# A Supplementary Materials

## A.1 Detailed Explanation of the Dynamic Scaling Factor Approach

We provide a detailed explanation of applying the dynamic scaling factor to the ProxyDR model. First, we define $d_{x,y} := \left\| \tilde{f}(x) - \tilde{W}_y \right\|$. We rewrite Eq. (1) as:

$$p(c|x) = \frac{d_{x,c}^{-s}}{d_{x,c}^{-s} + B_x}, \tag{7}$$

where $B_x = \sum_{y \neq c, y \in \mathcal{Y}} d_{x,y}^{-s}$. When using the NormFace model [36], *i.e.*, a proxy-based SD softmax model on spherical embedding, Zhang *et al.* [45] have found that $\theta_y$ was close to $\frac{\pi}{2}$ during the training process for $y \neq c$, *i.e.*, for different classes. By employing this, when assuming $\theta_y \approx \frac{\pi}{2}$ for $y \neq c$, we obtain $B_x \approx \sum_{y \neq c, y \in \mathcal{Y}} \left(\frac{\pi}{2}\right)^{-s} = (|\mathcal{Y}| - 1)\left(\frac{\pi}{2}\right)^{-s}$.
The expression $\frac{\partial^2 p(c|x)(\theta_c)}{\partial \theta_c^2}$ can be written as:

$$\frac{\partial^2 p(c|x)(\theta_c)}{\partial \theta_c^2} = \frac{B_x s \theta_c^{(s-2)} \psi_{\text{ProxyDR}}(s, \theta_c)}{\left(B_x \theta_c^s + 1\right)^3} \tag{8}$$

where $\psi_{\text{ProxyDR}}(s, \theta_c) = B_x(s+1)\theta_c^s - s + 1$.
We follow the estimation process of [45]. Specifically, we use $B_{x;\text{avg}}$ to estimate $B_x$ and $\theta_{c;\text{med}}$ to estimate $\theta_c$, where $B_{x;\text{avg}}$ is the average of $B_x$ in a mini-batch and $\theta_{c;\text{med}}$ is the median of the $\theta_c$ values in a mini-batch. We clipped the $\theta_{c;\text{med}}$ value to be in the range $\left[0, \frac{\pi}{4}\right]$. Unlike their approach of using an equation to get a rough approximation, we use the Adam optimizer [15] to update a scale factor $s$ that minimizes $\psi^2_{\text{ProxyDR}}(s, \theta_c)$, *i.e.*, $\psi_{\text{ProxyDR}}(s, \theta_c) \approx 0$.

Through a similar process, we can apply dynamic scale factor training [45] in the NormFace model [36]. Specifically, we define $B_x = \sum_{y \neq c, y \in \mathcal{Y}} \exp(s \cos \theta_y)$. Assuming $\theta_y \approx \frac{\pi}{2}$ for $y \neq c$, we obtain $B_x \approx \sum_{y \neq c, y \in \mathcal{Y}} \exp(s \cos \frac{\pi}{2}) = \sum_{y \neq c, y \in \mathcal{Y}} \exp(0) = |\mathcal{Y}| - 1$.
In NormFace model, the expression $\frac{\partial^2 p(c|x)(\theta_c)}{\partial \theta_c^2}$ can be written as:

$$\frac{\partial^2 p(c|x)(\theta_c)}{\partial \theta_c^2} = \frac{-s B_x \exp(s \cos \theta_c) \psi_{\text{NormFace}}(s, \theta_c)}{\left(\exp(s \cos \theta_c) + B_x\right)^3}, \tag{9}$$

where $\psi_{\text{NormFace}}(s, \theta_c) = \cos \theta_c \left(\exp(s \cos \theta_c) + B_x\right) + s \sin^2 \theta_c \left(\exp(s \cos \theta_c) - B_x\right)$.
In our implementation, we use the Adam optimizer [15] to update a scale factor $s$ that minimizes $\psi^2_{\text{NormFace}}(s, \theta_c)$. For $\psi_{\text{NormFace}}(s, \theta_c)$, we use $|\mathcal{Y}| - 1$ to estimate $B_x$ and $\frac{\pi}{4}$ to estimate $\theta_c$ to initialize the value of $s$.

**Table 1:** Summary of the studied datasets

| Dataset | Images per class | Images | Classes |
|---|---|---|---|
| CIFAR100 [16] | 600 | 60000 | 100 |
| NABirds [33] | 13 to 120 | 48562 | 555 |
| Small microplankton (MicroS) | 1 to 456 | 6738 | 109 |
| Large microplankton (MicroL) | 2 to 613 | 8348 | 102 |
| Mesozooplankton (MesoZ) | 3 to 486 | 6738 | 52 |

### A.2 Dataset Details

Table 1 summarizes the number of classes and images in each dataset. The three plankton datasets contain nonliving classes of artifacts and debris. In addition, the three plankton datasets have severe class imbalances. For example, each class in the small microplankton (MicroS) dataset contains 1 to 456 images. All plankton images were obtained using FlowCam (Yokogawa Fluid Imaging Technologies). FlowCam is a flow imaging microscope that captures particles flowing through glass flowcells with well-defined volumes. The three plankton datasets were obtained according to different types of samples (live and Lugol fixed whole seawater or 180 $\mu m$ WP2 plankton net samples) using three different FlowCams (FlowCam 8400, FlowCam VS, and FlowCam Macro) with various magnifications. Thus, the datasets include particles ranging from 5 to 2000 $\mu m$ in size, thus representing nano-, micro-, and mesozooplankton. The plankton samples were obtained from three coastal monitoring stations (IMR) along the Norwegian coast, including Holmfjord in the north, Austevoll in the west and Torungen in the south. In addition, for the nano- and microplankton, seawater samples were obtained from a tidal zone at a depth of 1 meter at the research station at Flødevigen in southern Norway, which is approximately 2 nautical miles from the southern monitoring station at Torungen. The sampling period for the three datasets covered all seasons over a period of approximately 2.5 years.

For the implementation of plankton datasets, we incorporate the available object size information. Specifically, when a data point $x$ has size $v_{\text{size}} = [\text{width}, \text{height}]^{\mathrm{T}}$, we take the elementwise logarithm $v'_{\text{size}} = [\log(\text{width}), \log(\text{height})]^{\mathrm{T}}$. By applying a linear transformation, we obtain an embedding vector $f_{\text{size}}(x)$ according to the size information, *i.e.*, $f_{\text{size}}(x) = W_{\text{size}}^{\mathrm{T}} v'_{\text{size}} + b_{\text{size}}$, where $W_{\text{size}}$ is a learnable matrix with shape $2 \times 128$ and $b_{\text{size}} \in \mathbb{R}^2$ is a learnable vector. Then, we add this vector to the original embedding vector, namely, $f(x) := f(x) + f_{\text{size}}(x)$.

**NABirds.** Data provided by the Cornell Lab of Ornithology, with thanks to photographers and contributors of crowdsourced data at `www.AllAboutBirds.`

org/Labs. This material is based upon work supported by the National Science Foundation under Grant No. 1010818.

**Small Microplankton (MicroS).** This dataset contains images of fixed and live seawater samples acquired at a depth of $5\,m$ at the three monitoring stations and a depth of $1\,m$ in the tidal zone (see above). The seawater samples were carefully filtered through a $80\,\mu m$ mesh to ensure that $100\,\mu m$ flowcell was not clogged and imaged using a $10\times$ objective. This FlowCam configuration results in a total magnification of $100\times$ and images particles ranging from 5 to $50\,\mu m$. Before resizing, one pixel in an image represented $0.7330\,\mu m$.

**Large Microplankton (MicroL).** This dataset contains images of fixed and live seawater samples acquired at a depth of $5\,m$ at the three monitoring stations and a depth of $1\,m$ in the tidal zone (see above). The seawater samples were not filtered and were imaged using a $2\times$ objective, targeting 35 to $500\,\mu m$ particles. Before resizing, one pixel in an image represented $2.9730\,\mu m$. Due to instrument repair and adjustments to improve image quality, the camera settings were modified during the 3 years of imaging to acquire this dataset. Therefore, the image appearance and quality are slightly variable.

**Mesozooplankton (MesoZ).** This dataset contains images of mesozooplankton samples acquired at the three coastal monitoring stations (see above) and a transect in the Norwegian Sea (Svinøysnittet). The samples were obtained using an IMR standard plankton net (WP2) or a multinet mammoth (both $180\,\mu m$ mesh) and fixed with 4% formaldehyde. The images were acquired by two FlowCam instruments (one in Bergen and one in Flødevigen), and the image appearance differs slightly between the two instruments. The FlowCam macro was equipped with a $0.5\times$ objective, resulting in a total magnification of 12.5 and imaging organisms ranging from 180 to $2000\,\mu m$. Before resizing, one pixel in an image represented $9.05\,\mu m$. All images in the mesozooplankton dataset are in grayscale.

**Hierarchical Structures of the Datasets.** Table 2 and Figs. 9, 10, and 11 show the hierarchical structures of the datasets. Because the NABirds dataset contains too many (555) classes to visualize, we do not show the hierarchical structures of the NABirds dataset [33]. We used the hierarchy provided by the Cornell Lab of Ornithology.

**Table 2:** The hierarchical structure of the CIFAR100 dataset used in our experiment

| Level 0 | Level 1 | Level 2 | Level 3 (class level) |
|---|---|---|---|
| Animals | Invertebrates | insects | bee, beetle, butterfly, caterpillar, cockroach |
| | | non-insect invertebrates | crab, lobster, snail, spider, worm |
| | Mammals | aquatic mammals | beaver, dolphin, otter, seal, whale |
| | | large carnivores | bear, leopard, lion, tiger, wolf |
| | | large omnivores and herbivores | camel, cattle, chimpanzee, elephant, kangaroo |
| | | medium-sized mammals | fox, porcupine, possum, raccoon, skunk |
| | | people | baby, boy, girl, man, woman |
| | | small mammals | hamster, mouse, rabbit, shrew, squirrel |
| | Non-mammal vertebrates | fish | aquarium fish, flatfish, ray, shark, trout |
| | | reptiles | crocodile, dinosaur, lizard, snake, turtle |
| Artificial | Artificial (indoor) | food containers | bottles, bowls, cans, cups, plates |
| | | household electrical devices | clock, computer keyboard, lamp, telephone, television |
| | | household furniture | bed, chair, couch, table, wardrobe |
| | Artificial (outdoor) | large man-made outdoor things | bridge, castle, house, road, skyscraper |
| | | vehicles 1 | bicycle, bus, motorcycle, pickup truck, train |
| | | vehicles 2 | lawn-mower, rocket, streetcar, tank, tractor |
| Nature (non-animal) | Nature (non-specific organism) | large natural outdoor scenes | cloud, forest, mountain, plain, sea |
| | Plants | flowers | orchids, poppies, roses, sunflowers, tulips |
| | | fruit and vegetables | apples, mushrooms, oranges, pears, sweet peppers |
| | | trees | maple, oak, palm, pine, willow |

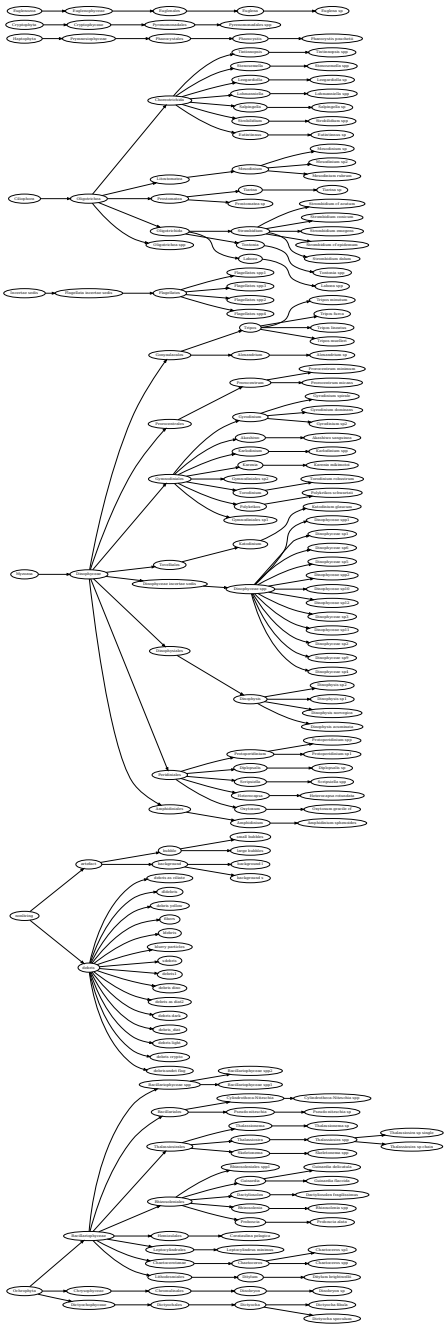


**Fig. 9:** The hierarchical structure of the MicroS dataset used in our experiment. Best viewed by zooming in.

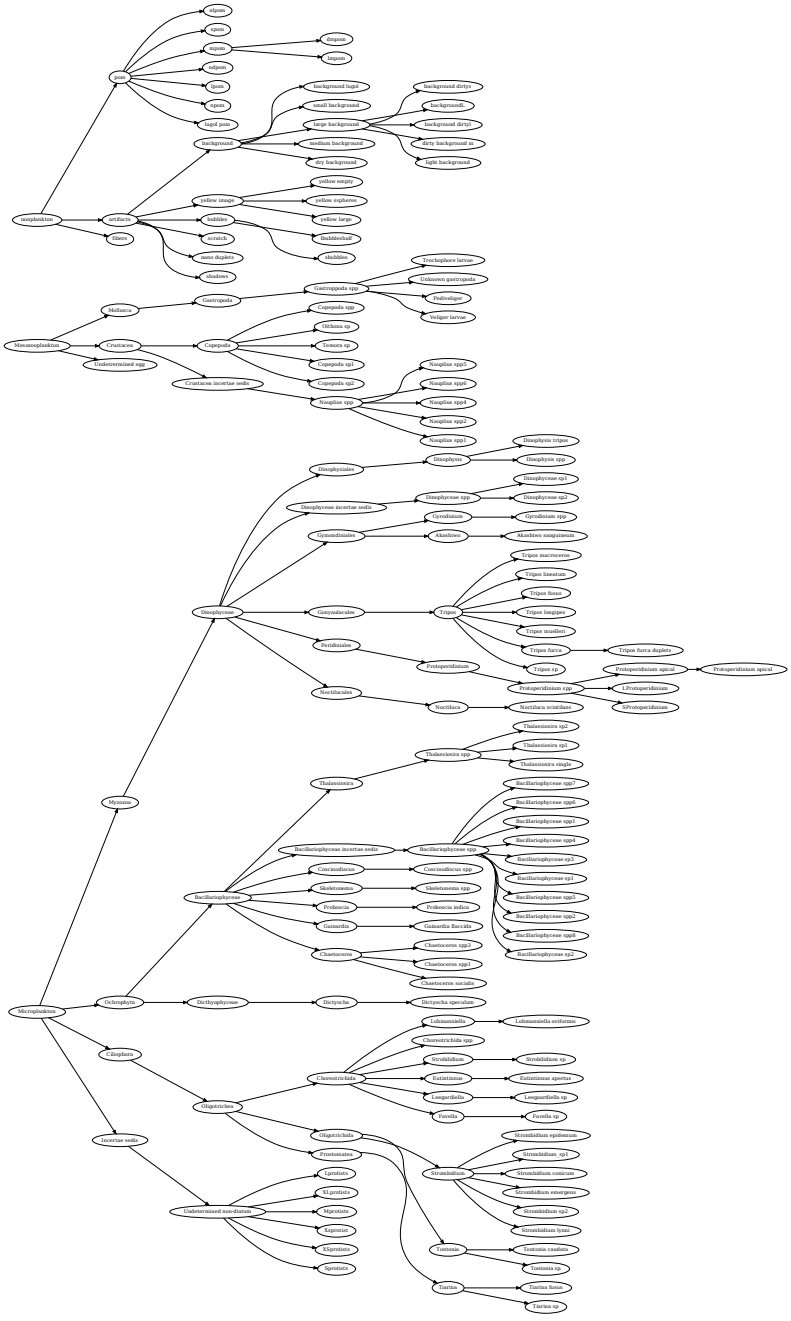



**Fig. 10:** The hierarchical structure of the MicroL dataset used in our experiment. Best viewed by zooming in.

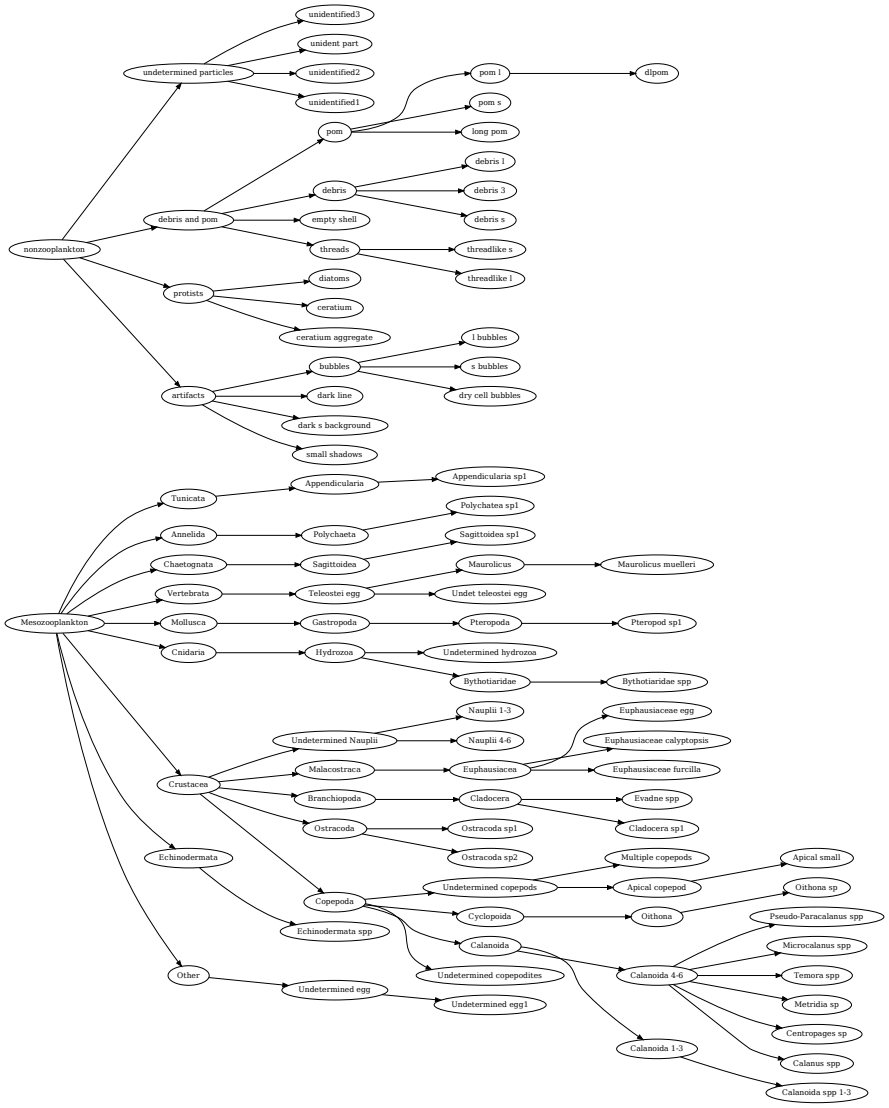



**Fig. 11:** The hierarchical structure of the MesoZ dataset used in our experiment. Best viewed by zooming in.

### A.3 Performance Measures Details

**Mean Correlations.** We introduce this measure to evaluate how well the learned class representatives capture the predefined semantic structure of classes. We obtain the class representatives (either proxies or prototypes) from the learned model using training data points. Then, we obtain the pairwise distance matrix $D_L$ based on the class representatives. We compare matrix $D_L$ with the distance matrix $D_H$ based on the predefined hierarchical structure. Specifically, for each class (row), we calculate Spearman's rank correlation coefficient according to the two matrices. Then, we determine the mean correlation using Fisher transformations. More precisely, we apply a Fisher transformation $\operatorname{arctanh}(\cdot)$ on each correlation coefficient, take the average of the transformed values, and apply $\tanh(\cdot)$ on the average value.
For the plankton datasets, the predefined hierarchical structures of the living classes are based on biological taxonomies. However, these datasets also contain some nonliving classes, such as "large bubbles" and "dark debris". Considering that the predefined hierarchical structures of the living classes are scientifically defined, for the plankton datasets, we report mean correlations among whole classes or only among living classes.

**AHD.** The average hierarchical distance of the top-$k$ predictions [3] was calculated as the average hierarchical distance $d_H$ between the labeled classes and each of the top-$k$ most likely classes. In contrast to Bertinetto *et al.* [3], who considered only misclassified cases, we consider all cases. Hence, in our calculations, the final denominators differ. When $k = 1$, the AHD measure is the same as the average hierarchical cost (AHC) measure defined by Garnot and Landrieu [9]. We report results for $k = 1, 5$.

**HP@$k$.** The hierarchical precision at $k$ [8] was taken as a performance measure. Specifically, let us denote a (hierarchical) neighborhood set of a class $c$ with the distance threshold $\epsilon$ as $N(c, \epsilon)$, *i.e.*, $n \in N(c, \epsilon) \iff d_H(c, n) \leq \epsilon$. We define $\text{hCorrectSet}(c, k)$ as the neighborhood set $N(c, \epsilon)$ with the smallest $\epsilon$ such that $|\text{hCorrectSet}(c, k)| \geq k$. Then, the hierarchical precision at $k$ is calculated as the fraction of the top-$k$ predictions in $\text{hCorrectSet}(c, k)$. We report the results for $k = 5$.

**HS@$k$.** The hierarchical similarity at $k$ is a measure that was introduced by Barz and Denzler [2] with the name "hierarchical precision at $k$", although this metric does not evaluate precision and instead assesses similarity. Here, we use a different measure with the same name that was defined by Frome *et al.* [8]. Hence, we renamed the measure "hierarchical similarity at $k$". When $c$ is a label of a query data point $x$, let $R = ((x_1, c_1), \cdots, (x_m, c_m))$ be the ordered list of image-label pairs based on the distance (sorted by ascending distance) to point $x$ in the normalized embedding space. Considering $\cos(\theta) = 1 - \frac{\|u_1 - u_2\|^2}{2}$, where

$u_1$ and $u_2$ are unit vectors and $\theta$ is the angle between $u_1$ and $u_2$, we defined the similarity between the $i$th and $j$th classes $s_H(i,j)$ as:

$$s_H(i,j) = 1 - \frac{d_T(i,j)^2}{2}, \tag{10}$$

where $i$ and $j$ are the indices for classes ($1 \leq i,j \leq |\mathcal{Y}|$).
The hierarchical similarity at $k$ is then defined as:

$$HS@k := \frac{\sum_{i=1}^{k} s_H(I(c), I(c_i))}{\max_\pi \sum_{i=1}^{k} s_H(I(c), I(c_{\pi_i}))}, \tag{11}$$

where $I(\cdot)$ is an index function that outputs the corresponding index (between 1 and $|\mathcal{Y}|$) for a class and $\pi$ is an index permutation that ranges from 1 to $m$. We report results for $k = 50, 250$.

**AHS@$K$.** The average hierarchical similarity at $K$ was introduced by Barz and Denzler [2] as the "average hierarchical precision at $K$". Due to similar reasons as for HS@$k$, we renamed the measure. The average hierarchical similarity at $K$ is defined as the area under the curve of HS@$k$ from $k = 1$ to $k = K$. We report results for $K = 250$.

**Top-$k$ Accuracy.** The classification accuracy was calculated, with correct classification defined as whether the labeled class is in the top-$k$ predictions (most likely classes), *i.e.*, the $k$ classes with the highest confidence values. We report results for $k = 1, 5$.

### A.4 Results for All Five Datasets

Figs. 12, 13, 14, 15, and 16 show the mean correlation values on the different datasets.

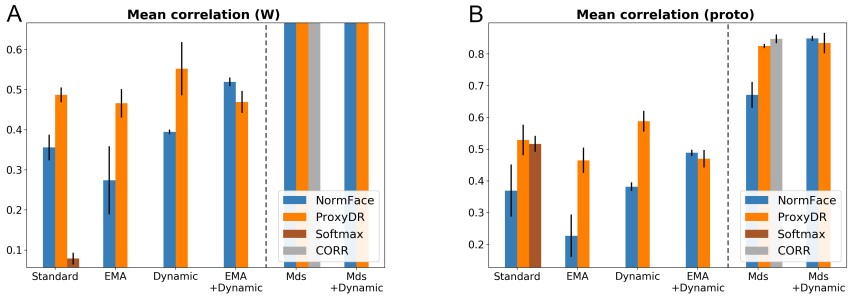



**Fig. 12:** Correlation measures on the CIFAR100 dataset. (A) Values using proxies. The mean correlation value for all MDS-based methods was 0.8580. (B) Values using prototypes.

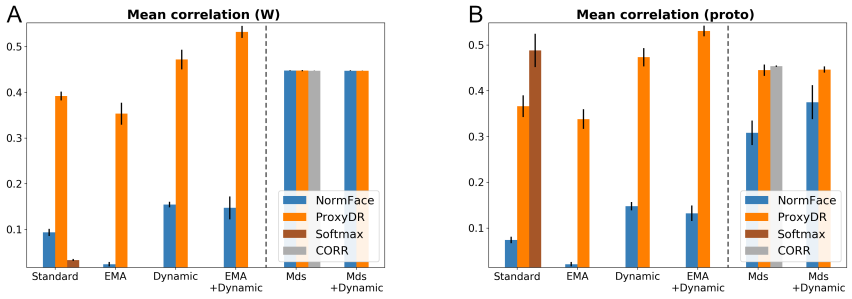



**Fig. 13:** Correlation measures on the NABirds dataset. (A) Values using proxies. The mean correlation value for all MDS-based methods was 0.4476 (this value is small as the dataset contains 555 classes and the embedding dimension is 128.). (B) Values using prototypes.

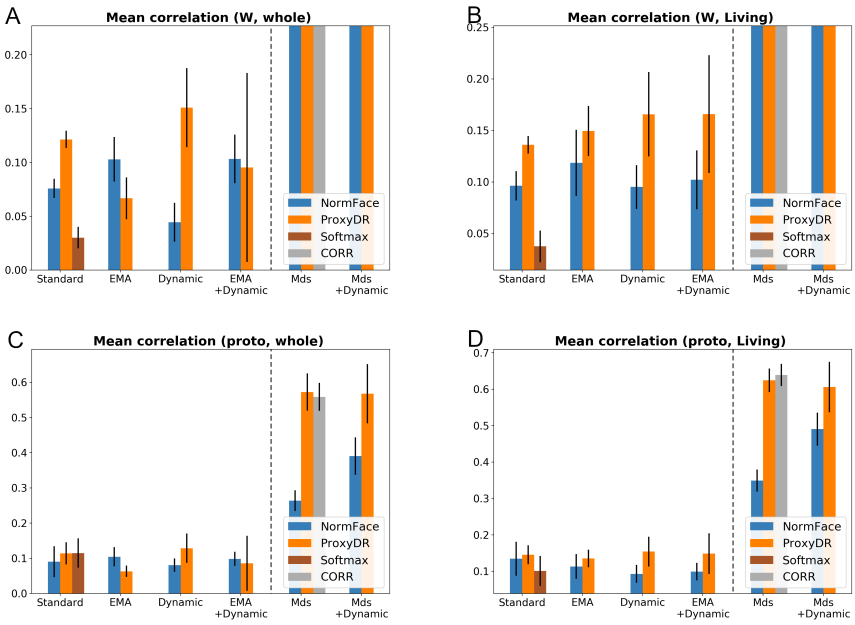



**Fig. 14:** Correlation measures on the MicroS dataset. 'Living' indicates that only biological classes were used (no nonliving classes). (Top) Values using proxies (A: *whole* classes, B: *living* classes). The mean correlation values were 0.9306 (*whole*) and 0.9011 (*living*) for all MDS-based methods. (Bottom) Values using prototypes (C: *whole* classes, D: *living* classes).

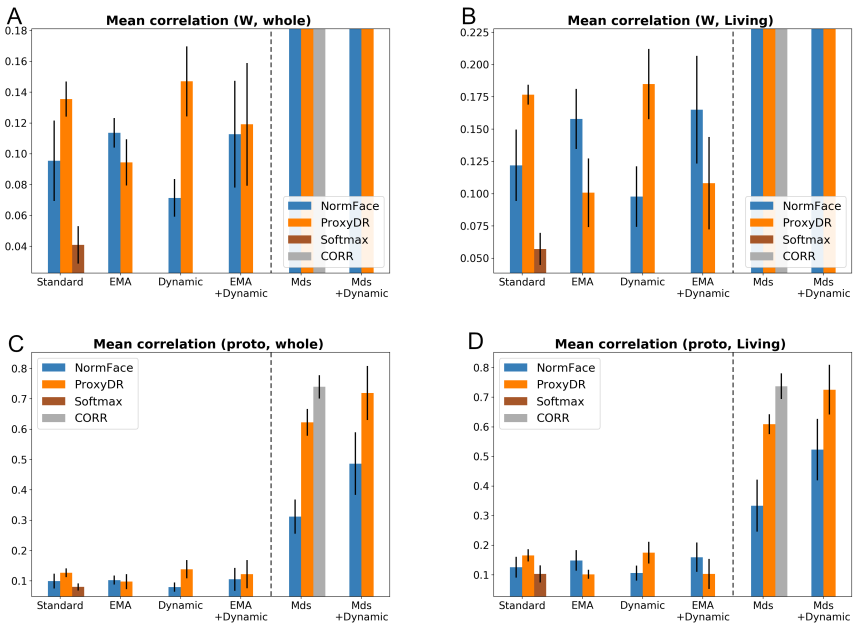



**Fig. 15:** Correlation measures on the MicroL dataset. (Top) Values using proxies (A: *whole* classes, B: *living* classes). The mean correlation values were 0.9543 (*whole*) and 0.9426 (*living*) for all MDS-based methods. (Bottom) Values using prototypes (C: *whole* classes, D: *living* classes).

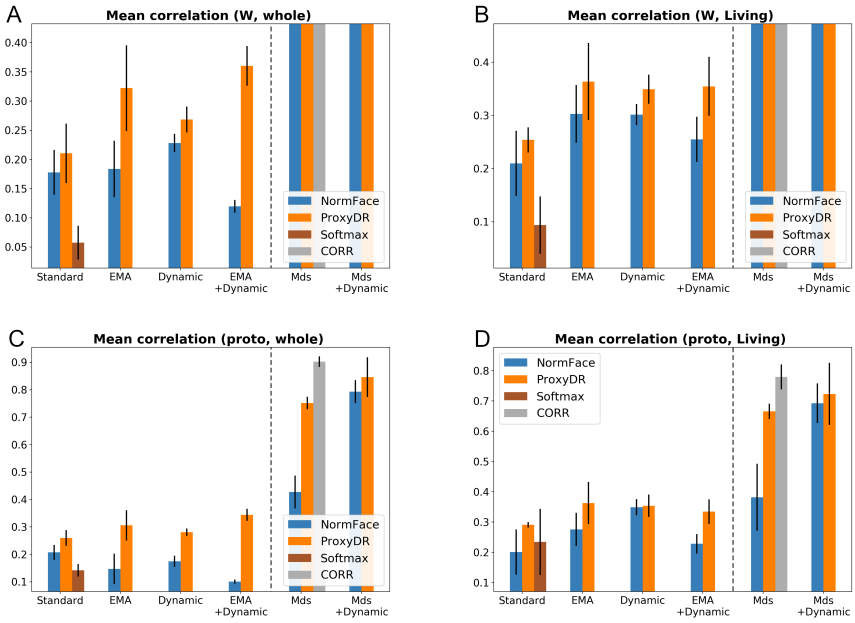



**Fig. 16:** Correlation measures on the MesoZ dataset. (Top) Values using proxies (A: *whole* classes, B: *living* classes). The mean correlation values were 0.9783 (*whole*) and 0.9602 (*living*) for all MDS-based methods. (Bottom) Values using prototypes (C: *whole* classes, D: *living* classes).

Figs. 17, 18, 19, 20, and 21 show the hierarchy-informed performance measures on the different datasets.

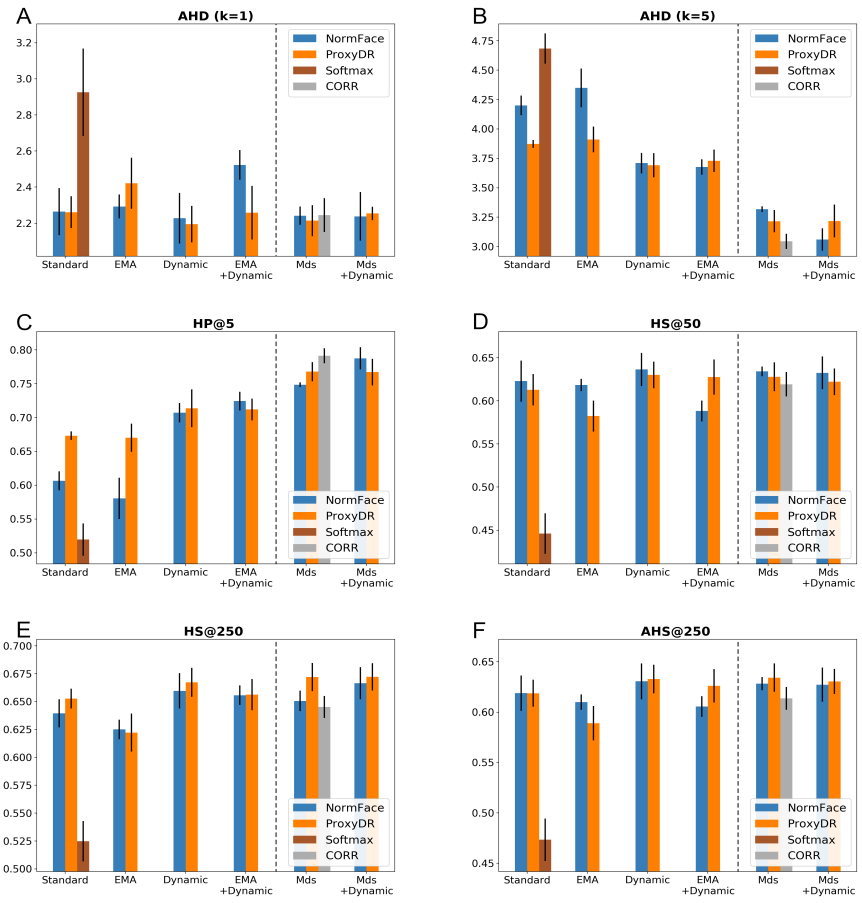



**Fig. 17:** Hierarchical performance measures on the CIFAR100 dataset. The symbol ↓ denotes that lower values indicate better performance. The symbol ↑ denotes that higher values indicate better performance. (A) AHD (k=1): ↓. (B) AHD (k=5): ↓. (C) HP@5: ↑. (D) HS@50: ↑. (E) HS@250: ↑. (F) AHS@250: ↑.

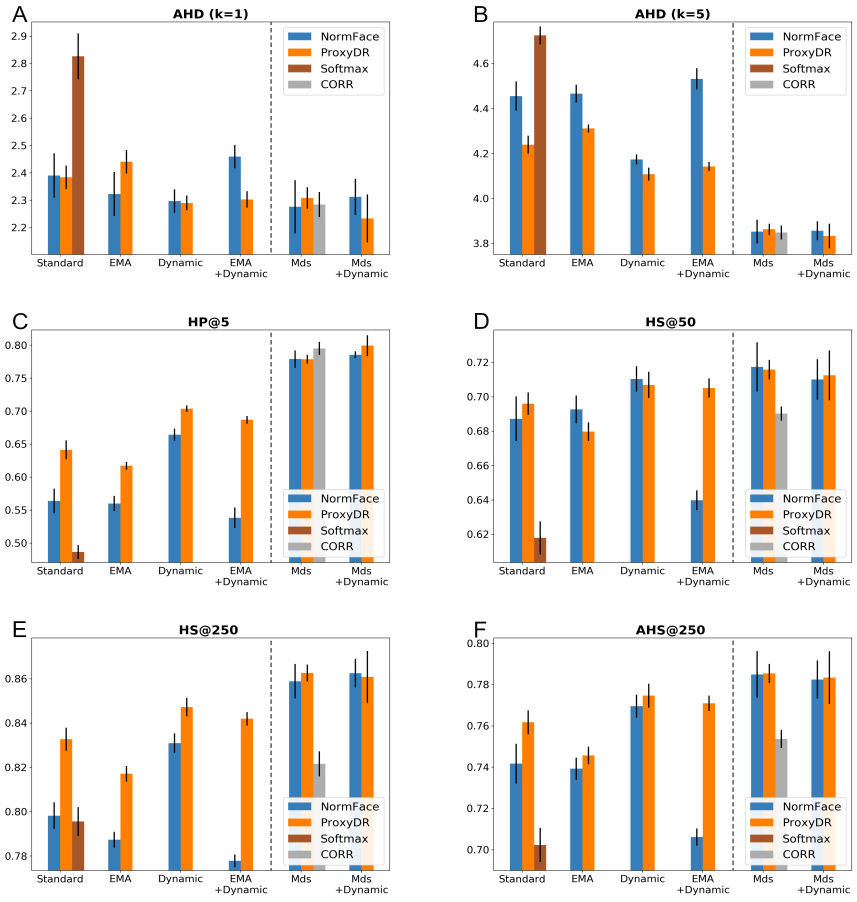



**Fig. 18:** Hierarchical performance measures on the NABirds dataset. The symbol ↓ denotes that lower values indicate better performance. The symbol ↑ denotes that higher values indicate better performance. (A) AHD (k=1): ↓. (B) AHD (k=5): ↓. (C) HP@5: ↑. (D) HS@50: ↑. (E) HS@250: ↑. (F) AHS@250: ↑.

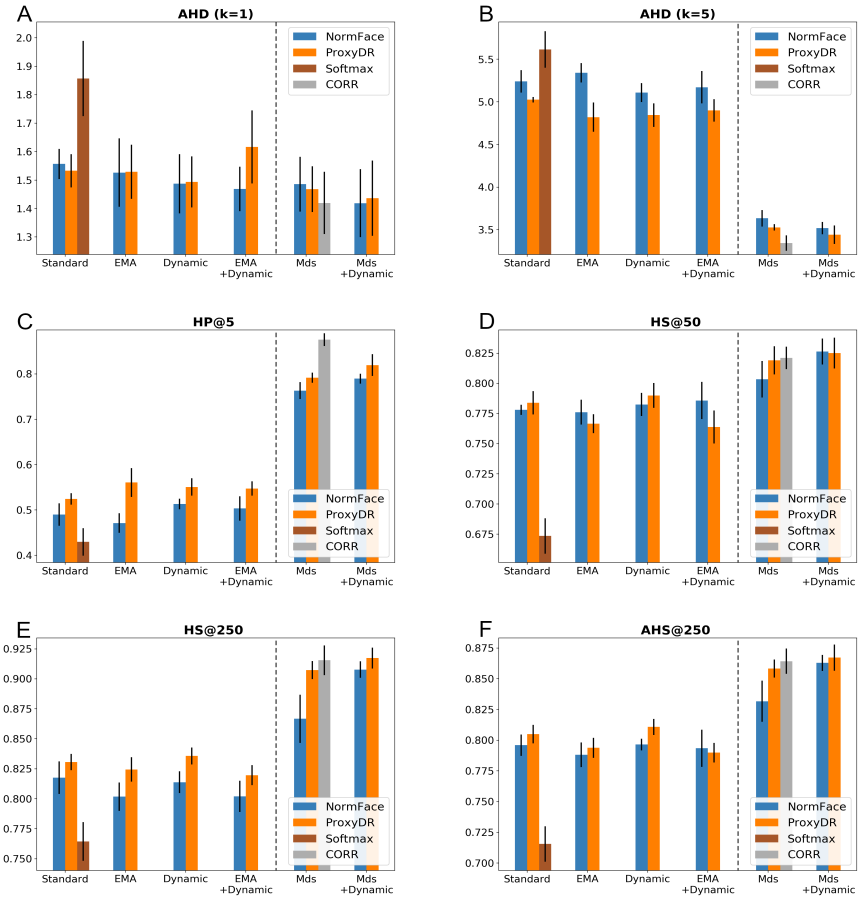



**Fig. 19:** Hierarchical performance measures on the MicroS dataset. The symbol ↓ denotes that lower values indicate better performance. The symbol ↑ denotes that higher values indicate better performance. (A) AHD (k=1): ↓. (B) AHD (k=5): ↓. (C) HP@5: ↑. (D) HS@50: ↑. (E) HS@250: ↑. (F) AHS@250: ↑.

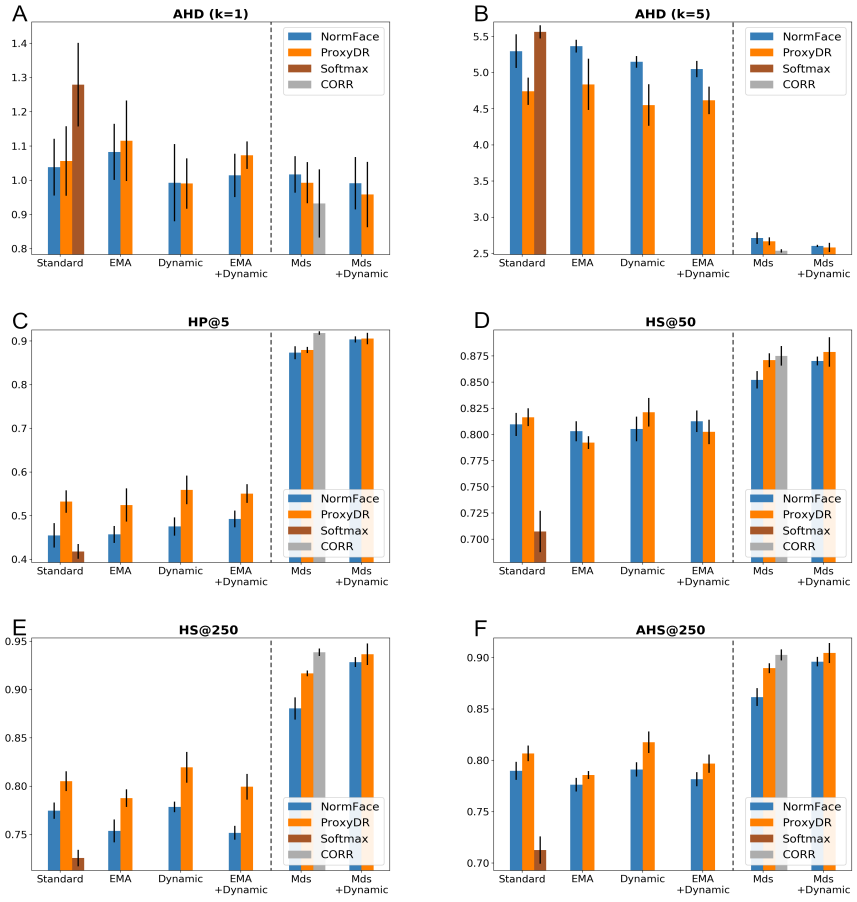



**Fig. 20:** Hierarchical performance measures on the MicroL dataset. The symbol ↓ denotes that lower values indicate better performance. The symbol ↑ denotes that higher values indicate better performance. (A) AHD (k=1): ↓. (B) AHD (k=5): ↓. (C) HP@5: ↑. (D) HS@50: ↑. (E) HS@250: ↑. (F) AHS@250: ↑.

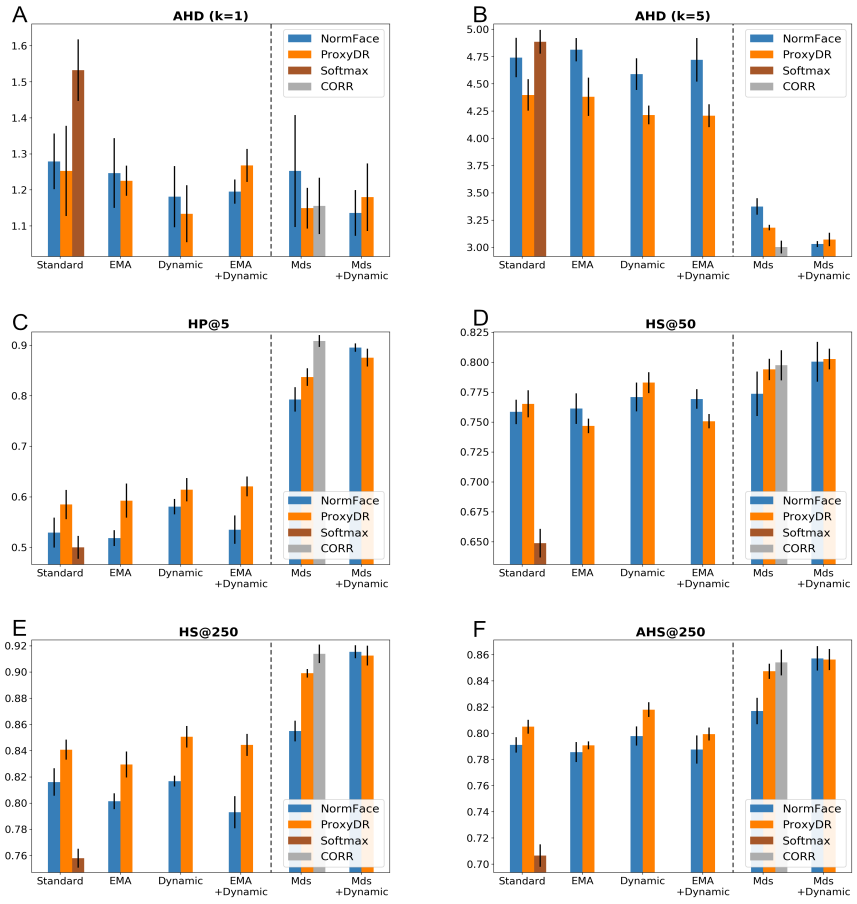



**Fig. 21:** Hierarchical performance measures on the MesoZ dataset. The symbol ↓ denotes that lower values indicate better performance. The symbol ↑ denotes that higher values indicate better performance. (A) AHD (k=1): ↓. (B) AHD (k=5): ↓. (C) HP@5: ↑. (D) HS@50: ↑. (E) HS@250: ↑. (F) AHS@250: ↑.

Figs. 22, 23, 24, 25, and 26 show the top-$k$ accuracies on the different datasets.

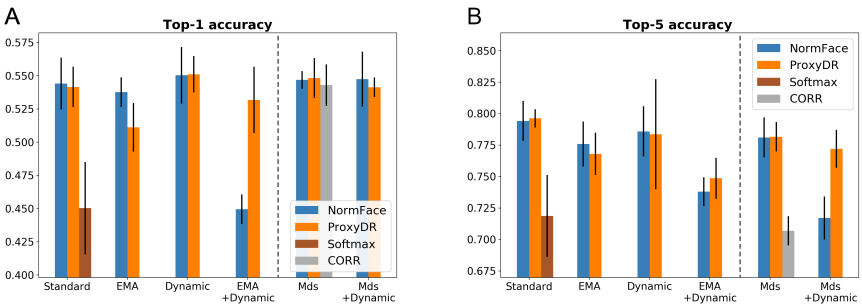



**Fig. 22:** Top-$k$ accuracy results (A: $k = 1$, B: $k = 5$) on the CIFAR100 dataset

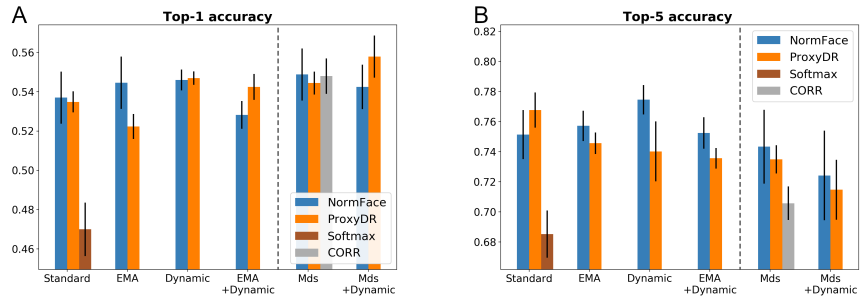



**Fig. 23:** Top-$k$ accuracy results (A: $k = 1$, B: $k = 5$) on the NABirds dataset

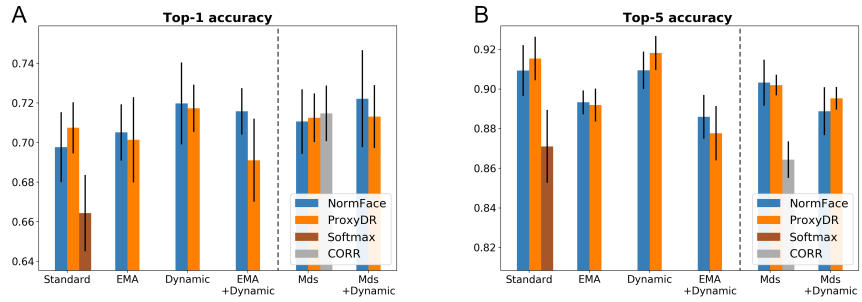



**Fig. 24:** Top-$k$ accuracy results (A: $k = 1$, B: $k = 5$) on the MicroS dataset

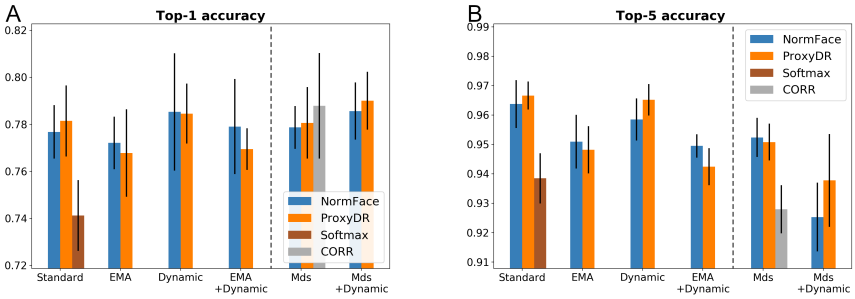



**Fig. 25:** Top-$k$ accuracy results (A: $k = 1$, B: $k = 5$) on the MicroL dataset

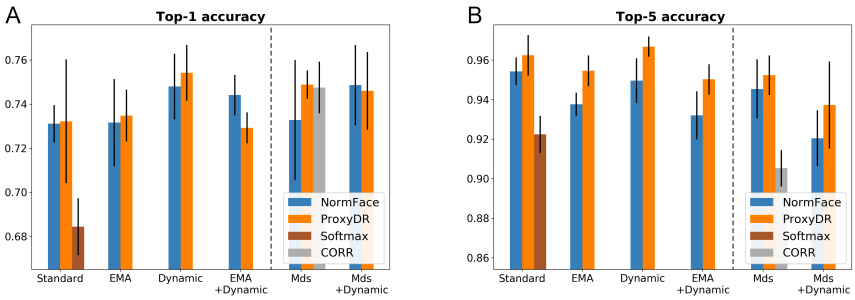
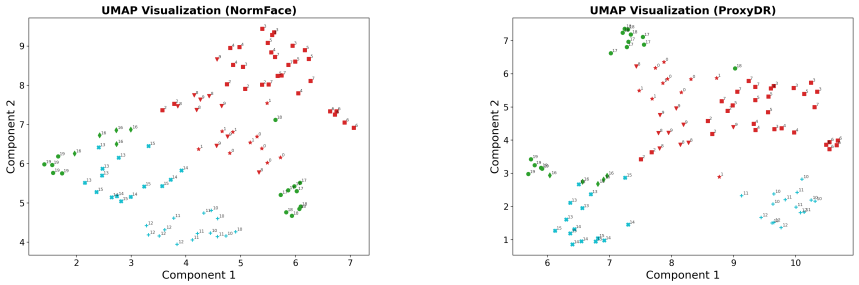



**Fig. 26:** Top-$k$ accuracy results (A: $k = 1$, B: $k = 5$) on the MesoZ dataset




**(a)** NormFace **(b)** ProxyDR

**Fig. 27:** UMAP visualization [20] of class proxies with dynamic scale factor training on the CIFAR100 dataset

### A.5 Visualization of Class Proxies

Fig. 27 visualizes learned class proxy representatives with dynamic scale factor training for the CIFAR100 dataset. Colors represent very coarse-level categorization (level 0 in Table 2). Shapes represent coarse-level categorization (level 1). Numbers represent less coarse-level categorization (level 2). These numbers are based on the represented ordering in Table 2. The visualization shows that ProxyDR allows for more varied proxy distances and learns more distinctive class structures, which aligns with the analysis in Secs. 3.3 and 4.2.