A Strong Transfer Baseline for RGB-D Fusion in Vision Transformers

Georgios Tziafas[1] and Hamidreza Kasaei[1]

***Abstract*— The Vision Transformer (ViT) architecture has recently established its place in the computer vision literature, with multiple architectures for recognition of image data or other visual modalities. However, training ViTs for RGB-D object recognition remains an understudied topic, viewed in recent literature only through the lens of multi-task pretraining in multiple modalities. Such approaches are often computationally intensive and have not yet been applied for challenging object-level classification tasks. In this work, we propose a simple yet strong recipe for transferring pretrained ViTs in RGB-D domains for single-view 3D object recognition, focusing on fusing RGB and depth representations encoded jointly by the ViT. Compared to previous works in multimodal Transformers, the key challenge here is to use the atested flexibility of ViTs to capture cross-modal interactions at the downstream and not the pretraining stage. We explore which depth representation is better in terms of resulting accuracy and compare two methods for injecting RGB-D fusion within the ViT architecture (i.e., early vs. late fusion). Our results in the Washington RGB-D Objects dataset demonstrates that in such RGB $\rightarrow$ RGB-D scenarios, late fusion techniques work better than most popularly employed early fusion. With our transfer baseline, adapted ViTs score up to 95.1% top-1 accuracy in Washington, achieving new state-of-the-art results in this benchmark. We additionally evaluate our approach with an open-ended lifelong learning protocol, where we show that our adapted RGB-D encoder leads to features that outperform unimodal encoders, even without explicit fine-tuning. We further integrate our method with a robot framework and demonstrate how it can serve as a perception utility in an interactive robot learning scenario, both in simulation and with a real robot.**

I. INTRODUCTION

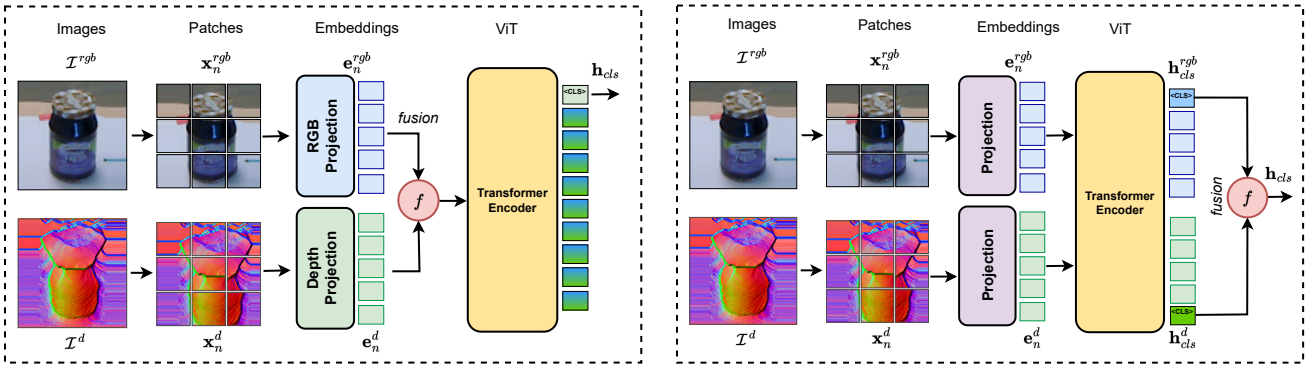
Transfer learning approaches for computer vision have a long standing tradition for image classification, most popularly using Convolutional Neural Networks (CNNs). More recently, the Vision Transformer (ViT) [12] architecture and its many variants [32, 44, 3, 22] have also shown promising transfer results, providing flexible representations that can be fine-tuned for downstream tasks, more recently also in few-shot settings [9]. This flexibility is due to the famous capability of the Transformer architecture to capture long-range dependencies in the input sequence, a trait that is missing from CNNs that are designed with inductive biases for high sensitivity in locality, through their pooling operations. This capability however comes at the cost of data inefficiency [28], as performance gains over CNNs are noticed in Transformers that are pretrained in large-scale datasets, such as ImageNet21k [40] and JFT-300M [42]. When moving from RGB-only to view-based 3D object recognition (RGB-D), a dataset of similar magnitude for pretraining is amiss, granting RGB-D representation learning a topic that has yet to grow. Recent alternative directions include transferring from models pretrained on collections of multimodal datasets [18, 31, 16, 17], although they focus on scene-level tasks, they are constrained to the use of the early fusion strategy and are often computationally intensive to fine-tune.

In this work we wish to explore such questions by revisiting the RGB-D object recognition task and study recipes for transferring an RGB-only pretrained vanilla ViT (i.e. in ImageNet1k [11]) into an RGB-D object-level dataset. We begin by exploring different representation formats for the input depth modality and design two variants that adapt ViT to fuse RGB and depth (see Fig. 1), namely: a) *Early* fusion, where RGB and depth are fused before the encoder and RGB-D patches are represented jointly in the sequence, and b) *Late* fusion, where we move the fusion operation after the encoder, leaving the patch embedders intact from their pretraining. Our hypothesis is that when fine-tuning in small (or moderate) sized datasets, the late fusion baseline is very likely to perform better, as it doesn't change the representation of the input compared to the pretraining stage, but casts the challenge as a distribution shift in the input images (i.e. both RGB and depth are processed by the same weights and must me mapped to the same label).

Experimental results with the Washington RGB-D Objects dataset [29] positively reinforce our hypothesis, as the late fusion baseline far outperforms the early variant. More interestingly, we show that with our late fusion recipe, ViTs achieve new state-of-the-art results in this benchmark, surpassing a plethora of methods that specifically study RGB-D fusion techniques for object recognition. In out experiments we further demonstrate the representational strength of our approach by evaluating using an open-ended lifelong learning scenario, where we show that our late fusion encoder outperforms unimodal versions of same scale, even without fine-tuning. Finally, we demonstrate the applicability of our approach in the robotics domain by integrating our method with a simulated and a real robot framework and show how the robot can be taught by a human user to recognize new objects, in order to perform a table cleaning task. In summary, our contributions are threefold, namely:

- We experimentally find that late works better than early fusion in RGB $\rightarrow$ RGB-D transfer scenario
- We achieve new state-of-the-art results for RGB-D object recognition in the Washington RGB-D Objects benchmark
- We show that our method can be applied in an online lifelong learning setup, including simulation and real robot demonstrations.

[1]Department of Artificial Intelligence, University of Groningen, The Netherlands {g.t.tziafas,hamidreza.kasaei}@rug.nl



(a) Early fusion

(b) Late fusion

Fig. 1: Two different baselines for fusing RGB-D representations in the ViT architecture. In *early* fusion (*left*), a separate projection is used for RGB and depth and the fused embeddings are fed to the encoder, providing a single <CLS> token. In *late* fusion (*right*), the same weights are used for projecting RGB and depth and the two modalities are fed separately to the encoder. The two final <CLS> tokens are fused to provide the final representation for classification.

II. RELATED WORKS

In this section we discuss previous works on RGB-D fusion with CNNs for view-based object recognition, multimodal Transformers and briefly discuss on the topic of lifelong learning, which we include as an evaluation scenario in our experiments.

A. RGB-D Fusion with CNNs

As in RGB image classification, multiple traditional CNN-based approaches have replaced conventional approaches [4, 43] for extending to the RGB-D modalities. The focus of such works lies in RGB-D fusion, where deep features extracted from CNNs are fused through a multimodal fusion layer [46] or custom networks [45]. Rahman et. al. [14] propose a parallel three-stream CNN which processes two depth encodings in two streams and RGB in the last one. Cheng et. al. [10] proposed to integrate Gaussian mixture models with CNNs through fisher kernel encodings. Zia et. al. [53] propose mixed 2D/3D CNNs which are initialized with 2D pretrained weights and extend to 3D to also incorporate depth. Such methods study how to inject fusion in the locally-aware CNN architecture. Instead, in our work we implement fusion as a simple operation on embeddings from the different modalities and opt to gain cross-modal alignment by virtue of the long-range context modeling capabilities of the Transformer architecture.

B. Multimodal Learning with Transformers

In the absence of a large-scale RGB-D dataset for pre-training, recent works try to alleviate this bottleneck by pretraining on collections of datasets from multiple modalities [18, 31, 16, 17] and rely on the flexibility of Transformers to capture cross-modal interactions. However, such methods focus on scene/action recognition or semantic segmentation tasks, leaving the RGB-D object recognition task unexplored. Furthermore, they employ an early fusion technique for converting heterogeneous modalities in the same sequence representation, leaving open questions of whether this is the best fusion technique in homogeneous modalities such as RGB-D, as well as if its the best fusion technique for directly transferring from one homogeneous modality to another, without the pretraining step. Finally, they rely heavily on model capacity and specialized Transformer architecture variants (e.g Swin [32]) in order to enable multimodal pretraining to boost performance in unimodal downstream tasks. Such models set a high computational resource entry point for practitioners, casting them not widely accessible for fine-tuning in arbitrary datasets.

C. Lifelong Learning

An emerging topic in deep learning literature, most commonly referred to as *Lifelong* or *Continual Learning*, studies the scenario of a learning agent continuously incorporating new experiences from an online data stream. In the context of image classification, the challenge is stated as learning to classify images from an ever-growning set of new data and tasks, while avoiding the effect of catastrophic forgetting [35, 8, 39, 50, 51, 41]. Even though works for using Transformers in lifelong learning are starting to grow [13, 47, 15], to the best of our knowledge, this is the first work that touches on lifelong learning with Transformers for RGB-D object recognition. We highlight however that the focus of this work is not on lifelong learning algorithms, but rather to establish a baseline in the Washington benchmark for future references.

III. APPROACH

Our goal is to have a single model that can be transferred to RGB-D downstream tasks, while being pretrained solely in RGB. Even though the two modalities are homogeneous, different depth representations might insert discrepancies in the size of the input depth image. To deal with this, we adopt the Transformer architecture, because the self-attention operation has shown a tremendous ability to model long-range dependencies of variable size inputs. Unlike standard fine-tuning strategies, we wish to enable ViT to learn from the depth modality, as well as learn how to successfully model

the correspondences between the two modalities. To that end, we explore two different RGB-D representation fusion techniques.

A. ViT Prerequisites

The ViT model handles the visual input as a sequence of image patches. The original $H \times W$ image $\mathscr{I}$ is split into patches of size $h \times w$, resulting in a total of $N = \frac{H \cdot W}{h \cdot w}$ patches. Each patch is flattened into a single vector representation $\mathbf{x}_n \in \mathbb{R}^{3*h*w}$ and projected into an embedding space through a linear map $\mathscr{E}(\mathbf{x}_n) = \mathbf{e}_n$, $\mathscr{E} : \mathbb{R}^{3*h*w} \rightarrow \mathbb{R}^D$. A trainable image-level embedding $\mathbf{e}_0$ (i.e. the <CLS> token representation) is stacked with the embeddings sequence and the patch embeddings are further added with positional encodings $\mathbf{p}_n$, either learned jointly or hand-crafted (e.g 2D sinusoid). The resulting sequence $[\mathbf{e}_0, \{\mathbf{e}_n + \mathbf{p}_n\}_{n=1}^N]$ is passed through L layers of Transformer encoder blocks, resulting in the sequence of hidden states $[\mathbf{h}_n^l]_{n=0}^N$, $l = 1, .., L$. For downstream classification tasks, the final hidden <CLS> state $\mathbf{h}_{cls} \doteq \mathbf{h}_0^L$ is fed into a linear layer over the number of target classes combined with a softmax loss.

B. Depth Representation

In order to make ViT compatible with the RGB-D modality, we need to express the input depth map with the same format as RGB. In the standard ViT pipeline, the input RGB image is first resized to a fixed resolution (in base configuration, $H = W = 224$), center-cropped and then normalized according to the mean and standard deviation of the training dataset. We experiment with three different types of depth representations, inspired by previous works, namely:

1) **Raw depth** maps, truncated to a pre-set maximum depth value (e.g. around 3.5 meters for Kinect) and clipped to $[0, 255]$ range. We stack three instances of the resulting map to "convert" it to RGB.
2) **HHA** transformations of the raw depth maps, which have shown to encode geometric properties, such as geocentric pose. To compute the transform, depth is first converted to a disparity map using the camera intrisics. The HHA is then build as an image with three channels at each pixel, including horizontal disparity, height above ground and the angle at the pixel's local surface normal.
3) **Surface Normal** reconstructions, which have shown to encode fine-grained 3D details about shape, texture and surface. These images are generated by rendering a 3D point-cloud from the RGB-D pair and estimating surface normals at each point. The 3D vectors are back-projected to the camera reference frame and colourized separately in a channel.

The resulting colourized depth image is fed into the same resize-crop-normalization preprocessing as in RGB.

C. RGB-D Fusion Techniques

We explore two different types of RGB-D fusion, aiming to asses which is the most accurate way to adapt pretrained RGB Transformers for RGB-D recognition tasks in the absence of large-scale RGB-D datasets.

a) ***Early Fusion****:* In early fusion, the RGB-D representations are fused before the Transformer encoder, and the encoder is fine-tuned as-is in the multi-modal representations. Following [18], we use a separate patch embedding layer for each modality, $\mathscr{E}^{rgb}$ and $\mathscr{E}^{d}$ and fuse the two representations before adding the position embeddings, using addition and L2 normalization:

$$\mathbf{e_n} = \frac{\mathscr{E}^{rgb}(x_n^{rgb}) + \mathscr{E}^{d}(x_n^{d})}{\left\| \mathscr{E}^{rgb}(x_n^{rgb}) + \mathscr{E}^{d}(x_n^{d}) \right\|_2} \quad (1)$$

We call this baseline the *dual*-embedder, as it separates embeddings for the two modalities. In our experiments we also implement the early fusion baseline with a *joint*-embedder, stacking the two modalities channel-wise and using a single projection to embed them jointly. In this architecture, a single <CLS> embedding is learned for the entire RGB-D pair. The key insight is that through the self-attention operation the joint RGB-D embedding $\mathbf{h}_{cls}$ will adapt to model the inter-modal alignment between the fused representations. To make the pretrained ViT checkpoint compatible with the adapted architecture, we copy the weights of the pretrained patch embedder in both RGB and depth embedders.

b) ***Late Fusion****:* In late fusion, we pass the two images from the ViT encoder separately and aggregate their final <CLS> embeddings, before passing it to the classifier. We experiment with different types of late fusion operations $f : \mathbb{R}^D \times \mathbb{R}^D \rightarrow \mathbb{R}^{D'}$, such as max pooling ($D' = D$), averaging ($D' = D$), addition ($D' = D$) and concatenation ($D' = 2 \cdot D$). In this baseline, the encoder has to learn how to classify both images of the same object view, while processing the two modalities separately. The final hidden representation used for classification is the fusion of the two hidden states for the two modalities: $\mathbf{h}_{cls} = f(\mathbf{h}_{cls}^{rgb}, \mathbf{h}_{cls}^{d})$

D. Implementation Details

In order to fine-tune the adapted ViT, we use a dataset-specific linear layer on top of the final <CLS> embedding with a softmax loss over the datasets' categorical distribution of labels. For the early fusion baseline, we replicate the implementation by [18] and develop the RGB and depth embedding layers as 2D convolution layers with D feature maps and kernel size and stride of N. The input channels are 3 for RGB and depth separately in the dual version and 6 for the joint embedder variant. For training the late fusion baseline, we generate two batches (RGB-labels, Depth-labels) and interleave the depth batch in-between the RGB images, so that the model processes the RGB-D pair in pairs of two, even in the case of distributed parallel training with multiple GPUs and/or nodes. We experiment with the default public configurations of ViT-x, where $x = \{$T,S,B,L$\}$ for $\{$*tiny*, *small*, *base*, *large*$\}$. We develop our method using PyTorch [36] and the Hugging Face API [48].

IV. EXPERIMENTS

This section describes our evaluation setup and presents our experimental results. It is organized as follows: First (Sec. IV-A), we give the specifics of the RGB-D dataset used

TABLE I: Ablation study of different model components. We report top-1 predicted accuracy (%) results of different evaluation scenarios for a ViT-T model on the first trial of the Washington RGB-D Objects dataset. Best results are highlighted in bold.

Method	k-NN	Lin.Eval	FT
RGB	79.5	80.2	83.0
Depth (Raw)	59.7	62.0	69.9
Depth (HHA)	65.1	67.8	73.2
Depth (SurfNorm)	66.3	69.9	76.7
RGB-D (Early w/ dual-emb)	-	-	82.1
RGB-D (Early w/ joint-emb.)	-	-	81.0
RGB-D (Late w/ avg)	82.0	82.1	85.5
RGB-D (Late w/ max)	**85.4**	85.7	88.4
RGB-D (Late w/ cat)	**85.4**	**87.4**	**90.0**

for training and the evaluation procedure. Then (Sec. IV-B), we perform ablation studies for different variants of the depth representation and the fusion approach. We select the best performing configuration of our ablation studies and scale it to compare with previous state-of-the-art for RGB-D object recognition in the Washington benchmark (Sec. IV-C). Finally (Sec. IV-D), we study the performance of our approach when evaluated in an online lifelong learning scenario and (Sec. IV-E) demonstrate how it can be integrated with a robot framework for interactive robot learning applications.

A. Dataset and Evaluation

The Washington RGB-D object dataset [29] is a well established benchmark for object recognition tasks in RGB-D domains. It contains up to 41 877 views from 300 object instances, organized into 51 categories, including common household objects (cups, bowls, mugs etc.) with variations in fine-grained attributes (e.g. color). Each view was taken from 30◦, 45◦ and 60◦ elevation angles of a Kinect sensor. For depth representations, we use the surface normals as extracted from [6] and manually perform HHA conversions. Regarding evaluation, we perform the suggested experiment as in the original paper [29]. In particular, the dataset provides 10 train/test splits, where in each split, one instance per object category is used for testing and the rest for training. For a single trial, a total of 51 category instances ($\sim$ 7k RGB-D pairs) are used for validation and the remaining 249 instances ($\sim$ 35k RGB-D pairs) are used for training. We use top-1 predicted accuracy as the evaluation metric and report averaged mean and standard deviation of accuracies across the 10 trials.

B. Ablation Studies

We ablate the following aspects of our approach: a) the format of the input depth image, as described in Sec. III-B, b) the type of RGB-D fusion used (*Early* vs. *Late*), c) the type of embedder used in the *Early* fusion baseline (*joint-* vs. *dual-emb.*) as well as d) the type of fusion method used in the *Late* fusion baseline, including averaging *(avg)*, max pooling *(max)* and concatenation *(cat)*. In order to reduce computational overhead, we experiment with ViT-T using only the first trial of the Washington RGB-D evaluation setup. In order to gain better insight in each configuration's contribution, we use three different evaluation scenarios, namely: a) k-nearest neighbour (*k-NN*) on top of frozen pretrained embeddings, b) training a linear head on top of frozen pretrained embeddings (*Lin.Eval*), and c) fine-tuning the Transformer end-to-end with a classification head (*FT*). We note that the early baseline does not include results with frozen representations, as the encoder cannot be used out-of-the-box for RGB-D embeddings at the input. For RGB-D methods we report results using the *SurfNorm* depth format, as we observe that achieves best results. For *k-NN*, we use $k = 3$ and cosine similarity as the distance metric, as we experimentally find that this the best performing setup. For *Lin.Eval*, we train using minibatch SGD with momentum value 0.9, batch size 128, a learning rate of $5 \cdot 10^{-4}$ and early stopping. For fine-tuning, we use AdamW [34], batch size 512, a learning rate of $9 \cdot 10^{-5}$ with no warmup and a linear decay over 10 epochs. We train on 2 Nvidia Titan Xp GPUs. Our results are summarized in Table I.

We observe that the *SurfNorm* depth representation leads to best depth-only performance in ViT, as it is found in previous works using CNNs. Regarding early fusion, we also confirm that using separate embeddings for the two modalities (*dual* baseline) leads to marginally better results than the *joint*. Regarding late fusion, the concatenation operation overperforms other studied fusion approaches, with the cost of doubling the hidden size of the classifier's input. When comparing the two fusion baselines, we observe that the late fusion baseline far outperforms the early one. In particular, the early baseline achieves worst results than RGB-only. We believe that this result reinforces our original hypothesis, namely that in the absence of large-scale multimodal datasets for pretraining, attempting to modify the input embeddings greatly disturbs the fine-tuning process, leading to overfitting. In contrast, in the late fusion baseline, the encoder "sees" the same representations but only learns to adapt the final layers to incorporate depth features. We confirm this statement by verifying that the weights in the adapted ViT have greater absolute difference on average in the early rather than the late baseline.

C. Offline RGB-D Object Recognition

Table II presents results for the 10 trials of the evaluation setup, compared with previous state-of-the-art methods, as reported in [6]. We use ViT-B and the best configuration from our ablation experiments (i.e. *Late* + *cat*). For reference, we include results over the 10 trials using the dual-embedder early fusion architecture. We also include another baseline, ViT-B Ensemble (*Ens.*), in which we have fine-tuned the ViT on RGB and depth separately and then use an ensemble of both fine-tuned models during inference. The two final representations from the fine-tuned encoders are fused and fed to the classifier, as in late fusion. We highlight that this

TABLE II: Mean and standard deviation results for top-1 predicted accuracy (%) over the 10 trials of the Washington RGB-D Objects dataset. Best results are highlighted in bold, and second best are underlined. Our approach achieves new state-of-the-art results in this benchmark. The method with † uses one encoder per modality, thus doubling the spatial requirements.

Method	RGB	Depth	RGB-D
Fusion 2D/3D CNNs [52]	89.0 ± 2.1	78.4 ± 2.4	91.8 ± 0.9
STEM-CaRFs [1]	88.0 ± 2.0	80.8 ± 2.1	92.2 ± 1.3
MM-LRF-ELM [30]	84.3 ± 3.2	82.9 ± 2.5	89.6 ± 2.5
VGG f-RNN [5]	89.9 ± 1.6	84.0 ± 1.8	92.5 ± 1.2
DECO [7]	89.5 ± 1.6	84.0 ± 2.3	93.6 ± 0.9
MDSI-CNN [2]	89.9 ± 1.8	84.9 ± 1.7	92.8 ± 1.2
HP-CNN [49]	87.6 ± 2.2	$\underline{85.0 \pm 2.1}$	91.1 ± 1.4
RCFusion [33]	89.6 ± 2.2	$\underline{85.9 \pm 2.7}$	$\underline{94.4 \pm 1.4}$
MMFLAN [38]	83.9 ± 2.2	84.0 ± 2.6	93.1 ± 1.3
DenseNet121-RNN [6]	91.5 ± 1.1	86.9 ± 2.1	93.5 ± 1.0
ResNet101-RNN [6]	$\underline{92.3 \pm 1.0}$	$\mathbf{87.2 \pm 2.5}$	94.1 ± 1.0
Ours (ViT-B Ens.)†	90.8 ± 1.9	83.7 ± 2.1	90.4 ± 1.5
Ours (ViT-B Early)	-	-	89.5 ± 1.5
Ours (ViT-B Late)	$\underline{92.6 \pm 1.1}$	83.6 ± 2.4	$\underline{94.8 \pm 1.5}$
Ours (ViT-L Late)	$\mathbf{92.9 \pm 1.3}$	83.5 ± 2.1	$\mathbf{95.1 \pm 1.3}$

method requires to keep two encoders in memory, one per modality. This baseline is included in order to verify whether joint fine-tuning achieves better performance than fine-tuning each modality separately and ensembling the resulting models. Finally, we further scale our model using the ViT-L model, in order to provide a fair comparison capacity-wise with the previous state-of-the-art model [6], which uses a ResNet101 architecture [21]. We train using the same hyper-parameters as the fine-tuning experiments of the previous section, but a learning rate of $3 \cdot 10^{-5}$ and batch sizes of 64 and 32 for ViT-B and ViT-L respectively [1].

We observe that our model achieves new state-of-the-art in the Washington benchmark, even when using the ViT-B model, which has less capacity than the ResNet101 of the previous state-of-the-art. When scaling to ViT-L, our model achieves a margin of 0.9% from previous best result in RGB-D. Compared to our ensemble baseline, we observe that joint fine-tuning indeed leads to better scores than ensembling two modality-specific encoders with late fusion.

D. Open-Ended Lifelong RGB-D Object Recognition

In this section we wish to evaluate our approach in an online fashion, where we assume that the learning agent is presented with novel object instances throughout a lifespan. In order to evaluate in such an open-ended scenario, we follow the evaluation protocol proposed in [19, 27, 26]. In particular, we develop a simulated user who gradually introduces new object categories to the agent by presenting an unseen view of an object category. An illustration is given in Fig. 2. After teaching each new category, the user

[1]The choice of batch size had to be compromised due to limited computational resources, we expect with larger batch sizes to further improve performance.

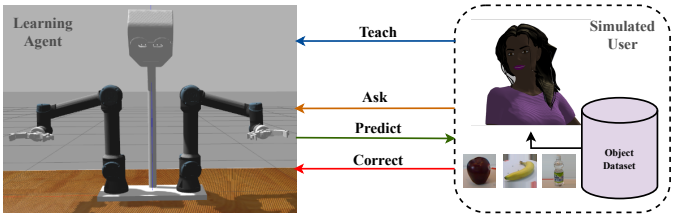


Fig. 2: Cartoon illustration of the evaluation protocol implemented for an open-ended lifelong learning scenario: The simulated user teaches a new category to the robot using three randomly selected instances, and then samples new instances to evaluate the robot on all learned categories, and make sure that interference has not happened after introducing the new category. Once a certain threshold of accuracy is met, the user introduces a new category. The user is also enabled to correct mis-classifications from the model.

examines the classification accuracy over a collected set of instances and evaluate whether the category has been learned and interference has not been happened. The user can also correct mis-classified examples in order to update the category models. Training and evaluating the agent is performed until a specific *protocol threshold* value is met (e.g for threshold 0.67, the accuracy rate must be at least double from the error rate), after when a new category is introduced, or the agent learned all existing categories. Random sampling is used to select new data points from each category from the evaluation dataset.

In order to measure the effect of *catastrophic forgetting* in the evaluation, the user tests the agent in all previous categories after each new introduction. The evaluation stops either when the agent has learned all categories without catastrophic forgetting (according to the protocol threshold) or is unable to do after attempting it for more than a specified number of *Question/Correction Iterations (QCI)*. Evaluation metrics include: *(i)*, *Average number of Learned Categories (ALC)*, *(ii)*, *Average number of stored Instances per Category (AIC)*, *(iii)*, *Global Classification Accuracy (GCA)* and *(iv)*, *Average Protocol Accuracy (APA)*. In order to asses the performance of the different methods with stricter teachers, we repeat the evaluation while setting the value of the protocol threshold to $\{0.7, 0.8, 0.9\}$. We report results in these metrics for Washington RGB-D Objects dataset, comparing with uni-modal baselines. In this experiment we use the k-NN classifier strategy on top of pretrained embeddings, which is compatible with the *Instance-Based Learning* (IBL) approach of the evaluation protocol. As before, we use $k = 3$ and cosine distance function. Results are summarized in Table III.

We observe that even without any RGB-D fine-tuning, fusing the embeddings generated by the ImageNet ViT checkpoint still provides accuracy benefits over RGB-only and Depth-only classification, in all protocol threshold settings.

E. Robot Demonstrations

We develop a simulation environment in Gazebo to evaluate the real-time performance of the proposed approach in the

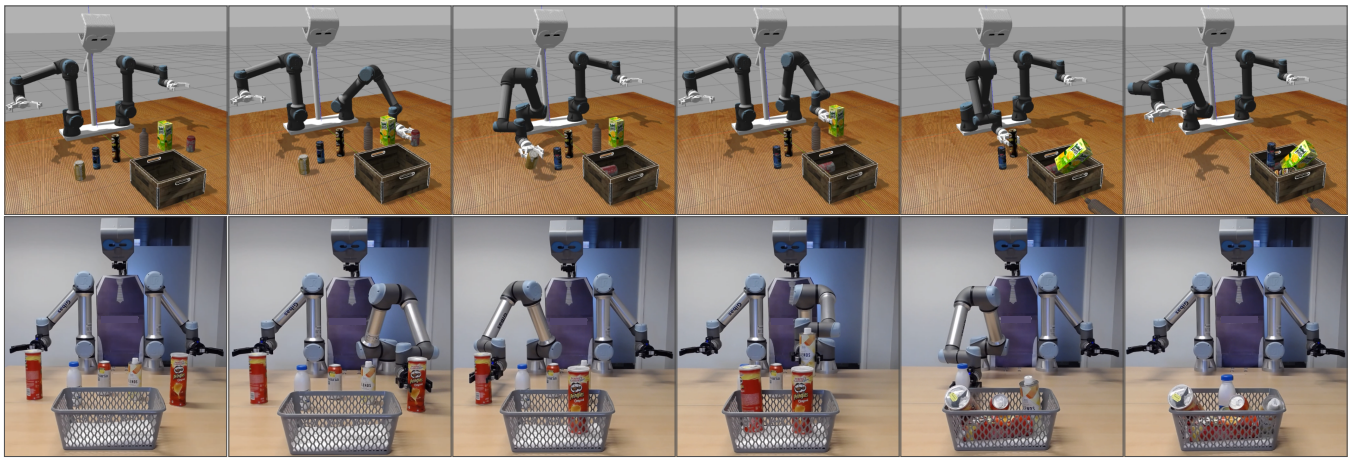


Fig. 3: A sequence of snapshots capturing the experimental setup and the behaviour of the robot in Gazebo (*top)* and in a real-world (*bottom)*. We randomly place objects and instruct the robot to recognize them and place them in the container.

TABLE III: Online lifelong learning evaluation on Washington RGB-D with a simulated teacher. Higher protocol threhsold values represent a "stricter" teacher, requiring more correct answers to consider a classification trial successful. We report results for ViT-B in RGB-only, Depth-only and our RGB-D late fusion baseline. The arrow demonstrates if better results are higher or lower for each metric (refer to the text for explanation of the evaluation metrics).

Threshold	Method	Washington RGB-D Objects				
		QCI↓	ALC↑	AIC↓	GCA↑	APA↑
0.7	ViT-B (RGB)	1325	**51**	6.4	86.9	88.2
	ViT-B (Depth)	1329	**51**	7.7	77.7	79.2
	ViT-B (RGB-D *Late*)	1325	**51**	**5.9**	**88.3**	**89.1**
0.8	ViT-B (RGB)	1369	**51**	**6.1**	88.6	89.4
	ViT-B (Depth)	2029	**51**	7.8	81.8	84.2
	ViT-B (RGB-D *Late*)	1370	**51**	6.3	**88.7**	**90.2**
0.9	ViT-B (RGB)	2368	**51**	7.3	90.2	93.1
	ViT-B (Depth)	2954	34	8.4	90.1	90.5
	ViT-B (RGB-D *Late*)	1695	**51**	**6.7**	**90.7**	**93.9**

context of a `clear_table` task (see Fig. 3). For this round of experiments, we integrate our work into the cognitive robotic system presented in [25][24]. We performed 10 clear_table experiments. At the beginning of each experiment, we randomly place four to six objects and a container on the table. The robot does have any knowledge about the objects, therefore, it recognizes all objects as "*unknown*". A human user teaches all object categories to the robot using a GUI and the robot recognizes all object instances before placing them into the container. Note that the pose of the container is known to the robot in advance. In all experiments, we observed that the robot could incrementally learn all object categories using a single instance for teaching, recognized them correctly, and completed the task successfully. A video of these experiments has been attached to the paper as supplementary material.

V. CONCLUSION

In this work we propose a simple yet strong recipe for fine-tuning ViTs in RGB-D domains. We experiment with two different types of fusion (early vs. late) and demonstrate that unlike most prior arts that use early fusion, the late fusion strategy transfers better in the low-data regime. By fine-tuning a ViT with our late fusion approach, we push the state-of-the-art in the Washington RGB-D Objects benchmark by 0.9%, using non optimal configurations due to computational restraints. We further show that our approach is more robust than unimodal approaches when the training-test paradigm is replaced with an open-ended lifelong learning scenario and demonstrate how it can serve as a perception utility for interactive lifelong robot learning, both in simulation and with a real robot. We hope that our approach will lead more research on efficiently transferring ViTs for robotics-specific domains.

This work leaves us with a multitude of potential future directions, regarding the sophistication of the RGB-D fusion, the efficiency of transferring and generalization to novel domains. For the first topic, the fusion method we present in this work is a single operation between modality-specific embeddings. Other approaches that entangle fusion within the encoder can be considered in the future, such as hierarchical feature fusion. Another limitation is that our method currently fine-tunes the entire pretrained model, setting a time and compute requirement that is still considerable. There is a broad literature in using adapters for efficient parameter-light fine-tuning of Transformers in NLP [23, 37, 20]. It would be interesting to explore adapters for even more efficient transfer of ViTs in RGB-D domains. Finally, regarding generalization, our model transfers from ImageNet and its high accuracy in the Washington benchmark is guaranteed due to overlap of existing object classes (as suggested by high k-NN RGB-only scores). We expect that this is not the case when moving in-the-wild. In the future we plan to investigate transferring ViTs that are pretrained with self-supervised objectives, such as masked autoencoding, and compare their transfer performance in-the-wild with supervised methods.

REFERENCES

[1] Umar Asif, Mohammed Bennamoun, and Ferdous A. Sohel. "A Multi-Modal, Discriminative and Spatially Invariant CNN for RGB-D Object Labeling". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40.9 (2018), pp. 2051–2065. DOI: 10.1109/TPAMI.2017.2747134.

[2] Umar Asif, Bennamoun, and Ferdous Sohel. "A Multi-Modal, Discriminative and Spatially Invariant CNN for RGB-D Object Labeling". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40 (2018), pp. 2051–2065.

[3] Hangbo Bao, Li Dong, and Furu Wei. "BEiT: BERT Pre-Training of Image Transformers". In: *CoRR* abs/2106.08254 (2021). arXiv: 2106.08254. URL: https://arxiv.org/abs/2106.08254.

[4] Liefeng Bo, Xiaofeng Ren, and Dieter Fox. "Depth kernel descriptors for object recognition". In: *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems*. 2011, pp. 821–826. DOI: 10.1109/IROS.2011.6095119.

[5] Ali Caglayan and Ahmet Burak Can. "Exploiting Multi-layer Features Using a CNN-RNN Approach for RGB-D Object Recognition". In: *ECCV Workshops*. 2018.

[6] Ali Caglayan et al. "When CNNs Meet Random RNNs: Towards Multi-Level Analysis for RGB-D Object and Scene Recognition". In: *CoRR* abs/2004.12349 (2020). arXiv: 2004.12349. URL: https://arxiv.org/abs/2004.12349.

[7] Fabio Maria Carlucci et al. "$(DE)^2$ CO: Deep Depth Colorization". In: *CoRR* abs/1703.10881 (2017). arXiv: 1703.10881. URL: http://arxiv.org/abs/1703.10881.

[8] Arslan Chaudhry et al. "Efficient Lifelong Learning with A-GEM". In: *CoRR* abs/1812.00420 (2018). arXiv: 1812.00420. URL: http://arxiv.org/abs/1812.00420.

[9] Yuzhong Chen et al. *Mask-guided Vision Transformer (MG-ViT) for Few-Shot Learning*. 2022. DOI: 10.48550/ARXIV.2205.09995. URL: https://arxiv.org/abs/2205.09995.

[10] Yanhua Cheng et al. "Convolutional Fisher Kernels for RGB-D Object Recognition". In: *2015 International Conference on 3D Vision*. 2015, pp. 135–143. DOI: 10.1109/3DV.2015.23.

[11] Jia Deng et al. "Imagenet: A large-scale hierarchical image database". In: *2009 IEEE conference on computer vision and pattern recognition*. Ieee. 2009, pp. 248–255.

[12] Alexey Dosovitskiy et al. *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. 2021. arXiv: 2010.11929 [cs.CV].

[13] Arthur Douillard et al. "DyTox: Transformers for Continual Learning with DYnamic TOken eXpansion". In: *arXiv preprint arXiv:2111.11326* (2021). URL: https://arxiv.org/abs/2111.11326.

[14] Andreas Eitel et al. "Multimodal Deep Learning for Robust RGB-D Object Recognition". In: *CoRR* abs/1507.06821 (2015). arXiv: 1507.06821. URL: http://arxiv.org/abs/1507.06821.

[15] Beyza Ermis et al. *Continual Learning with Transformers for Image Classification*. 2022. DOI: 10.48550/ARXIV.2206.14085. URL: https://arxiv.org/abs/2206.14085.

[16] Rohit Girdhar et al. *OmniMAE: Single Model Masked Pretraining on Images and Videos*. 2022. DOI: 10.48550/ARXIV.2206.08356. URL: https://arxiv.org/abs/2206.08356.

[17] Rohit Girdhar et al. *OmniMAE: Single Model Masked Pretraining on Images and Videos*. 2022. DOI: 10.48550/ARXIV.2206.08356. URL: https://arxiv.org/abs/2206.08356.

[18] Rohit Girdhar et al. "Omnivore: A Single Model for Many Visual Modalities". In: *CoRR* abs/2201.08377 (2022). arXiv: 2201.08377. URL: https://arxiv.org/abs/2201.08377.

[19] S. Hamidreza Kasaei, Lués Seabra Lopes, and Ana Maria Tomé. "Coping with Context Change in Open-Ended Object Recognition without Explicit Context Information". In: *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 2018, pp. 1–7. DOI: 10.1109/IROS.2018.8593922.

[20] Junxian He et al. "Towards a Unified View of Parameter-Efficient Transfer Learning". In: *CoRR* abs/2110.04366 (2021). arXiv: 2110.04366. URL: https://arxiv.org/abs/2110.04366.

[21] Kaiming He et al. "Deep Residual Learning for Image Recognition". In: *CoRR* abs/1512.03385 (2015). arXiv: 1512.03385. URL: http://arxiv.org/abs/1512.03385.

[22] Kaiming He et al. "Masked Autoencoders Are Scalable Vision Learners". In: *CoRR* abs/2111.06377 (2021). arXiv: 2111.06377. URL: https://arxiv.org/abs/2111.06377.

[23] Neil Houlsby et al. "Parameter-Efficient Transfer Learning for NLP". In: *Proceedings of the 36th International Conference on Machine Learning*. Ed. by Kamalika Chaudhuri and Ruslan Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. PMLR, Sept. 2019, pp. 2790–2799. URL: https://proceedings.mlr.press/v97/houlsby19a.html.

[24] Hamidreza Kasaei and Mohammadreza Kasaei. "MV-Grasp: Real-Time Multi-View 3D Object Grasping in Highly Cluttered Environments". In: *arXiv preprint arXiv:2103.10997* (2021).

[25] S Hamidreza Kasaei et al. "Towards lifelong assistive robotics: A tight coupling between object perception and manipulation". In: *Neurocomputing* 291 (2018), pp. 151–166.

[26] S. Hamidreza Kasaei. "Interactive Open-Ended Learning for 3D Object Recognition". In: *CoRR* abs/1912.09539 (2019). arXiv: 1912.09539. URL: http://arxiv.org/abs/1912.09539.

[27] S. Hamidreza Kasaei. "OrthographicNet: A Deep Learning Approach for 3D Object Recognition in Open-Ended Domains". In: *CoRR* abs/1902.03057 (2019). arXiv: 1902.03057. URL: http://arxiv.org/abs/1902.03057.

[28] Salman H. Khan et al. "Transformers in Vision: A Survey". In: *CoRR* abs/2101.01169 (2021). arXiv: 2101.01169. URL: https://arxiv.org/abs/2101.01169.

[29] Kevin Lai et al. "A large-scale hierarchical multi-view RGB-D object dataset". In: *2011 IEEE International Conference on Robotics and Automation*. 2011, pp. 1817–1824. DOI: 10.1109/ICRA.2011.5980382.

[30] Fengxue Li et al. "Multi-Modal Local Receptive Field Extreme Learning Machine for object recognition". In: *2016 International Joint Conference on Neural Networks (IJCNN)*. 2016, pp. 1696–1701. DOI: 10.1109/IJCNN.2016.7727402.

[31] Huayao Liu et al. "CMX: Cross-Modal Fusion for RGB-X Semantic Segmentation with Transformers". In: *arXiv e-prints*, arXiv:2203.04838 (Mar. 2022), arXiv:2203.04838. arXiv: 2203.04838 [cs.CV].

[32] Ze Liu et al. "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows". In: *CoRR* abs/2103.14030 (2021). arXiv: 2103.14030. URL: https://arxiv.org/abs/2103.14030.

[33] Mohammad Reza Loghmani et al. "Recurrent Convolutional Fusion for RGB-D Object Recognition". In: *CoRR* abs/1806.01673 (2018). arXiv: 1806.01673. URL: http://arxiv.org/abs/1806.01673.

[34] Ilya Loshchilov and Frank Hutter. "Decoupled Weight Decay Regularization". In: *International Conference on Learning Representations*. 2019. URL: https://openreview.net/forum?id=Bkg6RiCqY7.

[35] Zheda Mai et al. "Online Continual Learning in Image Classification: An Empirical Survey". In: *CoRR* abs/2101.10423 (2021). arXiv: 2101.10423. URL: https://arxiv.org/abs/2101.10423.

[36] Adam Paszke et al. "PyTorch: An Imperative Style, High-Performance Deep Learning Library". In: *Advances in Neural Information Processing Systems 32*. Ed. by H. Wallach et al. Curran Associates, Inc., 2019, pp. 8024–8035. URL: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

[37] Jonas Pfeiffer et al. "MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, Nov. 2020, pp. 7654–7673. DOI: 10.18653/v1/2020.emnlp-main.617. URL: https://aclanthology.org/2020.emnlp-main.617.

[38] Lingfeng Qiao et al. "Private and common feature learning with adversarial network for RGBD object classification". In: *Neurocomputing* 423 (2021), pp. 190–199. DOI: 10.1016/j.neucom.2020.07.129. URL: https://doi.org/10.1016/j.neucom.2020.07.129.

[39] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, and Christoph H. Lampert. "iCaRL: Incremental Classifier and Representation Learning". In: *CoRR* abs/1611.07725 (2016). arXiv: 1611.07725. URL: http://arxiv.org/abs/1611.07725.

[40] Tal Ridnik et al. "ImageNet-21K Pretraining for the Masses". In: *CoRR* abs/2104.10972 (2021). arXiv: 2104.10972. URL: https://arxiv.org/abs/2104.10972.

[41] Guangyuan Shi et al. "Overcoming Catastrophic Forgetting in Incremental Few-Shot Learning by Finding Flat Minima". In: *CoRR* abs/2111.01549 (2021). arXiv: 2111.01549. URL: https://arxiv.org/abs/2111.01549.

[42] Chen Sun et al. "Revisiting Unreasonable Effectiveness of Data in Deep Learning Era". In: *CoRR* abs/1707.02968 (2017). arXiv: 1707.02968. URL: http://arxiv.org/abs/1707.02968.

[43] Shuai Tang et al. "Histogram of Oriented Normal Vectors for Object Recognition with a Depth Sensor". In: *Computer Vision – ACCV 2012*. Springer Berlin Heidelberg, 2013, pp. 525–538. DOI: 10.1007/978-3-642-37444-9_41. URL: https://doi.org/10.1007%2F978-3-642-37444-9_41.

[44] Hugo Touvron et al. "Training data-efficient image transformers & distillation through attention". In: *CoRR* abs/2012.12877 (2020). arXiv: 2012.12877. URL: https://arxiv.org/abs/2012.12877.

[45] Anran Wang et al. "Large-Margin Multi-Modal Deep Learning for RGB-D Object Recognition". In: *IEEE Transactions on Multimedia* 17.11 (2015), pp. 1887–1898. DOI: 10.1109/TMM.2015.2476655.

[46] Anran Wang et al. "MMSS: Multi-modal Sharable and Specific Feature Learning for RGB-D Object Recognition". In: *2015 IEEE International Conference on Computer Vision (ICCV)* (2015), pp. 1125–1133.

[47] Zhen Wang et al. *Online Continual Learning with Contrastive Vision Transformer*. 2022. DOI: 10.48550/ARXIV.2207.13516. URL: https://arxiv.org/abs/2207.13516.

[48] Thomas Wolf et al. "Transformers: State-of-the-Art Natural Language Processing". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Online: Association for Computational Linguistics, Oct. 2020, pp. 38–45. URL: https://www.aclweb.org/anthology/2020.emnlp-demos.6.

[49] Hasan F. Zaki, Faisal Shafait, and Ajmal Mian. "Viewpoint Invariant Semantic Object and Scene Categorization with RGB-D Sensors". In: *Auton. Robots* 43.4 (Apr. 2019), pp. 1005–1022. ISSN: 0929-5593. DOI: 10.1007/s10514-018-9776-8. URL: https://doi.org/10.1007/s10514-018-9776-8.

[50] Hanbin Zhao et al. “Memory Efficient Class-Incremental Learning for Image Classification”. In: *IEEE Transactions on Neural Networks and Learning Systems* (2021), pp. 1–12. DOI: 10.1109/TNNLS.2021.3072041.

[51] Fei Zhu et al. “Class-Incremental Learning via Dual Augmentation”. In: *Advances in Neural Information Processing Systems*. Ed. by M. Ranzato et al. Vol. 34. Curran Associates, Inc., 2021, pp. 14306–14318. URL: https://proceedings.neurips.cc/paper/2021/file/77ee3bc58ce560b86c2b59363281e914-Paper.pdf.

[52] Saman Zia, Yücel Yemez, and Deniz Yuret. “RGB-D Object Recognition Using Deep Convolutional Neural Networks”. In: *The IEEE International Conference on Computer Vision (ICCV)*. Oct. 2017, pp. 896–903. URL: /bib/zia/zia2017rgbd/rgb-d-object.pdf,http://openaccess.thecvf.com/content%5C_ICCV%5C_2017%5C_workshops/w17/html/Zia%5C_RGB-D%5C_Object%5C_Recognition%5C_ICCV%5C_2017%5C_paper.html.

[53] Saman Zia et al. “RGB-D Object Recognition Using Deep Convolutional Neural Networks”. In: *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*. 2017, pp. 887–894. DOI: 10.1109/ICCVW.2017.109.