I Spy With My Little Eye
A Minimum Cost Multicut Investigation of Dataset Frames

— Supplemental Material —

A. Embedding Models

For all models, we use the recommended pre-training procedure without any modifications. The CLIP model is pre-trained on a web-scale dataset consisting of scraped image-text pairs. DINOv2 is pretrained on a combination of curated datasets. All details can be found in Tab. 3. If part of the architecture, the classification head is replaced by torch.nn.Identity().

Model	Architecture	Pre-training	Source
CLIP RN-50	ConvNet	openai	https://github.com/mlfoundations/open_clip
CLIP ViT-B/32	ViT	openai	https://github.com/openai/CLIP
ConvNeXt V2	ConvNet	ImageNet-1K	transformers.ConvNextV2ForImageClassification
DINOv2	ViT	custom	https://github.com/facebookresearch/dinov2
ResNet-50	ConvNet	IMAGENET1K_V2	torchvision.models.resnet50
ViT-B/32	ConvNet	IMAGENET1K_V1	torchvision.models.vit_b_32
VGG19-BN	ConvNet	IMAGENET1K_V1	torchvision.models.vgg19_bn
Inception-ResNet-V2	ConvNet	IMAGENET1K_V1	timm

Table 3. Details and sources for the employed embedding models.

A.1. Distributions

We ablate whether the mean of the normalized s_c distribution is decisive for selecting the optimal cal. During the normalization, the data mean may shift, depending on the distribution. Fig. 8 shows all normalized probability density plots and compares it to several data distributions. We compare the clusterings for the same cal as for ImageNette and $cal = 0.94 - 0.98 * \mu$. The normalized μ are visualized in Fig. 9.

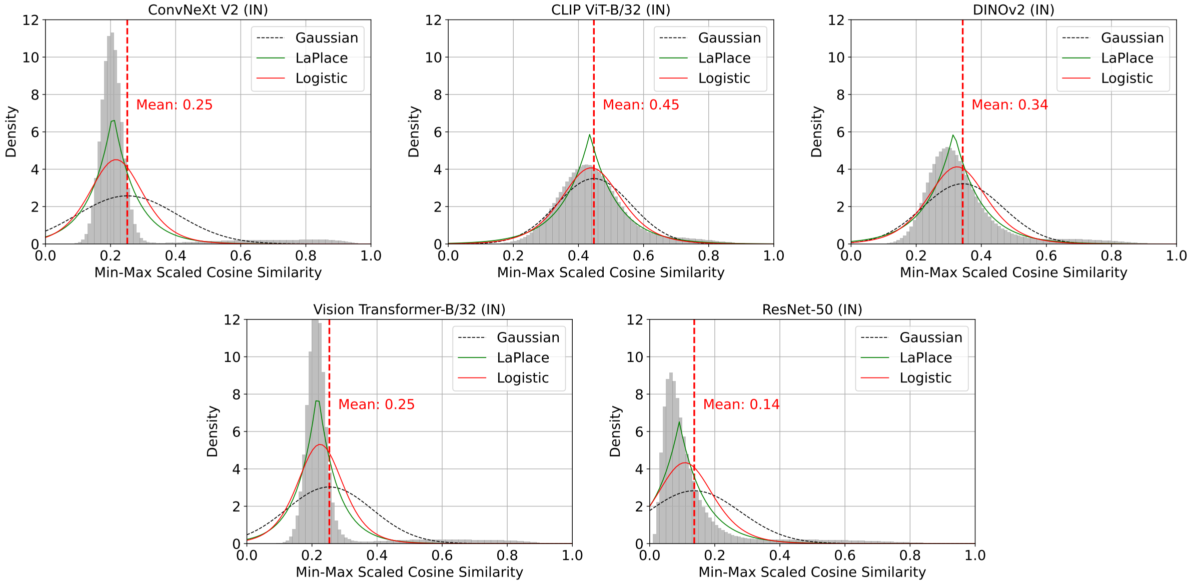


Figure 8. Distributions and mean for normalized ImageNette

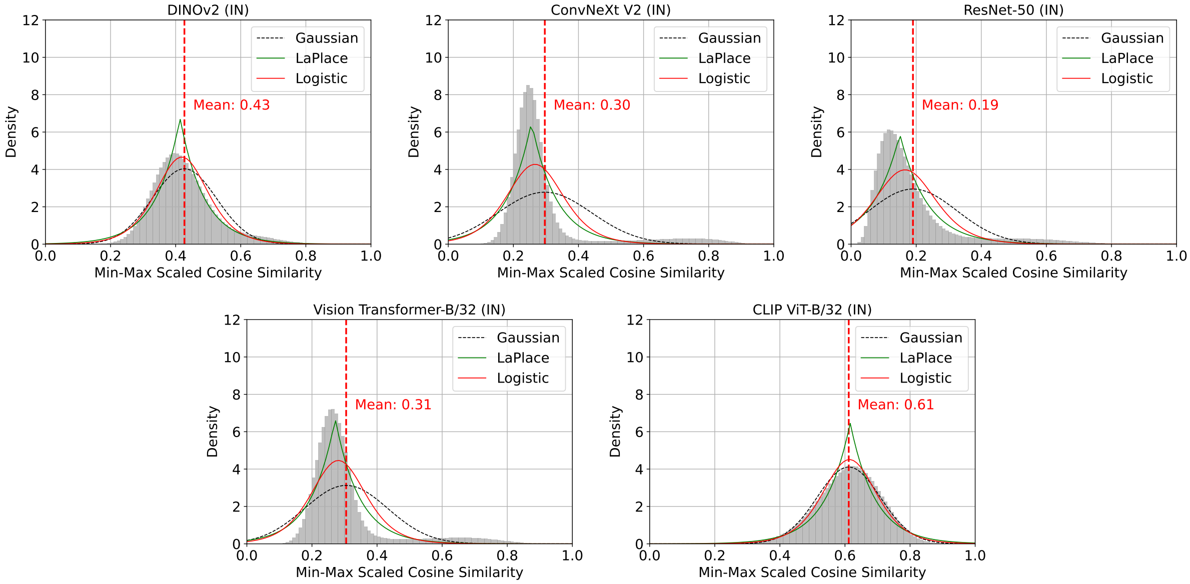


Figure 9. Distributions and mean for normalized ImageWoof

B. Embedding Space Analysis

Fig. 10 shows how the images' cosine similarities are distributed.

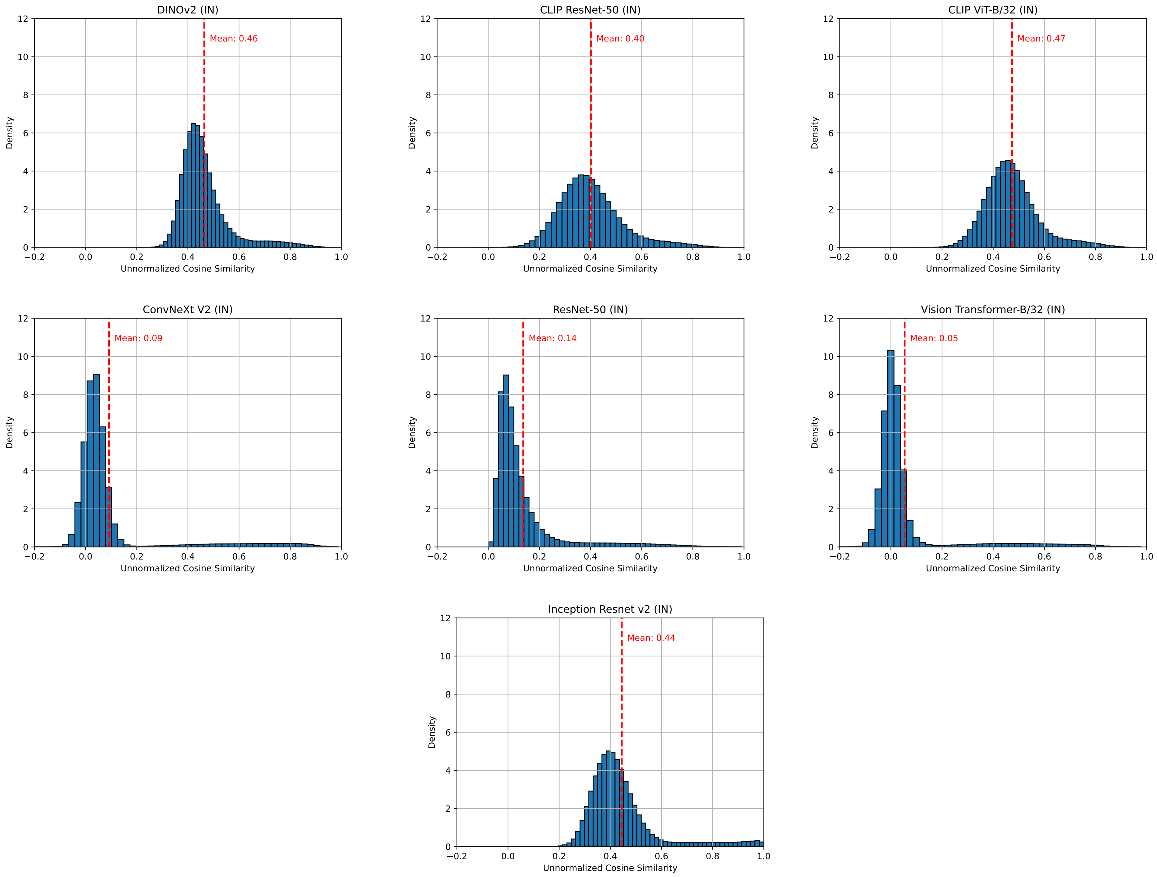


Figure 10. Probability density functions of un-normalized cosine similarity across embedding spaces for ImageNette. All distributions have a long tail towards the high cosine similarities. The *narrow cone effect* can be observed for all models in the top row.

C. Solver ablation

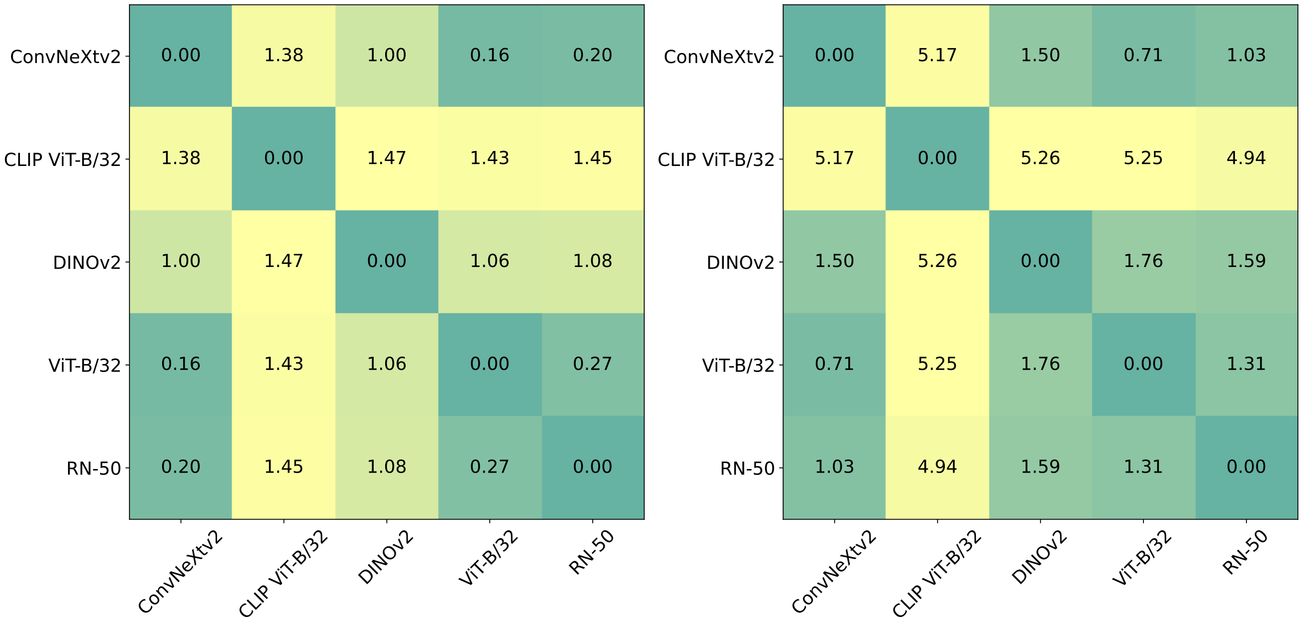
Embedding space	Costs		Number of clusters	
	GAEC	GAEC + KL	GAEC	GAEC + KL
ConvNeXt V2	-116×10^6	$\mathbf{-125 \times 10^6}$	325	**281**
CLIP ViT-B/32	-116×10^6	$\mathbf{-117 \times 10^6}$	1016	**939**
DINOv2	-141×10^6	$\mathbf{-147 \times 10^6}$	114	**102**
ViT-B/32	-150×10^6	$\mathbf{-159 \times 10^6}$	459	**404**
ResNet-50	-286×10^6	$\mathbf{-309 \times 10^6}$	265	**249**

Table 4. On ClimateTV, we ablate the efficacy of using KL in conjunction with GAEC. It yields lower costs and fewer clusters in every case.

D. Results

D.1. ImageNette and ImageWoof

ImageNette's cluster sizes vary greatly, with the largest spread of ConvNeXt V2, followed by ResNet-50. All model's median cluster sizes is below 5. Most models have few outliers outside the quartiles, except for ConvNeXt V2. All models with a narrow cone produce highly mixed clusters, while the other models have clusters sizes up to the class size (red line). While most model's mean cluster size is between 298 and 362, the CLIP model's is 119, as it contains the largest number of single image clusters *i.e*. 0.3% of all images. The majority (84%) of ConvNeXt V2 image clusters contain images from a single class. The mixed clusters are all combinations of 2 classes, which contain 1 or 2 mis-clustered samples. While 71% of all images are clustered in accordance with their class membership, few images have been added to incorrect classes. We evaluate the clusterings in terms of their VI. Fig. 11 compares the VI of the embedding clusterings to each other, indicating the disagreement between the different embedding spaces in terms of image similarity. For both datasets, the same trends are visible. CLIP Vit-B/32 is the most distinct from the other clusterings, while ConvNeXt V2 and ViT-B/32 share the highest similarity. Overall, the differences in VI are smaller for the more distinct classes in ImageNette compared to the fine-grained classification classes in ImageWoof.

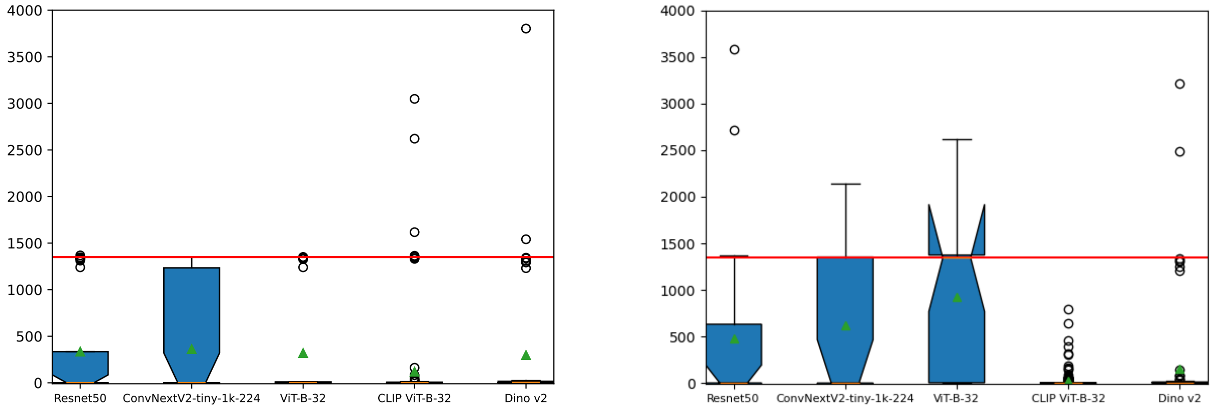


(a) VI per Embedding for ImageNette.

(b) VI per Embedding for ImageWoof.

Figure 11. When comparing the VI per clustering, the CLIP ViT-B/32 is most distinct from the other models in both cases. This effect is even stronger in the more divere ImageWoof dataset.

Fig. 12a shows the ImageNette cluster sizes for the optimal *cal*. In contrast to ImageWoof (Fig. 12b, the ViT-B/32 cluster sizes are much less distributed. For ImageWoof, cluster sizes are generally larger and also ResNet-50 and ConvNeXt V2 clusters now also exceed the number of images per class. The same trends can be observed for the other models.



(a) ImageNette cluster sizes are mainly smaller than the class size.

(b) ImageWoof cluster sizes highly vary.

Figure 12. Visualization of cluster sizes for embedding spaces employed (red line indicates class size). Overall, Resnet-50 and ConvNeXt V2 return larger clusters compared to CLIP ViT-B 32. Cluster sizes for the vision transformer (ViT-B 32) differ highly between datasets.

D.2. ClimateTV

When comparing the clustering performance for ClimateTV, the differences between the embedding models become larger in terms of VI. The greatest similarity can be found between the DINOv2 and ResNet-50 model. Again, the CLIP clustering appears to be most dis-similar to the other clusterings, as shown in Fig. 13.

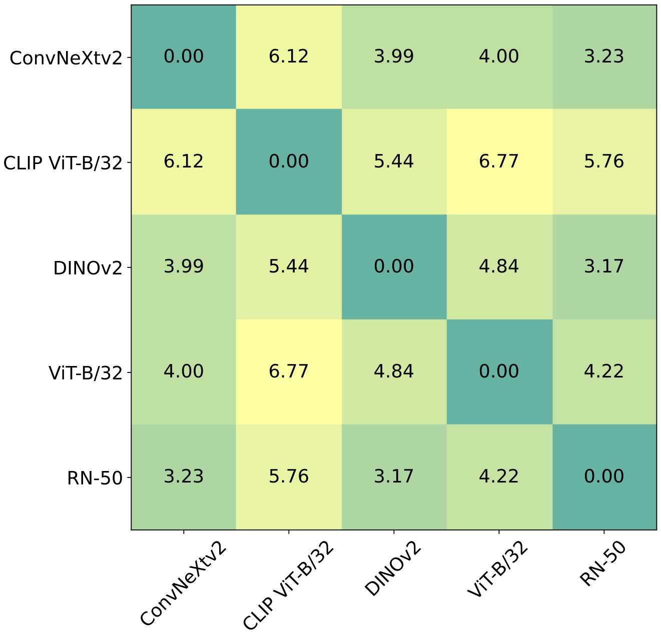


Figure 13. VI Heat Map for ClimateTV clusterings.

ClimateTV produces clusters of up to 17k images. All embedding clusterings have a handful of large clusters with a large majority below 500, as shown in Fig. 14. Tab. 5 contains the detailed statistics for all embedding clusterings. It stands out that CLIP ViT-B/32 has a large number of cluster of size 1. Since the dataset includes 37k images, this makes almost 0.4k images in single image clusters. Moreover, CLIP ViT-B/32 clusterings contain many small classes, as the median shows.

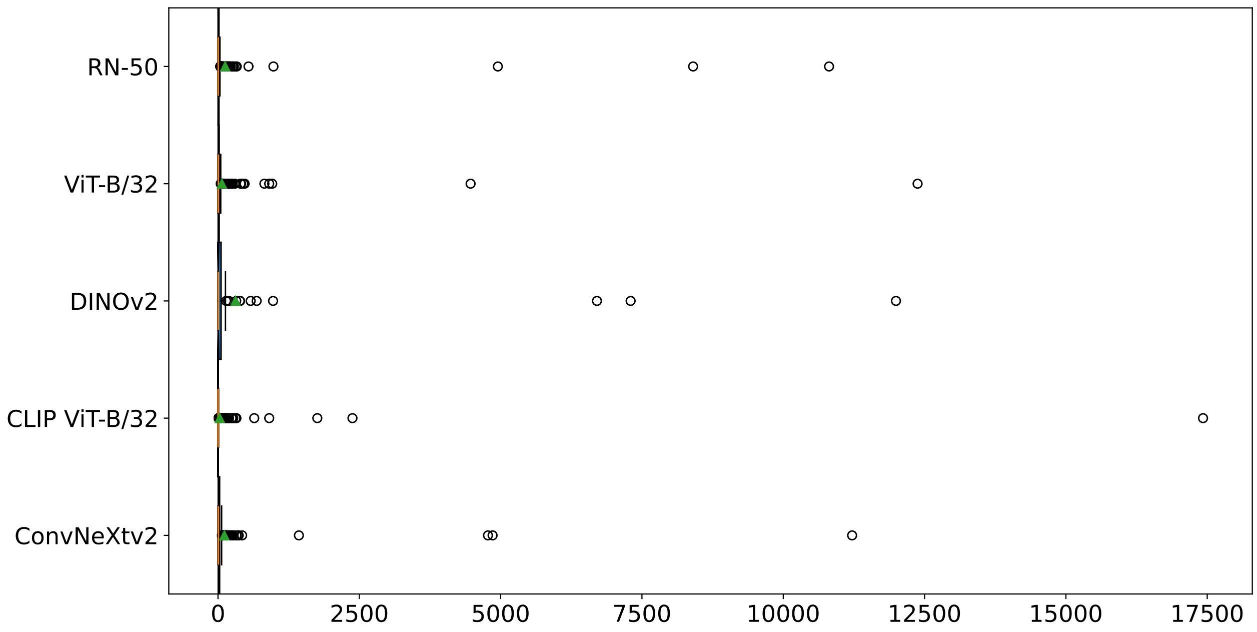


Figure 14. The comparison of ClimateTV cluster sizes shows that CLIP has by far the largest cluster. In contrast to the other models, CLIP's average cluster size is below 35.

Embedding	# clust	min. size	median size	mean size	max. size	% size 1
ConvNeXt V2	281	1	9	112.9	11,218	0.03%
CLIP ViT-B/32	1016	1	2	31.2	17,426	1.0%
DINOv2	102	1	6	310.9	11,994	0.06%
Vit-B/32	459	1	7	69.1	12,376	0.08%
ResNet-50	249	1	5	127.4	10,810	0.12%

Table 5. ClimateTV cluster statistics

We compare the largest clusters' contents for each embedding clustering in 6 It stands out that the largest CLIP ViT-B/32 class is very noisy. It contains nature or its products (*e.g*. fruits) in natural (*e.g*. arctic sea with iceberg), generate (*e.g*. visualization of the globe), industrial (*e.g*. wind energy plant), or catastropical (*e.g*. flooded city). It appeared highly unlikely that this many images share a common features. We are very satisfied with the ConvNeXt V2 clusters, as they contain specific semantic concepts. DINOv2 and CLIP embeddings also appear to contain semantic information, due to the higher specificity of their clusters compared to ResNet-50 and ViT-B/32. The ResNet-50 cluster content was most difficult to infer, as the models had visual similarity, but varied a lot. It appears that the ResNet-50 embeddings have a shape-bias. This could be an explanation for the *circles* cluster which contains numerous round object from buttons to the earth. The ViT-B/32 cluster content was quite abstract, with the common theme of *blue* in the largest cluster. One of CLIP embeddings strong suits is text understanding. One CLIP cluster which contains political news is combined with all other computer generated content by ConvNeXt V2. In contrast, ConvNeXt V2 can detect images that contain red soil or sand, which CLIP clusters in its large *nature* cluster. There are no two clusters of the two models which are alike, *i.e*. share 80% or more of images. The combination of the two clusterings appears less helpful. When looking at CLIP cluster for ConvNeXt V2's *polar bear* cluster, the single image clusters are less helpful in detecting outliers due to several of them containing polar bears. Moreover, the larger clusters are not semantically different from each other, so their benefit is small.

In the main paper we reported the biggest overlap between DINOv2 and ConvNeXt V2 clusters being *frogs*. We visualized the jointly clustered images in Fig. 15. Fig. 17 and Fig. 16 contain the images which have been included in the ConvNeXt V2 and DINOv2 cluster respectively.

	ConvNeXt V2	CLIP ViT-B/32	DINOv2	ViT-B/32	ResNet-50
#1 content (size)	comp. gen. (11,218)	nature (17,426)	humans (11,994)	blue color (12,376)	comp. gen. (10,810)
#2 content (size)	speaker (4,858)	Formal humans (2,379)	event info (7,300)	humans (4,469)	humans (8,405)
#3 content (size)	outdoor photo (4,776)	text w/ ”climate change” (1,757)	nature (6,704)	outdoor photo (959)	nature (4,951)
#4 content (size)	protest (1,431)	poster/presentation (906)	map/globe (974)	words (906)	circles (981)
#5 content (size)	portraits (427)	cold (641)	food (684)	humans w/ text (821)	nature w/ foreground (540)

Table 6. ClimateTV cluster size and content for its top 5 largest clusters per embedding model. comp. gen. = Computer generated content.

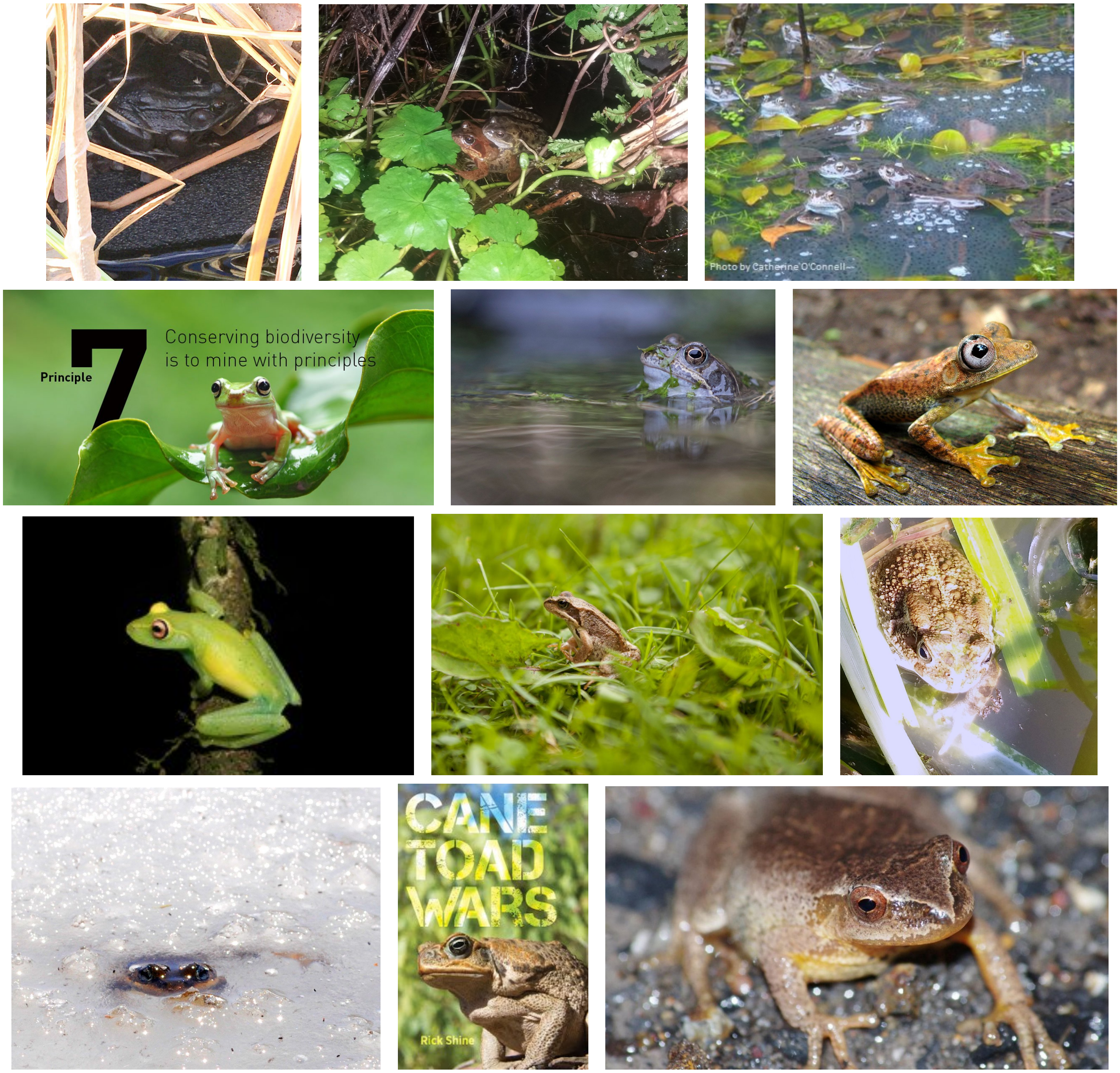


Figure 15. The *frog* cluster has the largest overlap between the clusterings of ConvNeXt V2 and DINOv2.

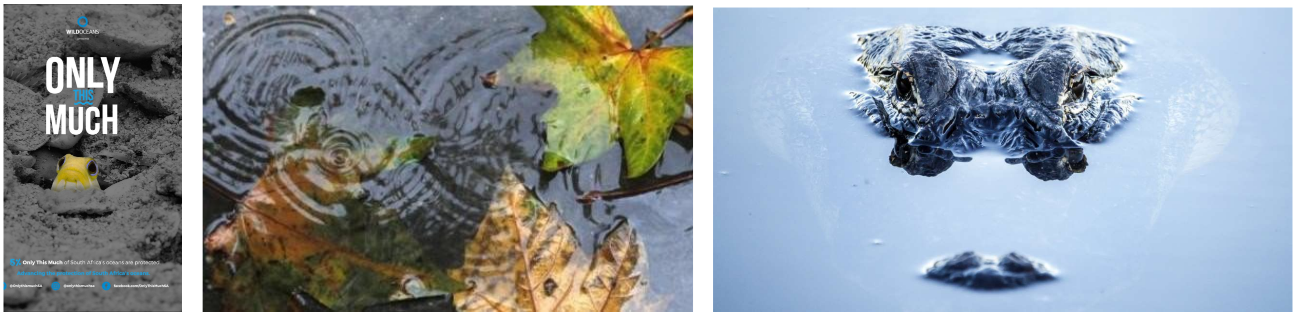


Figure 16. DINOv2's *frog* also contained 3 images that were not shared with the ConvNeXt V2 cluster.

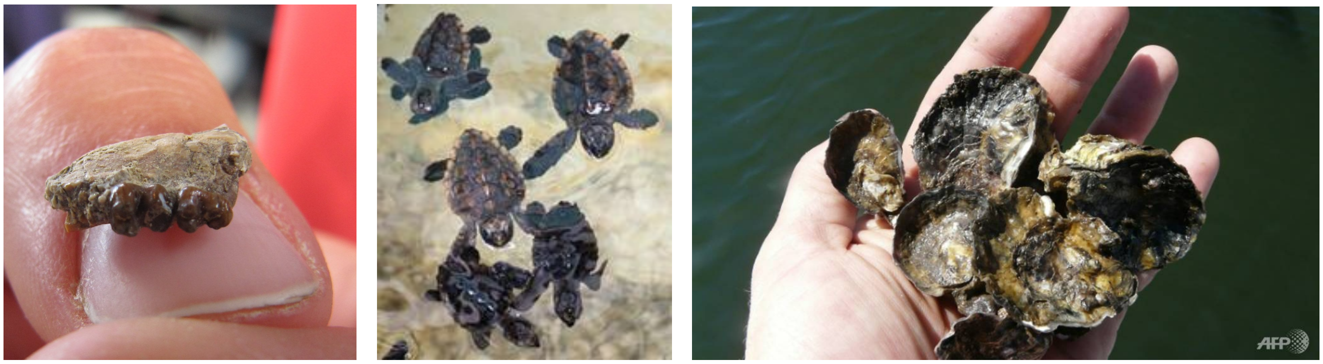


Figure 17. ConvNeXt V2's *frog* also contained 3 images that were not shared with the DINOv2 cluster.