Self-Supervised Incremental Learning of Object Representations from Arbitrary Image Sets
Supplementary Material

George Leotescu[1*] Alin-Ionut Popa[1*] Diana Grigore[1] Daniel Voinea[1] Pietro Perona[1,2]
[1]Amazon Inc. [2]California Institute of Technology
{leoteg,popaaln,digrigor,dvoinea}@amazon.com perona@caltech.edu

This material includes technical specifics like hyperparameter settings, data processing procedures, and training configurations. The goal is to give the audience a comprehensive understanding of the self-supervised training process used in our pipeline. Additionally, we present supplementary experimental results that complement and reinforce the arguments made in the main paper.

Implementation Details. All the experiments were performed on a machine with 8 NVidia A10G GPU cards with 24GB memory each, using PyTorch. For $\Phi^{\texttt{BKB}}$ design choice we employ variations of the ViT [3] backbone (ViT-Small, ViT-Base, ViT-Large, ViT-Large with Registers, ViT-Huge) initialised with weights pre-trained via the DINOv2 [7], DINO-REG [2] or MAWS [8] frameworks. $\Phi^{\texttt{CROSS}}$ is a multi-head cross attention block followed by 2 (up-sample and down-sample) linear projections with GELU activations applied in between. The $\Phi^{\texttt{CROSS}}$ layers are randomly initialised. We train for 800 epochs on ABO and 8000 epochs on iNaturalist dataset. We use a significantly larger number of epochs on iNaturalist due to the complex nature and fine-grained structure of the dataset. We train using the AdamW optimiser [6] and various batch sizes distributed over 8 GPUs which are conditioned by the size of the backbone (20 when using the ViT-Small backbone or 3 for ViT-Huge). We found the training to be highly sensitive in regard to the teacher momentum and learning rate combinations. The values of teacher momentum and learning rate for jobs using ViT-Base or larger are validated to 0.9996 and $5e-6$, respectively. For ViT-Small we set the teacher momentum to 0.996. The rest of the hyper-parameters follow the configurations described in [1, 7]. A warm-up scheduler is applied to the teacher temperature for 30 epochs from 0.04 to 0.07, and to the the learning rate for 10 epochs. The later follows a cosine decay down to $1e-8$. During training we use image sets with fixed length of 4 elements. Every epoch, we randomly sample 4 different images for each object, thus the ordering is irrelevant.

As an overall rule, we used the same model configuration for both evaluation setups (single-image versus MILE). Each backbone maintains a consisent input patch size, matching the pre-trained model configuration. Specifically, DINOv2-Small, DINOv2-Base, DINOv2-Large, DINO-Registers-Large, MAWS-Base and MAWS-Large all utilize a patch size of 14, while MAWS-Huge employs a patch size of 16.

We generate 2 global and 8 local crops for each image. The global crops are generated such that each crop covers an image ratio between 0.45 and 1.0. For the local crops, the image ratio is between 0.05 and 0.45. Additionally, we follow the data augmentation pipeline described in BYOL [4]. For the larger model variants such as ViT-Large and ViT-Huge we utilise the LORA [5] adapter. This is plugged over the visual encoding backbone targeting the key, query and value modules from the backbone with the matrix rank validated to the value of 48. This adds a double benefit to our learning procedure, as it enables robust usage of GPU resources, while considerably speeding up the training procedure.

To gain better insight into the mechanics of this training process, we performed a t-SNE analysis [9] on the embeddings retrieved at different epochs during training. These results are illustrated in Figure 2. The embeddings are from the MILE with MAWS-ViT-L model, taken at epochs 0, 500, and $1,000$. We selected 200 random classes from the *Aves* category of the iNaturalist dataset, due to its high intra-class diversity. We then generated random combinations of image sets for each of the selected classes. In the t-SNE plots, we color-coded 20 random classes and gray-plotted the rest. Initially, the embeddings are uniformly dispersed across the latent space. However, as the training loss stabilizes, the embeddings begin to group

[*] Authors contributed equally.

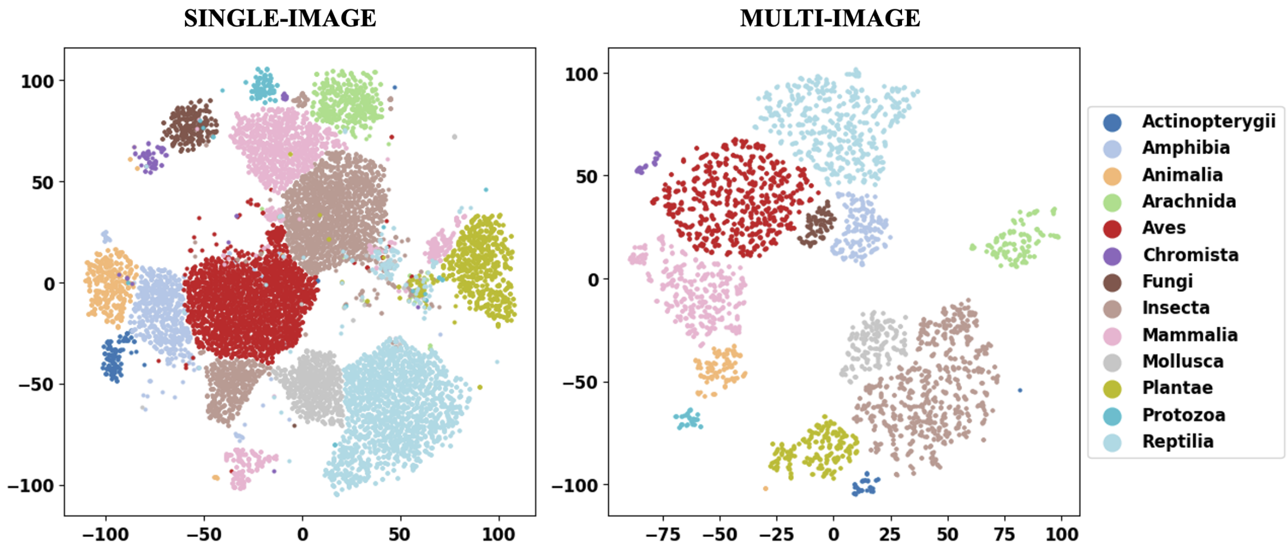

Figure 1. **t-SNE visualization of single-image vs. multi-image embeddings computed from image sets of length** 8 **with MAWS ViT-L backbone on the iNaturalist macro-categories.** Notice the clear class separability and implicit clustering achieved by our method. On the other hand, for the single-view, there are multiple collisions between the elements of similar classes.

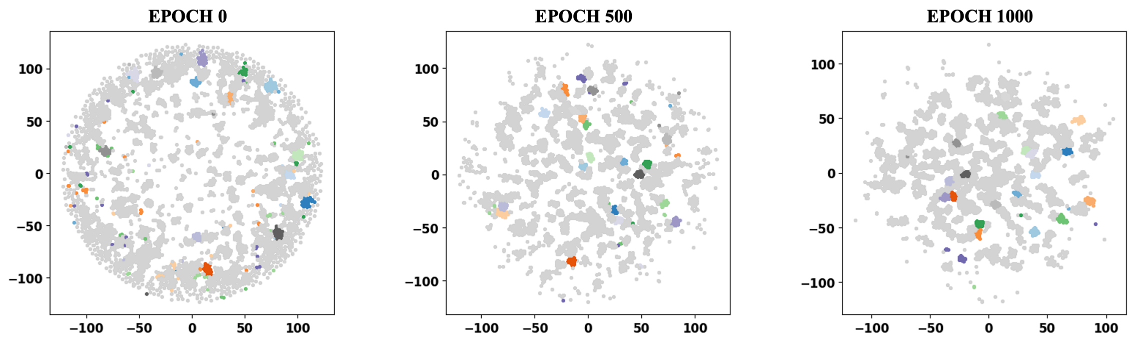

Figure 2. **Evolution of the multi-image latent embedding during the SSL training for iNaturalist with t-SNE.** We select random embeddings from 200 random *Aves* classes and color plotted a subset of 20. At epoch 0, the embeddings are randomly dispersed throughout the latent space. As the training converges the randomness is progressively replaced by implicit grouping centered around each class of interest (effect gradually highlighted by the coloured plots from epoch 500 and 1000).

into clusters associated with the classes of interest. This demonstrates the model's ability to learn meaningful class-specific representations through the self-supervised training process.

Additional Visual Results and Failure Cases. Additional visual results with the PCA color coded implicit region correspondence retrieved via the encoding backbone Φ^{BKB} are illustrated in Figure 3. Context-conditioned class-agnostic object segmentations are present in Figure 5 highlighting the implicit capability of our approach to better infer the contour of the common object based on the previously seen image context. Failure cases usually happen when the image information does not contain clear evidence towards a single object, or the surrounding background guide the focus towards other elements. Such examples are shown in Figure 6.

Furthermore, in Figure 7 we illustrate the incremental impact of the latent state over the conditional class-agnostic object segmentation. We select different multi-image samples from iNaturalist and consider one of the images as a reference image (*i.e.* marked with a green rectangle) which is positioned at the end of the input sequence. Next, we progressively increase the

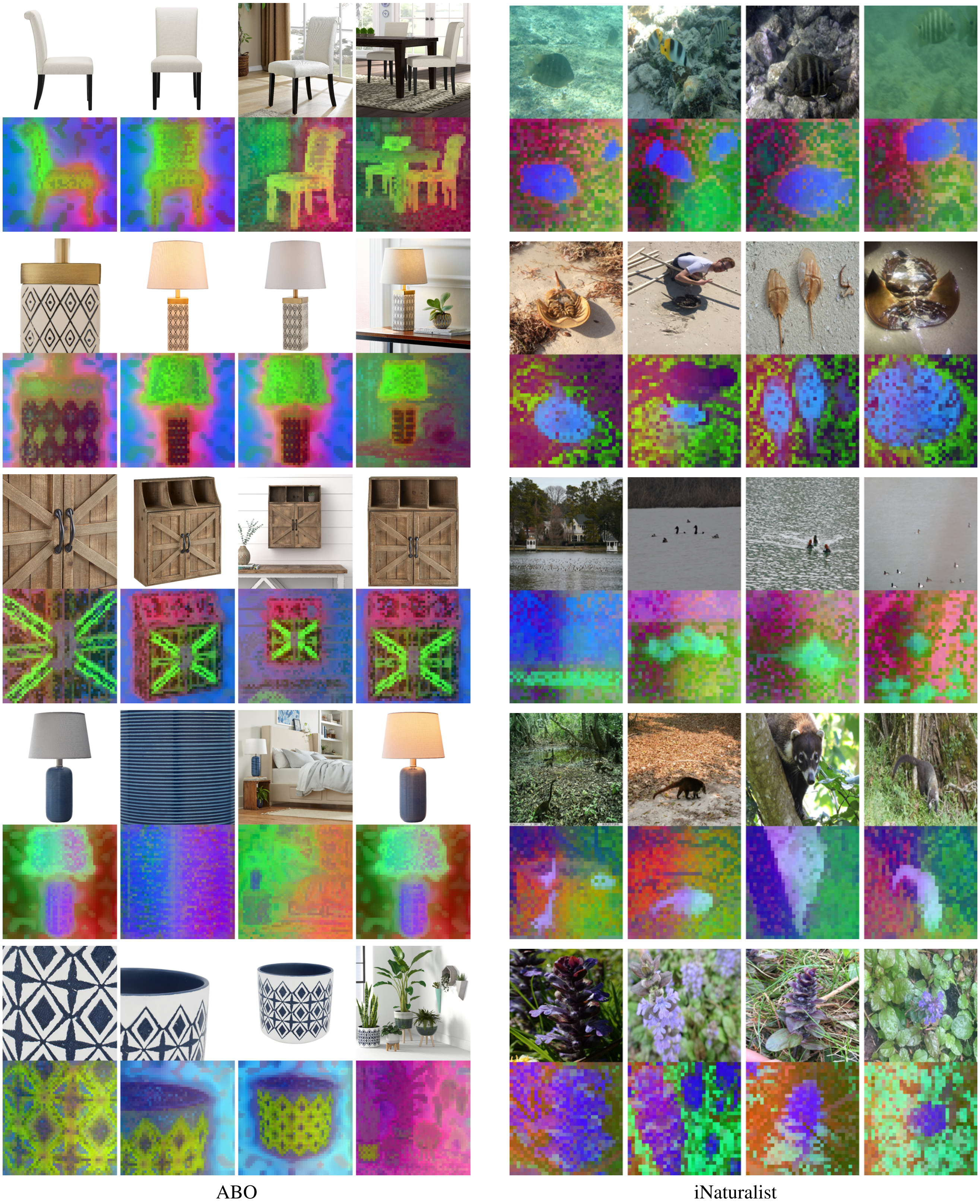

Figure 3. **Color coded region correspondence emphasised via PCA.** Notice the robustness of the encoding extraction backbone to the class variety, single versus multi object presence within individual images, occlusion with different background elements, partial versus full views of the objects or the high visual complexity of the portrayed scenes.

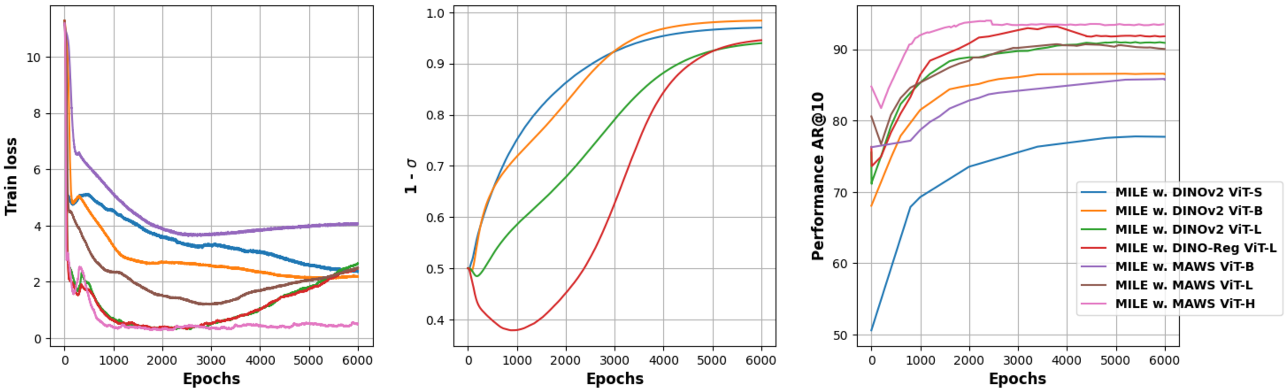

Figure 4. **SSL training logs from iNaturalist.** From left to right we illustrate: the evolution during training epochs for loss $\mathcal{L}_{\texttt{MI}}$, gating function $1-\sigma$ and the AR@10 performance for different $\Phi^{\texttt{BKB}}$ initialisations. An important factor in model performance (for both DINOv2 and MAWS variants) is dictated by the backbone size. Furthermore, notice the inverse correlation effect between loss and performance even though the learning and evaluation tasks are not directly linked. The $1 - \sigma$ plot shows how the pondering weight of the incremental states of $\mathbf{L}$ changes as training progresses. If initially the impact is dominated by the backbone encoding of $\mathbf{I}_0$, as the training progresses it learns to better assimilate the previously seen visual information. Even though the $\Phi^{\texttt{BKB}}$ configurations differs considerably, the $\Phi^{\texttt{CROSS}}$ manifests consistent behaviour by converging after similar number of epochs, thus proving the versatility of our approach.

image support for the latent state from a *single* image to *four* images. Noticeably, as more images depicting the category of interest are added, the attention is better focused towards the common depicted object across the multi-image input.

Computational Complexity. Another important observation regards the computational complexity of the comparing methods. For both MILE and all the single-image baselines, the computational complexity of building the embedding(s) of an object with $|\mathcal{I}|$ images is linear with $|\mathcal{I}|$ (e.g. $\mathcal{O}(|I| \times E)$, where $\mathcal{O}(E)$ represents the cost of running $\Phi^{\texttt{BKB}}$ on a single image). However, for retrieving an object from a gallery with N indexed objects, similarly with $|\mathcal{I}|$ images each, the computational complexity of the single-image baseline is quadratic w. $|\mathcal{I}|$ (e.g. $\mathcal{O}(|\mathcal{I}|^2 \times N \times S)$ where $\mathcal{O}(S)$ represents the cost of comparing two latent embeddings), while for MILE, the complexity is independent of $|\mathcal{I}|$ (*i.e.* $\mathcal{O}(N \times S)$). The resulting speed-up factor $|\mathcal{I}|^2$ is more impactful as the number of elements within the multi-image set increases.

Permutation invariant architecture design choice. We did not take into account a permutation invariant architecture for two reasons: **(1)** Complexity vs. performance trade-off. All reported metrics are averaged across all permutations of the input sequences, with average standard deviations of 0.23 and 0.27 on ABO and iNaturalist, respectively. Additionally, the studies shown in Figure 5 of the main paper indicate that the incremental embedding construction is robust to noise, with MILE scoring minimal performance drops when up to 50% of the input samples are replaced with random ones (as opposed to the single-image evaluation setup). Therefore, we did not identify a strong rationale for employing a different architecture beyond the lightweight cross-attention mechanism. **(2)** Building an incremental embedding acts as a filtering bottleneck, making MILE robust to noise and also producing useful byproducts, such as the discriminative elements of the common object.

Motivation behind multi-image object search design choice. The rationale behind the multi-image setup is to leverage the complementary visual attributes of the object, which are challenging to encapsulate through a single-image representation. This is applicable to gallery image data setup. Note that many major online retailers (*e.g.* eBay, Alibaba, Amazon, Target) utilize a multi-image setup for the product galleries on their platforms. In these cases, each item in the product catalog is typically represented by a collection of multiple images, rather than a single image. This allows users or buyers to search for a specific product by providing their own random photos of the item, leveraging the full visual information available for each product in the catalog, not just a single image. A similar scenario exists for biologists and naturalists who aim to identify specific animal or plant species. In these cases, each animal or plant species is often depicted through multiple photographic representations, enabling better visual matching and identification within the gallery.

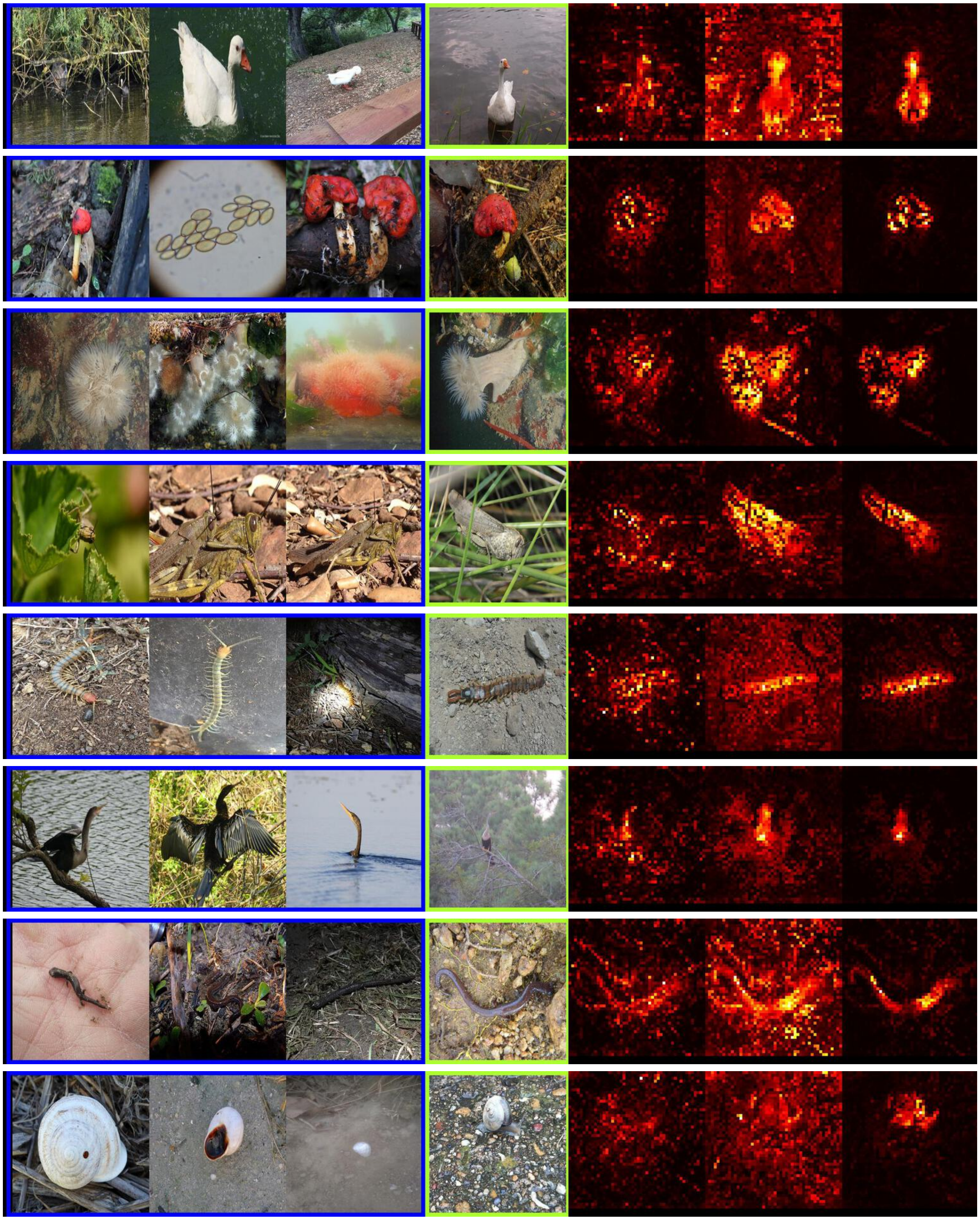

Figure 5. **Visualisations of latent cross-attentions on iNaturalist.** Different variants of attention maps are illustrated for the image marked with a green rectangle. The attention map to the left originates from the backbone encoder Φ^{BKB}. The middle and the right attention maps are the results of the cross-attention between the in-place `[CLS]` token and the latent associated with the images within the blue rectangle. Notice the improvement brought by the conditioning based on the prior imaging information (*i.e.* blue rectangle) shifting the attention towards the common-object of interest.

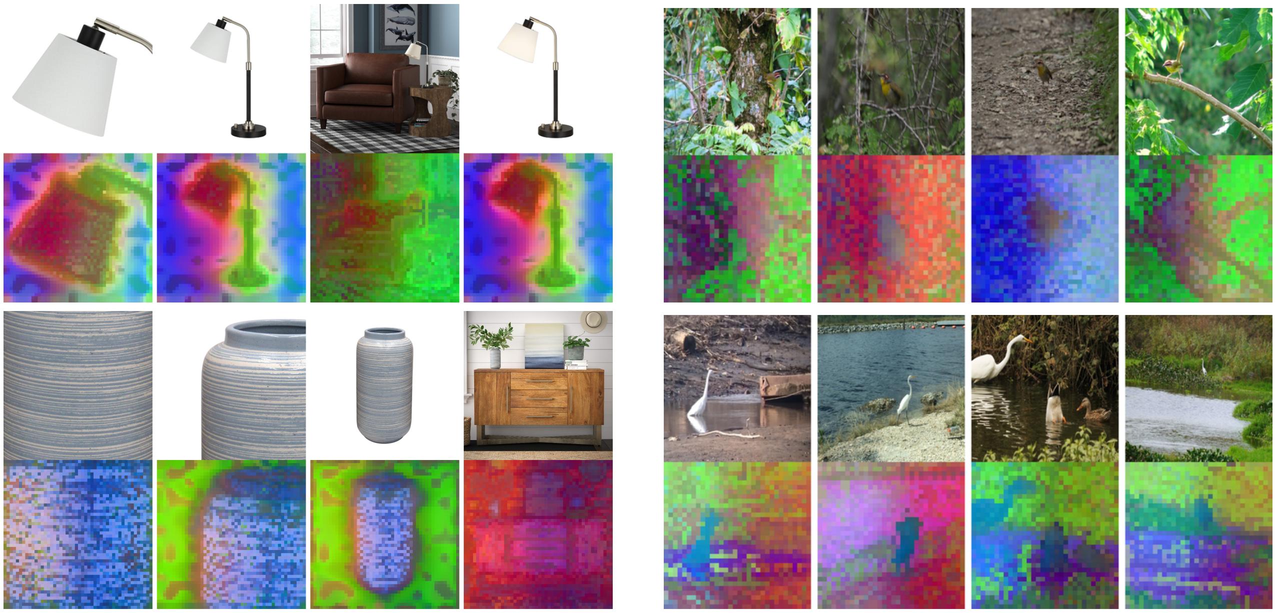

Figure 6. **PCA visualisations for failure cases.** Usually, our approach fails when the object of interest is either out of focus in the main scene, there are multiple dominant objects or the surrounding background is too complex, thus causing a high visual ambiguity degree.

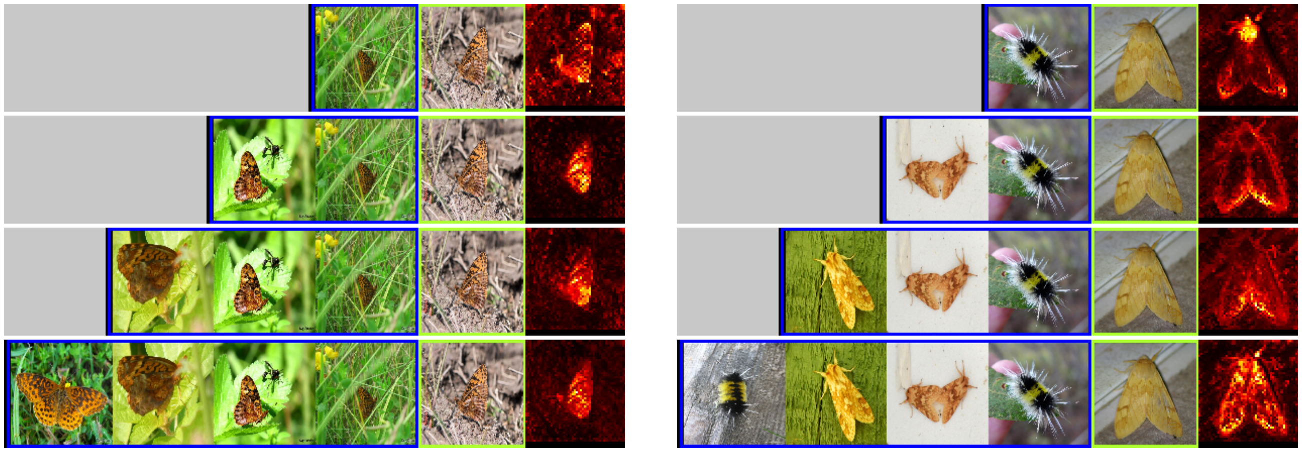

Figure 7. **Progressive improvement of conditional segmentation.** For different multi-image samples grouped in two columns, we illustrate the conditional segmentation for the last image (*i.e.* green rectangle) by progressively increasing the image support (*i.e.* blue rectangle) for the latent variable from 1 to 4 images. This procedure emphasizes from a visual perspective the incremental capability of **L** to assimilate discriminative elements of the object or category of interest with each processing step.

Robustness to noise. Another relevant study is shown in Figure 1. We sampled random image sets from the most representative sub-class (in terms of image samples) for each of the 13 categories from the iNaturalist dataset. We then computed the associated single-image and multi-image embeddings using MAWS ViT-Large backbones. Next, we created 2D t-SNE [9] projections of these embeddings. This analysis demonstrates the enhanced class separability achieved by our approach. Another key observation is that semantically similar classes, such as *Amphibia* and *Reptilia*, are positioned adjacent to each other in the projection.

Impact of SSL. The SSL paradigm played a key-role in the development of our framework, as it unlocked the immediate extension to large scale data corpuses. A detailed view of the SSL training loss and its impact on the recall metrics is illustrated in Figure 4. MILE performance scales up based on the $\Phi^{\texttt{BKB}}$ model size, while σ and $\Phi^{\texttt{CROSS}}$ bring consistent gains, despite their light weight size.

References

[1] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *CVPR*, 2021. 1

[2] Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. *arXiv preprint arXiv:2309.16588*, 2023. 1

[3] 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 *ICLR*, 2021. 1

[4] 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. *NeurIPS*, 33:21271–21284, 2020. 1

[5] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 1

[6] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. 1

[7] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 1

[8] Mannat Singh, Quentin Duval, Kalyan Vasudev Alwala, Haoqi Fan, Vaibhav Aggarwal, Aaron Adcock, Armand Joulin, Piotr Dollár, Christoph Feichtenhofer, Ross Girshick, et al. The effectiveness of mae pre-pretraining for billion-scale pretraining. *arXiv preprint arXiv:2303.13496*, 2023. 1

[9] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 1, 6