Towards Fine-grained Visual Representations by Combining Contrastive Learning with Image Reconstruction and Attention-weighted Pooling

Jonas Dippel [1] Steffen Vogler [1] Johannes Höhne [1]

[1]Bayer AG

Abstract

- This paper presents Contrastive Reconstruction, ConRec - a self-supervised learning algorithm that obtains image representations by jointly optimizing a contrastive and a self-reconstruction loss.
- state-of-the-art contrastive learning methods (i.e. SimCLR) have shortcoming with regard to fine-grained classification tasks.
- ConRec tackles these shortcomings and extends the SimCLR framework by adding (1) a self-reconstruction task, (2) an attention mechanism within the contrastive learning task.
- This is accomplished by applying a simple encoder-decoder architecture with two heads.
- We show that both extensions contribute towards an improved vector representation for images with fine-grained visual features.

Augmentations

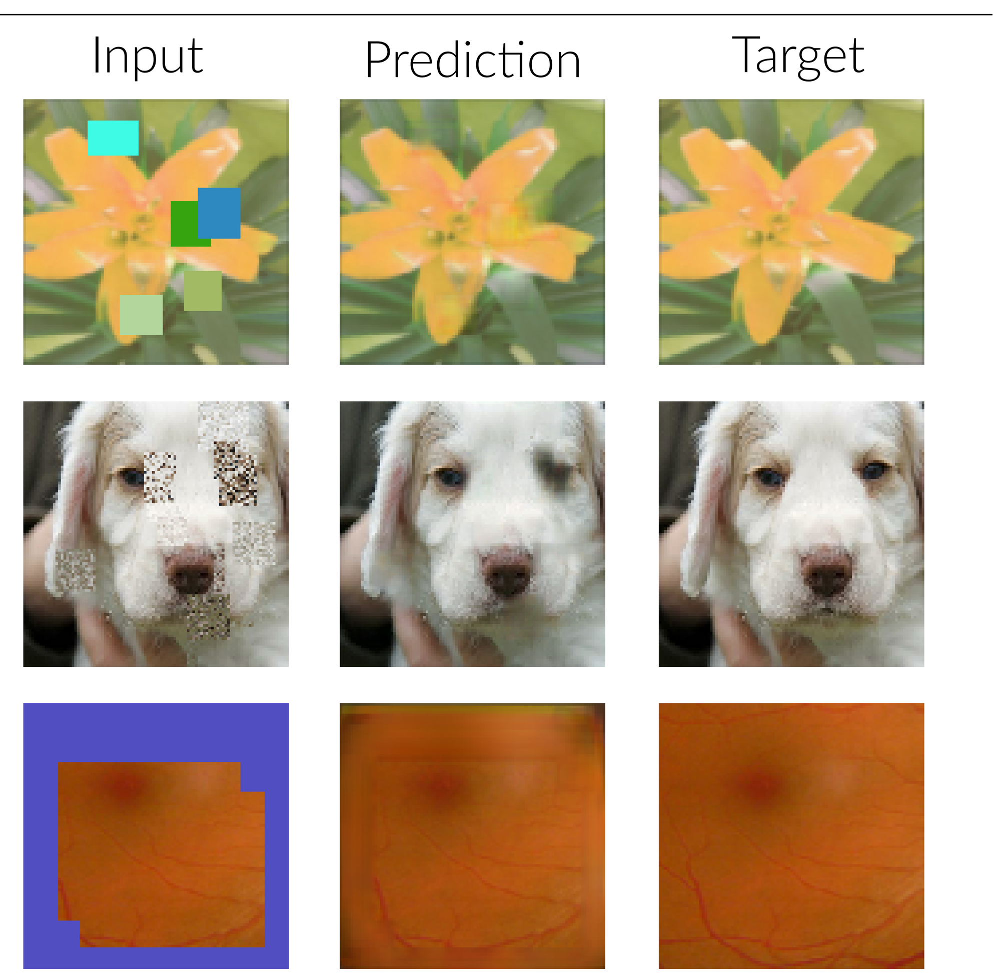


Figure 1:Augmented images and respective reconstruction predictions by our ConRec model.

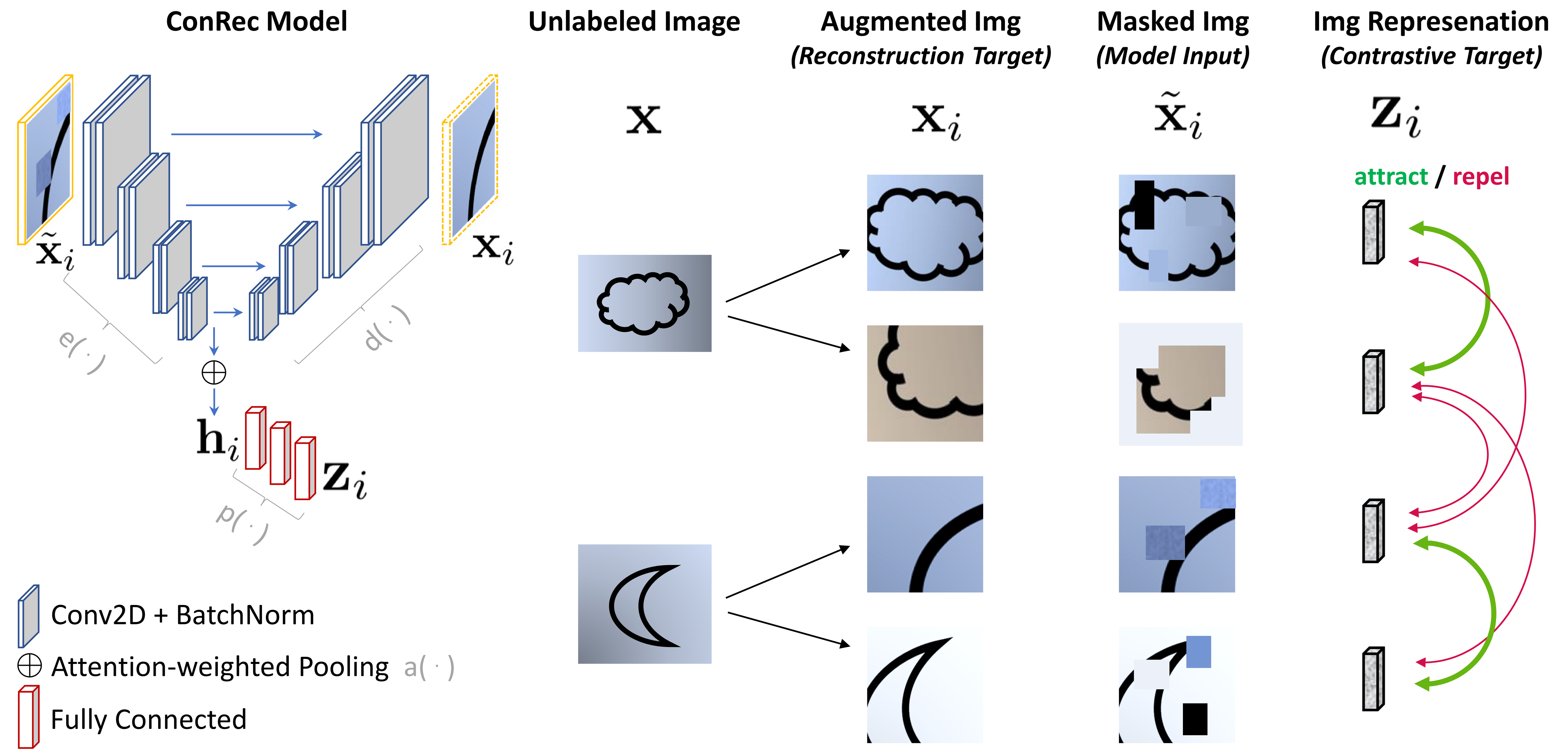


Figure 2:Learning Framework for Contrastive Reconstruction – ConRec. The ConRec model consists of a fully convolutional encoder-decoder architecture with skip connections as well as a projection head comprising fully connected layers. The model receives a masked image $\tilde{\mathbf{x}}_i$ and outputs the unmasked reconstruction target $\mathbf{x}_i$ as well as the contrastive image representation vector $\mathbf{z}_i$.

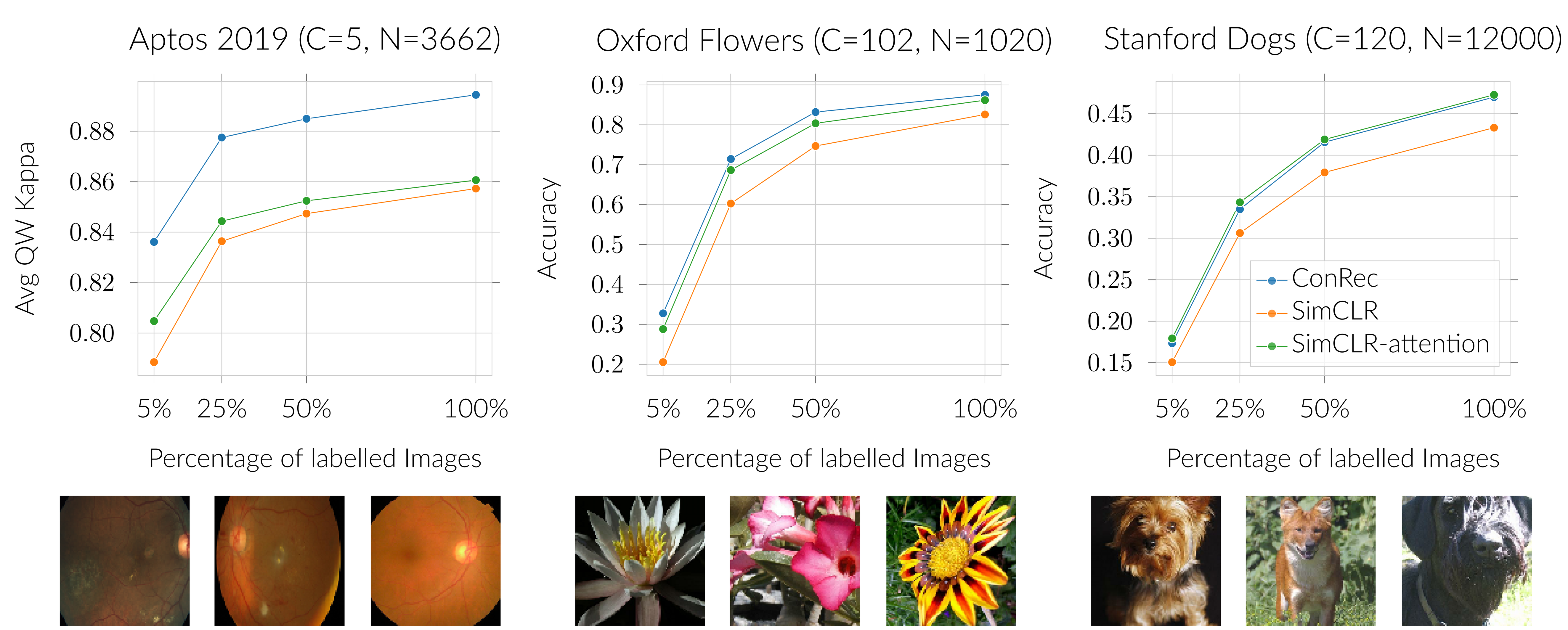


Figure 3:Model accuracies for training a linear classifier on a subset (1% to 100%) of the labeled representations with different number of classes C and number of samples N.

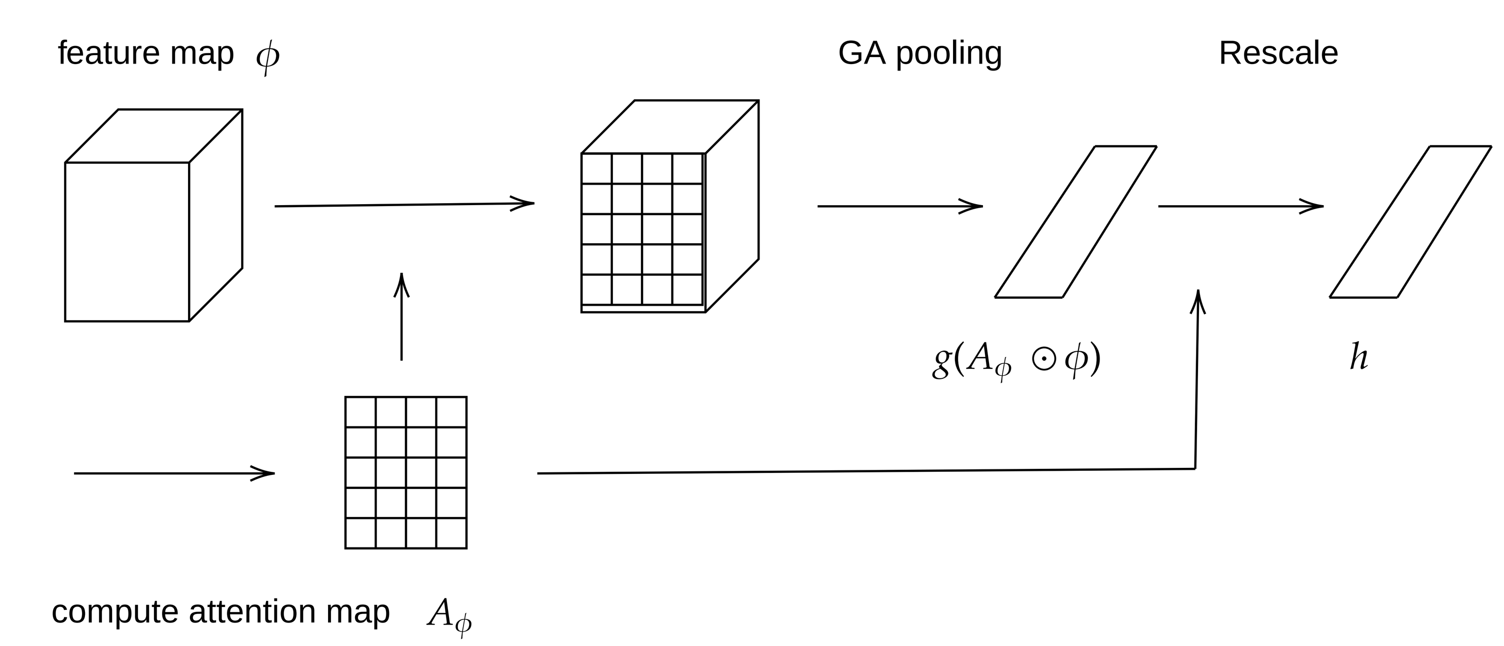
Method

In the training process, the model receives a masked image $\tilde{x}_i$ and outputs the reconstructed image $x_i = d(e(\tilde{x}_i))$ as well as the contrastive vector representation $\mathbf{z}_i = p(a(e(\tilde{x}_i)))$. The training loss is composed of two parts: the contrastive loss $\mathbf{L}_c$ and the reconstruction loss $\mathbf{L}_r$.

$$\mathbf{L}_{ConRec} = \mathbf{L}_c + \alpha * \mathbf{L}_r$$

Attention Pooling

- Global average pooling discards some local features in the encoder output activation map, which may carry relevant fine-grained information.
- We introduce an attention weighted pooling mechanism that aggregates the spatial content of the final feature map of the encoder in a parametric manner.



Results

Model	Frozen	Aptos	Flowers	Dogs	#Params
SimCLR U-net	✓	85.72	86.01	43.96	4.693M
SimCLR Attention U-net	✓	86.06	88.37	50.31	4.867M
ConRec U-net	✓	89.44	90.29	49.57	4.867M
DenseNet121 (ImageNet)	✓	86.70	(92.97)	(88.07)	8.062M
U-net (Random)		82.11	81.54	55.2	4.693M
DenseNet121 (Random)		64.53	82.03	57.63	8.062M

Table 1:Linear evaluation results and respective baselines. ImageNet results in parenthesis indicate flaws in the evaluation as the datasets were included in supervised ImageNet-pretraining.