On Equivariant and Invariant Learning of Object Landmark Representations

Zezhou Cheng Jong-Chyi Su Subhransu Maji
University of Massachusetts Amherst
{zezhoucheng, jcsu, smaji}@cs.umass.edu

Abstract

Given a collection of images, humans are able to discover landmarks by modeling the shared geometric structure across instances. This idea of geometric equivariance has been widely used for the unsupervised discovery of object landmark representations. In this paper, we develop a simple and effective approach by combining instance-discriminative and spatially-discriminative contrastive learning. We show that when a deep network is trained to be invariant to geometric and photometric transformations, representations emerge from its intermediate layers that are highly predictive of object landmarks. Stacking these across layers in a "hypercolumn" and projecting them using spatially-contrastive learning further improves their performance on matching and few-shot landmark regression tasks. We also present a unified view of existing equivariant and invariant representation learning approaches through the lens of contrastive learning, shedding light on the nature of invariances learned. Experiments on standard benchmarks for landmark learning, as well as a new challenging one we propose, show that the proposed approach surpasses prior state-of-the-art.

1. Introduction

Learning in the absence of labels is a challenge for existing machine learning and computer vision systems. Despite recent advances, the performance of unsupervised learning remains far below that of supervised learning, especially for few-shot image understanding tasks. This paper considers the task of unsupervised learning of object landmarks from a collection of images. The goal is to learn representations that can be used to establish correspondences across objects, and to predict landmarks such as eyes and noses when provided with a few labeled examples.

One way of inferring structure is to reason about the global appearance in terms of disentangled factors such as geometry and texture. This is the basis of alignment based [27, 37] and generative modeling based approaches

Code and data are available at https://people.cs.umass.edu/~zezhoucheng/contrastive_landmark/

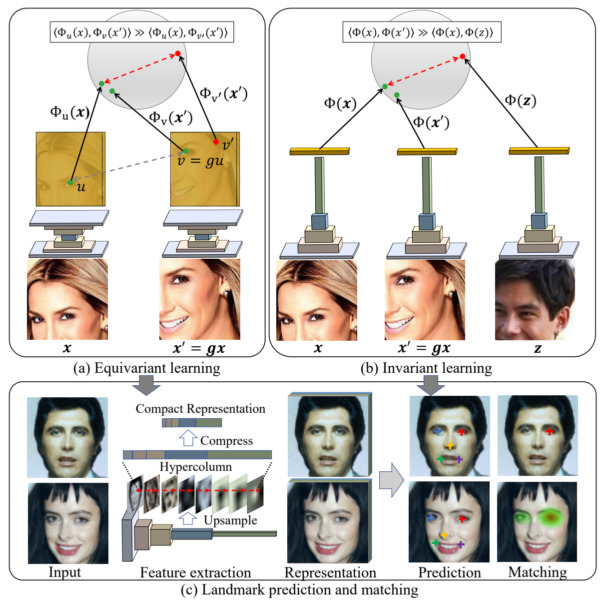


Figure 1. **Equivariant and invariant learning. (a)** Equivariant learning requires representations across locations to be invariant to a geometric transformation g while being distinctive across locations. **(b)** Invariant learning encourages the representations to be invariant to transformations while being distinctive across images. Thus both can be seen as instances of contrastive learning. **(c)** A hypercolumn feature and its compact representation are highly predictive of object landmarks.

for landmark discovery [28, 29, 47, 57, 61, 65]. An alternate is to learn a representation that geometrically transforms in the same way as the object, a property called *geometric equivariance* (Fig. 1a) [49–51]. However, useful invariances may not be learned (*e.g.*, the raw pixel representation itself is equivariant), limiting their applicability in the presence of clutter, occlusion, and inter-image variations.

A different line of work has proposed instance discriminative *contrastive learning* as an unsupervised objective [3, 5, 13, 21, 23, 25, 26, 40, 52, 58, 69]. The goal is to learn a representation Φ that has higher similarity between an image $\mathbf{x}$ and its transformation $\mathbf{x}'$ than with a different one $\mathbf{z}$, *i.e.*, $\langle\Phi(\mathbf{x}), \Phi(\mathbf{x}')\rangle \gg \langle\Phi(\mathbf{x}), \Phi(\mathbf{z})\rangle$, as illustrated in Fig. 1b. A combination of geometric (*e.g.*, cropping and scaling) and photometric (*e.g.*, color jittering and blurring) transformations are used to encourage the representation to be *invariant* to these transformations while being *distinctive* across images. Recent work [5–7, 23] has shown that contrastive learning is effective, even outperforming ImageNet [12] pre-training on various tasks. However, to predict landmarks a representation cannot be invariant to geometric transformations. This paper asks the question: *are equivariant losses necessary for unsupervised landmark discovery?* In particular, do representations predictive of object landmarks automatically emerge in intermediate layers of a deep network trained to be invariant to image transformations? While empirical evidence suggests that semantic parts emerge when deep networks are trained on supervised tasks [18, 68], is it also the case for unsupervised learning?

This work aims to address these by presenting a unified view of the equivariant and invariant learning approaches. We show that when a deep network is trained to be invariant to geometric and photometric transformations, its intermediate-layer representations are highly predictive of landmarks (Fig. 1b). The emergence of invariance and the loss of geometric equivariance is gradual in the representation hierarchy, a phenomenon that has been studied empirically [31, 63] and theoretically [1, 53, 54]. This observation motivates a *hypercolumn* representation [22], which we find to be more effective for landmark predictions (Fig. 1c).

We also observe that objectives used in equivariant learning can be seen as a contrastive loss between representations across locations within the *same image*, as opposed to invariant learning where the loss is applied *across images* (Fig. 1). This observation sheds light on the nature of the invariances learned by the two approaches. It also allows us to obtain a compact representation of the high-dimensional hypercolumns simply by learning a linear projection under the spatially contrastive objective. The projection results in spatially distinctive representations and significantly improves the landmark matching performance (Tab. 1 and Fig. 2).

To validate these claims, we perform experiments by training deep networks using Momentum Contrast (MoCo) [23] on several landmark matching and detection benchmarks. Other than commonly used ones, we also present a comparison by learning on a challenging dataset of birds from the iNaturalist dataset [55] and evaluating on the CUB dataset [56]. We show that the contrastive-learned representations (without supervised regression) can be predictive in landmark matching experiments. For landmark detection, we adapt the commonly used linear evaluation setting by varying the number of labeled examples (Fig. 3 & 4). Our approach is simple, yet it offers consistent improvements over prior approaches [28, 49–51, 65] (Tab. 2). While the hypercolumn representation leads to a larger embedding dimension, it comes at a modest cost as our approach outperforms the prior state-of-the-art [49], with as few as 50 annotated training examples on the AFLW benchmark [30] (Fig. 4).

Furthermore, we use dimensionality reduction based on the equivariant learning to improve the performance on landmark matching (Tab. 1), as well as landmark prediction in the low data regime (Tab. 4).

2. Related Work

Background. A representation $\Phi : \mathcal{X} \rightarrow \mathbb{R}^C$ is said to be equivariant (or covariant) with a transformation g for input $\mathbf{x} \in \mathcal{X}$ if there exists a map $M_g : \mathbb{R}^C \rightarrow \mathbb{R}^C$ such that: $\forall \mathbf{x} \in \mathcal{X} : \Phi(g\mathbf{x}) \approx M_g\Phi(\mathbf{x})$. In other words, the representation transforms in a predictable manner given the input transformation. For natural images, the transformations can be geometric (*e.g.*, translation, scaling, and rotation), photometric (*e.g.*, color changes), or more complex (*e.g.*, occlusion, viewpoint or instance variations). Note that a sufficient condition for equivariance is when Φ is invertible since $M_g = \Phi^{-1} \circ g$. Invariance is a special case of equivariance when M_g is the identity function, *i.e.*, $\phi(g\mathbf{x}) \approx \phi(\mathbf{x})$. There is a rich history in computer vision on the design of covariant (*e.g.*, SIFT [35]), and invariant representations (*e.g.*, HOG [11] and Bag-of-Visual-Words [48]).

Deep representations. Invariance and equivariance in deep network representations result from both the architecture (*e.g.*, convolutions lead to translational equivariance, while pooling leads to translational invariance), and learning (*e.g.*, invariance to categorical variations). Lenc *et al.* [31] showed that early-layer representations of a deep network are nearly equivariant as they can be "inverted" to recover the input, while later layers are more invariant. Similar observations have been made by visualizing these representations [36, 63]. The gradual emergence of invariance can also be theoretically understood in terms of a "information bottleneck" in the feed-forward hierarchy [1, 53, 54]. While equivariance to geometric transformations is relevant for landmark representations, the notion can be generalized to other transformation groups [9, 16].

Landmark discovery. Empirical evidence [41, 68] suggests that semantic parts emerge when deep networks are trained on supervised tasks. This has inspired architectures for image classification that encourage part-based reasoning, such as those based on texture representations [2, 8, 32] or spatial attention [15, 46, 60]. In contrast, our work shows that *parts also emerge when models are trained in an unsupervised manner.* When no labels are available, equivariance to geometric transformations provides a natural self-supervisory signal. The equivariance constraint requires $\Phi_u(\mathbf{x})$, the representation of $\mathbf{x}$ at location u, to be invariant to the geometric transformation g of the image, *i.e.*, $\forall \mathbf{x}, u : \Phi_{gu}(g\mathbf{x}) = \Phi_u(\mathbf{x})$ (Fig. 1a). This alone is not sufficient since both $\Phi_u(\mathbf{x}) = \mathbf{x}_u$ and $\Phi_u(\mathbf{x}) =$ constant satisfy this property. Constraints based on locality [49, 50] and diversity [51] have been proposed to avoid this pathology. Yet, inter-image invariance is not directly enforced.

Another line of work is based on a generative modeling approach [4, 28, 29, 34, 45, 47, 57, 61, 65]. These methods implicitly incorporate equivariant constraints by modeling objects as deformation (or flow) of a shape template together with appearance variation in a disentangled manner. In contrast, our work shows that learning representations invariant to both geometric and photometric transformations is an effective strategy. These invariances emerge at different rates in the representation hierarchy, and can be selected with a small amount of supervision for the downstream task.

Unsupervised learning. Recent work has shown that unsupervised objectives based on density modeling [3, 5–7, 13, 23, 26, 40, 52, 58] outperform unsupervised (or self-supervised) learning based on pretext tasks such as colorization [64], rotation prediction [17], jigsaw puzzle [38], and inpainting [42]. These contrastive learning objectives [21] are often expressed in terms of noise-contrastive estimation (NCE) [20] (or maximizing mutual information [26, 40]) between different views obtained by geometrically and photometrically transforming an image. The learned representations thus encode invariances to these transformations while preserving information relevant for downstream tasks. However, the effectiveness of unsupervised learning depends on how well these invariances relate to those desired for end tasks. Despite recent advances, existing methods for unsupervised learning significantly lack in comparison to their supervised counterparts in the few-shot setting [19]. Moreover, their effectiveness for landmark discovery has not been sufficiently studied in the literature.[1] In part, it is not clear why invariance to geometric transformations might be useful for landmark prediction since we require the representation to carry some spatial information about the image. Understanding these trade-offs and improving the effectiveness of contrastive learning for landmark prediction is one of the goals of the paper.

3. Method

Let $\mathbf{x} \in \mathbb{R}^{H \times W \times 3}$ denote an image of an object, and $u \in \Omega = \{0, \ldots, H-1\} \times \{0, \ldots, W-1\}$ denote pixel coordinates. The goal is to learn a function $\Phi_u(\mathbf{x}) : \Omega \rightarrow \mathbb{R}^C$ that outputs a pixel representation at spatial location u of input $\mathbf{x}$ that is predictive for object landmarks. We assume $C \gg 3$ aiming to learn a high-dimensional representation of landmarks. This is similar to [49] which learns a local descriptor for each landmark, and unlike those that represent them as a discrete set [67], or on a planar ($C = 2$) [51, 65] or spherical ($C = 3$) [50] coordinate system. In other words the representation should be predictive of landmarks or effective for matching, without requiring compactness or topology in the embedding space. Note that this is in contrast to some work on literature where a fixed set of landmarks are discovered (*e.g.*, [28, 51, 65]). One may obtain this, for instance, by clustering the landmark representations in the embedding space.

We describe commonly used equivariance constraints for unsupervised landmark discovery [49–51], followed by models based on invariant learning [23, 40]. We then present our approach that integrates the equivariant and invariant learning approaches.

3.1. Equivariant and invariant representations

Equivariant learning. The equivariance constraint requires $\Phi_u(\mathbf{x})$, the representation of $\mathbf{x}$ at location u, to be invariant to the geometric deformation of the image (Fig. 1a). Given a geometric warping function $g : \Omega \rightarrow \Omega$, the representation of $\mathbf{x}$ at u should be same as the representation of the transformed image $\mathbf{x}' = g\mathbf{x}$ at $v = gu$, that is, $\forall \mathbf{x}, u \in \Omega : \Phi_v(\mathbf{x}') = \Phi_u(\mathbf{x})$. This constraint can be captured by the loss:

$$\mathcal{L}_{equi} = \frac{1}{|\Omega|} \sum_{u \in \Omega} \|\Phi_u(\mathbf{x}) - \Phi_v(\mathbf{x}')\|^2. \quad (1)$$

A diversity (or locality) constraint is necessary to encourage the representation to be distinctive across locations. For example, Thewlis *et al*. [50] proposed the following:

$$\mathcal{L}_{div} = \frac{1}{|\Omega|} \sum_{u \in \Omega} \|gu - \operatorname*{argmax}_v \langle \Phi_u(\mathbf{x}), \Phi_v(\mathbf{x}')\rangle\|^2, \quad (2)$$

which they replaced by a probabilistic version that combines both the losses as:

$$\mathcal{L}'_{equi} = \frac{1}{|\Omega|^2} \sum_{u \in \Omega} \sum_{v \in \Omega} \|gu - v\| \; p(v|u; \Phi, \mathbf{x}, \mathbf{x}'). \quad (3)$$

Here $p(v|u; \Phi, \mathbf{x}, \mathbf{x}')$ is the probability of pixel u in image $\mathbf{x}$ matching v in image $\mathbf{x}'$ with Φ as the encoder shared by $\mathbf{x}$ and $\mathbf{x}'$ computed as below, and $\tau \in \mathbb{R}^+$ is a scale parameter:

$$p(v|u; \Phi, \mathbf{x}, \mathbf{x}') = \frac{\exp(\langle \Phi_u(\mathbf{x}), \Phi_v(\mathbf{x}')\rangle / \tau)}{\sum_{t \in \Omega} \exp(\langle \Phi_u(\mathbf{x}), \Phi_t(\mathbf{x}')\rangle / \tau)}. \quad (4)$$

Invariant learning. Contrastive learning is based on the similarity over pairs of inputs (Fig. 1b). Given an image $\mathbf{x}$ and its transformation $\mathbf{x}'$ as well as other images $\mathbf{z}_i$, $i \in \{1, 2 \ldots N\}$, the InfoNCE [40] loss minimizes:

$$\mathcal{L}_{inv} = -\log \frac{\exp\left(\langle \Phi(\mathbf{x}), \Phi(\mathbf{x}')\rangle\right)}{\sum_{i=1}^{N} \exp(\langle \Phi(\mathbf{x}), \Phi(\mathbf{z}_i)\rangle)}. \quad (5)$$

The objective encourages representations to be invariant to transformations while being distinctive across images. To address the computational bottleneck in evaluating the denominator, Momentum Contrast (MoCo) [23] computes

[1] Note that MoCo [23] was evaluated on pose estimation, however, their method was trained with 150K labeled examples and the entire network was fine-tuned.

the loss over negative examples using a dictionary queue and updates the parameters based on momentum.

Transformations. The space of transformations used to generate image pairs $(\mathbf{x}, \mathbf{x}')$ plays an important role in learning. A common approach is to apply a combination of *geometric transformations*, such as cropping, resizing, and thin-plate spline warping, as well as *photometric transformations*, such as color jittering and adding JPEG noise. Transformations can also denote channels of an image or modalities such as depth and color [52].

Hypercolumns. A deep network of n layers (or blocks[2]) can be written as $\Phi(\mathbf{x}) = \Phi^{(n)} \circ \Phi^{(n-1)} \circ \cdots \circ \Phi^{(1)}(\mathbf{x})$. A representation $\Phi(\mathbf{x})$ of size $H' \times W' \times C$ can be spatially interpolated to the input size $H \times W \times C$ to produce a pixel representation $\Phi_u(\mathbf{x}) \in \mathbb{R}^C$. The hypercolumn representation of layers $k_1, k_2, \ldots, k_n$ is obtained by concatenating the interpolated features from the corresponding layers, that is, $\Phi_u(\mathbf{x}) = \Phi_u^{(k_1)}(\mathbf{x}) \oplus \Phi_u^{(k_2)}(\mathbf{x}) \oplus \cdots \oplus \Phi_u^{(k_n)}(\mathbf{x})$.

3.2. Approach

Given a large unlabeled dataset, we first train representations using instance-discriminative contrastive learning framework of MoCo [23]. A combination of geometric and photometric transformations are applied to generate pairs $(\mathbf{x}, \mathbf{x}')$. We then extract single layer or hypercolumn representations from the trained network to represent landmarks (Fig. 1c). Subsequently, we incorporate a spatial contrastive learning to reduce dimensionality and induce spatial diversity by training a linear projector over the frozen landmark representation. Let $w \in \mathbb{R}^{C \times d}$, where $d \ll C$, used to project the landmark representation as $\Phi'_u(x) = w^T \Phi_u(x)$. The goal that the projected embeddings are spatially distinct within the same image, *i.e.*,

$$\forall u, v \in \Omega : u \neq v \Leftrightarrow \Phi'_u(\mathbf{x}) \neq \Phi'_v(\mathbf{x}), \tag{6}$$

is obtained by optimizing objective in Eqn. 3 with $\mathbf{x}' = \mathbf{x}$.

Discussion. Note that since the linear projection is location-wise, spatial equivariance is preserved but intra-image contrast is improved. The projected embeddings are equally effective as the hypercolumn representations for landmark regression, but are significantly better for landmark matching (Tab. 1). The intuition is that the hypercolumn features contain sufficient information about landmarks, but the projection step makes them spatially distinct which is more suitable for matching. Novotny *et al*. [39] proposed a similar approach to extract compact representations for cross-instance semantic matching from a network pre-trained with class labels. In comparison, we only use unsupervised representations. The idea of spatially contrastive learning has also been shown to be effective for learning scene-level representations [43].

[2]Due to skip-connections, we cannot decompose the encoding over layers, but can across blocks.

4. Experiments

We first outline the datasets and implementation details of the proposed method (§ 4.1). We then evaluate our model and provide comparisons to the existing methods qualitatively and quantitatively on landmark matching (§ 4.2) and landmark detection benchmarks (§ 4.3). We conclude with ablation studies and discussions (§ 4.4).

4.1. Benchmarks and implementation details

Human faces. We first compare the proposed model with prior art on the existing human face landmark detection benchmarks. Following DVE [49], we train our model on aligned CelebA dataset [33] and evaluate on MAFL [67], AFLW [30], and 300W [44]. The overlapping images with MAFL are excluded from CelebA. MAFL comprises 19,000 training images and 1000 test images with annotations on 5 face landmarks. Two versions of AFLW are used: AFLW$_M$ which contains 10,122 training images and 2995 testing images, which are crops from MTFL [66]; AFLW$_R$ which contains tighter crops of face images with 10,122 for training and 2991 for testing. 300W provides 68 annotated face landmarks with 3148 training images and 689 test images. We apply the same image pre-processing procedures as in DVE, the current state-of-the-art, for a direct comparison. We also train our model on the unaligned raw CelebA dataset to evaluate the efficiency of representation learning on in-the-wild unlabeled images.

Birds. We collect a challenging dataset of birds where objects appear in clutter, occlusion, and exhibit wider pose variation. We randomly select 100K images of birds from the iNaturalist 2017 dataset [55] under the "Aves" class to train unsupervised representations. For the performance in the few-shot setting, we collect a subset of CUB dataset [56] containing 35 species of *Passeroidea*[3] super-family, each annotated with 15 landmarks. We sample at most 60 images per class which results in 1241 images as our training set, 382 as validation set, and 383 as test set (see appendix for the details).

Evaluation. We use landmark matching and detection as the end tasks for evaluation. In landmark matching, following DVE [49], we generate 1000 pairs of images from the MAFL test set as the benchmark, among which 500 are pairs of the same identity obtained by warping images with thin-plate spline (TPS) deformation, and others are pairs of different identities. Each pair of images consists of a reference image with landmark annotations and a target image. We use the nearest neighbor matching with cosine distance between pixel representations for landmark matching, and report the mean pixel error between the predicted landmarks and the ground-truth landmarks.

In the landmark regression task, following [49, 50], we

[3]This is the biggest Aves taxa in iNaturalist.

Figure 2. **Visualization of landmark matching** with cosine distance using 3840-D hypercolumn features and 256-D features projected from hypercolumn. Failure cases of using hypercolumn includes (Left) mismatching between two eyes and (Middle) lack of robustness to large viewpoint or (Right) appearance changes across different identities. The proposed feature projection method alleviates these issues.

Method	Dim.	Aligned		In-the-wild	
		Same	Diff.	Same	Diff.
DVE	64	0.92	2.38	1.27	3.52
Ours	3840	**0.73**	6.16	**0.78**	5.58
Ours + proj.	256	**0.71**	**2.06**	**0.96**	**3.03**
Ours + proj.	128	**0.82**	**2.19**	**0.98**	**3.05**
Ours + proj.	64	**0.92**	2.62	**0.99**	**3.06**

Table 1. **Landmark matching results.** We report the mean pixel error between the predicted landmarks and the ground-truth across 1000 pairs of images from MAFL (*lower is better*). The test set consists of 500 same-identity and 500 different-identity pairs. We compare DVE [49] with Hourglass net and our models with ResNet50 trained from aligned or in-the-wild CelebA dataset. We also evaluate the effect of feature projection (+proj.) with different output dimensions. Our results better than DVE's [49] are marked in **bold**.

train a linear regressor to map the representations to landmark annotations while keeping the representations frozen. The landmark regressor is a linear regressor per target landmark. Each regressor consists of K filters of size $1\times1\times C$ on top of a C-dimensional representation to generate K intermediate heatmaps, which are then converted to spatial coordinates by `soft-argmax` operation. These K coordinates are finally converted to the target landmark by a linear layer (see appendix for details). We use $K = 50$ to keep the evaluation consistent with prior works [49, 50], but we find that this hyperparameter is not critical (see Sec. 4.4). We report errors in the percentage of inter-ocular distance on face benchmarks and the percentage of correct keypoints (PCK) on CUB. A prediction is considered correct according to the PCK metric if its distance to the ground-truth is within 5% of the longer side of the image. The occluded landmarks are ignored during evaluation. We did not find fine-tuning to be uniformly beneficial but include a comparison in the appendix.

Implementation details. We use MoCo [23] to train our models on CelebA or iNat Aves for 800 epochs with a batch size of 256 and a dictionary size of 4096. ResNet18 or ResNet50 [24] are used as our backbones. We extract hypercolumns [22] per pixel by stacking activations from the second (`conv2_x`) to the last convolutional block (`conv5_x`). We resize the feature maps from the selected convolutional blocks to the same spatial size as DVE [49] (*i.e*. 48×48). We also follow DVE (with Hourglass network) to resize the input image to 136×136 then center-crop the image to 96×96 for face datasets. Images are resized to 96×96 without any cropping on the bird dataset. For a comparison with DVE on the CUB dataset we used their publicly available implementation. More details are in the appendix.

4.2. Landmark matching

Quantitative results. Tab. 1 compares the proposed method with DVE [49] quantitatively. We train DVE and our models on both aligned and in-the-wild unaligned version of CelebA dataset, and report the mean pixel error on aligned face images from MAFL. Our hypercolumn representation has high performance in same-identity matching but is not robust to cross-identity variations. However, the proposed feature projection makes the hypercolumn more suitable for landmark matching. We experiment with different feature dimensions after projection and find that our method with 128 or higher dimensional features achieves the state-of-art. DVE outperforms ours with 64-D features when the representations are learned on the aligned CelebA dataset. This is because the architecture of the Hourglass network and the joint training of the backbone and feature extractor enables DVE to learn a more compact representation than our method. However, to lift the feature dimension from 64 to 256, DVE requires re-training the entire model while we only need to re-train a linear feature projector. Moreover, when the representation is learned from the in-the-wild CelebA dataset, our model *outperforms DVE by a large margin*. This suggests our representation is more invariant to nuisance factors than that of DVE. We also observe that our method with smaller networks (*e.g*. ResNet18) with 128-D projected features outperforms DVE, and both DVE and our methods outperform representations from ImageNet pretrained networks. (see appendix for more details).

Qualitative results. Fig. 2 presents the qualitative results of landmark matching. Our method with hypercolumn for

Method	# Params. Millions	Unsuper.	MAFL	AFLW$_M$	AFLW$_R$	300W	CUB
			Inter-ocular Distance (%) ↓				PCK ↑
TCDCN [67]	–	×	7.95	7.65	–	5.54	–
RAR [59]	–	×	–	7.23	–	4.94	–
MTCNN [66]	–	×	5.39	6.90	–	–	–
Wing Loss [14]	–	×	–	–	–	4.04	–
Generative modeling based							
Structural Repr. [65]	–	✓	3.15	–	6.58	–	–
FAb-Net [57]	–	✓	3.44	–	–	5.71	–
Deforming AE [47]	–	✓	5.45	–	–	–	–
ImGen. [28]	–	✓	2.54	–	6.31	–	–
ImGen.++ [29]	–	✓	–	–	–	5.12	–
Equivariance based							
Sparse [51]	–	✓	6.67	10.53	–	7.97	–
Dense 3D [50]	–	✓	4.02	10.99	10.14	8.23	–
DVE SmallNet [49]	0.35	✓	3.42	8.60	7.79	5.75	–
DVE Hourglass [49]	12.61	✓	2.86	7.53	6.54	4.65	61.91
Invariance based							
Ours (ResNet18)	11.24	✓	**2.57**	8.59	7.38	5.78	**62.24**
Ours (ResNet18 + proj.)	11.24	✓	**2.71**	**7.23**	**6.30**	5.20	58.49
Ours (ResNet50)	23.77	✓	**2.44**	**6.99**	**6.27**	5.22	**68.63**
Ours (ResNet50 + proj.)	23.77	✓	**2.64**	**7.17**	**6.14**	4.99	**62.55**

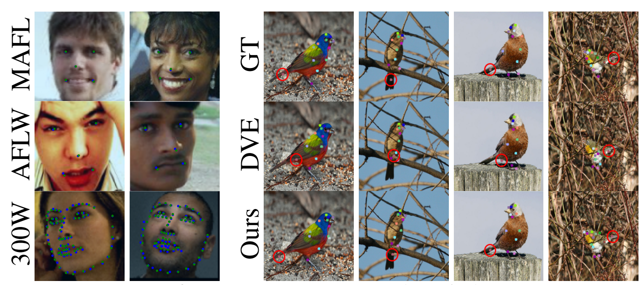
Table 2. **Results on landmark detection.** Comparison on face benchmarks, including MAFL, AFLW$_M$, AFLW$_R$, and 300W, and CUB dataset. We report the error in the percentage of inter-ocular distance on human face dataset (*lower is better*), and the percentage of correct keypoints (PCK) on CUB dataset (*higher is better*). We project the hypercolumn (*i.e.* + proj.) to 256-D features on face and 512-D on bird dataset and provide results of other dimensions in the appendix. Our results better than DVE's [49] are marked in **bold**.

matching is not robust to viewpoint and appearance changes, and frequently mismatches the left and right eyes. Incorporating the proposed feature projection adds diversity effectively and solves these issues.

4.3. Landmark detection

Quantitative results. Tab. 2 presents a quantitative evaluation of multiple benchmarks. On faces, our model with a ResNet50 achieves state-of-the-art results on all benchmarks except for 300W. On iNat Aves → CUB, out approach outperforms prior state-of-the-art [49] by a large margin, suggesting improved invariance to nuisance factors. Incorporating the feature projection results in small performance degradation in some cases but remains the state-of-art. Our method with ResNet18 is comparable with DVE and benefits from using a deeper network. We present more results of landmark detection under different configurations in the appendix.

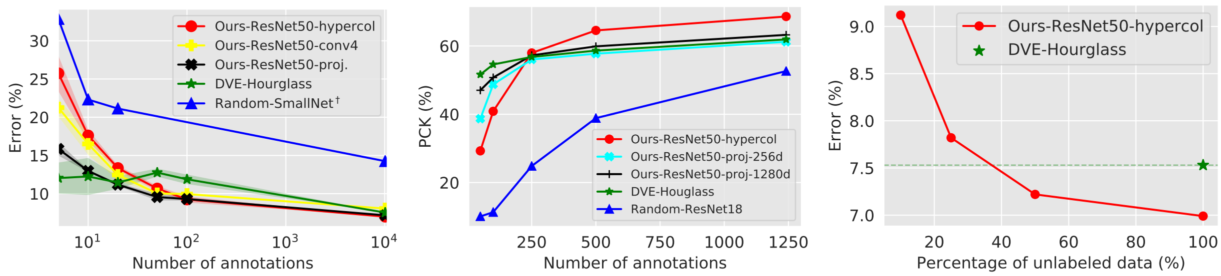
Qualitative results. Fig. 3 shows qualitative results of landmark regression on human faces and birds. We notice that both DVE and our model with hypercolumn representations are able to localize the foreground object accurately. However, our model localizes many keypoints better (*e.g.*, on the tails of the birds) and is more robust to the background clutter (*e.g.*, the last column of Fig. 3b).



(a) Human face (b) Birds

Figure 3. **Detected landmarks (a)** on faces (*blue:* predictions, *green:* ground truth) and **(b)** on CUB. Notice that our method localizes tails of birds (circled) much better. *Zoom in for details.*

Limited annotations. Fig. 4a and 4b compare our model with DVE [49] using a limited number of annotations on AFLW$_M$ and CUB dataset respectively. Without feature projection, our performance is better as soon as a few training examples are available (*e.g.*, 50 on AFLW$_M$ and 250 on CUB). This can be attributed to the higher dimensional embedding of the hypercolumn representation. The scheme can be improved by using a single-layer representation as shown in the yellow line. Our feature projection further improves the performance in the low-data regime as shown in the black line. Interestingly, this improvement is not solely due to the dimension reduction: increasing the dimension of projected



(a) Limited anno. on AFLW$_M$ (b) Limited anno. on CUB (c) Unlabeled CelebA images

Figure 4. **The effect of dataset size. (a)** A comparison of our model with DVE [49] by varying the number of annotations for landmark regression on AFLW$_M$ dataset. `Random-SmallNet`†: is a randomly initialized "small network" taken from [49]. `Ours-ResNet50`: are based on hypercolumn, or its compact representations, or fourth-layer features trained using contrastive learning. **(b)** Similar results on CUB dataset. `Random-ResNet18`: is trained from scratch on the CUB dataset. **(c)** Results of landmark regression on AFLW$_M$ using different numbers of *unlabeled* images from CelebA for training.

feature from 256 to 1280 improves the performance across different dataset sizes on CUB (see Fig. 4b). Note that all unsupervised learning models (including DVE and our model) outperform the randomly initialized baseline on both human face and bird datasets. Numbers corresponding to Fig. 4 are in the appendix.

Limited unlabeled data. Fig. 4c shows that our model with hypercolumn representation matches the performance of DVE on AFLW$_M$ using only 40% of the images on the CelebA dataset. This suggests that invariances are acquired more efficiently in our framework.

4.4. Ablation studies and discussions

Hypercolumns. Tab. 3 compares the performance of using individual layer and hypercolumn representations. The activations from the fourth convolutional block consistently outperforms those from the other layers. For an input of size 96×96, the spatial dimension of the representation is 48× 48 at Layer #1 and 3×3 at Layer #5, reducing by a factor of two at each successive layer. Thus, while the representation loses geometric equivariance with depth, contrastive learning encourages invariance, resulting in Layer #4 with the optimal trade-off for this task. While the best layer can be selected with some labeled validation data, the hypercolumn representation provides further benefits everywhere except the very small data regime (Tab. 3 and Fig. 4a).

Dimensionality and linear regressor. In Tab. 4, we reduce the size of the landmark regressor to evaluate its effect on the landmark regression performance. We chose 50 intermediate landmarks to keep the evaluation consistent with DVE. However, the choice is not critical as seen by the performance of a smaller linear regressor. There is a small drop in performance, while it remains comparable to DVE. The proposed feature projection with equivariant learning is more effective than non-negative matrix factorization (NMF), a classical dimension reduction method.

Method	C	K	#P	MAFL	AFLW$_M$	AFLW$_R$	300W
DVE	64	50	17	2.86	7.53	6.54	4.65
Ours	3840	50	961	**2.44**	**6.99**	**6.27**	5.22
Ours	3840	10	192	**2.40**	**7.27**	**6.30**	5.40
Ours+proj.	256	50	65	**2.64**	**7.17**	**6.14**	4.99
Ours+proj.	256	10	13	**2.67**	**7.24**	**6.23**	5.07
Ours+proj.	64	50	17	**2.77**	**7.21**	**6.22**	5.19
Ours+NMF	64	50	17	**2.80**	7.60	6.69	5.62

Table 4. **The effect of landmark regressor on landmark regression**. We vary the number of parameters (#P in thousands) in the landmark regressor by changing the number of intermediate landmarks (K) and feature dimensions (C). We compare the proposed feature projection (*i.e.* +proj.) with non-negative matrix factorization (NMF) for dimension reduction. Our results better than DVE's [49] are marked in **bold**.

Effectiveness of unsupervised learning. Tab. 5 compares representations using the linear evaluation setting for randomly initialized, ImageNet pretrained, and contrastively learned networks using a hypercolumn representation. Contrastive learning provides significant improvements over ImageNet pretrained models, which is less surprising since the domain of ImageNet images is quite different from faces. Interestingly, random networks have competitive performances with respect to some prior work in Tab. 2. For example, [50] achieve 4.02% on MAFL, while a randomly initialized ResNet18 with hypercolumns achieves 4.00%.

Are the learned representations semantically meaningful? We found that parts can be reliably distilled from the learned representation using non-negative matrix factorization (NMF) (see [10] for another application of NMF for visualizing semantic parts from deep network activations). Fig. 5 shows two such components and a "map" of several components which are indicative of parts (left) and are robust to image transformations (right). Additionally, Fig. 2 shows that the correspondence obtained using nearest neighbor matching are semantically meaningful. Furthermore, our method can be naturally extended to figure-ground seg-

Dataset	Single layer					Hypercolumn			
	#1 (64)	#2 (256)	#3 (512)	#4 (1024)	#5 (2048)	#4 - #5 (3072)	#3 - #5 (3584)	#2 - #5 (3840)	#1 - #5 (3904)
MAFL	5.77	4.58	3.03	**2.73**	3.66	2.73	2.65	**2.44**	2.51
AFLW$_M$	24.20	21.34	11.95	**8.83**	11.55	8.14	8.31	**6.99**	7.40
AFLW$_R$	16.27	14.15	9.66	**7.37**	8.83	6.95	**6.24**	6.27	6.34
300W	16.45	13.08	7.66	**6.01**	7.70	5.68	5.28	5.22	**5.21**

Table 3. **Landmark detection using single layer and hypercolumn representations.** The error is reported in the percentage of inter-ocular distance using linear regression over individual layers (left) and combinations (right), with a ResNet50. The embedding dimension for each is shown in parenthesis. Layer #4 performs the best across datasets, while hypercolumns offer an improvement.

Network	Supervision	MAFL	AFLW$_M$	AFLW$_R$	300W
Res. 18	Random	4.00	14.20	10.11	9.88
	ImageNet	2.85	8.76	7.03	6.66
	Contrastive	2.57	8.59	7.38	5.78
Res. 50	Random	4.72	16.74	11.23	11.70
	ImageNet	2.98	8.88	7.34	6.88
	Contrastive	**2.44**	**6.99**	**6.27**	**5.22**

Table 5. **Effectiveness of unsupervised learning.** Error using randomly initialized, ImageNet pretrained, and contrastively trained ResNet50 for landmark detection. Frozen hypercolumn representations with a linear regression were used for all methods. The results with feature projection are included in the appendix.

Figure 5. **Semantic parts distillation.** The object parts distilled from our representation using NMF are semantically meaningful and consistent across different instances (left). The parts are also robust to geometric transformations (right).

mentation. In appendix we show that the proposed model outperforms an ImageNet pretrained model in few-shot settings (*e.g.* given only 10 annotated images).

Commonalities and differences. Equivariance is necessary but not sufficient for an effective landmark representation. It also needs to be distinctive or invariant to nuisance factors. This is enforced in the equivariance objective (Eqn. 3) as a contrastive term over locations within the same image, as the loss is minimized when $p(v|u;\Phi,\mathbf{x},\mathbf{x}')$ is maximized at $v = gu$. This encourages intra-image invariance, unlike the objective of contrastive learning (Eqn. 5) which encourages inter-image invariance. However, a single image may contain enough variety to guarantee some invariance. This is supported by its empirical performance and recent work showing that representation learning is possible even from a single image [62]. However, our experiments suggest that inter-image invariance can be more effective on datasets with greater clutter, occlusion, and pose variations.

Is there any advantage of one approach over the other? Our experiments show that for a deep network of the same size, invariant representation learning can be just as effective (Tab. 2). However, invariant learning is conceptually simpler and scales better than equivariance approaches, as the latter maintains high-resolution feature maps across the hierarchy. Using a deeper network (*e.g.*, ResNet50 vs. ResNet18) gives consistent improvements, outperforming DVE [49] on four out of five datasets, as shown in Tab. 2. A drawback of our approach is that the hypercolumn representation is not directly interpretable or compact, which results in lower performance in the extreme few-shot case. However, as seen in Fig. 4a, the advantage disappears with as few as 50 training examples on the AFLW benchmark. This problem can be effectively alleviated by learning a compact representation using equivariant learning which further reduces the number of required training examples to 20. Invariant learning is also more data-efficient and can achieve the same performance with half the unlabeled examples, as seen in Fig. 4c.

5. Conclusion

We show that intermediate layer representations of a deep network trained using instance-discriminative contrastive learning outperform landmark representation learning approaches that are based on unsupervised equivariant learning alone. We also show that equivariant learning approaches can be viewed through the lens of a (spatial) contrastive learning, resulting in weaker generalization than inter-image invariances for landmark recognition tasks. However, these two forms of contrastive learning are complementary and we use the latter to learn a compact representation which is better suited for landmark matching tasks. We illustrate our results on existing benchmarks and a new challenging one where there is a larger variation in pose and viewpoint, where the improvements using our approach are more pronounced.

Acknowledgements. We would like to thank Erik Learned-Miller, Daniel Sheldon, Rui Wang, Huaizu Jiang, Gopal Sharma, and Zitian Chen for discussion and feedback. The project is supported by NSF grants #1661259 and #1749833. Our experiments were performed on the UMass GPU cluster funded by the Mass. Tech. Collab.

References

[1] Alessandro Achille and Stefano Soatto. Emergence of invariance and disentanglement in deep representations. *The Journal of Machine Learning Research*, 2018. 2

[2] Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Netvlad: Cnn architecture for weakly supervised place recognition. In *CVPR*, 2016. 2

[3] Philip Bachman, R Devon Hjelm, and William Buchwalter. Learning representations by maximizing mutual information across views. In *NeurIPS*, 2019. 1, 3

[4] Iaroslav Bespalov, Nazar Buzun, and Dmitry V Dylov. Brul\'e: Barycenter-regularized unsupervised landmark extraction. *arXiv preprint arXiv:2006.11643*, 2020. 3

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *ICML*, 2020. 1, 2, 3

[6] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. Big self-supervised models are strong semi-supervised learners. *NeurIPS*, 2020. 2, 3

[7] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 2, 3

[8] Mircea Cimpoi, Subhransu Maji, and Andrea Vedaldi. Deep filter banks for texture recognition and segmentation. In *CVPR*, 2015. 2

[9] Taco Cohen and Max Welling. Group equivariant convolutional networks. In *ICML*, 2016. 2

[10] Edo Collins, Radhakrishna Achanta, and Sabine Susstrunk. Deep feature factorization for concept discovery. In *ECCV*, 2018. 7

[11] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005. 2

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 2

[13] Alexey Dosovitskiy, Jost Tobias Springenberg, Martin Riedmiller, and Thomas Brox. Discriminative unsupervised feature learning with convolutional neural networks. In *NeurIPS*, 2014. 1, 3

[14] 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 *CVPR*, 2018. 6

[15] Jianlong Fu, Heliang Zheng, and Tao Mei. Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition. In *CVPR*, 2017. 2

[16] Robert Gens and Pedro M Domingos. Deep symmetry networks. In *NeurIPS*, 2014. 2

[17] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. In *ICLR*, 2018. 3

[18] Abel Gonzalez-Garcia, Davide Modolo, and Vittorio Ferrari. Do semantic parts emerge in convolutional neural networks? *IJCV*, 2018. 2

[19] Priya Goyal, Dhruv Mahajan, Abhinav Gupta, and Ishan Misra. Scaling and benchmarking self-supervised visual representation learning. In *ICCV*, 2019. 3

[20] Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *AISTATS*, 2010. 3

[21] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *CVPR*, 2006. 1, 3

[22] Bharath Hariharan, Pablo Arbeláez, Ross Girshick, and Jitendra Malik. Hypercolumns for object segmentation and fine-grained localization. In *CVPR*, 2015. 2, 5

[23] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020. 1, 2, 3, 4, 5

[24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 5

[25] Olivier J Hénaff, Aravind Srinivas, Jeffrey De Fauw, Ali Razavi, Carl Doersch, SM Eslami, and Aaron van den Oord. Data-efficient image recognition with contrastive predictive coding. *arXiv preprint arXiv:1905.09272*, 2019. 1

[26] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. In *ICLR*, 2019. 1, 3

[27] Gary Huang, Marwan Mattar, Honglak Lee, and Erik G Learned-Miller. Learning to align from scratch. In *NeurIPS*, 2012. 1

[28] Tomas Jakab, Ankush Gupta, Hakan Bilen, and Andrea Vedaldi. Unsupervised learning of object landmarks through conditional image generation. In *NeurIPS*, 2018. 1, 2, 3, 6

[29] Tomas Jakab, Ankush Gupta, Hakan Bilen, and Andrea Vedaldi. Self-supervised learning of interpretable keypoints from unlabelled videos. In *CVPR*, 2020. 1, 3, 6

[30] Martin Koestinger, 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 *ICCVW*, 2011. 2, 4

[31] Karel Lenc and Andrea Vedaldi. Understanding image representations by measuring their equivariance and equivalence. In *CVPR*, 2015. 2

[32] Tsung-Yu Lin, Aruni RoyChowdhury, and Subhransu Maji. Bilinear cnn models for fine-grained visual recognition. In *ICCV*, 2015. 2

[33] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *ICCV*, 2015. 4

[34] Dominik Lorenz, Leonard Bereska, Timo Milbich, and Bjorn Ommer. Unsupervised part-based disentangling of object shape and appearance. In *CVPR*, 2019. 3

[35] David G Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 2004. 2

[36] Aravindh Mahendran and Andrea Vedaldi. Visualizing deep convolutional neural networks using natural pre-images. *IJCV*, 2016. 2

[37] Erik G Miller, Nicholas E Matsakis, and Paul A Viola. Learning from one example through shared densities on transforms. In *CVPR*, 2000. 1

[38] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *ECCV*, 2016. 3

[39] David Novotny, Diane Larlus, and Andrea Vedaldi. Anchornet: A weakly supervised network to learn geometry-sensitive features for semantic matching. In *CVPR*, 2017. 4
[40] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 1, 3
[41] Maxime Oquab, Léon Bottou, Ivan Laptev, and Josef Sivic. Is object localization for free?-weakly-supervised learning with convolutional neural networks. In *CVPR*, 2015. 2
[42] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *CVPR*, 2016. 3
[43] Pedro O Pinheiro, Amjad Almahairi, Ryan Y Benmaleck, Florian Golemo, and Aaron Courville. Unsupervised learning of dense visual representations. *NeurIPS*, 2020. 4
[44] Christos Sagonas, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *ICCVW*, 2013. 4
[45] Enrique Sanchez and Georgios Tzimiropoulos. Object landmark discovery through unsupervised adaptation. In *NeurIPS*, 2019. 3
[46] Pierre Sermanet, Andrea Frome, and Esteban Real. Attention for fine-grained categorization. *ICLRW*, 2015. 2
[47] Zhixin Shu, Mihir Sahasrabudhe, Riza Alp Guler, Dimitris Samaras, Nikos Paragios, and Iasonas Kokkinos. Deforming autoencoders: Unsupervised disentangling of shape and appearance. In *ECCV*, 2018. 1, 3, 6
[48] Josef Sivic and Andrew Zisserman. Video google: A text retrieval approach to object matching in videos. In *ICCV*, 2003. 2
[49] James Thewlis, Samuel Albanie, Hakan Bilen, and Andrea Vedaldi. Unsupervised learning of landmarks by descriptor vector exchange. In *ICCV*, 2019. 1, 2, 3, 4, 5, 6, 7, 8
[50] James Thewlis, Hakan Bilen, and Andrea Vedaldi. Unsupervised learning of object frames by dense equivariant image labelling. In *NeurIPS*, 2017. 1, 2, 3, 4, 5, 6, 7
[51] James Thewlis, Hakan Bilen, and Andrea Vedaldi. Unsupervised learning of object landmarks by factorized spatial embeddings. In *ICCV*, 2017. 1, 2, 3, 6
[52] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. *ECCV*, 2020. 1, 3, 4
[53] Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000. 2
[54] Naftali Tishby and Noga Zaslavsky. Deep learning and the information bottleneck principle. In *2015 IEEE Information Theory Workshop (ITW)*, 2015. 2
[55] 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 *CVPR*, 2018. 2, 4
[56] 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. 2, 4
[57] Olivia Wiles, A Koepke, and Andrew Zisserman. Self-supervised learning of a facial attribute embedding from video. In *BMVC*, 2018. 1, 3, 6
[58] Zhirong Wu, Yuanjun Xiong, Stella Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *CVPR*, 2018. 1, 3
[59] Shengtao Xiao, Jiashi Feng, Junliang Xing, Hanjiang Lai, Shuicheng Yan, and Ashraf Kassim. Robust facial landmark detection via recurrent attentive-refinement networks. In *ECCV*, 2016. 6
[60] Tianjun Xiao, Yichong Xu, Kuiyuan Yang, Jiaxing Zhang, Yuxin Peng, and Zheng Zhang. The application of two-level attention models in deep convolutional neural network for fine-grained image classification. In *CVPR*, 2015. 2
[61] Yinghao Xu, Ceyuan Yang, Ziwei Liu, Bo Dai, and Bolei Zhou. Unsupervised landmark learning from unpaired data. *arXiv preprint arXiv:2007.01053*, 2020. 1, 3
[62] Asano YM., Rupprecht C., and Vedaldi A. A critical analysis of self-supervision, or what we can learn from a single image. In *ICLR*, 2020. 8
[63] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *ECCV*. Springer, 2014. 2
[64] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *ECCV*, 2016. 3
[65] Yuting Zhang, Yijie Guo, Yixin Jin, Yijun Luo, Zhiyuan He, and Honglak Lee. Unsupervised discovery of object landmarks as structural representations. In *CVPR*, 2018. 1, 2, 3, 6
[66] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. Facial landmark detection by deep multi-task learning. In *ECCV*, 2014. 4, 6
[67] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. Learning deep representation for face alignment with auxiliary attributes. *IEEE TPAMI*, 2015. 3, 4, 6
[68] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *CVPR*, 2016. 2
[69] Chengxu Zhuang, Alex Lin Zhai, and Daniel Yamins. Local aggregation for unsupervised learning of visual embeddings. In *ICCV*, 2019. 1