Enhancing Semantic Segmentation with Continual Self-Supervised Pre-training

Anonymous authors
Paper under double-blind review

Abstract

Self-supervised learning (SSL) has emerged as a central paradigm for training foundation models by leveraging large-scale unlabeled datasets, often producing representations with strong generalization capabilities. These models are typically pre-trained on general-purpose datasets such as ImageNet and subsequently adapted to various downstream tasks through finetuning. While recent advances have explored parameter-efficient strategies for adapting pre-trained models, extending SSL pre-training itself to new domains—particularly under limited data regimes and for dense prediction tasks—remains underexplored. In this work, we address the problem of adapting vision foundation models to new domains in an unsupervised and data-efficient manner, specifically targeting downstream semantic segmentation. We propose GLARE (Global Local and Regional Enforcement), a novel continual self-supervised pre-training task designed to enhance downstream segmentation performance. GLARE introduces patch-level augmentations to encourage local consistency and incorporates a regional consistency constraint that leverages spatial semantics in the data. For efficient continual pre-training, we initialize Vision Transformers (ViTs) with weights from existing SSL models and update only lightweight adapter modules—specifically UniAdapter—while keeping the rest of the backbone frozen. Experiments across multiple semantic segmentation benchmarks on different domains demonstrate that GLARE consistently improves downstream performance with minimal computational and parameter overhead.

1 Introduction

Self-supervised learning (SSL) has revolutionized the training of foundation models by enabling the extraction of rich, generalizable features from vast unlabeled datasets (Caron et al., 2021; Chen et al., 2021; Oquab et al., 2024; Assran et al., 2023). This paradigm has proven to be particularly valuable in computer vision, where the abundance of unlabeled images can be leveraged to learn robust visual representations without the need for expensive manual annotations. Recent advances in SSL have introduced sophisticated techniques, spanning from innovative data augmentation strategies (Grill et al., 2020) and masked image modeling (He et al., 2022) to enhanced feature matching through self-attention mechanisms (Su et al., 2023; Su & Ji, 2024) and refined loss functions.

While these developments have led to powerful models pre-trained on extensive generic datasets like ImageNet (Deng et al., 2009), they often fall short when confronted with specialized technical domains. This limitation becomes particularly apparent in real-world scenarios where domain-specific data is scarce, especially labeled data. The challenge is further compounded when considering dense prediction tasks like semantic segmentation, which demand more fine-grained semantic understanding compared to classification tasks. These issues have therefore sparked significant interest in adaptation techniques, particularly finetuning and continual learning to bridge the gap between generic pre-training and domain-specific applications.

Our research focuses on enhancing foundation model features for task-specific limited data scenarios through continual pre-training within a pure SSL framework. While previous research has explored this direction by training batch normalization layers using conventional SSL approaches for classification tasks (Reed et al., 2022), these studies do not fully investigate other aspects of the SSL paradigm, such as augmentations and

matching, in limited data scenarios. Furthermore, while extensive research exists on continual pre-training for classification datasets (Lin et al., 2022; Cheng et al., 2023; Reed et al., 2022; Tang et al., 2024), the distinct challenges posed by semantic segmentation warrant separate investigation, as features of foundation models often perform differently between classification and segmentation tasks (Su & Ji, 2024; Caron et al., 2021; Oquab et al., 2024).

Therefore, we focus our work on identifying an appropriate data-efficient technique for the continual SSL pre-training of foundation models with limited unlabeled data, tailored for downstream semantic segmentation. Our key contributions include:

- We explore the benefits of SSL-based continual pre-training for semantic segmentation under limited data conditions. To this aim, we employ a trainable adapter (Lu et al., 2023) after the self-attention layers in ViTs, in order to mitigate catastrophic forgetting during continual SSL pre-training.
- We propose a data-efficient augmentation strategy centered on patch-wise strong blurring, instead of the typical patch-wise or block-wise masking approaches. This seemingly simple modification proves particularly effective in facilitating the learning of inter-patch relationships, especially in limited data settings.
- We propose explicit local and regional consistency enforcement to improve learning dense features with limited data. These constraints help to learn more distinct spatial features crucial for segmentation tasks. Specifically, we propose a regional consistency mechanism by penalizing feature inconsistency between the student (base encoder) and teacher (momentum encoder) within spatial patch groups, processed through a cross-attention layer. In this context, we sample the region features based on attention maps to leverage the semantic content already learned by the original model. Alongside the local patch-wise consistency, the regional consistency helps in learning better spatially distinct features suitable for semantic segmentation.

We provide model-specific experiments with existing state-of-the-art foundation models, while comparing our proposed SSL with standard SSL approaches. We validate our approach through experiments across four datasets: ADE20k (Zhou et al., 2017), Pascal Context (Mottaghi et al., 2014), Cityscapes (Cordts et al., 2016) and LoveDA (Wang et al., 2021a) (satellite images), showing performance improvements on downstream segmentation task after applying the proposed SSL approach in continual pre-training.

2 Related Works

2.1 SSL for global image understanding

SSL methods for vision aim to learn rich visual representations from large-scale unlabeled data by enforcing various learning objectives and paradigms. One prominent paradigm is joint embedding networks, exemplified by methods like MoCo (He et al., 2020) and DINO (Caron et al., 2021). MoCo employs contrastive learning with a memory bank for gathering negative examples (He et al., 2020; Wu et al., 2018), while later approaches, such as SimCLR (Chen et al., 2020b) and MoCo-V3 (Chen et al., 2021), eliminate the memory bank and use training batch samples as negatives. In contrast, methods like DINO (Caron et al., 2021), BYOL (Grill et al., 2020), and SimSiam (Chen & He, 2021) forego negative pairs altogether, focusing instead on positive pair similarity. Techniques for achieving this include contrastive learning (e.g., ZeroCL (Zhang et al., 2022)), clustering (e.g., SwAV (Caron et al., 2020), SeLa (Asano et al., 2020), MSN (Assran et al., 2022)), and alternative objectives such as redundancy reduction in Barlow Twins (Zbontar et al., 2021).

Another paradigm, inspired by Natural Language Processing (NLP) (Radford et al., 2018; Devlin et al., 2019), is generative SSL. Examples include iGPT (Chen et al., 2020a), which reconstructs masked pixels, and vision-specific approaches like BEiT (Bao et al., 2022) and MAE (He et al., 2022), which reconstruct masked patches. More recently, I-JePa (Assran et al., 2023) advanced this concept by predicting embeddings of masked patches using contextual and target encoders. Hybrid approaches, such as iBoT (Zhou et al., 2022), DINOv2 (Oquab et al., 2024), combine joint embedding learning with masked image modeling techniques like those in MAE (He et al., 2022).

2.2 SSL for dense prediction

SSL methods tailored at dense prediction tasks, including semantic segmentation, have gained special attention in recent years. These methods can be grouped into three main categories:

Learning local features This group emphasizes learning representations at the pixel or patch level. Methods like PixPro (Xie et al., 2021c) and DenseCL (Wang et al., 2021b) employ contrastive learning to enforce consistency between pixel representations across different views. DetCo (Xie et al., 2021a) integrates instance-patch and patch-level contrastive losses, achieving strong object detection performance without compromising image classification results. LOCA (Caron et al., 2024) clusters matching patch features between query and reference views, encouraging consistent distributions of patch-level clusters. By back-tracking random augmentations (Pinheiro et al., 2020), LOCA establishes patch correspondences. In our approach, we adopt a similar strategy for local consistency, but avoid LOCA's constraints of smaller query images relative to the reference, allowing for richer information. Furthermore, instead of LOCA's clustering approach, we enforce an inter-view local consistency using a DINO-like objective.

Enhancing spatial awareness The second category focuses on providing location awareness during pre-training. LOCA (Caron et al., 2024) predicts the positions of query patches within a reference image. Similarly, UP-DETR (Dai et al., 2022) extends the DETR (Carion et al., 2020) framework to localize random patches. Other approaches solve spatial reasoning tasks, such as rearranging jigsaw puzzles (Zhai et al., 2022) or identifying incorrect patch positions (Sameni et al., 2023). More recently, ADCLR (Zhang et al., 2023b) introduces query patch tokens, treating cropped image regions as additional class tokens, enhancing spatial awareness during pre-training.

Maximizing regional or object-level similarity This group of methods focuses on similarity within regions or at the object level. ReSim (Xiao et al., 2021) aligns representations of overlapping sliding window regions across views. At the object level, methods like SoCo (Wei et al., 2021), ORL (Xie et al., 2021b), and SCRL (Roh et al., 2021) employ techniques such as selective search which are very expensive. SelfPatch (Yun et al., 2022) defines regions as the set of patches in the direct neighborhood of a central patch and enforces similarity between these neighbors. FLSL (Su et al., 2023) on the other hand introduces intra- and inter-view clustering, attracting representations of the same concept while repelling clusters of different concepts across augmentations. Finally, the most recent work UDI (Su & Ji, 2024) encourages multimodal local predicitions by adding an additional class token and solve the semantic misalignment problem.

In this work, we use UDI pre-trained model to initialize our backbones for continual pre-training as UDI (Su & Ji, 2024) showcases strong results in SSL for dense prediction tasks. We then leverage the best performing model and show consistent improvements using our pre-training strategy in continual pre-training for semantic segmentation.

2.3 Continual pre-training

Recent works have explored unsupervised continual pre-training as a means to adapt large-scale pre-trained models to new domains. These approaches aim to bridge the gap between general-purpose pre-training and downstream domain-specific tasks by refining representations learned on source domains. Hierarchical Pre-Training (HPT) (Reed et al., 2022) introduces a framework for self-supervised continual pre-training, where a model trained on source domain data is further pre-trained on target domain data by finetuning all model weights in a sequential framework. Other approaches adapts ViTs using masked image modeling on target domain (Mendieta et al., 2023). More recent efforts have explored parameter-efficient adaptation strategies, such as incorporating lightweight modules like adapters (e.g., LoRA) to reduce compute cost (Scheibenreif et al., 2024; Khanna et al., 2025). While these methods demonstrate the promise of continual pre-training, they are primarily developed for downstream classification tasks and rely on large-scale datasets (typically exceeding 100k images). In contrast, our work focuses on designing a continual self-supervised pre-training pipeline tailored to dense prediction tasks, specifically semantic segmentation, in data-scarce target domains.

3 Preliminaries and Problem Formulation

In this section, we describe the learning premise with detailed description of the tools and the problem formulation.

3.1 Vision Transformers

Let $X \in \mathbb{R}^{C\times H\times W}$ be an image, where H and W represent the height and width of the image, respectively, and C the number of channels. A Vision Transformer model (ViT) (Dosovitskiy et al., 2020) considers the image x as a set of N non overlapping patches $x_i \in \mathbb{R}^{CP^2}$, of resolution $P \times P$ and with C channels. These patches are then projected through a linear layer to a space of dimension D, such that $z_i = Wx^{(i)} + E^i_{pos}$, where $W \in \mathbb{R}^{D\times CP^2}$ is a linear projection and $E^i_{pos} \in \mathbb{R}^D$ is the positional embedding for the patch at index i. A learnable token $z_{[CLS]} \in \mathbb{R}^D$, referenced as [CLS], is prepended to the sequence of patches to extract global information from the image. The resulting input sequence is thus defined as $z = [z_{[CLS]}, z_1, z_2, \ldots, z_N]$. Then, ViTs take the input e to produce global level ($e_{[CLS]}$) and patch level (e_i) representations by using its encoder. In the same way as in (Yun et al., 2022; Zhang et al., 2023b), we refer to the encoder as f_θ with parameters θ and we use Equation (2) to represent the whole process of a ViT:

$$f_\theta(x) = f_\theta([z_{[CLS]}, z_1, z_2, \ldots, z_N]) \tag{1}$$

$$= [f^{[CLS]}_\theta(x), f^{(1)}_\theta(x), f^{(2)}_\theta(x), \ldots, f^{(N)}_\theta(x)], \tag{2}$$

with $f^{[CLS]}_\theta(x)$ and $f^{(i)}_\theta(x)$ being the final representations of the global image token [CLS] and the i-th patch, respectively.

3.2 Continual pre-training with adapters

Our interest is to improve the SSL ViT features via a task-agnostic SSL framework in the continual learning setup. Therefore, we limit our work's investigation to finding the right SSL framework for continual learning rather than exploring different continual learning paradigms. Among the various continual learning approaches, adapter-based (Ding et al., 2022; Lu et al., 2023; Hu et al., 2022), and memory or replay based (Winter et al., 2024; Reed et al., 2022; Wang et al., 2024) strategies are considered to be state-of-the-art. However, due to their simplicity and compatibility with a wide range of architectures, adapter-based methods are widely used in continual learning (Scheibenreif et al., 2024; Khanna et al., 2025) and finetuning. We therefore consider a simple adapter named UniAdapter (Lu et al., 2023) composed of two linear layers with activation for adapting features after every self-attention layer in ViT. Therefore, given a pre-trained SSL model, our investigation involves training only the adapter layers while freezing the model weights, following the standard practice. Note that, by using the parameter-efficient adapter used typically for finetuning models, our problem formulation departs from previously used SSL continual learning, for which all network layers are trained (Reed et al., 2022). Our choice is mainly motivated by the need to isolate natural catastrophic forgetting from the suitability of the SSL framework (see Table 2).

Specifically, in this work we adopt an adapter with a down-projection layer $W_{down} \in \mathbb{R}^{D\times r}$, a nonlinear activation function σ (notably ReLu (Agarap, 2018)), and an up-projection layer $W_{up} \in \mathbb{R}^{r\times D}$, where D and r are the embedding and bottleneck dimensions, respectively. The adapter blocks are employed after each attention layer. More specifically, with x being the output of a ViT-block, we have the corresponding output of the adapter defined as:

$$x' = \text{Adapter}(x) = x + s\ \sigma(xW_{down})W_{up}, \tag{3}$$

where $s > 0$ is a scaling factor.

3.3 Problem formulation

Given an encoder trained via SSL (Su & Ji, 2024; Caron et al., 2021), we are interested in improving the output feature embedding e by training only the adapter parameters θ_A via SSL. Thus we write the new ViT parameters as θ_A, resulting in a different feature embedding e' given an input image. Here θ_A denotes

the trainable parameters of the Adapter layers. Given a new target dataset $\mathcal{T}$, our problem is to study and develop the most suitable SSL framework that maximizes the performance of e'. We measure the performance of the feature embedding e' by finetuning our model for semantic segmentation together with an additional decoder. In other words, we use semantic segmentation finetuning as the representative metric for gauging feature embeddings e'.

4 Method: GLARE pre-training

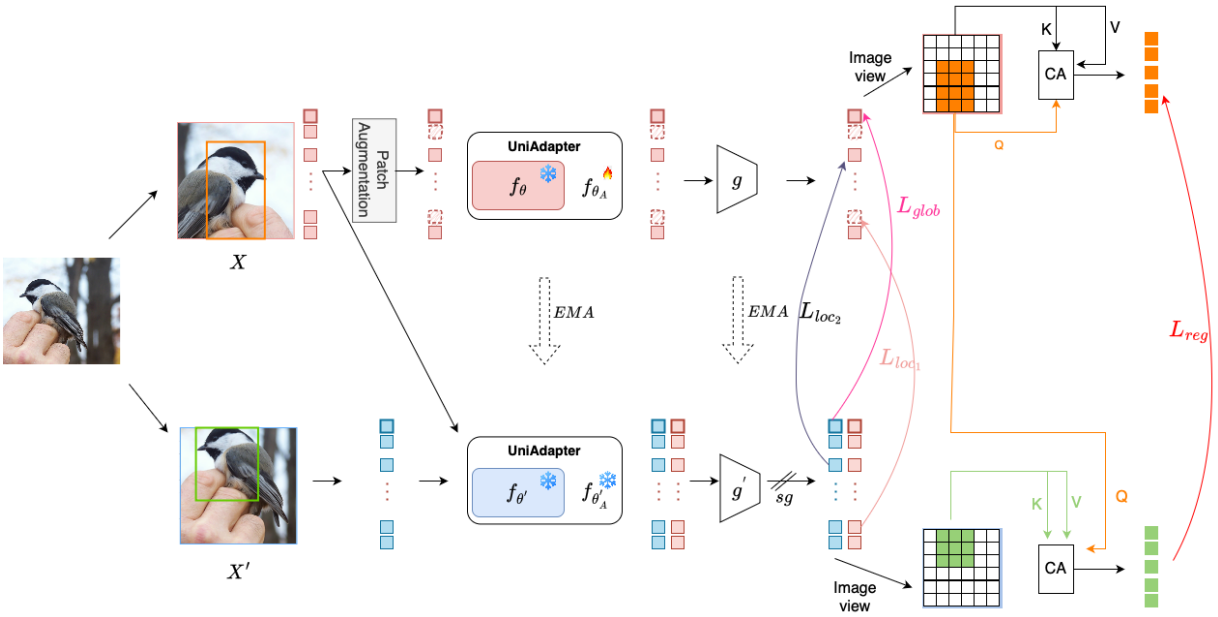

Figure 1: Overview of GLARE continual pre-training framework. Given an image, two views X and X' are generated with image-level augmentation. Each view goes through the base and momentum encoders f_θ, θ_A and f_{θ',θ'_A}. All the parameters are frozen except for those of the adapter on the base encoder. GLARE applies three levels of feature consistency during the pre-training. Firstly, global consistency is considered on [CLS] tokens (L_{glob}) of the two views Section 4.1. Secondly, regional consistency is applied on sampled regions, with their representations obtained using a cross-attention module to calculate L_{reg} Section 4.2. Finally, we enforce local consistency focusing on patch-augmentation consistency with distorted vs. not distorted patches of the same view (L_{loc_1}) and inter-view local consistency on matching patches from the two views (L_{loc_2}). Section 4.3

In this section, we describe our method for pre-training the model parameters θ_A on the target dataset $\mathcal{T}$ using our proposed SSL strategy.

GLARE (Global Local and Regional Enforcement) is an SSL framework focused on learning representations at different levels during the pre-training process. When we as humans look at a picture, in practice we focus our attention to varying levels of detail (Navon, 1977; Shi et al., 2014). We can summarize these levels as 1) the image as a *whole*, 2) as a set of *regions/objects*, and 3) as a collection of specific details on these regions, which can be encoded into *patches* or *pixels*. GLARE, inspired by this concept, is designed as a three-level pre-training strategy, combining the enforcement of **global**, **regional**, and **local** consistency for learning coarse to fine-grained representations, which are crucial to develop a more in-depth understanding of image data. We adopt the usual student-teacher framework for self-supervised learning described in Caron et al. (2021); He et al. (2020); Chen et al. (2021); Yun et al. (2022); Zhang et al. (2023b). Following the same naming convention as in Chen et al. (2021), we consider two ViT encoders: a *base encoder* f_{θ,θ_A} and a *momentum encoder* f_{θ',θ'_A}, parameterized by (θ, θ_A) and (θ', θ'_A), respectively. The parameters related to the momentum encoder (θ', θ'_A) are updated through an exponential moving average (EMA) of those of the base encoder (θ, θ_A), as in He et al. (2020). Figure 1 describes GLARE pre-training strategy. Specifically, in our

continual pre-training setup, we train only the adapter parameters θ_A of the base encoder and update those of the momentum encoder θ'_A through EMA as stated in Section 3.3. In the following sections, we describe more in detail the three levels of consistency enforced.

4.1 Global feature consistency

This pre-training objective focuses on understanding the *overall picture* of an image, also referred to as image-level understanding. Early pre-training methods such as Caron et al. (2021); He et al. (2020); Chen et al. (2021) have tackled this problem using different techniques. The common idea is to have the model learn representations that are invariant to transformations on the image level by maximizing the similarity of representations between augmented views of the same image.

Given an image I, a positive pair of views (X, X') is generated by applying random augmentations. In this work, we consider enforcing global consistency by maximizing the similarity of the representations of this positive pair, specifically employing the DINO loss (Caron et al., 2021):

$$L_{glob} = H\left(g_\lambda\left(f^{[CLS]}_{\theta,\theta_A}(x)\right), sg\left(g_{\lambda'}\left(f^{[CLS]}_{\theta',\theta'_A}(x)\right)\right)\right) \tag{4}$$

where $H(a, b) = -a \log b$ is the cross-entropy loss, $sg(\cdot)$ is the stop gradient operation, and g_λ is an MLP projection head commonly used in most SSL methods (Caron et al., 2021; Chen et al., 2021; Yun et al., 2022; Zhang et al., 2023b; Su et al., 2023). The parameters λ' and θ'_A are updated with an exponential moving average of λ and θ_A.

4.2 Regional level consistency

The next pre-training level involves learning region/object representations: we want the model to extract semantic information from regions, whether them being some specific objects (animal, person, etc.) or the background. We aim to enforce the consistency of the representations between two correspondent regions of two views that contain the same semantics. In the context of self-supervised learning, we do not have access to any explicit annotation such as bounding boxes or segmentation masks. We therefore approximate candidate regions through *sampling operations*.

Region Sampling This method refers to providing region proposals, similarly to what was done in early object detection models like R-CNN (Girshick et al., 2014), Fast-RCNN (Girshick, 2015), which use selective search to find candidate regions. This process is quite expensive, especially in a self-supervised learning scenario. For that reason, we consider two strategies:

- **Random sampling**: for each region, we randomly sample a starting patch and a number of rows and columns of patches corresponding to the size of our candidate region within an interval $[min_p, max_p]$, where min_p, max_p define the minimum and maximum of patches to consider.

- **Attention–aware region sampling**: the attention map of a block of a self-supervised ViT encoder often contain several insights on an image. In fact, over the different heads of the last block, the attention is directed towards different regions, as presented in Figure 2. Consequently, to enforce more semantically-rich regions, we use the attention from the different heads of the encoder to generate the starting patch for candidate regions. In this case, a starting patch is the one getting the most attention on a specific head. From the starting patch, we define the region using a similar process as in *random sampling*. This is feasible primarily in the context of continual pre-training, as it leverages publicly available pre-trained models that already exhibit a useful signal for attention-aware region sampling—something not achievable when training models from scratch.

All results presented in this work employ attention-aware region sampling, as we determined it to be more effective. For more details, see the supplementary material.

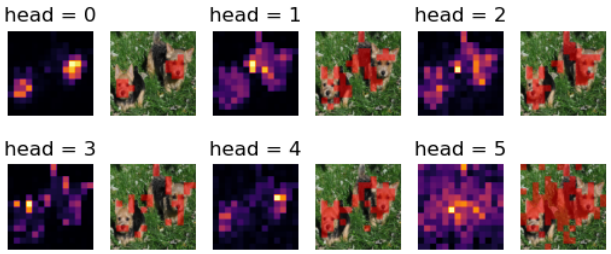

Figure 2: Attention map of the last block of a DINO (Caron et al., 2021) pre-trained model over different heads on an image. The different heads have their attention directed toward specific regions in the image. Some heads focus more on the dog on the left, others on the dog on the right and also on the background.

Region correspondence Let R be a candidate sample region from the view X of the student network, with z_r being the representation of a patch in R. By back-tracking the cropping augmentations, we can find the correspondent region R' on the view X' of the teacher network, with patch representations defined as z'_r. To encourage the model to learn region-semantic information which aligns with the context, we first extract the semantic context of the query region R with respect to the view X, through a cross-attention module:

$$\tilde{z}_r = \text{CA}(z_r, Z, Z) = (W_v Z)\text{softmax}(\tau(W_k Z)^T(W_q z_r)), \tag{5}$$

where W_q, W_k, and W_v are learnable matrices of the cross-attention module, Z the representation of the view X and τ a scaling factor. We then proceed by enabling semantics sharing which extracts the semantics from the correspondent region R' through the query region R, by using the same cross-attention module. This helps the model to extract only relevant information from R', since we know from construction that there are inherent differences between R and R'. We obtain the new representation defined as:

$$\tilde{z}'_r = \text{CA}(z'_r, Z_R, R) = (W_v Z_R)\text{softmax}(\tau(W_k Z_R)^T(W_q z'_r)), \tag{6}$$

where Z_R is the representation of the region R. Hence, region consistency is enforced with the following loss function applied on the obtained representations $\tilde{z}_r$ and $\tilde{z}'_r$:

$$L_{\text{reg}} = \sum_{\tilde{z}_r \in Z_R, \tilde{z}'_r \in Z_{R'}} H\left(g_\lambda\left(\tilde{z}_r\right), sg\left(g_{\lambda'}\left(\tilde{z}'_r\right)\right)\right). \tag{7}$$

4.3 Local consistency

The final level of pre-training that we consider is the *local consistency*. Under limited data, there are not enough examples to guide SSL to preserve consistency of features between corresponding patches in the base and momentum encoder. We propose and combine two approaches to encourage local consistency. First, we apply random patch blurring, to encourage local patch features to be consistent to each other in the single base encoder view. We call this *patch-augmentation consistency*. Additionally, for all patch features, we also enforce local consistency between the base and momentum encoder. We call this *inter-view consistency*. Note that patch features are the smallest spatial distinction of features in a ViT model. The goal is to extract local semantics across views and between patches within a view to enhance the ability of the model to capture smaller details.

Patch-augmentation consistency Firstly, we aim to alleviate the problem of limited data via separate patch-level augmentations for a fraction of the total image patches. Specifically, we consider strong blurring on 30% of random patches on the base encoder views. This process creates incomplete information on the augmented patches. However, unlike the standard practice in Masked Image Modeling (MIM) (Zhou et al., 2022; Zhang et al., 2023a; Oquab et al., 2024), we do not feed empty tokens for the blurred patches. Thus, we expect the following: *i)* the attention between the local patches in the ViT model are used to "complete" the local patch feature and *ii)* additional scale robustness is imposed on the image patch feature via inter-patch

consistency. The main reason for not using MIM in this case, especially for out-of-domain datasets, is that it is known to be data-intensive, requiring a large number of samples and training time to learn meaningful representations (He et al., 2022). This challenge becomes even more pronounced with new domains that exhibit high variability, unlike standard datasets like ImageNet (Deng et al., 2009), and with few samples as typically encountered in continual pre-training. We also investigate other patch-level augmentations such as *rotation* and *noising*, and blurring turns out to be the best-performing approach.

Once, a patch-level augmentation has been defined, we randomly apply it to a set of patches of the view X and we distill the knowledge from the non-distorted patches using the same objective L_{loc_1} as in Zhou et al. (2022); Oquab et al. (2024). Let X_m be a distorted version of the view X:

$$L_{loc_1} = \sum_{mk} H\left(g_\lambda\left(f^{[mk]}_{\theta,\theta_A}(x_m)\right), sg\left(g_{\lambda'}\left(f^{[mk]}_{\theta',\theta'_A}(x)\right)\right)\right), \tag{8}$$

with $f^{[mk]}_{\theta,\theta_A}$ corresponding to the mask patches.

Inter-view local consistency Secondly, we learn local semantic information across different views of an image through correspondence. Similarly, to what is done in Section 4.2, we apply a matching algorithm that back-tracks the augmentation process to find correspondence between a patch from the student view X, and the ones in the teacher view X'. We then proceed by enforcing the consistency between the correspondent patches in the two views X and X'. Let $x^{(s)}$ be a patch of the student view X and $C(x^{(s)}) = \{t \in h(x^{(s)} \mid X')\}$ the set of indices of the correspondent patches of $x^{(s)}$ in the teacher view X', with $h(x^{(s)} \mid X')$ being a function which maps a patch in student view to the ones in the teacher by providing the correspondent indices. We can then write the loss function for a given $x^{(s)}$ as:

$$L^{(s)}_{loc_2} = \sum_{t \in C(x^{(s)})} H\left(g_\lambda\left(f^{(s)}_{\theta,\theta_A}(x)\right), sg\left(g_{\lambda'}\left(f^{(t)}_{\theta',\theta'_A}(x')\right)\right)\right), \tag{9}$$

and we then have $L_{loc_2} = \sum_{x^{(s)}} L^{(s)}_{loc_2}$.

Therefore, the local level consistency loss is then defined by $L_{loc} = L_{loc_1} + L_{loc_2}$. Finally, the overall GLARE objective including all pre-training levels is then defined as:

$$L = L_{\text{glob}} + L_{\text{reg}} + L_{\text{loc}}. \tag{10}$$

5 Experiment Setup

We use ViT-S/16 (Dosovitskiy et al., 2020) for our experiments due to its cost-efficiency balance between training cost and performance, as in Su & Ji (2024). We apply continual pre-training for 100 epochs on the target dataset, after observing lower performance when training longer. We also used a register token in our ViT encoder as in Darcet et al. (2024); Su & Ji (2024). We use a batch size of 512, and a shared projection head across the different pre-training levels as done in Su & Ji (2024); Zhang et al. (2023b); Zhou et al. (2022) with the output dimension of $K = 8192$. The learning rate is linearly increased for the first epoch to its base value calculated using the linear scaling rule in Chen et al. (2020b), which is $lr = 1.5 \cdot 10^{-4}$. After warmup, the learning rate is decreased using a cosine scheduler (Loshchilov & Hutter, 2022). The weight decay is set to 0.1 and we use AdamW optimizer (Loshchilov & Hutter, 2017). We follow the data augmentations of BYOL (Grill et al., 2020) which was also used in Caron et al. (2021) (i.e. color jittering, gaussian blur, solarization) on random resized crops. Specifically, given an input image we generate 2 global views of resolution 224×224 and 10 local views of resolution 96×96 as in DINO (Caron et al., 2021).

We compare GLARE with SOTA SSL methods described in Su et al. (2023); Su & Ji (2024) on ViT-S by analyzing the performances of these pre-training methods against GLARE when they are used from scratch or in a continual pre-training setup. In this work, we initialize our model using the pre-trained weights from UDI (Su & Ji, 2024), which currently represents the state-of-the-art in leveraging self-supervised learning for downstream segmentation tasks. We then train only the parameters of the adapter layer as shown in Section 3.2.

All our experiments are performed employing 8 NVIDIA® A100 GPUs, with 80 GB of memory each.

6 Main Results

In the following, we use the notation $ssl_A \rightarrow ssl_B$ to present the results of our experiments. This means that starting from the pre-trained model obtained using the pre-training strategy ssl_A on ImageNet-1k (Deng et al., 2009), we continue the pre-training using the strategy ssl_B on a specific target dataset.

6.1 Semantic segmentation performance on different benchmarks

In Table 1, we report the performance on semantic segmentation of models obtained by continual pre-training with different methods starting from UDI weights (Su & Ji, 2024). We also include the performance when the model is randomly initialized without any pre-training knowledge. We report the average mIoU scores and standard deviation after finetuning 3× the pre-trained models on semantic segmentation using FPN (Kirillov et al., 2019). We consider 3 classes of datasets, namely *general domain*: ADE20k (Zhou et al., 2017), Pascal Context (Mottaghi et al., 2014), *driving*: Cityscapes (Cordts et al., 2016) and *aerial*: LoveDA (Wang et al., 2021a), which contain respectively $20k$, 4998, 2975 and 2522 images in their training set, making them suitable for continual pre-training in limited data scenarios ($<= 20k$ images). For each experiment, we first initialize the weights of the backbone with those of the starting model (UDI), then we do continual SSL pre-training with an adapter (UniAdapter) on the target dataset by training only the adapter parameters. The obtained model is then used for finetuning for segmentation. We observe that GLARE reports consistent improvement for continual pre-training from UDI weights over all the datasets. On ADE20k, we obtain an improvement of +0.4 over UDI and on LoveDA, which is an out-of-domain dataset with respect to the original pre-training dataset of UDI (ImageNet (Deng et al., 2009)), we achieve an improvement of +0.6. This shows that GLARE is able to take advantage of existing encoded features and new data distribution to improve semantic understanding, which transfer in semantic segmentation.

Table 1: Comparison of SSL pre-trained models and continual pre-trained models starting from UDI (Su & Ji, 2024) on 4 semantic segmentation benchmarks. We report mIoU on the validation sets. We use the FPN (Kirillov et al., 2019) framework with 20k iterations and 2k for LoveDA. GLARE continual pre-training from UDI consistently shows improvements over the other pre-training strategies.

Method	Backbone	ADE20k	P.Cont	Cityscapes	LoveDA
Random Init.	ViT-S/16	10.0 (± 0.03)	19.0 (± 0.00)	42.4 (± 0.25)	29.8 (± 0.04)
UDI (Su & Ji, 2024)	ViT-S/16	41.2 (± 0.11)	49.1 (± 0.04)	74.7 (± 0.01)	50.9 (± 0.02)
UDI → UDI	ViT-S/16	41.1 (± 0.11)	49.2 (± 0.04)	74.9 (± 0.17)	51.1 (± 0.01)
UDI → FLSL	ViT-S/16	41.2 (± 0.06)	48.7 (± 0.04)	74.2 (± 0.28)	49.9 (± 0.12)
UDI → GLARE	ViT-S/16	**41.6** (± 0.13)	**49.3** (± 0.01)	**75.3** (± 0.03)	**51.5** (± 0.01)

6.2 Comparison with other continual pre-training methods

In this section, we compare two different continual pre-training methods: Hierarchical Pre-Training (HPT) (Reed et al., 2022) and our adapter-based strategy, based on UniAdapter (Lu et al., 2023). We continue the pre-training using these strategies, and then we evaluate the quality of the new features by finetuning on semantic segmentation. We consider LoveDA (Wang et al., 2021a) for this experiment as its out-of-domain nature can help better assess the quality of the strategy. Table 2 presents the results on three metrics: (a) mean intersection over union (mIoU) averaged over all semantic categories, (b) all pixel accuracy (aAcc), and (c) mean class accuracy (mAcc). We observe that overall the adapter-based strategy provides a better improvement for continual SSL compared to HPT. In fact, HPT often results in performance degradation in semantic segmentation, in contrast to what is usually observed for classification tasks.

6.3 Influence of the SSL backbone

An important question is the role of the SSL backbone used for the continual pre-training of GLARE. Specifically, we want to examine whether the performance gains observed when starting from UDI (Su &

Table 2: Comparison of different continual pre-training framework. Experiments performed starting from UDI (Su & Ji, 2024) pre-trained on ImageNet (Deng et al., 2009). We show the segmentation performance of continual pre-trained models on LoveDA with a finetuning for 2k iterations.

Framework	Pre-training	mIoU	aAcc	mAcc
-	UDI (Su & Ji, 2024)	50.9	70.0	63.8
HPT	UDI → UDI	50.4	69.5	62.1
	UDI → FLSL	50.9	70.1	63.1
	UDI → GLARE	12.7	41.7	23.0
UniAdapter	UDI → UDI	51.1	70.1	63.9
	UDI → FLSL	49.0	68.5	60.9
	UDI → GLARE	**51.5**	**70.2**	**64.3**

Ji, 2024) also hold when initializing from other SSL backbones, in particular FLSL (Su et al., 2023) and DINO (Caron et al., 2021), all pre-trained on ImageNet-1k (Deng et al., 2009). To investigate this, Table 3 reports the results of continual pre-training with GLARE starting from these different backbones, using both HPT (Reed et al., 2022) and UniAdapter (Lu et al., 2023) frameworks. We use LoveDA (Wang et al., 2021a) as the target domain and evaluate performance after fine-tuning for semantic segmentation. We observe that GLARE consistently improves upon the original models. Moreover, starting from a stronger backbone (e.g., UDI) leads to larger gains. We further hypothesize that the original pre-training strategy influences how effectively GLARE can leverage a given backbone.

Table 3: Segmentation performance (mIoU) on LoveDA using different SSL backbones with and without continual pre-training. We compare original backbones against continual pre-trained ones with (Reed et al., 2022) and UniAdapter (Lu et al., 2023), using the initial pre-training method vs GLARE. We show results after 2k finetuning iterations.

Init.	Cont. pre-training	Original	Framework	
			HPT(34M)	UniAdapter(12M)
DINO	→ DINO	50.3	28.7	28.4
	→ GLARE			**50.6**
FLSL	→ FLSL	50.3	50.2	**50.6**
	→ GLARE			50.5
UDI	→ UDI	50.9	50.4	51.1
	→ GLARE			**51.5**

6.4 Do the continual pre-trained models forget?

In this section, we investigate how much our continual adapter-based pre-trained models forget their previously learned knowledge. In particular, we consider GLARE continual pre-trained model as well as UDI continual pre-trained model on LoveDA. We compare their performances against the original performances of the UDI starting model on the datasets ADE20k (Zhou et al., 2017), Pascal Context (Mottaghi et al., 2014), Cityscapes (Cordts et al., 2016), LoveDA (Wang et al., 2021a). Table 4 reports the finetuning performances. We observe that instead of a decrease in performance relative to the original model, we maintain or outperform it. This suggests that continual pre-training using adapters helps the model to get and maintain more semantic insight from one dataset to another without degrading previous performances.

Table 4: Evaluation of forgetfulness of our continual pre-trained models. We report the mIoU of the finetuned models (using FPN (Kirillov et al., 2019)) on which we initially performed continual pre-training on LoveDA.

Pre-training	Ptr. Data	ADE20k	P.Cont	Citys.
UDI (Su & Ji, 2024)	ImageNet (Deng et al., 2009)	41.1	**49.2**	74.7
UDI → UDI	LoveDA (Wang et al., 2021a)	41.1	48.8	**75.2**
UDI → GLARE		**41.3**	49.0	75.0

6.5 Results in Classification

While our continual pre-training framework is primarily evaluated on semantic segmentation, it is not limited to this task. To demonstrate its broader applicability, we also apply GLARE continual pre-training to classification. Specifically, we experiment with two out-of-domain datasets: Derm7pt (Kawahara et al., 2019), a dermoscopic benchmark for skin lesion analysis containing ∼800 samples, and COVIDx (Wu et al., 2023), a large-scale x-ray benchmark for COVID-19 detection, where we subsample 20% of the data for continual pre-training and fine-tuning to maintain a low-data setting. Models are initialized either randomly, from a DINO (Caron et al., 2021) pre-trained backbone, or from GLARE continual pre-training starting from DINO, and finally finetuning of a classification head. We report in Table 5 the top-1 accuracy of models trained from scratch, initialized from DINO, or from DINO → GLARE. We observe that DINO → GLARE consistently outperforms the other initializations, demonstrating the effectiveness of GLARE continual pre-training even for tasks beyond semantic segmentation.

Table 5: Top-1 classification accuracy on Derm7pt (Kawahara et al., 2019) and COVIDx (Wu et al., 2023) datasets using a ViT-S backbone. We compare models trained from scratch, DINO (Caron et al., 2021) pre-training, and continual pre-training with DINO → GLARE.

Dataset	Pre-training	acc@1 (%)
Derm7pt	Random Init.	39.6
	DINO (Caron et al., 2021)	48.0
	DINO → GLARE	**49.1**
COVIDx	Random Init.	62.5
	DINO (Caron et al., 2021)	60.7
	DINO → GLARE	**69.2**

7 Ablation Study

In this section, we investigate the effect of the different components of GLARE and their contribution to its performance. We also provide additional ablations in the supplementary material. Unless stated otherwise, we report the finetuning results of UDI → GLARE on LoveDA using FPN (Kirillov et al., 2019).

7.1 Different levels of understanding in GLARE

Our work proposes a pre-training strategy operating at different level of details. In Table 6, we show how global, local, and region understanding interact with each other for downstream semantic segmentation. For this experiment, we run the continual pre-training on 20% of ADE20k and report the performance of the finetuned model when considering some or all of the objectives. We observe that combining all levels of details in the pre-training is crucial for the performance of the continual pre-trained model. In fact, GLARE obtains +0.8 compared to only global consistency pre-training.

Table 6: Impact of the different pre-training levels in GLARE. We report the mIoU of finetuned continual pre-trained models when trained with different levels of GLARE on 20% of ADE20k

Global	Regional	Local	mIoU
✓	-	-	40.9
✓	-	✓	41.1
✓	✓	-	41.5
-	✓	✓	41.1
✓	✓	✓	**41.7**

Table 7: Ablation of patch-level augmentations. We consider continual pre-training with single or combinations of different patch augmentations on LoveDA. We report the mIoU of the finetuned model on LoveDA using FPN (Kirillov et al., 2019)

Masking		Blurring		mIoU
random	block	random	block	
-	✓	-	-	50.9
✓	-	-	-	50.9
-	-	✓	-	**51.5**
-	-	-	✓	51.5
-	-	-	-	51.3

7.2 Influence of patch-level augmentations

A crucial aspect of GLARE is its ability to learn fine-grained details of the image during the pre-training through local consistency enforcement. As explained in Section 4.3, we introduce patch-level augmentation as a mean to increase local semantics during continual pre-training. In this section, we evaluate patch-level *masking* (typically used in iBoT (Zhou et al., 2022), DINOv2 (Oquab et al., 2024)) and *blurring*. Table 7 shows the results with the different augmentations when finetuned on LoveDA (Wang et al., 2021a). We experiment with block-wise vs random application of masking and blurring during the pre-training. We consider a prediction ratio r set as 0 with a probability of 0.5 and uniformly sampled from range $[0.1, 0.5]$ as in Zhou et al. (2022). We observe that applying *random blurring* provides the best result and we use it for our continual pre-training setup on LoveDA. We hypothesize that in low-data regimes, blurring serves as a more effective augmentation than stronger perturbations such as masking. Unlike aggressive masking, blurring retains essential information, allowing the model to learn from partially distorted patches while preserving semantic context.

8 Conclusion

In this work, we explore continual self-supervised learning, specifically for downstream semantic segmentation. While traditional SSL methods are effective for general-purpose pre-training, we find that they struggle to adapt to new domains when used for continual pre-training, particularly on out-of-domain datasets. To address this, we use an adapter for efficient knowledge transfer and propose GLARE, an SSL framework that learns representations at multiple levels: (i) global consistency at the image level, (ii) regional consistency via attention-based candidate regions, and (iii) local consistency through patch-wise augmentation and inter-view patch consistency. This multi-level approach equips the continual pre-trained model with semantically rich representations that improve transferability for segmentation. Experiments on diverse datasets, including both general and out-of-domain (satellite) images, demonstrate GLARE's effectiveness in continual pre-training for semantic segmentation. Our findings advance continual SSL for dense prediction tasks and offer practical insights for adapting foundation models to specialized domains.

References

Abien Fred Agarap. Deep learning using rectified linear units (relu). *arXiv preprint arXiv:1803.08375*, 2018. URL https://api.semanticscholar.org/CorpusID:4090379.

Yuki Markus Asano, Christian Rupprecht, and Andrea Vedaldi. Self-labelling via simultaneous clustering and representation learning. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=Hyx-jyBFPr.

Mahmoud Assran, Mathilde Caron, Ishan Misra, Piotr Bojanowski, Florian Bordes, Pascal Vincent, Armand Joulin, Michael Rabbat, and Nicolas Ballas. Masked siamese networks for label-efficient learning. In *European conference on computer vision*, pp. 456–473. Springer, 2022.

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15619–15629, 2023.

Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. BEit: BERT pre-training of image transformers. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=p-BhZSz59o4.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pp. 213–229. Springer, 2020.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in Neural Information Processing Systems*, 33:9912–9924, 2020.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jegou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9630–9640, 2021. doi: 10.1109/ICCV48922.2021.00951.

Mathilde Caron, Neil Houlsby, and Cordelia Schmid. Location-aware self-supervised transformers for semantic segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 117–127, 2024.

Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1691–1703. PMLR, 13–18 Jul 2020a. URL https://proceedings.mlr.press/v119/chen20s.html.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1597–1607. PMLR, 13–18 Jul 2020b. URL https://proceedings.mlr.press/v119/chen20j.html.

Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15745–15753, 2021. doi: 10.1109/CVPR46437.2021.01549.

Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9640–9649, 2021.

Haoyang Cheng, Haitao Wen, Xiaoliang Zhang, Heqian Qiu, Lanxiao Wang, and Hongliang Li. Contrastive continuity on augmentation stability rehearsal for continual self-supervised learning. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5684–5694, 2023. doi: 10.1109/ICCV51070.2023.00525.

Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3213–3223, 2016.

Zhigang Dai, Bolun Cai, Yugeng Lin, and Junying Chen. Unsupervised pre-training for detection transformers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–11, 2022. ISSN 1939-3539. doi: 10.1109/tpami.2022.3216514. URL http://dx.doi.org/10.1109/TPAMI.2022.3216514.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers, 2024. URL https://arxiv.org/abs/2309.16588.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423/.

Ning Ding, Yujia Qin, Guang Yang, Fuchao Wei, Zonghan Yang, Yusheng Su, Shengding Hu, Yulin Chen, Chi-Min Chan, Weize Chen, Jing Yi, Weilin Zhao, Xiaozhi Wang, Zhiyuan Liu, Hai-Tao Zheng, Jianfei Chen, Yang Liu, Jie Tang, Juanzi Li, and Maosong Sun. Delta tuning: A comprehensive study of parameter efficient methods for pre-trained language models. *arXiv preprint arXiv:2203.06904*, 2022. URL https://arxiv.org/abs/2203.06904.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020.

Ross Girshick. Fast R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1440–1448, 2015.

Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 580–587, 2014.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in Neural Information Processing Systems*, 33:21271–21284, 2020.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9726–9735, 2020. doi: 10.1109/CVPR42600.2020.00975.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16000–16009, 2022.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. LoRA: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

Jeremy Kawahara, Sara Daneshvar, Giuseppe Argenziano, and Ghassan Hamarneh. Seven-point checklist and skin lesion classification using multitask multimodal neural nets. *IEEE Journal of Biomedical and Health Informatics*, 23(2):538–546, 2019. doi: 10.1109/JBHI.2018.2824327.

Samar Khanna, Medhanie Irgau, David B. Lobell, and Stefano Ermon. Explora: Parameter-efficient extended pre-training to adapt vision transformers under domain shifts, 2025. URL https://arxiv.org/abs/2406.10973.

Alexander Kirillov, Ross Girshick, Kaiming He, and Piotr Dollár. Panoptic feature pyramid networks. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6399–6408, 2019.

Zhiwei Lin, Yongtao Wang, and Hongxiang Lin. Continual contrastive learning for image classification. In *2022 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6. IEEE, July 2022. doi: 10.1109/icme52920.2022.9859995. URL http://dx.doi.org/10.1109/ICME52920.2022.9859995.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations*, 2022.

Haoyu Lu, Yuqi Huo, Guoxing Yang, Zhiwu Lu, Wei Zhan, Masayoshi Tomizuka, and Mingyu Ding. UniAdapter: Unified parameter-efficient transfer learning for cross-modal modeling. *arXiv preprint arXiv:2302.06605*, 2023.

Matias Mendieta, Boran Han, Xingjian Shi, Yi Zhu, and Chen Chen. Towards geospatial foundation models via continual pretraining, 2023. URL https://arxiv.org/abs/2302.04476.

Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. The role of context for object detection and semantic segmentation in the wild. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 891–898, 2014. doi: 10.1109/CVPR.2014.119.

David Navon. Forest before trees: The precedence of global features in visual perception. *Cognitive Psychology*, 9(3):353–383, 1977. ISSN 0010-0285. doi: https://doi.org/10.1016/0010-0285(77)90012-3. URL https://www.sciencedirect.com/science/article/pii/0010028577900123.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research Journal*, pp. 1–31, 2024.

Pedro O. Pinheiro, Amjad Almahairi, Ryan Y. Benmalek, Florian Golemo, and Aaron Courville. Unsupervised learning of dense visual representations. *Advances in Neural Information Processing Systems*, 33:4489–4500, 2020.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. 2018.

Colorado J Reed, Xiangyu Yue, Ani Nrusimha, Sayna Ebrahimi, Vivek Vijaykumar, Richard Mao, Bo Li, Shanghang Zhang, Devin Guillory, Sean Metzger, et al. Self-supervised pretraining improves self-supervised pretraining. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2584–2594, 2022.

Byungseok Roh, Wuhyun Shin, Ildoo Kim, and Sungwoong Kim. Spatially consistent representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1144–1153, 2021.

Sepehr Sameni, Simon Jenni, and Paolo Favaro. Representation learning by detecting incorrect location embeddings. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 9704–9713, 2023.

Linus Scheibenreif, Michael Mommert, and Damian Borth. Parameter efficient self-supervised geospatial domain adaptation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27841–27851, 2024. doi: 10.1109/CVPR52733.2024.02630.

Tianlin Shi, Ming Liang, and Xiaolin Hu. A reverse hierarchy model for predicting eye fixations. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2822–2829, 2014. doi: 10.1109/CVPR.2014.361.

Qing Su and Shihao Ji. Unsqueeze [CLS] bottleneck to learn rich representations. In *European Conference on Computer Vision*, pp. 19–37. Springer, 2024.

Qing Su, Anton Netchaev, Hai Li, and Shihao Ji. FLSL: Feature-level self-supervised learning. *Advances in Neural Information Processing Systems*, 36:6568–6581, 2023.

Chi Ian Tang, Lorena Qendro, Dimitris Spathis, Fahim Kawsar, Cecilia Mascolo, and Akhil Mathur. Kaizen: Practical self-supervised continual learning with continual fine-tuning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2841–2850, 2024.

Junjue Wang, Zhuo Zheng, Ailong Ma, Xiaoyan Lu, and Yanfei Zhong. LoveDA: A remote sensing land-cover dataset for domain adaptive semantic segmentation. In J. Vanschoren and S. Yeung (eds.), *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1. Curran Associates, Inc., 2021a. URL `https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/file/4e732ced3463d06de0ca9a15b6153677-Paper-round2.pdf`.

Wenhao Wang, Muhammad Ahmad Kaleem, Adam Dziedzic, Michael Backes, Nicolas Papernot, and Franziska Boenisch. Memorization in self-supervised learning improves downstream generalization. In *International Conference on Learning Representations*, 2024.

Xinlong Wang, Rufeng Zhang, Chunhua Shen, Tao Kong, and Lei Li. Dense contrastive learning for self-supervised visual pre-training. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pp. 3024–3033, 2021b.

Fangyun Wei, Yue Gao, Zhirong Wu, Han Hu, and Stephen Lin. Aligning pretraining for detection via object-level contrastive learning. *Advances in Neural Information Processing Systems*, 34:22682–22694, 2021.

Philip Matthias Winter, Maria Wimmer, David Major, Dimitrios Lenis, Astrid Berg, Theresa Neubauer, Gaia Romana De Paolis, Johannes Novotny, Sophia Ulonska, and Katja Bühler. PARMESAN: Parameter-free memory search and transduction for dense prediction tasks. *CoRR*, 2024.

Yifan Wu, Hayden Gunraj, Chi en Amy Tai, and Alexander Wong. Covidx cxr-4: An expanded multi-institutional open-source benchmark dataset for chest x-ray image-based computer-aided covid-19 diagnostics, 2023. URL `https://arxiv.org/abs/2311.17677`.

Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 3733–3742, 2018.

Tete Xiao, Colorado J Reed, Xiaolong Wang, Kurt Keutzer, and Trevor Darrell. Region similarity representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10539–10548, 2021.

Enze Xie, Jian Ding, Wenhai Wang, Xiaohang Zhan, Hang Xu, Peize Sun, Zhenguo Li, and Ping Luo. DetCo: Unsupervised contrastive learning for object detection. In *IEEE/CVF International Conference on Computer Vision*, pp. 8372–8381, 2021a. doi: 10.1109/ICCV48922.2021.00828.

Jiahao Xie, Xiaohang Zhan, Ziwei Liu, Yew Soon Ong, and Chen Change Loy. Unsupervised object-level representation learning from scene images. In *Neural Information Processing Systems*, 2021b. URL `https://api.semanticscholar.org/CorpusID:235593250`.

Zhenda Xie, Yutong Lin, Zheng Zhang, Yue Cao, Stephen Lin, and Han Hu. Propagate yourself: Exploring pixel-level consistency for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pp. 16684–16693, 2021c.

Sukmin Yun, Hankook Lee, Jaehyung Kim, and Jinwoo Shin. Patch-level representation learning for self-supervised vision transformers. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8344–8353, 2022. doi: 10.1109/CVPR52688.2022.00817.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stephane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12310–12320. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/zbontar21a.html.

Shuangfei Zhai, Navdeep Jaitly, Jason Ramapuram, Dan Busbridge, Tatiana Likhomanenko, Joseph Y Cheng, Walter Talbott, Chen Huang, Hanlin Goh, and Joshua M Susskind. Position prediction as an effective pretraining strategy. In *International Conference on Machine Learning*, pp. 26010–26027. PMLR, 2022.

Shaofeng Zhang, Feng Zhu, Junchi Yan, Rui Zhao, and Xiaokang Yang. Zero-CL: Instance and feature decorrelation for negative-free symmetric contrastive learning. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=RAW9tCdVxLj.

Shaofeng Zhang, Qiang Zhou, Zhibin Wang, Fan Wang, and Junchi Yan. Patch-level contrastive learning via positional query for visual pre-training. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 41990–41999. PMLR, 23–29 Jul 2023a. URL https://proceedings.mlr.press/v202/zhang23bd.html.

Shaofeng Zhang, Feng Zhu, Rui Zhao, and Junchi Yan. Patch-level contrasting without patch correspondence for accurate and dense contrastive representation learning. In *The Eleventh International Conference on Learning Representations*, 2023b. URL https://openreview.net/forum?id=10R_bcjFwJ.

Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5122–5130, 2017. URL https://api.semanticscholar.org/CorpusID:5636055.

Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. Image BERT pre-training with online tokenizer. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=ydopy-e6Dg.

A Details about Datasets

In this paper, we perform continual learning and report results on the 4 semantic segmentation datasets: ADE20k (Zhou et al., 2017), Pascal Context (P.Cont) (Mottaghi et al., 2014), Cityscapes (Citys.) (Cordts et al., 2016) and LoveDA (Wang et al., 2021a). We use the training set of these datasets to apply our continual learning setup with different pre-training methods such as UDI (Su & Ji, 2024), FLSL (Su et al., 2023), GLARE. We then finetune the resulting models for segmentation using FPN (Kirillov et al., 2019) and report the performances of the models on the validation sets. In the following we provide more details on these datasets.

ADE20k (Zhou et al., 2017). This dataset contains various scenes which are potentially cluttered with many objects. It includes fine-grained labels with 150 semantic classes. The training set is composed of 20,210 images and the validation set contains 2,000 images.

Pascal Context (Mottaghi et al., 2014). This is a segmentation dataset with denser annotations, which includes background classes like sky and grass in addition to foreground objects. The training set is composed of 4,998 images and the validation set contains 5,105 images. This dataset has 60 semantic classes.

Cityscapes (Cordts et al., 2016). This dataset is designed specifically for urban street scenes, showcasing objects in driving scenarios. It includes 19 semantic categories for segmentation. Its training set is composed of 2,975 images and its validation set is composed of 500 images.

LoveDA (Wang et al., 2021a). This focuses on different geographical environments between urban and rural. It contains high spatial resolution (0.3m) remote sensing images containing objects at different scales, complex backgrounds and inconsistent class distributions. Its training set is composed of 2,522 images, its validation set is composed of 1,669 images and test set of 1,796 images.

Derm7pt (Kawahara et al., 2019). This is a dermoscopic benchmark for skin lesion analysis containing ∼800 images. The dataset focuses on the classification of pigmented skin lesions into seven diagnostic categories and is commonly used to evaluate models in low-data medical imaging scenarios.

COVIDx (Wu et al., 2023). This is a large-scale chest X-ray benchmark designed for COVID-19 detection. It contains images labeled into three categories: normal, pneumonia, and COVID-19. The dataset aggregates X-ray scans from multiple sources and institutions, making it a diverse benchmark for evaluating models in medical image classification tasks.

Moreover, these datasets differ between each other by their size and their domain. Indeed, by evaluating our continual pre-training on these datasets we showcase the ability of our setup to work in scenarios with limited data and also out-of-domain with respect to the dataset used to pre-train the initial weights (especially in the case of LoveDA). This also stems the contrast between the dataset sizes that we consider in our setup ($<= 20k$) compared to what is used for standard pre-training which are in the order of millions of images.

B Other Experimental Details

B.1 Image-level data augmentation

The augmentation settings in GLARE are based on the augmentation pipeline of BYOL (Grill et al., 2020). In our approach, we begin by sampling two random crops from the input image using a large crop ratio (e.g., $0.25 \sim 1.0$) of size 224×224. We then proceed by sampling 10 other crops with a smaller crop ratio (e.g., $0.05 \sim 0.25$) of size 96×96. We use an asymmetric training process where the larger crops, usually referred to as global crops, are passed to the momentum encoder and then all crops (both global and local, which are the smaller ones) are passed to the base encoder. The distortions that we apply are:

- color jittering, with a probability of 0.8, brightness of 0.4, contrast of 0.4, saturation of 0.2 and hue of 0.1;

- gray scaling, with a probability of 0.2, gaussian blurring and solarization with probabilities of (1.0, 0.0), (0.1, 0.2) and (0.5, 0.0) for the first, second global crops and the local crops, respectively;
- color normalization, with mean $(0.485, 0.456, 0.406)$ and std. dev. $(0.229, 0.224, 0.225)$.

B.2 Evaluation details

For evaluation, we perform semantic segmentation on our four segmentation datasets, described in Appendix A. We follow the configurations of the package *mmsegmentation*[1] for finetuning, within the FPN (Kirillov et al., 2019) framework. We consider two configurations for the finetuning: we use 20k iterations schedule for all datasets except for LoveDA for which we use 2k iterations schedule. The resolution of input images during the experiments is 512×512. Then, as performance metrics we calculate the mean intersection over union (mIoU), the all pixel accuracy (aAcc) and the mean class accuracy (mAcc), for each dataset after finetuning.

B.3 Computation Time

Table 8 reports the continual pre-training time required of three configurations UDI $\rightarrow$ UDI, UDI $\rightarrow$ FLSL, and UDI $\rightarrow$ GLARE on 100 epochs T_{100}. The experiments are done on LoveDA (Wang et al., 2021a). We observe that GLARE continual pre-training has higher requirement in terms of computation time. Nevertheless, the continual pre-training is still relatively fast (30 min) since we are only training the adapter parameters for continual pre-training. As for downstream segmentation finetuning, the computation requirement is the same among all the configurations.

Table 8: Time requirements of continual pre-training.

Method	T_{100}
UDI $\rightarrow$ UDI	27 min
UDI $\rightarrow$ FLSL	18 min
UDI $\rightarrow$ GLARE	30 min

C Additional Ablations

C.1 Ablation of region sampling

One of the advantages of GLARE continuous pre-training pipeline is its ability to benefit from previously pre-trained models by leveraging learned semantics. This is done for example with region-level understanding which leverages the attention of the pre-trained model to guide region consistency enforcement. In this section, we ablate the use of attention to guide the region sampling compared to random region sampling: *attention-aware region sampling* and *random sampling*. Table 9 presents the results of finetuning the continual pre-trained model on LoveDA on either of these strategies. We experiment with 3 and 6 randomly sampled regions and with 6 regions sampled using attention-awareness. We observe that attention-aware sampling shows an improvement of +0.29% compared to random sampling, which aligns with the hypothesis of having more semantically meaningful regions using the attention map.

C.2 Effect of blurring strategy

In this section, we study how the blurring is applied during the pre-training process. There are two possible strategies which can be used: *random* and *block-wise* blurring applied on the patches. This is similar to what can be done in masking (Zhou et al., 2022). Table 10 presents the results of models undergone continual pre-training with GLARE using these two different strategies and finetuned on semantic segmentation. We use LoveDA as our reference dataset. We observe that applying *random blurring* leads to the best performance. Therefore, we decided to use that strategy for our main experiments in this work, unless stated otherwise.

[1] https://github.com/open-mmlab/mmsegmentation

Table 9: Ablation of the region sampling strategy. Experiments performed on UDI-GLARE with a finetuning of 2k iterations on LoveDA. M represents the number of sampled regions considered.

Strategy	mIoU	aAcc	mAcc
random ($M = 3$)	51.3	70.0	63.8
random ($M = 6$)	51.4	70.2	64.0
attention-aware ($M = 6$)	51.5	70.3	64.3

Table 10: Ablation of the block-wise vs. random blurring. Experiments performed with UDI → GLARE with a finetuning of 2k iterations on LoveDA.

Method	mIoU	aAcc	mAcc
random init.	19.3	37.4	34.4
block-wise	51.5	70.2	64.1
random	51.5	70.2	64.3

C.3 Effect of dataset scale

In this section, we evaluate the performance of our continual pre-training pipeline across different dataset scales. Specifically, we conduct experiments on ADE20K and LoveDA using subsets of 10%, 20%, 50% and 100% of the data. The results are summarized in Figure 3. We observe that even with a small dataset of 2k images (corresponding to 100% of LoveDA and 10% of ADE20K), our continual pre-training approach yields improvements. However, for highly out-of-domain datasets like LoveDA, we hypothesize that performing continual pre-training on a smaller set of unlabeled data can be detrimental. In contrast, ADE20K, which consists of images more closely aligned with ImageNet-1K, does not exhibit the same issue.

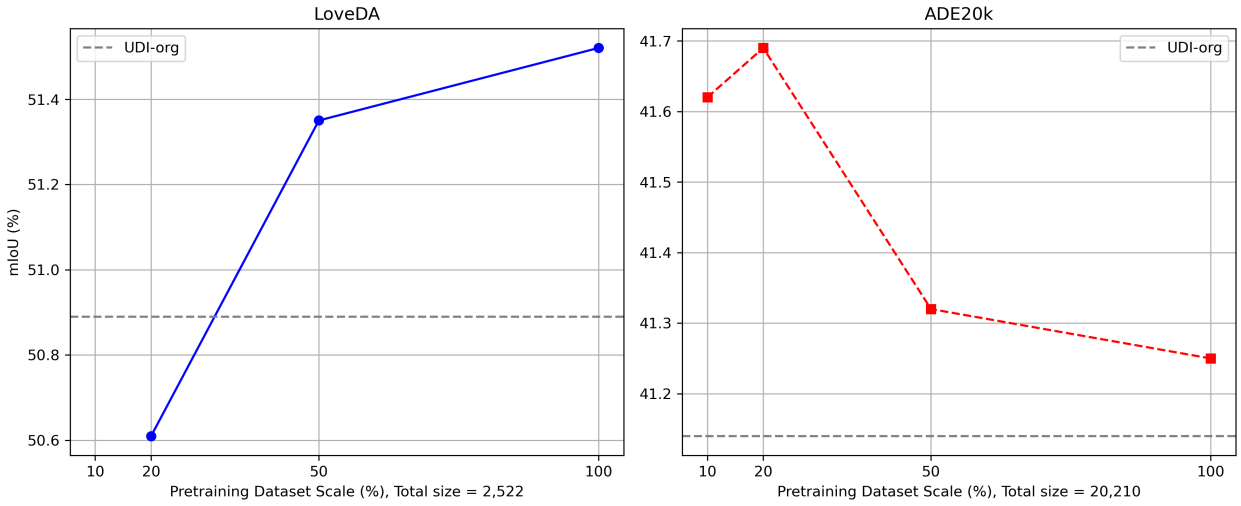

Figure 3: Effect of dataset scale on the performance of GLARE continual pre-training applied on LoveDA and ADE20k. The dashed gray line represent the baseline performance of UDI pre-trained model on the respective dataset.

D Visualization of Attention Maps

In this section we visualize some attention maps from the last block of the ViT encoder, using the [CLS] token as the query token. Figure 4 shows the attention from DINO (Caron et al., 2021), the original pre-trained weights from UDI, and a GLARE continual pre-trained model starting from UDI (Su & Ji, 2024) weights, finetuned on ADE20k. We observe similarities in how the attention is distributed across the images, focusing on various details such as foreground objects, object parts, and the background. GLARE continual pre-trained model demonstrates reduced noise relative to UDI, with its attention more precisely directed toward specific objects or regions. When using GLARE for continual pre-training, the model leverages what has been learned before and learns supplementary semantics specific to the dataset.

E PCA Visualization of embeddings

In this section, we provide an additional qualitative analysis of the learned representations through PCA visualizations of the embeddings. Specifically, we project the patch representations into three principal components using PCA and visualize them in RGB space. We compare DINO (Caron et al., 2021) pre-trained model against our continual pre-trained variant, DINO $\rightarrow$ glare. As shown in Figure 5, GLARE produces less noisy and more semantically coherent embeddings on both the COVIDx (Wu et al., 2023) and Derm7pt (Kawahara et al., 2019) images.

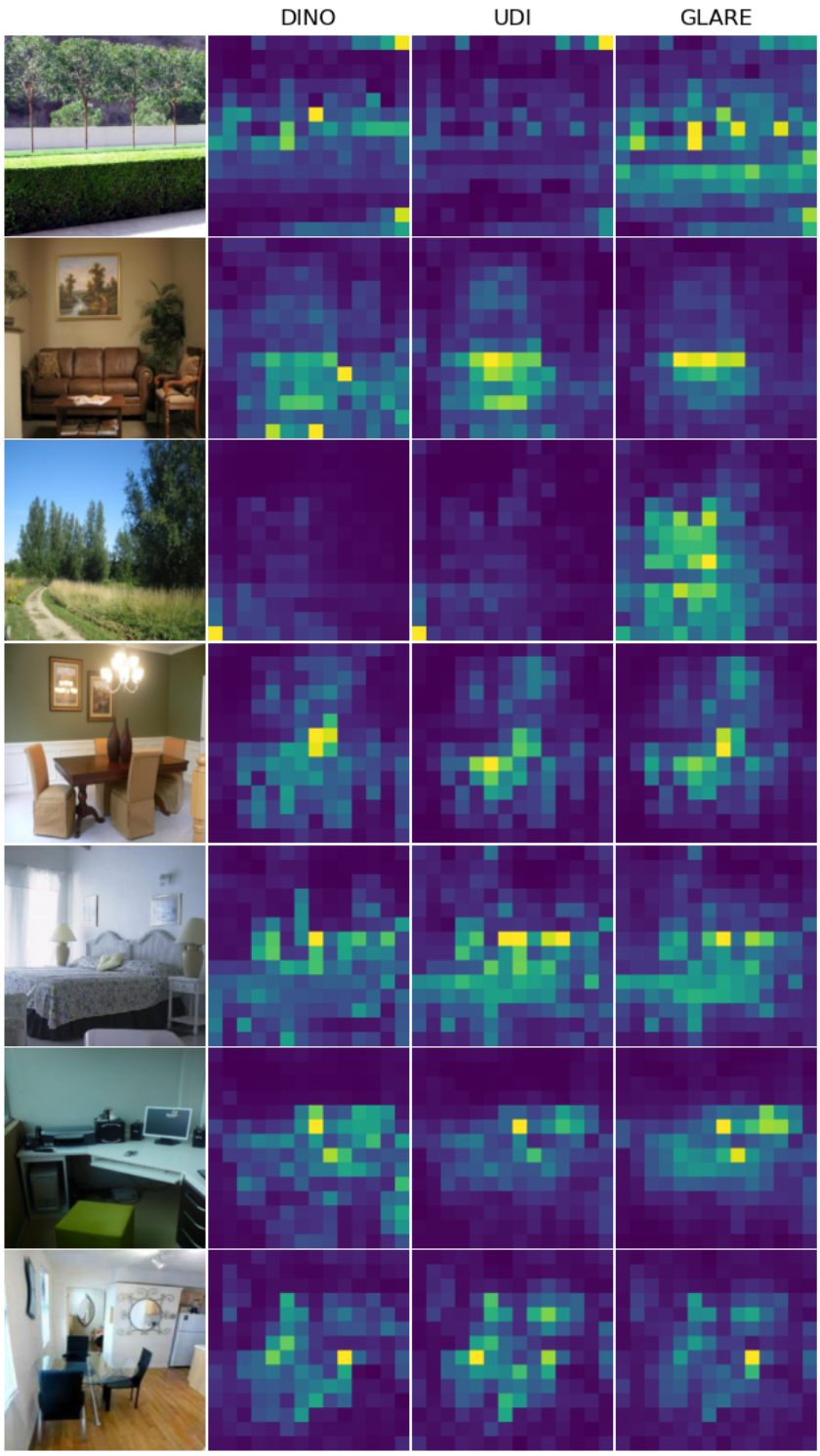

Figure 4: Visualization of self-attention maps obtained from DINO, UDI and GLARE continual pre-trained models from the last block of the ViT encoder starting from UDI.

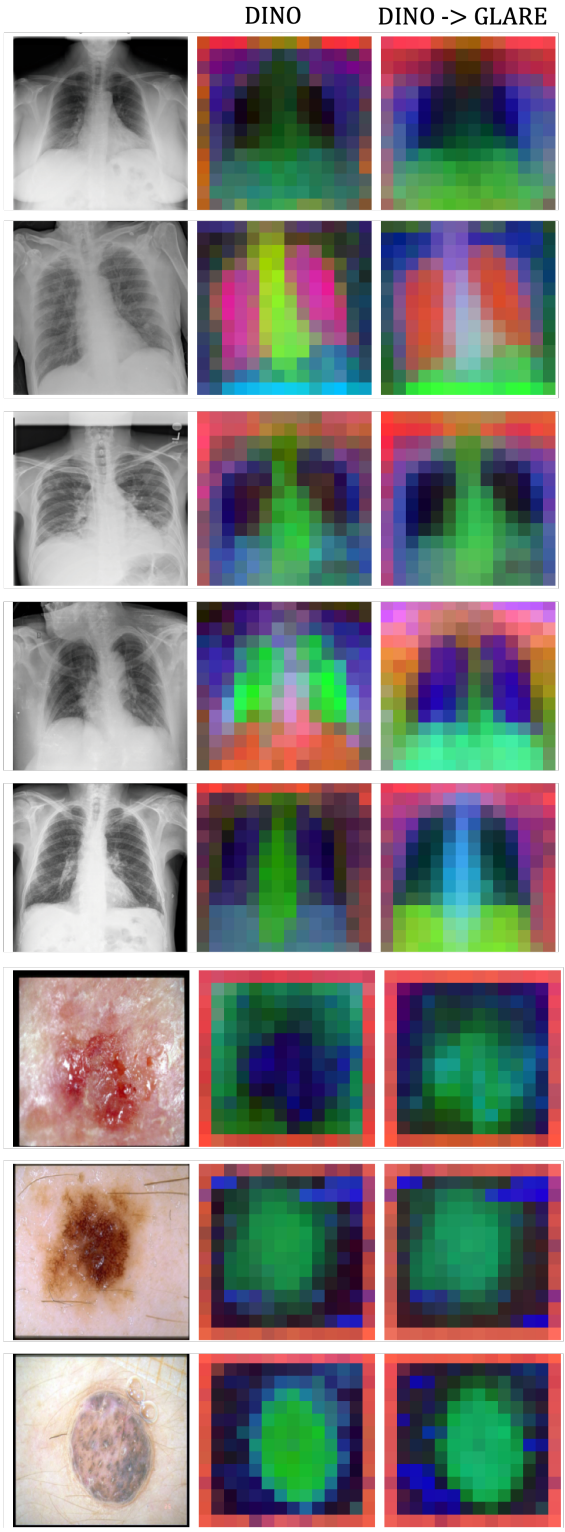

Figure 5: PCA Visualizations of the embeddings on images from COVIDx (Wu et al., 2023) and Derm7pt (Kawahara et al., 2019) datasets. We take the first 3 principal components and show the results for DINO and DINO → GLARE continual pre-trained model.