Supplementary Material
LEAD: Self-Supervised Landmark Estimation by Aligning Distributions of Feature Similarity

In this work, we present landmark prediction using our new method LEAD on a variety of face datasets. Two main assumptions in the problem setting that we presented were: 1) Availability of large-scale unannotated facial dataset for self-supervised pretraining, and 2) Availability of annotations for small scale supervised training. The reasoning behind such a setting is abundance of unannotated images available on the internet which can be leveraged to learn better features, while annotated images are hard to procure for training. Out of these two assumptions, we addressed the second one in the paper. In this work, we address the first assumption. Additionally, we examine the features from the lens of interpretability and present results on the "bird" category.

The rest of supplementary work is organized as following: we first present the results of using in-the-wild face images for self-supervised pretraining (in Sec. 1). Additionally, we show LEAD's efficacy on a more challenging bird landmark prediction (in Sec. 2). We then present the interpretability of features learnt by our model by part-discovery (in Sec. 3), followed by the implementation details in Sec. 4. Lastly we discuss interpetability of denser intermediate outputs of our model by clustering (in Sec. 5), followed by some additional visuals for the scale ablation as performed in the main paper (Sec. 6).

1. In-the-wild pretraining

Figure 1. Samples from CelebA In-the-wild

While we use large-scale unannotated dataset for pretraining, it is still a well-cropped data where the training images were completely occupied by the face. In this experiment, we use uncropped in-the-wild images from the CelebA [5] dataset. Samples from the in-the-wild CelebA dataset are shown in fig. 1.

We refer the reader to table 1 for the quantitative results in this setting. The pretraining on this dataset proves to be beneficial in some evaluation datasets, while still giving improvements over others. Pretraining on CelebA In-the-wild shows impressive results on AFLW$_R$ and 300W, which are better than those obtained by CelebA pretraining. For the remaining 2 datasets, the inter-ocular distance slightly increases compared to that of celebA pretraining, but is still better than the prior arts. This performance can be reasoned as the in-the-wild images have large variations between different regions in the image. As every time when the crops of image are taken as augmentation, each crop provides the network with a variety of image distributions and statistics to learn from. While in case of 'cropped celebA' pretraining, a large area of the image is occupied face, hence there is less variety in the kinds of crops it results in.

We also evaluate this model on unaligned face images. These images use the same split as MAFL, but from CelebA in-the-wild. We call is MAFL in-the-wild. For evaluation in this setting we randomly sample a percentage between 10 to 20 to increase the height and width of the original face bounding box for each image, this bigger bounding box is then used to crop the respective images. This helps in breaking the alignment of faces, that is present in the cropped MAFL dataset. We report the results on this evaluation in Table 2 and Fig. 2. The inter-ocular distance on this dataset is compared with both DVE and ContrastLandmarks. Our method gives *9.75%* and *2.88%* better relative IOD then DVE and ContrastLandmarks respectively.

Table 1. Effect of pretraining on in-the-wild face dataset.

Method	Feat. dim	MAFL	AFLW_M	AFLW_R	300W
DVE [7]	64	3.23	8.52	7.38	**5.05**
CL [1]	64	**3.00**	7.87	6.92	5.59
LEAD (ours)	64	3.01	**6.81**	**6.42**	5.34
CL [1]	128	**2.88**	7.81	6.79	**5.37**
LEAD (ours)	128	3.07	**6.79**	**6.38**	5.56
CL [1]	256	**2.82**	7.69	6.67	**5.27**
LEAD (ours)	256	3.09	**6.85**	**6.22**	5.53
CL [1]	3840	**2.46**	7.57	6.29	5.04
LEAD (ours)	3840	**2.46**	**6.48**	**5.64**	**4.47**

Table 2. Evaluation on MAFL in-the-wild.

Method	Feat. Dim.	IOD
DVE [7]	64	4.51
LEAD (Ours)	64	**4.07**
CL [1]	3840	3.12
LEAD (Ours)	3840	**3.03**

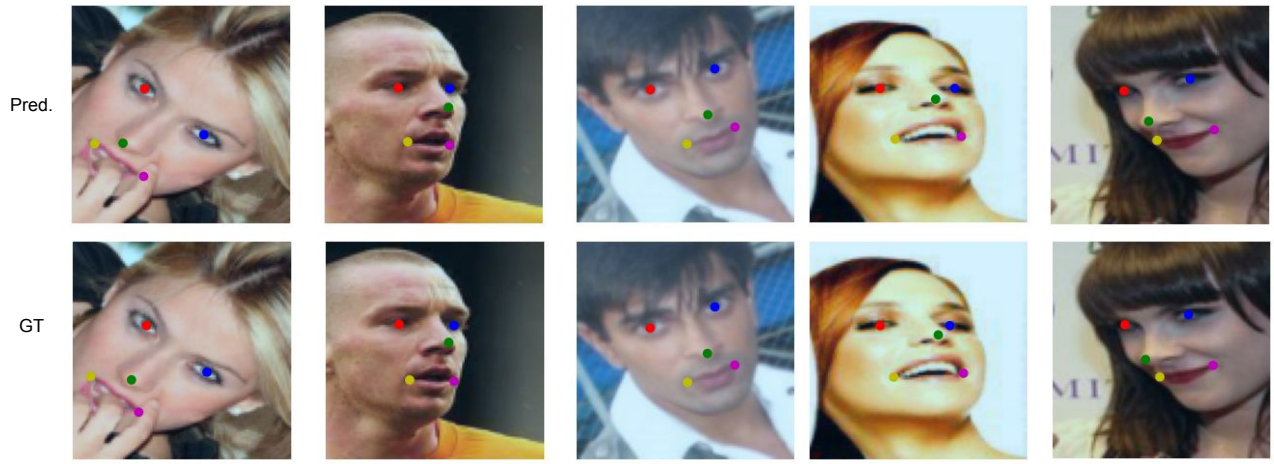

Figure 2. Landmark prediction on unaligned MAFL (in-the-wild)

2. Bird Landmark Prediction

We demonstrate the efficacy of our method on challenging bird landmark prediction task. We show the results in Fig. 3. In this task, we train the instance-level model on iNat17-Aves dataset, which contains in-the-wild images of "Aves" class from iNaturalist 2017 dataset [8]. This is followed by supervised training on a subset of CUB [9] dataset, containing 35 species from Passeroidea super-family. CUB images have 15 annotated landmarks. For both the datasets, we use the same split as [1]. We compare the percentage of correct keypoints. For PCK computation, a prediction is considered to be correct if its distance from the ground-truth keypoint is within 5% of the longer side of the image. Occluded keypoints are ignored during the evaluation. We report a competitive PCK of 67.3%, which we compare against ContrastLandmarks [1] in Table 3.

Table 3. PCK (Percentage of correct keypoints) on CUB dataset

Method	Feat. Dim.	PCK
CL[1]	3840	68.63
LEAD (Ours)	3840	67.31

3. Part Discovery

We perform deep feature factorization [2] on the representation obtained from stage 1 of training for part discovery. It can be noted in Fig. 4 that parts discovered by this method are consistent across instances. Intensity of the color denotes the

Figure 3. Landmark prediction on birds from CUB dataset

presence of the part. It is interesting to observe the intensity of the color corresponding to the hair is very less in case of the fourth and fifth instance in Fig. 4, where the person is wearing a cap and has less hair respectively.

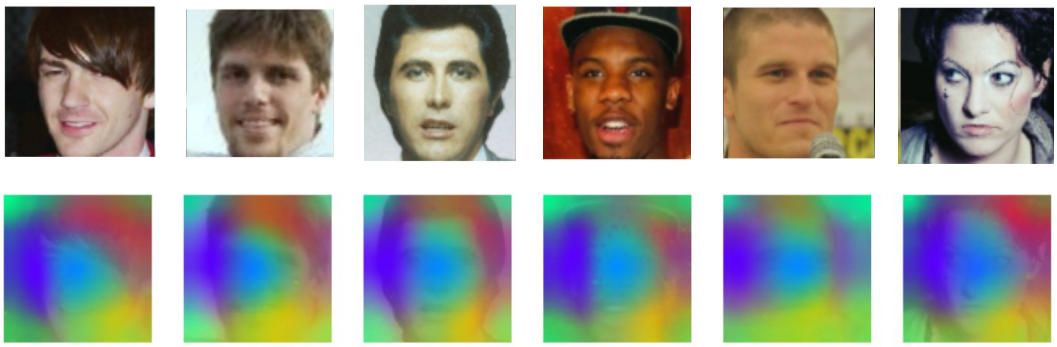

Figure 4. NMF part clustering [2] of learnt embeddings. Each color represents a discovered part.

4. Implementation details

Stage 1: Instance-level training. We use BYOL [3] to train the unsupervised representations. We train BYOL for 200 epochs with a batch size of 256 and use a cosine learning rate scheduler [6], with a warm-up period of 2 epochs. We follow the the augmentation pipeline as proposed in BYOL, wherein we use solarization as an augmentation only for the target encoder. We use the publicly available BYOL implementation from OpenSelfSup[1]. For comparison with DVE [7] and ContrastLandmarks [1], we use the released pretrained models from the respective official code repositories.

Table 4. Comparison of supervised training speeds at differ feature dimensions. Note that hypercolumn features (3840 feat. dim.) are 55× slower.

Feat. Dim.	FLOPS
3840	2.21
256	0.16
128	0.08
64	0.04

Stage 2: Dense training. Here we train an FPN decoder [4] while keeping the learned backbone encoder to be frozen. We train the decoder for 10 epochs with a batch size of 256, using a cosine learning rate scheduler with a warm-up of 2 epochs, similar to stage 1. We again follow the BYOL augmentation pipeline for training. We set the temperature τ to be 0.05 (Refer to Eq. 2 in the main paper).

Training of Supervised Landmark Regressors. For a fairer comparison we train the supervised regressors with frozen feature extractor exactly as proposed in [1]. We also show a comparison of supervised training speeds against different feature dimensions in Table 4.

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

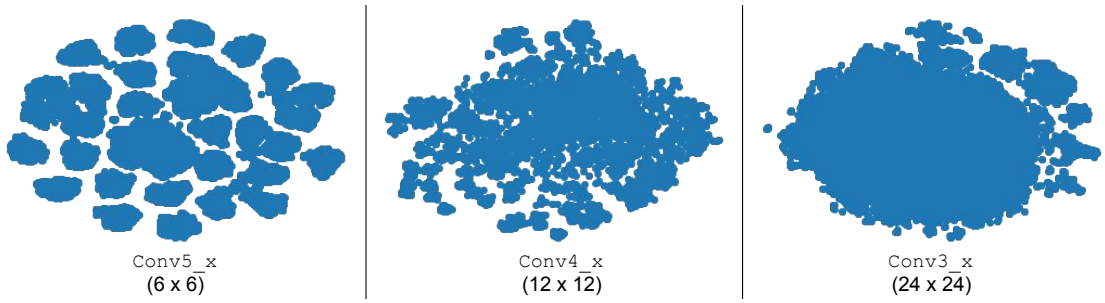

Figure 5. t-SNE plots of the intermediate layers' feature maps obtained after training (stage 1 of our model). Spatial dimension of the feature map indicated in brackets.

Figure 6. t-SNE embeddings tend to cluster part-wise. Each row shows patches from a cluster obtained from higher spatial resolution features (6×6 in this case). The 30 clusters shown corresponds to 30 clusters in Fig. 5. Each row denotes a cluster, which contains patch of semantically meaningful part of face. (Mouth: clusters 5, 13; Left and right eye: cluster 7 and 23, Nose: clusters 10, 15; Left and right jaw: clusters 18 and 12. Other clusters contain less discriminative parts like cheeks and forehead.)

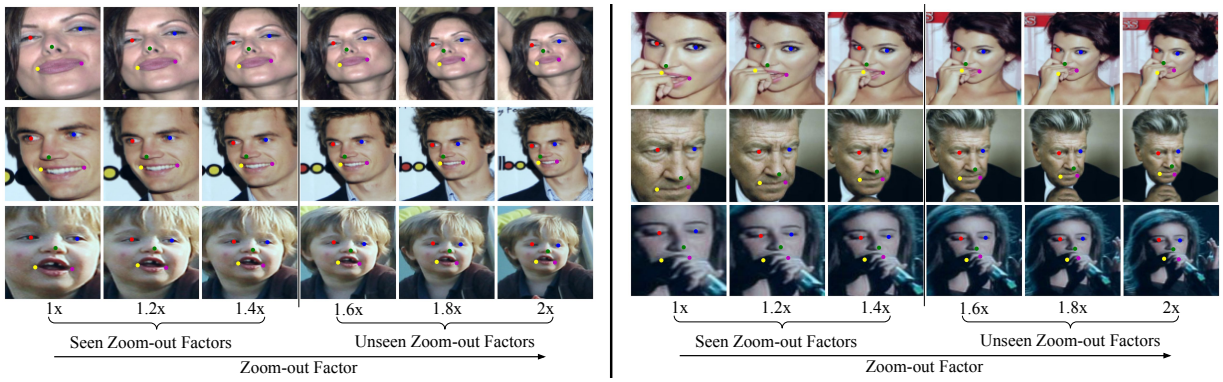

Figure 7. Additional Results on Scale ablation on seen scale (zoom-out factor ∈ 1-1.5x) and unseen scale (zoom-out factor ∈ 1.5-2x) variations of the faces from unaligned-MAFL dataset

5. Interpretability Analysis

We further examine the interpretabilty of the t-SNE embeddings obtained by clustering intermediate representations of the resnet model thereby capturing denser grids ($6 \times 6, 12 \times 12$ and 24×24). Fig. 5 shows emergence of semantically meaningful clusters in the 6×6 grid, Fig. 6 shows the corresponding parts captured by these clusters. It can also be observed that for further denser grids (12×12 and 24×24) we see the emergence of one big cluster which cannot be easily split into semantically meaningful regions. We further observed emergence of a single big cluster even after stage-2 training of our method, the possible cause of which may be the reduced dimensionality after stage 2 hurting the expressiveness as t-SNE embedding compared to the original hypercolumn ($3840D$).

6. Additional Visuals for Scale Ablations

We present additional examples of generalization of LEAD to seen and unseen scales of input face. These are shown in Fig. 7

References

[1] Zezhou Cheng, Jong-Chyi Su, and Subhransu Maji. Unsupervised discovery of object landmarks via contrastive learning. *arXiv preprint arXiv:2006.14787*, 2020.

[2] Edo Collins, Radhakrishna Achanta, and Sabine Susstrunk. Deep feature factorization for concept discovery. In *The European Conference on Computer Vision (ECCV)*, 2018.

[3] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised learning, 2020.

[4] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 936–944, 2017.

[5] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

[6] Ilya Loshchilov and F. Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv: Learning*, 2017.

[7] James Thewlis, Samuel Albanie, Hakan Bilen, and Andrea Vedaldi. Unsupervised learning of landmarks by descriptor vector exchange. In *International Conference on Computer Vision*.

[8] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8769–8778, 2018.

[9] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.