Cross-Domain Forgetting

We present a detailed analysis of forgetting in terms of segmentation in Table V as supplementary information to the main results presented in Table II. With no exception, memory replay performs better on source environments than finetuning. We note that the effect of forgetting is even stronger on the NYU data than in the deployment environments.

C. Details on the Continual-Learning Ablation Study

For both distillation and EWC, we use the same learning parameters as the experiments with replay buffers. In the following, we denote with $\mathbf{X}$ and $\mathbf{M}$ respectively an image and the corresponding mask from the training dataset $\mathcal{D}$. When $\mathbf{X}$ is a pseudo-label image, a pixel in $\mathbf{M}$ is *masked* if the corresponding pixel in $\mathbf{X}$ has an associated pseudo-label (background/foreground) and *not masked* if the corresponding pixel has unknown label; if $\mathbf{X}$ is an image replayed from NYU, all pixels in $\mathbf{X}$ are masked. For a given stage-1 experiment (i.e., in which we deploy the model pretrained on NYU in a new environment, cf., e.g., Tab. V), we denote the output prediction of the model pretrained on NYU as $\mathbf{y}_0(\mathbf{X})$ and the output prediction of the current stage-1 model as $\mathbf{y}(\mathbf{X})$; to indicate the predicted score associated to each class $c \in \{b, f\}$ (b = background, f = foreground) we write $\mathbf{y}_0(\mathbf{X})[c]$ and $\mathbf{y}(\mathbf{X})[c]$. Finally, we denote with $M(\mathbf{X}, \mathbf{M})$ a function that maps an input image $\mathbf{X}$ and its corresponding mask $\mathbf{M}$ to a vectorized version of $\mathbf{X}$ that contains only the pixels that are masked in $\mathbf{M}$.

The generic distillation loss reads as follows:

$$\mathcal{L} = \mathcal{L}_{ce} + \lambda \mathcal{L}_d, \tag{1}$$

where λ is a hyper-parameter and $\mathcal{L}_{ce}$ is the cross-entropy loss (cf. Sec. IV-G).

For output distillation, the regularization loss $\mathcal{L}_d$ is a cross-entropy loss between the prediction of the previous and the current model, masked by the input mask of each image, i.e.,

$$\mathcal{L}_d = -\sum_{(\mathbf{X},\mathbf{M})\in\mathcal{D}} \sum_{c\in\{b,f\}} \frac{M(\mathbf{y}_0(\mathbf{X}),\mathbf{M})[c] \cdot \log(M(\mathbf{y}(\mathbf{X}),\mathbf{M}))[c]}{|\mathcal{D}|}. \tag{2}$$

For feature distillation, similarly to [39] we consider the features outputted by the network at a selected layer and minimize the ℓ_2 norm between these as returned by the pre-trained model and by the current model. In particular, we consider the layer that precedes the final classification module in the Fast-SCNN architecture [44] and denote its output as $\mathbf{l}_0(\mathbf{X})$ and $\mathbf{l}(\mathbf{X})$, respectively for the pre-trained and for the current model. The regularization loss can therefore be expressed as:

$$\mathcal{L}_d = \frac{\|\mathbf{l}_0(\mathbf{X}) - \mathbf{l}(\mathbf{X})\|_2^2}{|\mathcal{D}|}. \tag{3}$$

For Elastic Weight Consolidation (EWC), we adopt the original loss introduced in [29], which is of the form:

$$\mathcal{L} = \mathcal{L}_{\text{main}} + \lambda \sum_i F_i(\theta_i - \theta_{i,0})^2, \tag{4}$$

where the sum is computed over the trainable parameters θ_i and $\theta_{i,0}$ respectively of the current and of the pre-trained model, and F_i is the element on the diagonal of the Fisher information matrix associated with the i-th parameters. $\mathcal{L}_{\text{main}}$ represents the main loss optimized in the given task, which in our case is the background-foreground cross-entropy loss $\mathcal{L}_{ce}$.

D. Localisation Parameters

In general, we run point-to-plane ICP with 3 nearest neighbors and initialise on the previously solved pose. We apply multiple filters to the input scan, even after the semantic filtering:

- We require the scan to have at minimum 500 points (i.e., rejecting scans where the segmentation classifies nearly everything as foreground).
- We subsample the scan to a maximum density of 10,000 pts/m^3.
- After nearest neighbor association, we reject the 20% points that are further away from the map.
- We reject associations where the estimated surface normals (estimated based on the 10 nearest neighbors) have a larger angle deviation than 1.5 rad.

As mentioned, in order to localise without segmentation and generate pseudolabels in the very cluttered office environment, we enforce additional filters:

- We only localise in 4 degrees of freedom (x, y, z, yaw).

Stage	Source → target	Segmentation quality [% mIoU]													
		NYU		Garage				Construction				Office			
		GT		Pseudo		GT		Pseudo		GT		Pseudo		GT	
		RB	FT	RB	FT	RB	FT	RB	FT	RB	FT	RB	FT	RB	FT
0	Pretraining on NYU	–	86.4	–	(22.5)	–	(33.9)	–	(22.7)	–	(27.6)	–	(39.6)	–	(46.5)
1	NYU → Garage	**68.3**	36.4	95.4	96.3	62.8	61.8	–	–	–	–	–	–	–	–
1	NYU → Construction	**78.6**	36.6	–	–	–	–	77.0	79.5	48.2	48.9	–	–	–	–
1	NYU → Office	**81.0**	66.2	–	–	–	–	–	–	–	–	69.7	70.9	53.9	51.2
2	Garage→ Construction	**70.3**	30.7	**91.8**	77.1	60.8	55.1	77.4	78.5	48.6	49.4	–	–	–	–
2	Garage→ Office	**70.9**	42.7	**92.8**	71.7	62.6	61.0	–	–	–	–	69.9	72.2	47.2	47.4
2	Construction→ Office	**78.6**	48.9	–	–	–	–	**71.3**	55.9	50.3	45.4	70.3	72.2	47.6	49.4
2	Construction→ Garage	**70.5**	36.7	94.4	95.6	62.2	62.0	**61.4**	43.3	49.3	42.3	–	–	–	–
2	Office → Garage	**68.7**	36.4	95.3	96.4	62.1	61.0	–	–	–	–	**61.2**	46.9	47.8	40.2
2	Office → Construction	**77.7**	38.8	–	–	–	–	73.1	73.0	49.9	49.1	**63.4**	44.7	47.5	33.3

TABLE V: Bold shows how the replay buffer (RB) prevents degradation of performance on the datasets on which the model has previously been trained, as opposed to simple fine-tuning (FT).

- We estimate normal directions based on 30 nearest neighbors and only associate points to the map if the angle between the normals is below 0.8 rad.

E. Pseudolabel Parameters

We empirically set the distance threshold to $\delta = 0.1$m and discard superpixels with a depth variance that surpasses 0.5m. We smooth the images with a Gaussian kernel ($\sigma = 0.2$) and oversegment them into approximately 400 superpixels with SLIC parameter $\texttt{compactness} = 10^5$. On the data captured from the garage, we use a different superpixel algorithm (SCALP [22]) that we later discard because of long runtimes. We do not notice qualitative differences between the created superpixels. In the office environment, we increase the standard deviation threshold to 1m due to large amounts of clutter.

To get an estimate of the quality of the pseudolabels themselves, we match frames where we have both manual ground-truth annotations and pseudolabels. Unfortunately, we could not recover pseudolabels for the images that were used to generate ground-truth in the office environment. When evaluating the pseudolabels, we also ignore all pixels that are not labelled (due to high variance or no reprojected LiDAR points in that superpixel). Therefore, the evaluation is strongly biased in favor of the pseudolabels. We measure 68.4% mIoU on the garage pseudolabels. For the same pixels (only those where pseudolabels are not ignored), our trained models get 64.3% mIoU. In the construction site environment, we measure 49.5% mIoU for the pseudolabels and 54.3% mIoU for our trained model.

F. Example of segmentation predictions

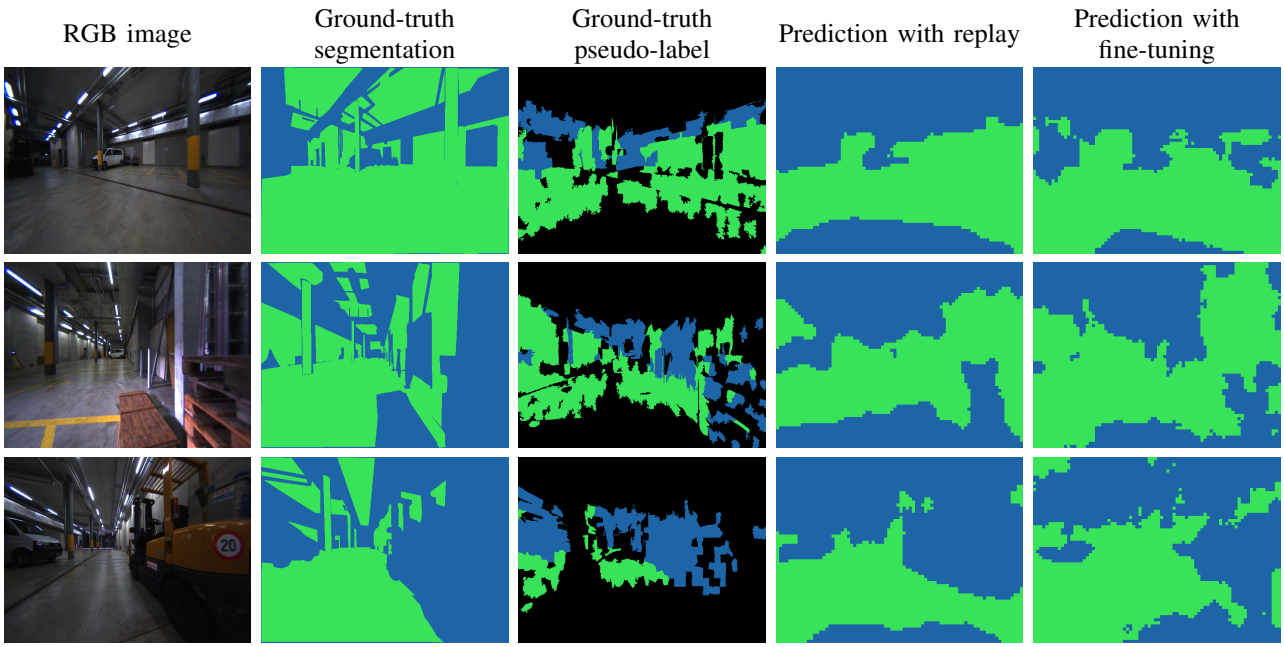
Figures 6, 7, and 8 show examples of segmentation masks produced by the network on the source environment in the experiments with transfer from a first to a second environment. We report a selection of frames for which we have available ground-truth segmentation and show the predictions obtained both with a model trained with simple finetuning and with one trained with replay from the source and the pre-training datasets.

In the qualitative outputs, we observe that the models learn biases towards regions that are generally unlabeled at training time. In particular, areas in the upper and lower part of the image are commonly classified as foreground, and show a curvature that roughly reflects the regions in the training pseudolabels where information is missing due to the reprojection of the LiDAR measurements into the camera view. This is in line with our discussion of the FoV mask, as supervision through pseudolabels is missing in those parts of the image; indeed, the learned biases in these unobserved regions often do not match the ground-truth class in these areas (cf., e.g., Fig. 6a, columns `Ground-truth segmentation` and `Prediction with replay`), and the evaluation would reflect this negatively if these areas were considered. We stress that the masked FoV region is most relevant for our application, as it represents the overlap of camera and LiDAR scans that we aim to filter and improve localization with. However, we also provide numbers when evaluating whole camera images instead of FoV masks in Table VI. As expected, the results outside of the LiDAR FoV are more noisy. From the qualitative examples and comparison with the FoV evaluation we know that this is due to wrong biases in image regions where no pseudolabels are available.

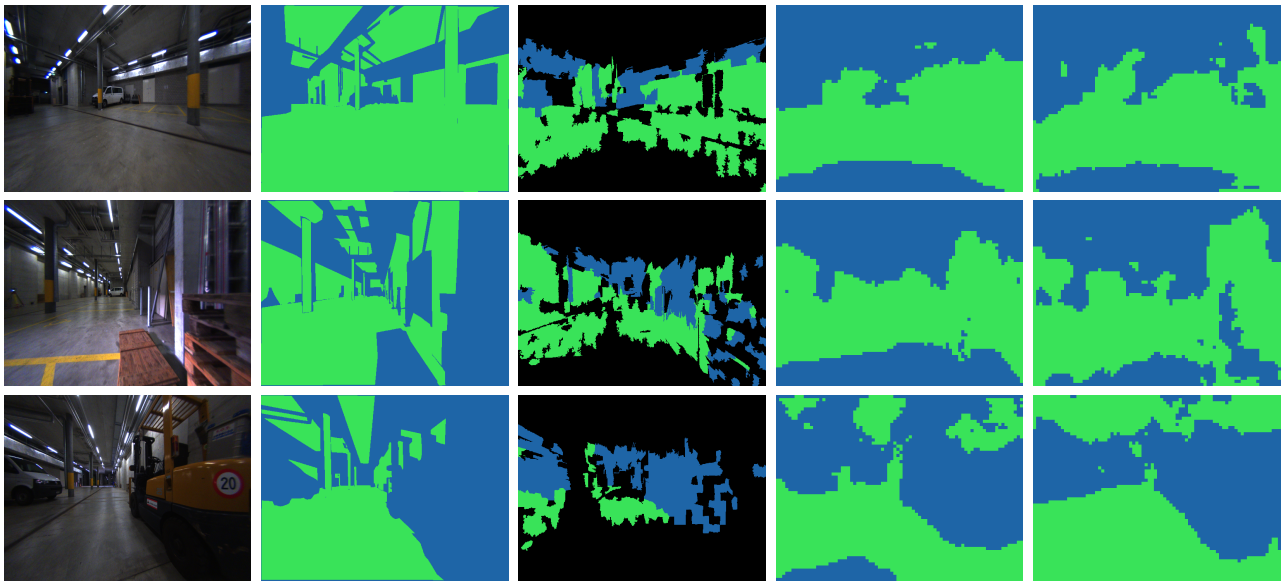
[5]This procedure is suggested by the skimage implementation that we use.

Stage	Source → target	Segmentation quality [% mIoU]													
		NYU		Garage				Construction				Office			
		GT (no mask)		Pseudo		GT (no mask)		Pseudo		GT (no mask)		Pseudo		GT (no mask)	
		RB	FT	RB	FT	RB	FT	RB	FT	RB	FT	RB	FT	RB	FT
0	Pretraining on NYU	–	86.4	–	(22.5)	–	(40.3)	–	(22.7)	–	(29.4)	–	(39.6)	–	(46.3)
1	NYU → Garage	**68.3**	36.4	95.4	96.3	44.5	43.3	–	–	–	–	–	–	–	–
1	NYU → Construction	**78.6**	36.6	–	–	–	–	77.0	79.5	32.7	32.7	–	–	–	–
1	NYU → Office	**81.0**	66.2	–	–	–	–	–	–	–	–	69.7	70.9	53.2	51.7
2	Garage→ Construction	**70.3**	30.7	**91.8**	77.1	43.8	46.0	77.4	78.5	34.7	34.6	–	–	–	–
2	Garage→ Office	**70.9**	42.7	**92.8**	71.7	45.3	48.0	–	–	–	–	69.9	72.2	52.1	50.3
2	Construction→ Office	**78.6**	48.9	–	–	–	–	**71.3**	55.9	34.7	36.4	70.3	72.2	46.6	47.5
2	Construction→ Garage	**70.5**	36.7	94.4	95.6	43.7	44.2	**61.4**	43.3	33.1	31.0	–	–	–	–
2	Office → Garage	**68.7**	36.4	95.3	96.4	43.3	42.9	–	–	–	–	**61.2**	46.9	46.8	42.7
2	Office → Construction	**77.7**	38.8	–	–	–	–	73.1	73.0	34.1	33.7	**63.4**	44.7	46.6	36.7

TABLE VI: While we in general evalaute segmentation quality only in the overlapping field of view of cameras and LiDAR, this table serves as a comparison as to how Table V would look when evaluating the whole camera images, including regions where the segmentation never has training signals because pseudolabels cannot be generated. We observe similar trends also in this table, while the results are more noisy.

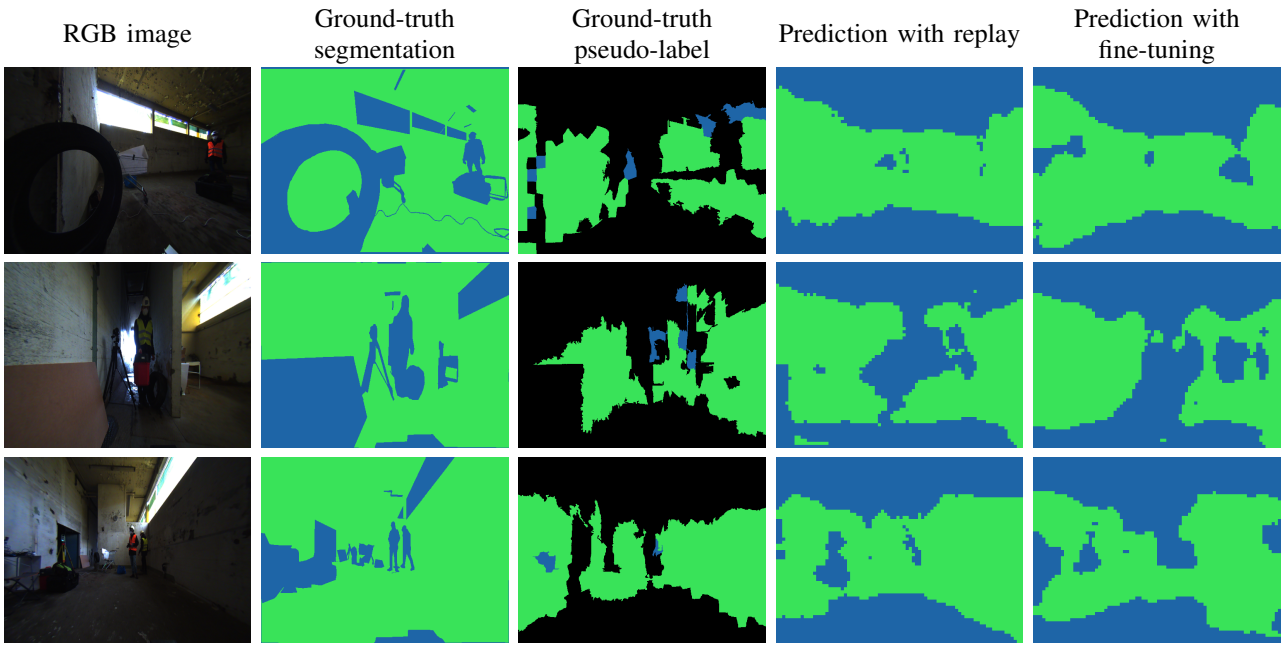


(a) `Garage→Construction`

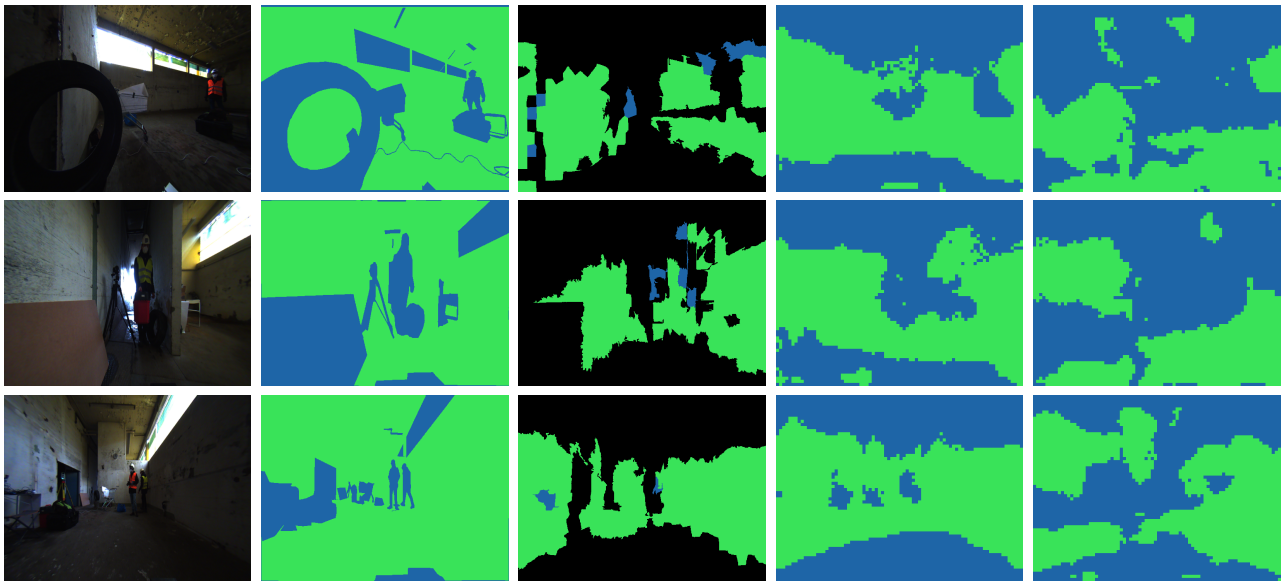


(b) `Garage→Office`

Fig. 6: Illustrations of (prevention of) forgetting for the parking garage as source environment. Green is *background*, blue is *foreground* and black pseudolabels are ignored in training.

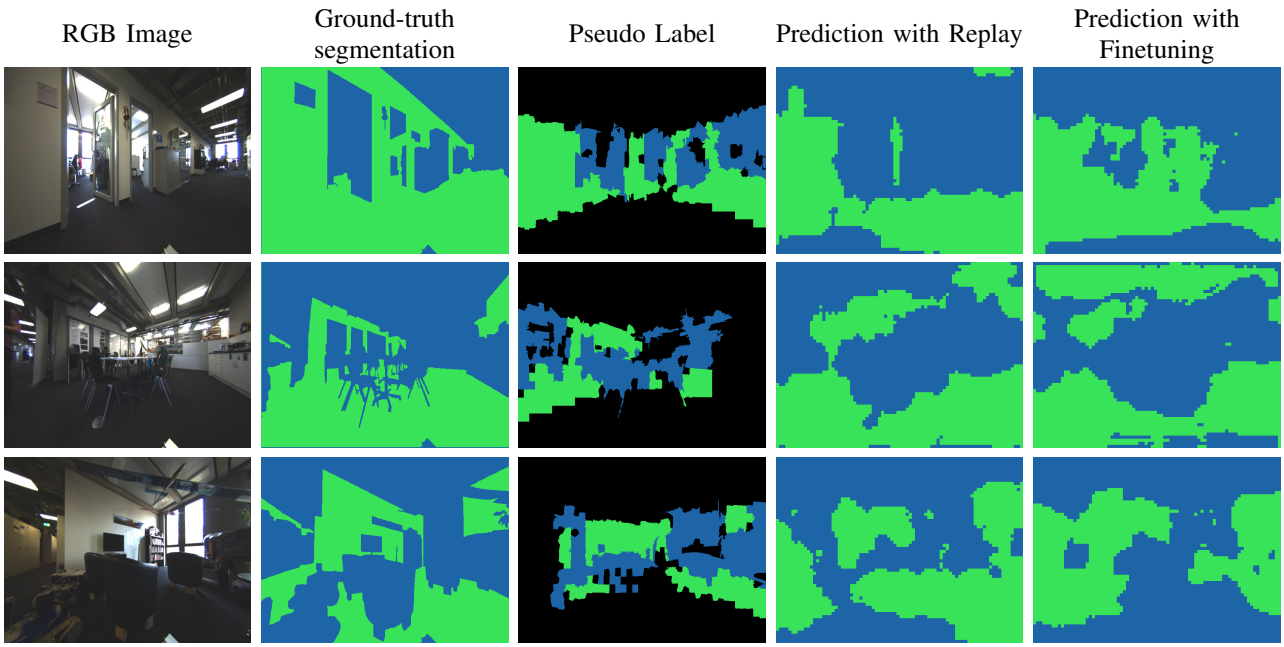


(a) `Construction→Garage`

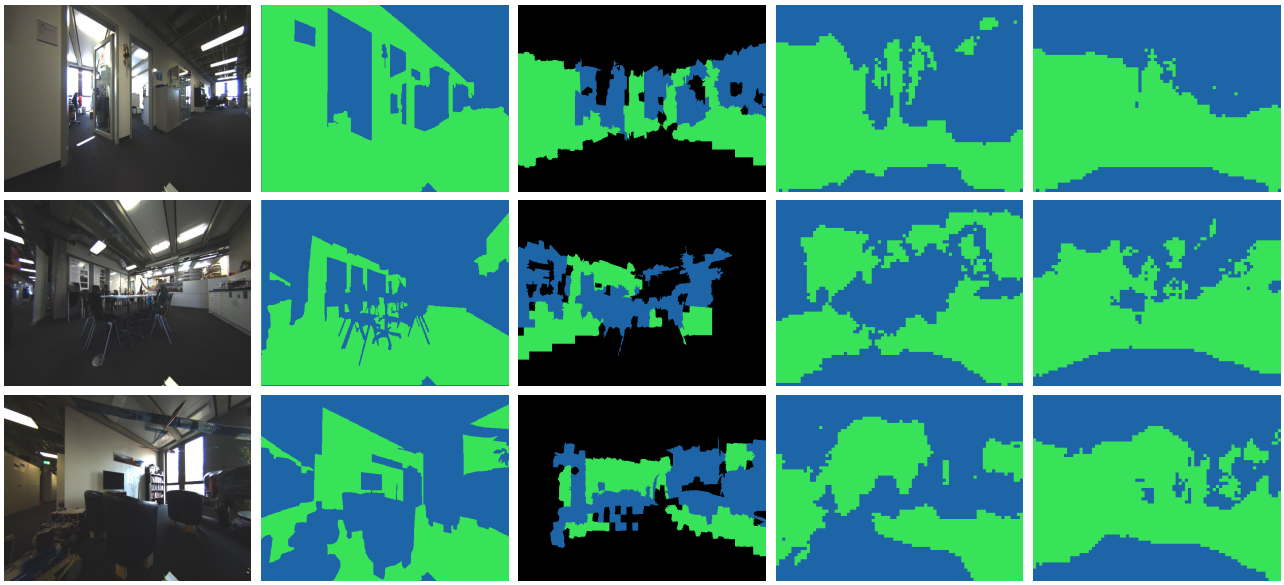


(b) `Construction→Office`

Fig. 7: Illustrations of (prevention of) forgetting for the construction site as source environment. Green is *background*, blue is *foreground* and black pseudolabels are ignored in training.

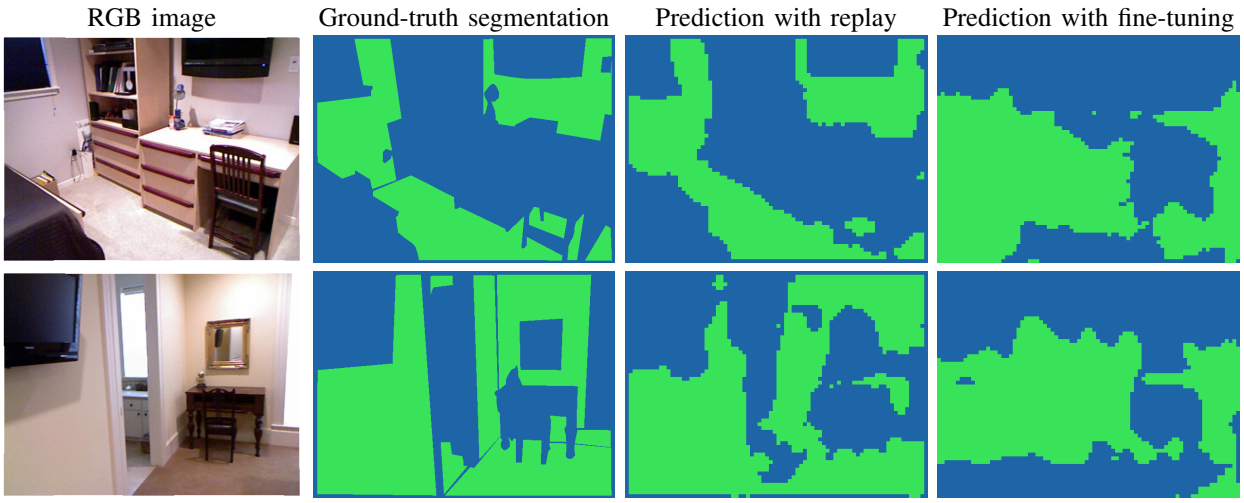


(a) `Office→Garage`

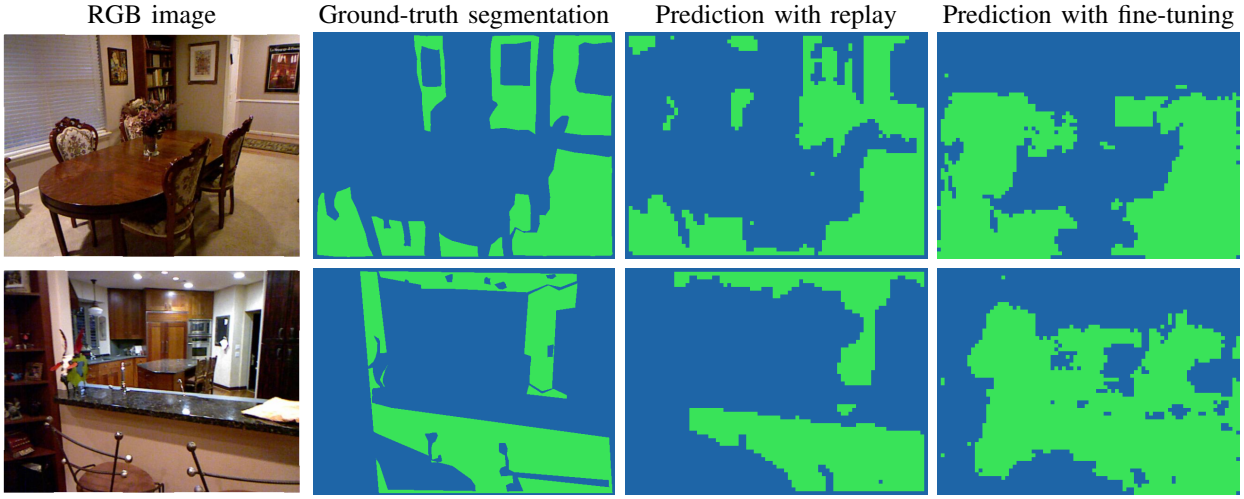


(b) `Office→Construction`

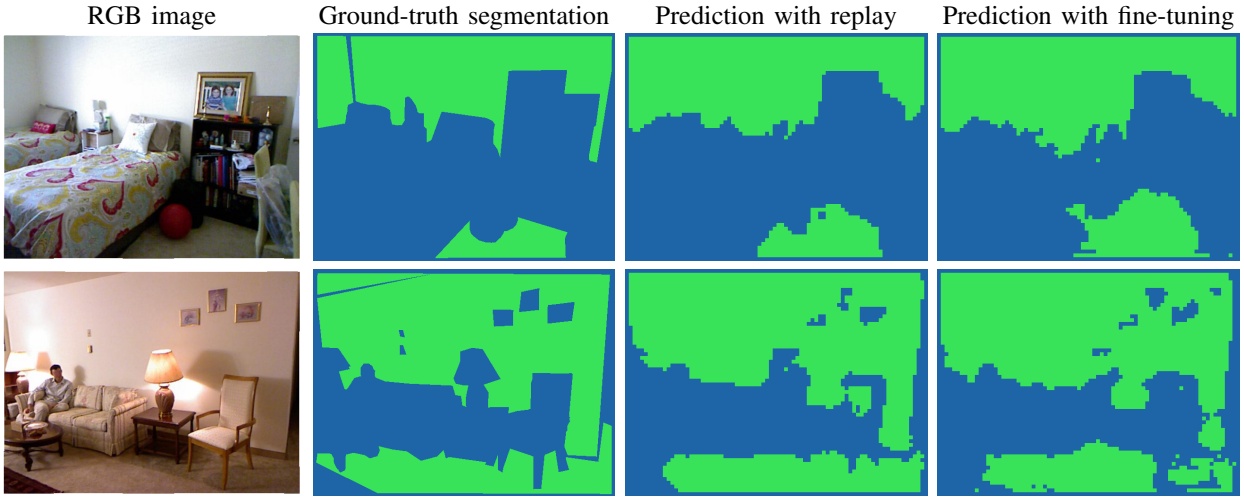
Fig. 8: Illustrations of (prevention of) forgetting for the office as source environment. Green is *background*, blue is *foreground* and black pseudolabels are ignored in training.



(a) NYU→`Garage`



(b) NYU→`Construction`



(c) NYU→Office

Fig. 9: Illustrations of (prevention of) forgetting for the NYU dataset as source environment. Green is *background*, blue is *foreground*.