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

Tejan Karmali[1]* Abhinav Atrishi[1]* Sai Sree Harsha[1]
Susmit Agrawal[1] Varun Jampani[2] R. Venkatesh Babu[1]
[1]Indian Institute of Science, Bengaluru [2]Google Research

Abstract

In this work, we introduce LEAD, an approach to discover landmarks from an unannotated collection of category-specific images. Existing works in self-supervised landmark detection are based on learning dense (pixel-level) feature representations from an image, which are further used to learn landmarks in a semi-supervised manner. While there have been advances in self-supervised learning of image features for instance-level tasks like classification, these methods do not ensure dense equivariant representations. The property of equivariance is of interest for dense prediction tasks like landmark estimation. In this work, we introduce an approach to enhance the learning of dense equivariant representations in a self-supervised fashion. We follow a two-stage training approach: first, we train a network using the BYOL [13] objective which operates at an instance level. The correspondences obtained through this network are further used to train a dense and compact representation of the image using a lightweight network. We show that having such a prior in the feature extractor helps in landmark detection, even under drastically limited number of annotations while also improving generalization across scale variations.

1. Introduction

Image landmarks are distinct locations in an image that can provide useful information about the object, like its shape and pose. They can be used to predict camera pose using Structure-from-Motion [14]. Landmark detection is a well studied problem in computer vision [53, 46, 52, 51, 11, 23] that was initially accomplished using annotated data. Landmark annotation requires a person to accurately label the pixel location where the landmark is present. This makes annotation a laborious, biased, and ambiguous task, motivating the need for newer paradigms such as few-shot learning [54, 48, 35, 43] and self-supervised learning [10, 49, 28, 15, 13].

Prior works in self-supervised landmark detection rely on the principles of reconstruction [51, 26] and equivariance [37, 39]. These methods are trained using dense objectives that are satisfied by every pixel (or by every patch of pixels, due to downsampling). This tends to capture only local information around each pixel, and is unaffected by structural changes in the image (like patch shuffling).

Most of the existing research in the field of self-supervised learning is focused towards the task of instance-level classification. Amongst the proposed pretext tasks for self-supervision, instance-discriminative methods [15, 4, 5, 7, 3], are known to be superior for the purpose of pre-training. Recent methods utilize these objectives for dense prediction tasks as well, where a distinct label is predicted either for every pixel (segmentation, landmark detection) or patch of pixels (detection) [31, 30, 47, 42]. The power of contrastive training is leveraged for landmark detection by Cheng et al. [8] to achieve state-of-the-art performance using Momentum Contrast (MoCo)-style [15] pre-training. This work demonstrates equivariant properties in the network when trained with a contrastive objective. This property is realised by extracting a hypercolumn-style feature map from the image. But using such a high-dimensional feature map (3840d for ResNet50 due to stacking up of features), which is $60\times$ larger than existing approaches, to represent an image is not scalable to large images.

Our key insight is based on the observation that self-supervised training on category-specific datasets (dataset that consists of images that belong to only single category) leads to meaningful part-clustering in feature space. We further utilize this finding to propose a dense self-supervised objective for landmark prediction. Specifically, LEAD involves two stages: (1) Global representation learning, and (2) Correspondence-guided dense and compact representation learning. The network from stage 1 leads to meaningful part clustering in the feature space, and hence can be used to draw correspondences between two images. This can be used for pixel/patch level training to learn compact descriptors that represent the spatial information of the image. We

*Equal contribution.

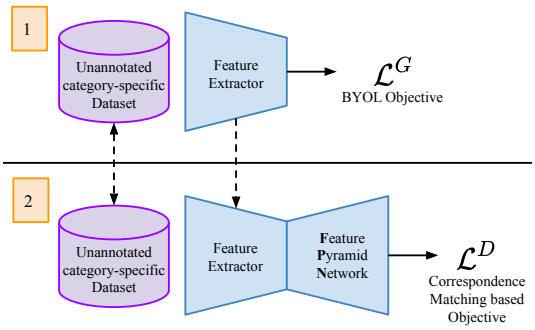

Figure 1. **LEAD Framework overview.** Two stage process for self-supervised landmark detection. **First**, an *instance-level* feature extractor is trained on a large Unannotated category-specific Dataset with the BYOL [13] objective. **Second**, using the correspondence matching property of the instance-level feature extractor, a *pixel-level* FPN [24] based feature extractor is trained on the same dataset. Finally, the pixel-level feature extractor is used to train a supervised regressor on limited data of landmark annotations.

illustrate the high-level idea in Fig. 1, and include a detailed architecture in Fig. 2.

We measure the performance of LEAD using percentage of inter-ocular distance (IOD). Landmarks estimated using our feature extractor show $\sim$10% improvement over prior art on facial landmark estimation, along with a boost in performance in the setting of severely limited annotations. We further obtain improved generalization to alignment and scale changes in the input images.

In summary, our contributions are:

- We show the emergence of high-fidelity landmarks in Bootstrap-Your-Own-Latent (BYOL) [13] style instance-level feature learning framework. (Sec. 3.2)
- We utilise this property to guide the learning of dense and compact feature maps of the image via a novel dimensionality reduction objective. (Sec. 3.3)
- Our evaluations show significant improvements over prior art on challenging datasets and across degrees of annotations, both qualitatively and quantitatively. (Sec. 4)

2. Related Works

Unsupervised Landmark prediction: The landmark prediction task has traditionally been studied in a supervised learning setting. Given the annotation-heavy nature of the problem, recent approaches have emphasized on unsupervised pretraining to learn information-rich features. These approaches can be divided based on two principles: equivariance and image generation.

Thewlis et al. [39] proposed an approach that uses equivariance of the feature descriptors across image warps as an objective for supervision. Suwajanakorn et al. [36] extended this idea for 3D landmark discovery from multi-view image pairs. This idea has also been used to model symmetrically deformable objects [40], and to learn object frames [38]. Further, Thewlis et al. [37] supplemented it using the principle of transitivity, which ensured that the descriptors learnt are robust across images.

Generative objective for landmark detection was initially used by Zhang et al. [51] and Lorenz et al. [26]. The main idea is to learn an image autoencoder with a landmark discovery bottleneck. Jakab et al. [17] coupled it with conditional image generation which could decouple the appearance and pose over an image pair. The key downside of these methods is that, the discovered landmarks are not interpretable. This was addressed by [18] where the landmark bottleneck is interpretable, due to availability of unpaired poses. [27] detects more semantically meaningful landmarks using self-training and deep clustering.

Self-supervised learning: Self-supervised learning follows the paradigm of training a network using a pretext task on a large-scale unlabeled dataset, followed by training a shallow network using limited annotated data. Initial works explored pretext tasks like classification of image orientation [12], patch-location prediction [29, 9], image colorization [49, 50], and clustering [2, 1]. While transformation invariant representation learning of an image [55, 22, 34, 20] has been extensively studied in supervised learning, the idea has outperformed prior pretext tasks when modelled as a contrastive learning problem [15, 4, 5, 7, 3] in the self-supervised learning setting. Here, the main idea is to push the embeddings of the *query* image and its augmentation ("positive" image) closer, while repelling it against the embeddings of the "negative" images (all other images). This is achieved using the InfoNCE [41] loss. A key disadvantage of these methods is the use of a large number of "negative" images which leads to high memory requirements. This was mitigated by methods like [13] and [6], which achieve competitive performance without "negative" images. While both of these seminal works concentrate on the classification task, there are some advances in adapting these techniques for dense prediction tasks like detection and segmentation [31, 30, 47, 42] as well. The only work that adopted the contrastive learning objective for the task of landmark prediction is ContrastLandmarks (CL) [8], where they train the network with the InfoNCE [41] objective. To adapt the output feature map to the resolution of the image, they use a hypercolumn representation from features across different layers. The key differences between this work and LEAD are: 1) We learn dense and compact descriptors via a novel correspondence matching guided dimensionality reduction objective while CL uses the objective proposed by Thewlis et al. [38], and 2) We do not use any "negative" images, as landmarks are ubiquitous in a category-specific dataset.

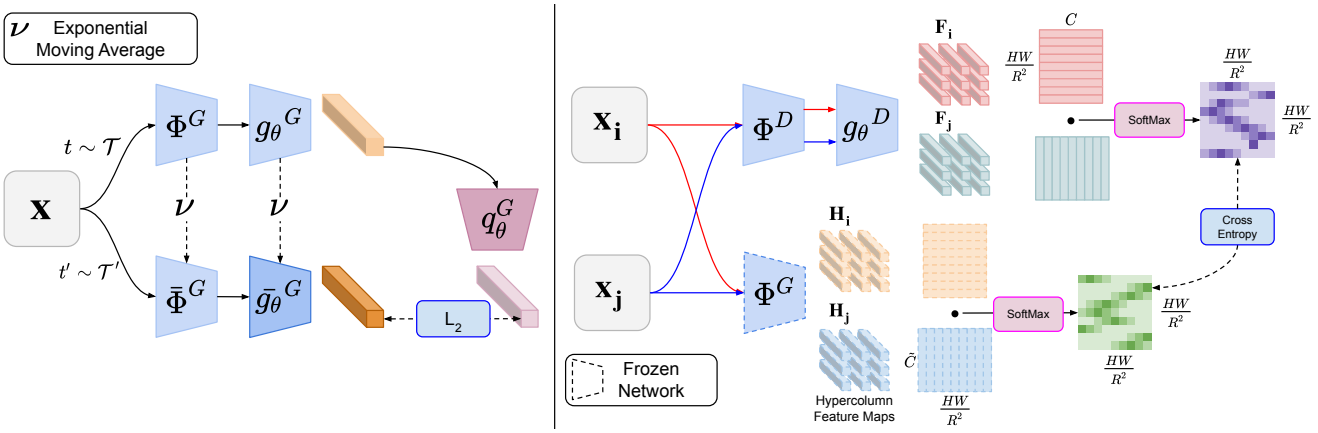

Figure 2. **LEAD training overview. Left:** Stage 1 of the training feature extractor Φ^G with BYOL objective, where the representation of key augmentation is predicted from query augmentation. **Right:** Stage 2 involves using frozen Φ^G to obtain dense correspondences, which are used to guide trainable network Φ^D to obtain dense and compact image representation. The correspondences, which also describe similarity between features, are converted to a probability distribution over spatial grid, by using a softmax (ref. Fig. 3). Distribution of Feature Similarity from Φ^D is guided by that from Φ^G using a cross-entropy loss.

3. Method

3.1. Background

Let $\mathcal{X} = \{\boldsymbol{x} \in \mathbb{R}^{H \times W \times 3}\}$ be a large-scale unannotated category-specific dataset. Our goal is to learn a feature extractor Φ, which, given $\boldsymbol{x} \in \mathcal{X}$ as input gives a feature map as output. As a pretext task, prior works have attempted to enforce instance-level representations to be invariant to transformations [8], and impose consistency on the dense pixel-level representations. In our approach LEAD, we use two stages. First, we learn a global representation of the image that leads to its part-wise clustering as described in Sec. 3.2. Then, we make use of this prior to guide the learning of a dense and compact representation of the image by a novel dimensionality reduction objective, which matches the distributions of feature similarity across two images, as described in Sec. 3.3.

3.2. Global Representation Learning

We follow the algorithm proposed in BYOL [13] to learn an instance-level representation of the image. BYOL uses an online network Φ^G and a target network $\bar{\Phi}^G$. Φ^G and $\bar{\Phi}^G$ share the same architecture, but the weights of $\bar{\Phi}^G$ are obtained using a momentum average of weights of Φ^G across multiple training iterations. These backbone networks are followed by projection heads g_θ^G and $\bar{g}_\theta{}^G$. Similar to the weights of the backbone, the weights of $\bar{g}_\theta{}^G$ are obtained using a momentum average. The necessity for the projection heads in self-supervised training has been discussed extensively in SimCLR [3], where the authors find the representations of last layer before the projection head to be most useful. Additionally, the online network has a prediction head q_θ^G (Fig. 2).

The training objective is to predict the representation of one view of the image from another using q_θ^G. Given an image $\boldsymbol{x}$, its two views $\boldsymbol{x_1}$ and $\boldsymbol{x_2}$ are generated by applying augmentations. We refer to $\boldsymbol{x_1}$ as the *query image* and $\boldsymbol{x_2}$ as the *key image*. Φ^G and $\bar{\Phi}^G$ generate features corresponding to $\boldsymbol{x_1}$ and $\boldsymbol{x_2}$ respectively. These feature maps are then projected using g_θ^G and $\bar{g}_\theta{}^G$ respectively to obtain the instance-level representations $\boldsymbol{z_1}$ and $\boldsymbol{z_2}$. Since both the views belong to the same instance, the predictor q_θ^G is trained to predict $\boldsymbol{z_2}$ given $\boldsymbol{z_1}$. The squared L_2 loss shown below is minimized for training:

$$\mathcal{L}^G = \|q_\theta^G(\boldsymbol{z_1}) - \boldsymbol{z_2}\|_2^2 \tag{1}$$

As shown in CL [8], the self-supervised contrastive objective produces hypercolumn based feature maps that have semantic understanding of the correspondences at pixel level between two images. In addition, we find that the BYOL objective gives significantly better correspondences than the MoCo objective, as shown in the Fig. 3. Hypercolumns are used here, since the self-supervised networks downsample the input image largely to obtain an instance-level representation. Creating a hypercolumn based feature map involves concatenating the intermediate feature maps along the channel dimension. Since the intermediate feature maps have lower spatial resolution than the original input image, they are upsampled to match the resolution of the input image. This has been illustrated in Fig. 3. However, hypercolumns incur a large cost in terms of memory. In the next section, we improve upon this by injecting pixel-level information into the network, thereby learning a dense and compact representation of the image.

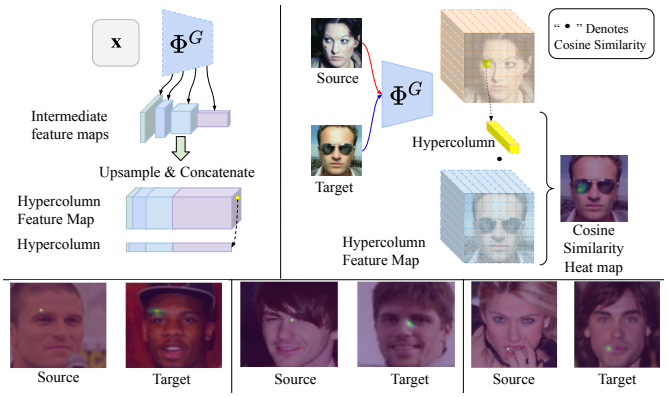

Figure 3. **Correspondence matching** performance using the hypercolumn representation. **Top Left:** Procedure to create hypercolumn from intermediate feature maps, by upsampling and concatenating them. Each feature vector across the spatial dimension denotes a hypercolumn. **Top Right:** Correspondence matching is performed from a point in the source image to the target image by taking cosine similarity of the hypercolumn corresponding to the source point and target's hypercolumn feature map, followed by softmax to obtain a heat map. **Bottom:** Examples of correspondence matching. Note that the resultant distribution peaks around the tracked point.

3.3. Dense and Compact Representation Learning

The bottleneck in framing the dense feature map learning problem is pixel-level correspondences. In the case of global feature vector learning, the image to form the positive pair is drawn by applying augmentation to the input image. But in the case of dense feature map learning, the correspondences between points in the query and the key images are not known. But since we have a trained BYOL network that can find *reasonable* (ref. Fig. 3) correspondences across images, we use it to guide the learning of dense and compact feature maps of images.

For the hypercolumn feature vector (or hypercolumn, for short), the ability to track a semantic point across two image depends on the distance between them in the $\tilde{C}$-d feature space. In this space, the features are clustered according to their semantic meaning. We aim to learn a compact feature space which has this property.

We now elaborate on the training method followed to learn such a low-dimensional feature space (Fig. 2). We train an encoder-decoder network $\Phi^D : \mathbb{R}^{H\times W\times 3} \rightarrow \mathbb{R}^{\frac{H}{R}\times\frac{W}{R}\times C}$. The encoder is initialised with Φ^G trained in Sec. 3.2. The output of the encoder goes to the projection head g_θ^D. We aim to retain the relationship defined by the cosine similarity between the hypercolumn feature maps from two images in their compact feature maps which are to be learnt. Let $\boldsymbol{x}_i$, $\boldsymbol{x}_j \in \mathcal{X}$ be two images, whose hypercolumn feature maps are $\boldsymbol{H}_i$, $\boldsymbol{H}_j \in \mathbb{R}^{H\times W\times 3} \rightarrow \mathbb{R}^{\frac{H}{R}\times\frac{W}{R}\times\tilde{C}}$ respectively. Note that, $\tilde{C} \gg C$, which makes the hypercolumn representation memory-intensive during inference. Let $\boldsymbol{F}_i = g_\theta^D(\Phi^D(\boldsymbol{x}_i))$ be compact feature maps of the respective images. Let $\boldsymbol{f}_i^{uv} \in \boldsymbol{F}_i$ be a feature vector at spatial location (u, v) in feature map $\boldsymbol{F}_i$. Similarly, let $\boldsymbol{h}_i^{uv} \in \boldsymbol{H}_i$ be a feature vector at spatial location (u, v) in the hypercolumn feature map $\boldsymbol{H}_i$. Since the aim is to retain the inter-feature relationship, we use cosine similarity as the measure of relationship between two feature vectors. To cover the whole feature space, we take cosine similarity with all the feature vectors. This relationship between the feature vector and the feature space as a probability distribution indicates which subspace of the feature space the feature vector is most similar to:

$$q_{ij}^{uv}[k,l] = \frac{\exp(\boldsymbol{f}_i^{uvT}\boldsymbol{f}_j^{kl}/\tau)}{\sum_{m,n=0}^{\frac{H}{R},\frac{W}{R}}\exp(\boldsymbol{f}_i^{uvT}\boldsymbol{f}_j^{mn}/\tau)} \tag{2}$$

where τ is temperature, which is a hyperparameter controlling the concentration level of the probability distribution q_{ij}^{uv} [45].

Similarly, such a relationship can be defined for $\boldsymbol{h}_i^{uv}$ with $\boldsymbol{H}_j$ as well. We denote this probability distribution as p_{ij}^{uv}. This ultimately leads us to optimize q_{ij}^{uv} to mimic p_{ij}^{uv}. We use cross-entropy between the both of them to achieve this objective:

$$\mathcal{L}^D = \sum_{u,v=0}^{\frac{H}{R},\frac{W}{R}}\sum_{k,l=0}^{\frac{H}{R},\frac{W}{R}} -p_{ij}^{uv}[k,l]\cdot\log(q_{ij}^{uv}[k,l]) \tag{3}$$

3.4. Landmark Detection

At this stage we have a feature extractor that is learned in a self-supervised fashion. To obtain the final landmark prediction, a limited amount of annotated data is used. Feature extractor is frozen and a lightweight predictor Ψ is trained over it. Ψ gives landmark heatmaps as output ($\in \mathbb{R}^{H\times W\times K}$) where K is the number of landmarks present). Expected location of the landmark k, weighed by the heatmap gives its final position $(\hat{x}^k, \hat{y}^k)$. It is supervised by the annotated location of the landmark (x^k, y^k) with an l_2 loss.

4. Experiments

Dataset: We evaluate LEAD on human faces. Following prior works, we use the CelebA [25] dataset containing 162,770 images for pretraining the network. To evaluate the learnt representation, four datasets are used. We firstly use MAFL which is a subset of CelebA. Two variants of AFLW [19] are used: the first being AFLW$_M$ which is the partition of AFLW with crops from MTFL [52]. It contains 10,122 training images and 2,995 test images. The second variant is AFLW$_R$, in which tighter crops of the face are used. This comprises of 10,122 training images and 2,991 testing images. We further use the 300-W [32] dataset which has 68 annotated landmarks, with 3148 training and 689 testing images. All the datasets are publicly available.

Implementation Details: We use a ResNet50 [16] backbone to train instance-level BYOL representation in stage 1.

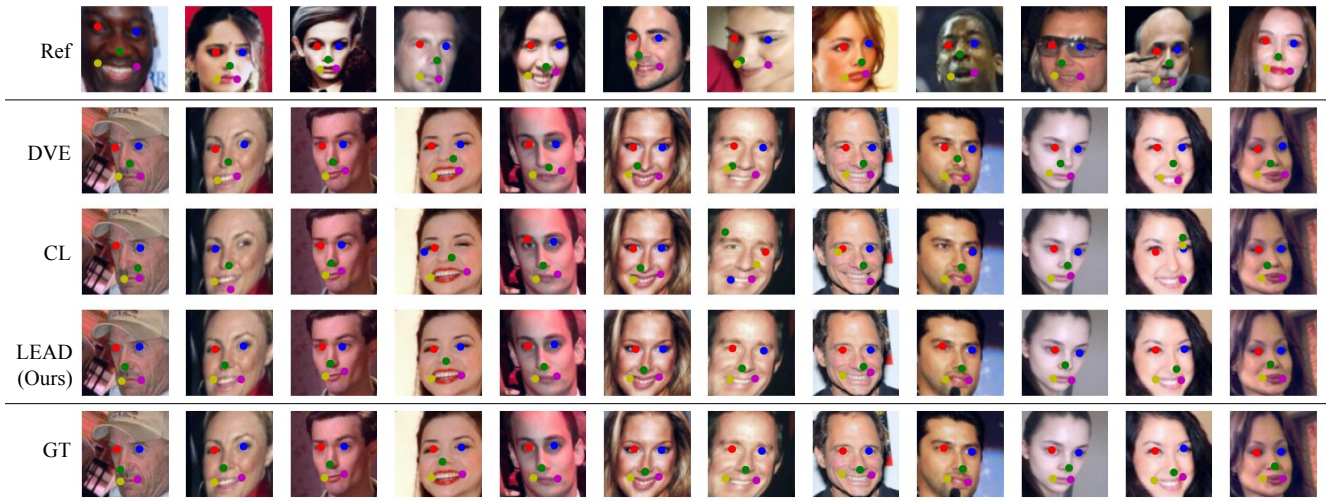

Figure 4. **Landmark Matching:** We observe that LEAD is able to predict the landmarks in the Query images (middle rows) using reference annotated image (first row). We compare our performance against DVE [37] and ContrastLandmarks [8] on a spectrum of head rotations.

In stage 2, the feature extractor of the trained ResNet in stage 1 is used as weight initialization for the encoder. The decoder is made up of FPN [24]. It is a lightweight network following the encoder which incorporates features from multiple scales of the encoder to create the final dense feature representation. This idea is similar to the creation of a hypercolumn feature map. FPN builds the final representation from features at 1/4, 1/8, 1/16 and 1/32 scales, using upsampling blocks as proposed in [30]. The final dense representation has a feature dimension of 64 and spatial downscaling of 1/4. The feature projection head is composed of 2 linear layers with BatchNorm and ReLU.

We use BYOL for stage 1 training with a batch size of 256 for 200 epochs using the SGD optimizer. The learning rate is set to $3\mathrm{e}{-2}$ with a cosine decay for stage 1 training. For stage 2, we train with a batch size of 256 for 20 epochs on the CelebA dataset. We set the temperature τ to be 0.05. For a fair comparison we train the supervised regressors with frozen feature extractor as proposed in [8]. The regressor initially comprises of 50 filters (to keep evaluations consistent with [8, 37]) of dimension $1 \times 1 \times K$ which transforms the input feature maps to heatmaps of intermediate virtual keypoints. These heatmaps are converted to $2K$ x-y pairs using a $softargmax$ layer, which are further linearly regressed to estimate manually annotated landmarks. Here K represents the number of annotated keypoints in the dataset. Following DVE, we resize the input image to (136×136) and then take a (96×96) central crop for performing the evaluations. For stage 1 training we take two (96×96) sized random crops. We perform all of our experiments on 2 Tesla V100 GPUs.

Evaluation: Following prior works, we use percentage of inter-ocular distance (IOD) as the error. We evaluate on two tasks, landmark matching and landmark regression. We describe each of the evaluation tasks next.

Table 1. **Landmark matching** performance comparison against prior art on MAFL dataset. The error is reported as a percentage of inter-ocular distance.

Method	Feat. dim.	Same	Different
DVE [37]	64	0.92	**2.38**
CL [8]	64	0.92	2.62
BYOL + NMF	64	0.84	5.74
LEAD (ours)	64	**0.51**	2.60
CL [8]	256	0.71	**2.06**
BYOL + NMF	256	0.91	4.26
LEAD (ours)	256	**0.48**	2.50
CL [8]	3840	0.73	6.16
LEAD (ours)	3840	**0.49**	**3.06**

4.1. Landmark Matching

In the landmark matching task, we are given two images. One is a reference image for which the landmarks are known and the other is a query image, for which the landmarks are to be predicted. Prediction is done by choosing the feature descriptor of a landmark in the reference image, and finding the location of the most similar feature descriptor to it in the feature map of the query image using cosine similarity. In line with DVE [37], we evaluate on a dataset consisting of 500 same identity and 500 different identity pairs taken from MAFL. Qualitative results of matching are shown in Fig. 4, while quantitative results are presented in Table 1. Also shown in Table 1 is the Non-negative Matrix Factorization (NMF [21], which gives low-rank approximation of non-negative matrix) baseline, wherein we apply NMF over the learned hypercolumn thereby showing that our dimensionality reduction objective is superior to naively applying NMF over the learned hypercolumn. Similar to the trends from correspondence matching using hypercolumn in Fig. 3,

Table 2. **Landmark regression** performance comparison against prior art. The error is reported as a percentage of inter-ocular distance. We achieve state-of-the-art result on the challenging AFLW datasets with ∼10% relative gain, while obtaining competitive results on MAFL and 300W.

Method	Unsupervised	MAFL	AFLW$_M$	AFLW$_R$	300W
TCDCN [53]	✗	7.95	7.65	-	5.54
RAR [46]	✗	-	7.23	-	4.94
MTCNN [52, 51]	✗	5.39	6.90	-	-
Wing Loss [11]	✗	-	-	-	4.04
Dense objective based					
Sparse [39]	✓	6.67	10.53	-	7.97
Structural Repr. [51]	✓	3.15	-	6.58	-
FAb-Net [44]	✓	3.44	-	-	5.71
Def. AE [33]	✓	5.45	-	-	-
Cond. Im. Gen [17]	✓	2.86	-	6.31	-
Int. KP. [18]	✓	-	-	-	5.12
Dense3D [38]	✓	4.02	10.99	10.14	8.23
DVE SmallNet [37]	✓	3.42	8.60	7.79	5.75
DVE Hourglass [37]	✓	2.86	7.53	6.54	**4.65**
Global Objective based					
ContrastLandmarks [8]	✓	<u>2.44</u>	6.99	6.27	5.22
LEAD (ours)	✓	**2.39**	**6.23**	**5.65**	<u>4.66</u>

the final dense model with 64 dimensional features is able to meaningfully match the landmarks from reference image to query image. This is verified across a head rotation ranging from left-facing to frontal faces and right-facing images. The matching is consistent across genders, showing no bias for any gender.

4.2. Landmark Regression

In the task of landmark regression, a lightweight regressor is trained on top of the features extracted by the pretrained network. This is done using supervised learning on the evaluation dataset. We report the inter-ocular distance on landmark regression in Table 2. Our model trained using the BYOL objective achieves results which are ∼10% better than the prior-art on a relative scale, on 2 out of 4 evaluation datasets, while maintaining a competitive performance on the 300-W dataset. Regression performance is qualitatively verified in the Fig. 7. We refer the reader to the supplementary material for additional datasets and visualizations.

4.3. Interpretability

We observe that the t-SNE embeddings obtained from our model trained with BYOL objective are interpretable. It divides the face spatially into 9 parts, where each clusters corresponds to one of the 9 parts. t-SNE clustering is visualized in Fig. 5 and interpretability of the clusters is verified in Fig. 6. We also compare our t-SNE plots against that of CL [8], wherein we see that CL embeddings are not well clustered when compared to LEAD which shows distinct clusters. We provide further feature clustering analysis in the supplementary material.

4.4. Ablation Studies

We ablate LEAD on factors like feature dimension, contribution from each stage, projection head, degree of annotation availability, and sensitivity to scale variations.

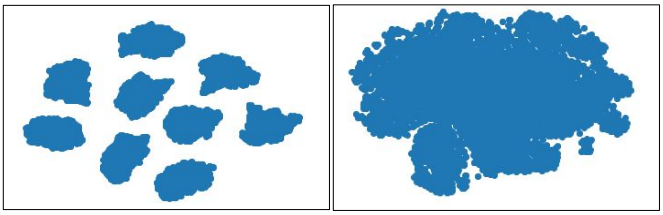

Figure 5. **t-SNE plots of output feature maps. Left:** LEAD stage 1 features **Right:** CL stage 1 features

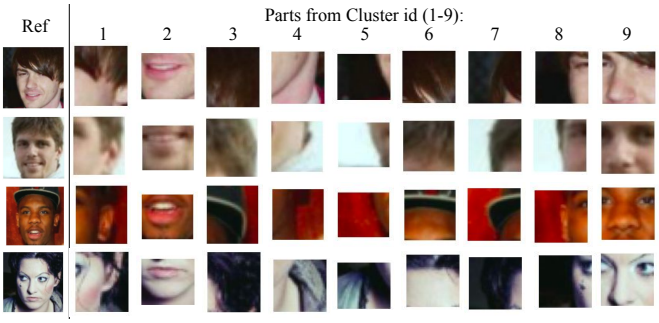

Figure 6. **t-SNE embeddings tend to cluster part-wise**. The 9 parts (along row) shown for each reference figure (along column) here belong to the 9 clusters in Fig. 5. Each cluster denotes a semantic part of the face.

Feature Dimensions. Feature dimension plays a significant role in the landmark regression task. Since the regressor takes features as input, its capacity depends on the dimensions of the feature, i.e. a higher dimensional feature implies that the regressor has more capacity to learn, resulting in better predictions. Our experiments in Table 4 indicate a superior performance on the challenging AFLW$_M$ dataset, while achieving competitive performance on MAFL and AFLW$_R$. Surprisingly, we find a large deviation in performance trends on the 300W dataset compared to the results obtained using hypercolumn feature maps (ref. Tab. 2) as guidance for the compact feature maps.

How much does stage 2 objective contribute? To answer this question, we run experiments on 2 different pretraining (stage 1) objectives, followed by 2 different dimensionality reduction (stage 2) objectives. To compare directly, we take CL's [8] pretraining and dimensionality reduction objectives and our objectives for the same. We keep the architectures same as LEAD and only vary the objective function for a fair comparison. We report our findings in Table 6. Irrespective of the stage 1 training, LEAD's dimensionality reduction procedure improves the IOD.

Is the projection head necessary in stage 2? Necessity of the projection head in self-supervised learning has been empirically shown to lead to meaningful representations [3]. We use it in our stage 1 training. However in stage 2, where we aim to get higher resolution feature maps as output, is the projection head still required? We use a projection head g_θ^D on the final feature map as given by Φ^D to apply the loss

Table 3. **Effect of projection head on landmark matching**. Projection head affects the matching on different identity. On increasing the dimension of the projection head's output, improvement is observed. Further gains are observed on increasing the final representation's (Φ^D's output) dimension.

Feat. dim.	Proj. dim.	Same	Different
64	✗	**0.48**	2.79
64	64	0.51	2.64
64	256	0.51	**2.60**
128	256	0.47	2.58
256	256	**0.48**	**2.50**

Table 4. Effect of feature dimension on landmark regression task

Method	Feat. dim.	MAFL	AFLW$_M$	AFLW$_R$	300W
DVE [37]	64	2.86	7.53	6.54	**4.65**
CL [8]	64	**2.77**	7.21	**6.22**	5.19
LEAD (ours)	64	2.93	**6.61**	6.32	5.32
CL [8]	128	**2.71**	7.14	**6.14**	**5.09**
LEAD (ours)	128	2.91	**6.60**	6.21	5.41
CL [8]	256	**2.64**	7.17	6.14	**4.99**
LEAD (ours)	256	2.87	**6.51**	**6.12**	5.37

Table 5. **Number of annotations:** LEAD consistently produces the lowest inter-ocular distance under the presence of different levels of annotations on the AFLW$_M$ dataset. The relative improvement is as high as **45%** over previous best (in case of '5 annotations' training setting)

Method	Feat. dim.	Number of annotations					
		1	5	10	20	50	100
DVE [37]	64	**14.23 ± 1.45**	**12.04 ± 2.03**	12.25 ± 2.42	11.46 ± 0.83	12.76 ± 0.53	11.88 ± 0.16
CL [8]	64	24.87 ± 2.67	15.15 ± 0.53	13.62 ± 1.08	11.77 ± 0.68	11.57 ± 0.03	10.06 ± 0.45
LEAD (Ours)	64	21.8 ± 2.54	13.34 ± 0.43	**11.50 ± 0.34**	**10.13 ± 0.45**	**9.29 ± 0.45**	**9.11 ± 0.25**
CL [8]	128	27.31 ± 1.39	18.66 ± 4.59	13.39 ± 0.30	11.77 ± 0.85	10.25 ± 0.22	9.46 ± 0.05
LEAD (ours)	128	**21.20 ± 1.67**	**13.22 ± 1.43**	**10.83 ± 0.65**	**9.69 ± 0.41**	**8.89 ± 0.2**	**8.83 ± 0.33**
CL [8]	256	28.00 ± 1.39	15.85 ± 0.86	12.98 ± 0.16	11.18 ± 0.19	9.56 ± 0.44	9.30 ± 0.20
LEAD (ours)	256	**21.39 ± 0.74**	**12.38 ± 1.28**	**11.01 ± 0.48**	**10.06 ± 0.59**	**8.51 ± 0.09**	**8.56 ± 0.21**
CL [8]	3840	42.69 ± 5.10	25.74 ± 2.33	17.61 ± 0.75	13.35 ± 0.33	10.67 ± 0.35	9.24 ± 0.35
LEAD (ours)	3840	**24.41 ± 1.38**	**14.11 ± 1.30**	**11.45 ± 0.88**	**10.21 ± 0.44**	**8.43 ± 0.25**	**8.09 ± 0.28**

Table 6. **Dimensionality reduction objective.** LEAD's proposed dimensionality reduction objective significantly improves the performance irrespective of the global representation learning objective. Results are reported on AFLW$_M$ dataset.

Global Rep. Obj. (Stage 1)	Dim. Red. Obj. (Stage 2)	Feat. dim.		
		64	128	256
CL	CL	7.86	7.81	7.31
CL	LEAD	**6.66**	**6.58**	**6.69**
LEAD	CL	7.89	7.86	7.41
LEAD	LEAD	**6.61**	**6.60**	**6.51**

during training. Eventually the g_θ^D is discarded and only Φ^D is utilised. Here, we ablate on the performance shown by Φ^D in the absence of projection head as well as the on the output dimension of the g_θ^D. Since we discard g_θ^D, we are allowed to keep its output's dimension as high as required. In our ablation (ref. Table 3), it is observed that for landmark matching on the same identity, there are marginal changes upon having g_θ^D as well as varying its output dimension. But the projection head emerges as a distinguishing component in case of matching on different identity. Consistent improvements are observed on increasing the feature dimension of the projection head. It can be seen that this leads to slight degradation of performance on the same identity. We also observe the effect of increasing the feature dimension by keeping the projection dimension fixed where we note a further improvement on matching.

How sensitive is it to the alignment and scale variations? At inference stage, the landmark regressor can encounter images which may have different alignments or scales when compared to the data it was trained on. To check the sensitivity of LEAD to these changes we use features from CelebA trained LEAD to train a landmark regressor on an unaligned-MAFL dataset. We create this dataset by taking images from MAFL subset of CelebA-in-the-wild [25] dataset cropped by the bounding box annotations. Furthermore, before taking a crop, we also randomly scale up the side length of the bounding box a factor uniformly randomly sampled between 1-1.5×. This results in zooming out of the image (ref. Fig. 10). We refer to this factor as "Zoom-out factor" We evaluate the regressor on the test split which is created by scaling up the side length of the bounding box by a zoom-out factor of 1-2× before cropping. We use 64d feature for this experiment. In Fig. 9, we observe that across the range

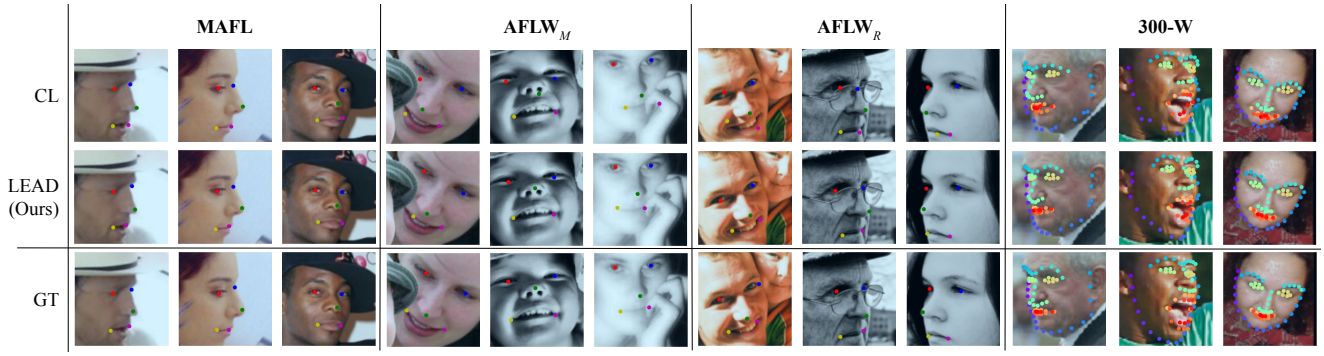

Figure 7. **Landmark regression:** We observe that features generated by pretraining using LEAD can easily be used to train a lightweight regressor to predict landmarks with high precision. Furthermore, the model is robust to aspects such as face orientation, lighting and minor occlusions.

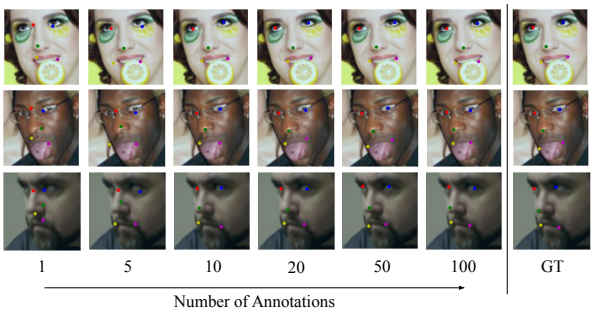

Figure 8. **Number of annotations.** Landmark prediction under different number of annotated images used for supervised training (mentioned below every column) on AFLW$_M$.

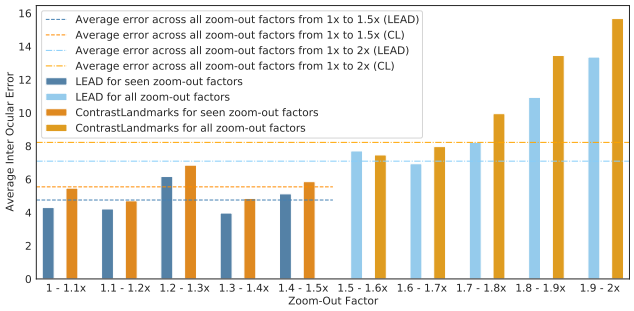

Figure 9. **Sensitivity to scale variations:** Sensitivity to seen scales (Zoom-out factor $\in$ 1-1.5x) vs unseen scales (Zoom-out factor $\in$ 1.5-2x) on unaligned-MAFL. LEAD performs better across scale changes, and is also less sensitive to *unseen* scales of face.

of evaluated scales, LEAD outperforms CL [8][1]. The gap between the two methods widens for larger zoom-out factors, which are unseen during training. We visualize the landmark regression against scale changes in Fig. 10.

How many annotated images are required for supervised training during evaluation? Since the evaluation of our method depends on the annotated samples, we run an ablation on the number of annotations required. We report the quantitative results in the Table 5, along with qualitative

[1]Same training and evaluation protocol was followed for both.

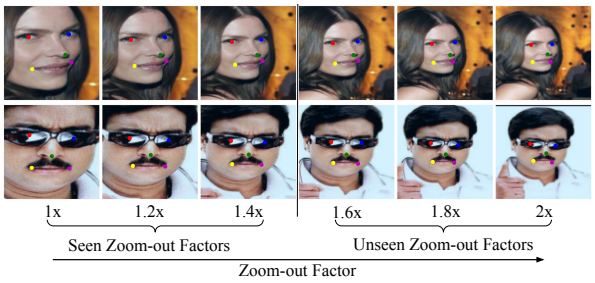

Figure 10. **Scale variation.** LEAD Landmark regression visualization across differently scaled (seen and unseen) images of unaligned-MAFL.

annotation-wise comparisons in Fig. 8. We test by varying the number of annotations to 1, 5, 10, 20, 50, and 100. We observe a consistent and significant gain in the performance with increasing number of annotations over the competent methods, a trend which even continues at different dimensions of features.

5. Conclusions

In this work, we demonstrate the superiority of the LEAD framework to learn representation at instance level from a category specific dataset. We further utilize this prior to train a dense and compact representation of the image, guided by the correspondence matching property of the learnt representation. Our experiments demonstrate the superiority of the BYOL objective over contrastive tasks like MoCo on category specific data for landmark detection. Our proposed dimensionality reduction method improves the results on both feature extractors. A future research direction could be the usage of this correspondence matching property to learn a variety of dense prediction tasks.

Acknowledgements This work was supported by MeitY (Ministry of Electronics and Information Technology) project (No. 4(16)2019-ITEA), Govt. of India.

References

[1] Yuki M. Asano, Christian Rupprecht, and Andrea Vedaldi. Self-labelling via simultaneous clustering and representation learning. In *International Conference on Learning Representations (ICLR)*, 2020.

[2] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[3] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, 2020.

[4] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. Big self-supervised models are strong semi-supervised learners. *arXiv preprint arXiv:2006.10029*, 2020.

[5] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020.

[6] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning, 2020.

[7] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers, 2021.

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

[9] Carl Doersch, Abhinav Gupta, and Alexei A. Efros. Unsupervised visual representation learning by context prediction. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*.

[10] Alexey Dosovitskiy, Jost Tobias Springenberg, Martin Riedmiller, and Thomas Brox. Discriminative unsupervised feature learning with convolutional neural networks. In *Advances in Neural Information Processing Systems*, 2014.

[11] Zhen-Hua Feng, Josef Kittler, Muhammad Awais, Patrik Huber, and Xiao-Jun Wu. Wing loss for robust facial landmark localisation with convolutional neural networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[12] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. In *International Conference on Learning Representations*, 2018.

[13] 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.

[14] Richard Hartley and Andrew Zisserman. *Multiple View Geometry in Computer Vision*. 2 edition, 2004.

[15] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[17] Tomas Jakab, Ankush Gupta, Hakan Bilen, and Andrea Vedaldi. Unsupervised learning of object landmarks through conditional image generation. In *Advances in Neural Information Processing Systems*, 2018.

[18] Tomas Jakab, Ankush Gupta, Hakan Bilen, and Andrea Vedaldi. Self-supervised learning of interpretable keypoints from unlabelled videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[19] Martin Köstinger, Paul Wohlhart, Peter M. Roth, and Horst Bischof. Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*.

[20] Dmitry Laptev, Nikolay Savinov, Joachim M. Buhmann, and Marc Pollefeys. Ti-pooling: Transformation-invariant pooling for feature learning in convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[21] Daniel D Lee and H Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

[22] Jianshu Li, Jian Zhao, Fang Zhao, Hao Liu, Jing Li, Shengmei Shen, Jiashi Feng, and Terence Sim. Robust face recognition with deep multi-view representation learning. In *Proceedings of the 24th ACM International Conference on Multimedia*, 2016.

[23] Jianshu Li, Pan Zhou, Y. Chen, Jian Zhao, S. Roy, Shuicheng Yan, Jiashi Feng, and T. Sim. Task relation networks. *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*.

[24] 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)*.

[25] 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)*, 2015.

[26] Dominik Lorenz, Leonard Bereska, Timo Milbich, and Björn Ommer. Unsupervised part-based disentangling of object shape and appearance. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[27] Dimitrios Mallis, Enrique Sanchez, Matt Bell, and Georgios Tzimiropoulos. Unsupervised learning of object landmarks via self-training correspondence. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

[28] Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In *CVPR*, 2020.

[29] M. Noroozi and P. Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *ECCV*, 2016.

[30] Pedro O O. Pinheiro, Amjad Almahairi, Ryan Benmalek, Florian Golemo, and Aaron C Courville. Unsupervised learning of dense visual representations. In *Advances in Neural Information Processing Systems*, 2020.

[31] Byungseok Roh, Wuhyun Shin, Ildoo Kim, and Sungwoong Kim. Spatilly consistent representation learning. In *CVPR*, 2021.

[32] Christos Sagonas, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *2013 IEEE International Conference on Computer Vision Workshops*.

[33] Zhixin Shu, Mihir Sahasrabudhe, Riza Alp Guler, Dimitris Samaras, Nikos Paragios, and Iasonas Kokkinos. Deforming autoencoders: Unsupervised disentangling of shape and appearance. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[34] Kihyuk Sohn and Honglak Lee. Learning invariant representations with local transformations. In *ICML*, 2012.

[35] Jong-Chyi Su, Subhransu Maji, and Bharath Hariharan. When does self-supervision improve few-shot learning? In *ECCV*, 2020.

[36] Supasorn Suwajanakorn, Noah Snavely, Jonathan J Tompson, and Mohammad Norouzi. Discovery of latent 3d keypoints via end-to-end geometric reasoning. In *Advances in Neural Information Processing Systems*, 2018.

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

[38] James Thewlis, Hakan Bilen, and Andrea Vedaldi. Unsupervised learning of object frames by dense equivariant image labelling. In *Advances in Neural Information Processing Systems*, 2017.

[39] James Thewlis, Hakan Bilen, and Andrea Vedaldi. Unsupervised learning of object landmarks by factorized spatial embeddings. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

[40] James Thewlis, Hakan Bilen, and Andrea Vedaldi. Modelling and unsupervised learning of symmetric deformable object categories. In *Advances in Neural Information Processing Systems*, 2018.

[41] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding, 2019.

[42] Xinlong Wang, Rufeng Zhang, Chunhua Shen, Tao Kong, and Lei Li. Dense contrastive learning for self-supervised visual pre-training. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2021.

[43] Zhecan Wang and Jian Zhao. Conditional dual-agent gans for photorealistic and annotation preserving image synthesis. 2017.

[44] O. Wiles, A.S. Koepke, and A. Zisserman. Self-supervised learning of a facial attribute embedding from video. In *British Machine Vision Conference*, 2018.

[45] Zhirong Wu, Yuanjun Xiong, Stella X. Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[46] Shengtao Xiao, Jiashi Feng, Junliang Xing, Hanjiang Lai, Shuicheng Yan, and Ashraf Kassim. Robust facial landmark detection via recurrent attentive-refinement networks. In *Computer Vision – ECCV 2016*.

[47] Zhenda Xie, Yutong Lin, Zheng Zhang, Yue Cao, Stephen Lin, and Han Hu. Propagate yourself: Exploring pixel-level consistency for unsupervised visual representation learning. 2021.

[48] Li Zhang, Tao Xiang, and Shaogang Gong. Learning a deep embedding model for zero-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[49] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *ECCV*, 2016.

[50] Richard Zhang, Phillip Isola, and Alexei A Efros. Split-brain autoencoders: Unsupervised learning by cross-channel prediction. In *CVPR*, 2017.

[51] Yuting Zhang, Yijie Guo, Yixin Jin, Yijun Luo, Zhiyuan He, and Honglak Lee. Unsupervised discovery of object landmarks as structural representations. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[52] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. Facial landmark detection by deep multi-task learning. In *Computer Vision – ECCV 2014*.

[53] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and X. Tang. Learning deep representation for face alignment with auxiliary attributes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38:918–930, 2016.

[54] F. Zhao, Jian Zhao, Shuicheng Yan, and Jiashi Feng. Dynamic conditional networks for few-shot learning. In *ECCV*, 2018.

[55] Jian Zhao, Jianshu Li, F. Zhao, Xuecheng Nie, Y. Chen, Shuicheng Yan, and Jiashi Feng. Marginalized cnn: Learning deep invariant representations. In *BMVC*, 2017.