From Keypoints to Object Landmarks via Self-Training Correspondence: A novel approach to Unsupervised Landmark Discovery

Dimitrios Mallis, Enrique Sanchez, Matt Bell and Georgios Tzimiropoulos

Abstract—This paper proposes a novel paradigm for the unsupervised learning of object landmark detectors. Contrary to existing methods that build on auxiliary tasks such as image generation or equivariance, we propose a self-training approach where, departing from generic keypoints, a landmark detector and descriptor is trained to improve itself, tuning the keypoints into distinctive landmarks. To this end, we propose an iterative algorithm that alternates between producing new pseudo-labels through feature clustering and learning distinctive features for each pseudo-class through contrastive learning. With a shared backbone for the landmark detector and descriptor, the keypoint locations progressively converge to stable landmarks, filtering those less stable. Compared to previous works, our approach can learn points that are more flexible in terms of capturing large viewpoint changes. We validate our method on a variety of difficult datasets, including LS3D, BBCPose, Human3.6M and PennAction, achieving new state of the art results. Code and models can be found at https://github.com/malldimi1/KeypointsToLandmarks.

Index Terms—Unsupervised Landmark Discovery, Self-Training, Clustering, Correspondence, Keypoints

1 INTRODUCTION

OBJECT parts, also known as landmarks, convey information about the shape and spatial configuration of an object in 3D space, especially for deformable objects like the human face, body and hand. Landmarks represent the locations of the specific parts with particular semantic meaning and thus follow an indexed configuration that is often manually designed.

The goal of landmark detection is to have a model that, for a particular instance of an object can estimate the locations of its parts or landmarks. Research in this field is mainly driven by supervised approaches, where sufficient amount of human-annotated data is provided. Common object categories used in part-based detection are faces [8], [17] or human bodies [38], [66], where thousands of annotated images with landmarks are available. However, as in many other Computer Vision disciplines, relying on human annotations to develop novel detectors is costly, and hence alternative methods based on unsupervised learning are being explored.

Unsupervised learning of object landmarks from a first glance seems an impossible task. A human annotator has understanding of the notion of objects and their parts, viewpoint invariance, occlusion and self-occlusion as well as examples of which landmarks to annotate in their disposal. On the contrary, unsupervised learning often relies on an auxiliary or proxy task, whereby the target task naturally arises as a latent process. Some techniques are either based on learning strong representations that can be mapped to manual landmarks using few images [61] or on discovering the landmarks from raw images through auxiliary proxy losses, such as equivariance [60], [61], [62], or tasks such as image generation [22], [53], [75]. Methods based on the *principle of equivariance* observe that a detector must be consistent under known synthetic image deformations and

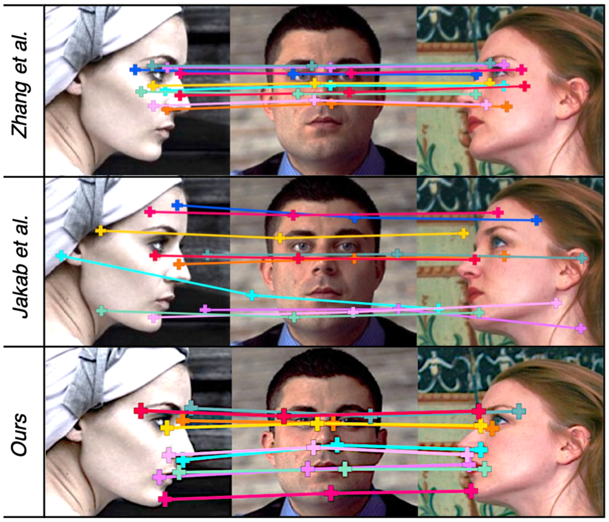


Fig. 1: Contrary to previous works that fail to cope with large viewpoint changes [22] or that fail to deal with object symmetries [75], our method finds correspondence across large viewpoint changes, leading to the discovery of landmarks that better represent the object's geometry.

- *Dimitrios Mallis is with the Computer Vision Lab, University of Nottingham, NG8 1BB, UK.*
  E-mail: dimitrios.mallis@nottingham.ac.uk
- *Enrique Sanchez is with Samsung AI Center Cambridge, CB1 2RE, UK.*
  E-mail: kike.sanc@gmail.com
- *Matt Bell is with the Department of Animal and Agriculture, Hartpury University, GL19 3BE , UK.*
  E-mail: matt.bell@hartpury.ac.uk
- *Georgios Tzimiropoulos is with the School of Electronic Engineering and Computer Science, Queen Mary University of London, E1 4NS , UK.*
  E-mail: g.tzimiropoulos@qmul.ac.uk

Manuscript submitted on May 2022.

attempt to optimise such objective. Methods based on image generation rely on reconstructing a deformed image through a generator that is conditioned on the detector's output; the detector and generator communicate through a bottleneck aimed to *distill the object's geometry*. For the generator to recover the input image from a deformed version of itself, the detector needs to produce meaningful landmarks.

While these approaches have shown good performance in limited scenarios where objects showcase little rigid deformation (frontal faces or bodies, shoes, cat faces, etc), they are limited, by definition, in two critical aspects. First, a proxy task does not enforce the explicit learning of object landmarks, and thus are prone to generate landmarks that would unlikely be selected by a human annotator. Second, these methods require synthetically generated deformations since local correspondences for unpaired images are not known in the unsupervised case. Learning from pairs of images where one is a synthetic deformation of the other results in representations with limited robustness to intraclass variation that may not generalise well to highly articulated objects like the human body, complicated backgrounds or large viewpoint changes (i.e. 3D rotations).

In this paper, we observe that, while landmark detectors are difficult to train in an unsupervised manner, generic keypoint detectors, on the contrary, are much simpler to obtain and thus propose a novel method that can convert the latter into the former. Generic keypoints, often also referred to as salient or interest points, are simply points in an image representing the locations where "something occurs", i.e. where there is a variation on the appearance, an edge, etc. Beyond representing a geometric position in an image, keypoints are represented by a feature descriptor, which is often used to find correspondences across different images (e.g. parts of two different images corresponding to different views of a building). Generic keypoints can be directly computed using Sobel filters (e.g. SIFT) or by training a detector on synthetic image deformations and homographic recovery (e.g. SuperPoint [16]).

Based on the similarities and differences between keypoints and landmarks, our goal in this paper is to convert a series of keypoints automatically detected for a given object category into semantically coherent landmarks that describe the object parts, filtering and refining during the training process the corresponding landmark locations. To this end, we propose a novel approach that a) discovers landmarks through self-training instead of auxiliary objectives and b) captures intraclass variation from random image pairs.

Our main starting point consists of populating a dataset of images belonging to a target object category (e.g. faces, birds) with a set of keypoints. It is expected that some of these points will show some consistency and will systematically overlap with what we would refer to as landmarks. From this initial setup, our goal is to develop a self-training approach that can be used to learn a landmark detector in a fully unsupervised manner. In particular, this paper proposes a network akin to that of SuperPoint [16] (i.e. with a detector head and a descriptor head) that learns iteratively, through self-training, to locate a set of keypoints and to assign to each a distinctive descriptor that is landmark-consistent. Our goal is then to turn a keypoint detector into a landmark detector where the points capture the semantic meaning of a particular object in an unsupervised manner and re-label the training data accordingly. Then, a simple landmark detector based on heatmap regression can be trained as the final network. To this end, we propose to iteratively alternate between **pseudo-labelling of keypoints along with correspondence recovery, through descriptor clustering**, and **model self-training with produced pseudo-labels**.

We observe that, compared to previous works, our proposed approach is capable of learning landmarks that are more flexible in terms of capturing changes in 3D viewpoint. See for example Fig. 1. We demonstrate some of the favourable properties of our method on a variety of difficult datasets including LS3D [7], BBCPose [12], Human3.6M [20] and PennAction [74], notably without utilizing temporal information.

This manuscript extends and modifies our prior work [36] both methodologically and experimentally. In particular, while in [36] the number of landmarks to be discovered was part of the algorithm, we opt for keeping them fixed as in prior work [22], [53], [62], [72] by using a two-way K-means clustering algorithm (Sec. 3.6). In addition, we observe that the negative pair selection in [36] might lead to the sampling of negative pairs that only differ in their cluster assignment because they encode different viewpoints of the same landmark. To avoid such an effect, we modify the negative pair selection to account only for samples that come from the same image, ensuring negative pairs refer not only to different clusters, but also to negative landmarks. Finally, rather than originally populating the descriptors with those of the keypoint detector, we opt for a warm-up strategy that removes the dependency of our method in the quality of the initial descriptors. Experimentally, we conduct a thorough ablation study and include results in the challenging human pose dataset PennAction [74] as well as CatFaces [73] and Caltech-UCSD Birds [65]. The contributions of our work can be summarised as follows:

- We propose a novel view on the unsupervised discovery of geometrically meaningful landmarks that, instead of relying on proxy or auxiliary losses, uses a self-training strategy that refines an initial set of unindexed keypoints to endow them with geometrically-aware descriptors.
- To the best of our knowledge, our approach, which alternates between correspondence recovery for pseudo-labelling and a contrastive loss for feature learning, is the first to directly propose a geometrically aware objective for unsupervised discovery through pseudo-labelling.
- Contrary to previous works, our method can deal with viewpoint changes thanks to an over parameterisation of the feature space that accounts for viewpoint-specific descriptors of the same landmark.
- We conduct extensive ablation studies and deliver competitive results in various challenging tasks and object categories.

2 Related Work

This paper brings the reasoning behind **clustering algorithms** for self-supervised representation learning to iteratively refine **generic keypoints**, and endow them with semantic meaning, in a process commonly known as **unsupervised landmark discovery**. As such, we provide a brief review on these three topics, departing from the latter, as it constitutes the main goal of this paper.

2.1 Landmark Discovery

Our goal in this paper is to build a landmark detector Ψ that can be learned without human supervision. Landmarks convey semantic information about a particular object and serve the modelling of rigid and non-rigid deformations. Because of this, a landmark detector must be *equivariant* to geometric transformations g, i.e. if an image $\mathbf{x}$ undergoes an image deformation defined by $g(\mathbf{x})$, the detector must follow suit: $\Psi(g(\boldsymbol{x})) = g(\Psi(\boldsymbol{x}))$. Such a simple yet essential requirement was the driving force behind the first method on unsupervised landmark discovery [62], where a network is trained to produce K heatmaps from which the corresponding landmark locations are derived through a differentiable *softargmax* operator [70]. By imposing the equivariant constraint on images and known deformations, as well as by adding auxiliary losses to avoid trivial solutions, the network can discover a set of K meaningful landmarks. The concept of equivariance can also be extended and used to learn networks that are designed to output dense feature maps rather than heatmaps [61]. While such extension does not aim at "discovering" object landmarks, it is possible to learn, on a few-shot basis, a per-landmark regressor, i.e. a regressor from feature maps to landmarks from a handful set of annotated samples. A similar approach was also extended to learn object symmetries without regard to the specific task of landmark discovery [63]. The equivariance constraint was also used to learn dense feature representations that cope with intra-class variation by exchanging features [60] between images before applying equivariance.

The use of equivariance as a proxy task to learn landmark detectors is usually prone to finding landmarks that do not have a proper semantic meaning (e.g. in the background). To avoid this issue, a different alternative consists of considering the proxy task of *image generation*, whereby a landmark detector is a necessary intermediate step to capture the geometry of an object for a decoder to generate a version of the input image [22], [53]. These frameworks share a common structure, consisting of a landmark detector, a "geometry distillation" bottleneck, and a conditional image generation. The detector and the bottleneck are meant to represent the object's geometry, which is forwarded to the conditional image generator along with a deformed version of the image. The whole pipeline is trained end-to-end with an image reconstruction loss. An alternative version [72] advocates for a differentiable autoencoder framework. Similar methods have also appeared, combining both equivariance and image generation for object feature representation [14], [28], [55], [57], or that attempt to disentangle pose from appearance [32], [56], which do not explicitly aim at learning object landmarks. These methods also suffer from the drawback of not being explicitly designed to produce semantically meaningful landmarks. On the contrary, our framework sets a novel direction whereby generic keypoints are transformed into semantically meaningful landmarks.

2.2 Keypoint detection

Keypoints, also known as salient or fiducial points, are used to represent the locations in an image that are of interest without regard to any semantic meaning. Keypoint detection is a critical step for any sparse image matching algorithm (Structure-from-Motion, Simultaneous Localisation and Mapping, 3D reconstruction, etc). Keypoints are accompanied by descriptors that allow their matching across different images, i.e. that allow *correspondence recovery*. Early works in keypoint detection and description were primarily based on computing local image variations, such as the histograms of the magnitude and orientation of image gradients (e.g. HOG [34], SIFT [35], SURF [5], and variants [1], [37], [52]) or the binary comparisons between neighbouring pixels(e.g. LBP [45]).

Lately, a there is an increasing interest in "learning" keypoint detectors and descriptors, using CNN-based approaches that can produce dense features [16], [29], [46]. Given that (in most cases) there is no concept of "ground-truth" keypoints, learning-based approaches work on an unsupervised setting, defining a proper proxy or auxiliary objective, e.g. invariance to viewpoint changes [29], [70], or feature discriminativeness [46]. In this paper, we study the feasibility of the keypoints detected by some of these methods to be converted into landmarks, observing that the strongest initialisation comes from those given by SuperPoint [16], which uses a three-stage approach with synthetic pre-training, homographic recovery, and discriminative matching.

2.3 Self-training via clustering

Self-training refers to a set of methods where a model's own predictions are used as pseudo-labels for model training.

Common methods for self-training can include converting the highly confident predictions into hard-labels [58], [67], the opposite [47], or applying a model ensemble [39]. Most self-training approaches focus on the task of image classification [47], [58], [67] whereby each training image is considered a particular class. Self-training is also applied for unsupervised segmentation [15], [25], foreground-background segmentation [18], [59] and salience object detection [71].

A recent line of methods for self-training rely on the concept of **clustering** to generate pseudo ground-truth annotations [3], [10], [11], [23], [30], [41], [68], [78]. These approaches are based on computing a set of clusters that can be used to "label" the training images. An optimization objective can be derived from these pseudo-labels, e.g. the typical cross-entropy [10], [40], a cluster identification [30], or even optimal transport problem [3], [11]. In all these cases, the ultimate goal is to learn a network that produces strong feature representations in an unsupervised setting to be applied to a downstream task thereafter. The generated pseudo-labels are not expected to convey any meaning or be kept after the training. To the best of our knowledge, we are the first to propose a self-clustering approach from pseudo-labels that are driven towards having a semantic meaning so as to populate the training set with the corresponding target labels.

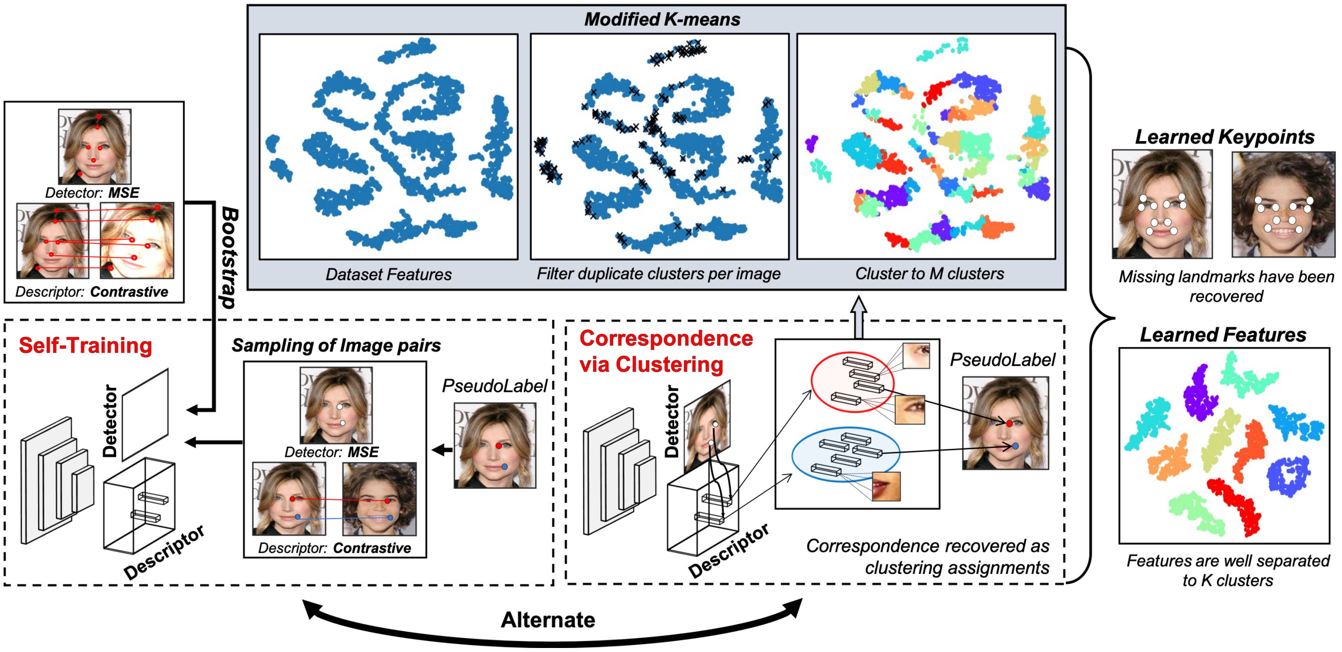


Fig. 2: Stage 1 of our proposed framework. A neural network is learned with two separate output heads (detector and descriptor head). During training, we alternate between correspondence recovery via clustering and self-training using the recovered correspondences. Training is bootstrapped by generic keypoints. In contrast to recent approaches, our framework enables learning of local features from unpaired image data. Correspondence is recovered via clustering following our Modified-KMeans algorithm. Our method is able to recover missing landmark locations and converge to well-separated features that can be used for accurate correspondence recovery. Dataset feature visualisation created through t-SNE [64].

3 METHOD

In this section we describe the different components of our approach.

3.1 Problem statement

Let $\mathcal{X} = \{\mathbf{x} \in \mathbb{R}^{W \times H \times 3}\}$ be a set of N images of a specific object category (e.g. faces, human bodies etc.). After running a generic keypoint detector on $\mathcal{X}$, our training set $\mathcal{X}$ becomes $\{\mathbf{x}_j, \{\mathbf{p}_i^j\}_{i=1}^{N_j}\}$, where $\mathbf{p}_i^j \in \mathbb{R}^2$ is a keypoint and N_j the number of detected keypoints in image $\mathbf{x}_j$. The original keypoints $\mathbf{p}^j$ for the j-th image are not ordered or in any correspondence with object landmarks. Also, multiple object landmarks will not be included in $\mathbf{p}^j$. Finally, some keypoints will be outliers corresponding to irrelevant background. Using only $\mathcal{X}$, our goal is to train a neural network $\mathbf{\Psi} : \mathcal{X} \rightarrow \mathcal{Y}$, where $\mathcal{Y} \in \mathbb{R}^{H_o \times W_o \times K}$ is the space of output heatmaps representing confidence maps for each of the K object landmarks we wish to detect. Note that the structure of $\mathcal{Y}$ implies that both order and landmark correspondence is recovered.

We will break down our problem into two stages. In the first stage, we will train a network $\mathbf{\Phi}$ producing a set of keypoints with landmark-aware descriptors, which aims to establish landmark correspondence, recover missing object landmarks and filter out irrelevant background keypoints. Then, we will use the output of this stage to train $\mathbf{\Psi}$ in a "supervised" way, using the pseudo-labels produced by $\mathbf{\Phi}$. Sections 3.2, 3.3, 3.4 and 3.5 are devoted to describing the first stage (**Stage1**) of our method, also depicted in Fig. 2. Section 3.6 describes the second stage (**Stage2**), and Section 3.7 introduces our flipping augmentation strategy.

3.2 Network Architecture

Our first stage comprises learning a network $\mathbf{\Phi}$ in a similar fashion to those of keypoint detectors, with a shared backbone $\mathbf{\Phi}_b : \mathcal{X} \rightarrow \mathcal{F}$ producing a set of intermediate features $\mathcal{F}$ and two heads: one for detecting the object landmarks $\mathbf{\Phi}_d$ and one for landmark-distinctive feature descriptor $\mathbf{\Phi}_f$.

The **detector head** $\mathbf{\Phi}_d$ will produce, for image $\mathbf{x}_j$, a single-channel spatial confidence map $H_j = \mathbf{\Phi}_d(\mathbf{\Phi}_b(\mathbf{x}_j)) \in \mathbb{R}^{H_o \times W_o \times 1}$ representing the presence/absence of an object landmark at a given location, without regard to any order or correspondence. We use non-maximum suppression to extract from H_j the landmark locations $\mathbf{p}_i^j$. The main purpose of $\mathbf{\Phi}_d$ is to recover the originally missed object landmarks, as well as to assign to each subsequent pseudo-label a corresponding spatial location.

The **feature extractor head** $\mathbf{\Phi}_f$ will produce for image $\mathbf{x}_j$ a dense feature map $\mathbf{F}_j = \mathbf{\Phi}_f(\mathbf{\Phi}_b(\mathbf{x}_j)) \in \mathbb{R}^{H_o \times W_o \times d}$ that will be used for **recovering correspondence**. At each landmark position $\mathbf{p}_i^j$ activated by the detector head, we will extract a d-dimensional feature descriptor $\mathbf{f}_i^j$ from $\mathbf{F}$. We use local features for recovering the correspondence of each individual keypoint through clustering.

3.3 Correspondence recovery

After applying $\mathbf{\Phi}$ on the training set, $\mathcal{X}$ becomes $\{\mathbf{x}_j, \{\mathbf{p}_i^j, \mathbf{f}_i^j\}_{i=1}^{N_j}\}$. Then, our first step in the iterative algorithm becomes using the features $\mathbf{f}$ to assign each keypoint a pseudo-label. We refer to this operation as correspondence recovery, as it allows us to identify correspondence of object parts across different images. To assign to each detected keypoint a pseudo-label, we follow [10] and perform K-means clustering on the collection of features $\mathbf{f}$. However, different from [10] where the clusters are used to make similar images have similar descriptors in an unsupervised way, our cluster assignment is indeed assigning a meaning label to a given keypoint. For this reason, we observe that it is important not to assign two different keypoints on a given image to the same cluster.

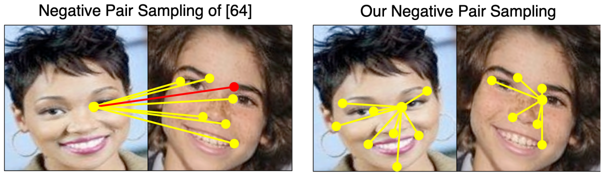
The clustering operation is then defined as:

$$\min_{\mathbf{C}\in\mathbb{R}^{d\times M}} \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N_j}\min_{\mathbf{y}_i^j\in\{0,1\}^M}\|\mathbf{f}_i^j-\mathbf{C}\mathbf{y}_i^j\|_2^2$$
$$\text{s.t.}\quad \mathbf{1}_M^T\mathbf{y}_i^j=1 \;\text{ and }\; \|\sum_j \mathbf{y}_i^j\|_0 = N_j, \qquad (1)$$

where M is the number of clusters, $\mathbf{y}_j^i$ is the cluster assignment for landmark $\mathbf{p}_j^i$ and $\boldsymbol{C}$ is the $d \times M$ centroid matrix. While in [36] the cluster assignment was performed using the Hungarian algorithm [27], here we opt for a simpler solution that does not compromise performance. For a given image i, we find $\{\mathbf{y}_i^j\}$, $\{\mathbf{f}_i^j\}$ by simply keeping, for each cluster k, the keypoint whose descriptor is closest to the centroid, i.e. we remove duplicate occurrences of the same cluster k on a single image. Enforcing a single keypoint per cluster for each image also provides a natural way of filtering out noisy keypoints. Given that a keypoint with a more representative feature has already been found for a cluster k in a particular image, it is likely that the second occurrence would be a noisy point.

We use this modified K-means formulation to recover correspondences for each detected keypoint. We note that it is crucial to use $M \gg K$: the resulting over-segmentation of the feature space enables the possibility of having several clusters per landmark, which is necessary for cases where viewpoint changes introduce large appearance changes. This differentiates our approach from prior works, which do not account for large out-of-plane rotations.

In addition to this modification, we also constrain our method to detect at most K landmarks per image. While in [36] the number of object landmarks was automatically discovered after progressive merging of similar clusters, here we restrict the detection to at most K clusters per image, in accordance with other recent unsupervised landmark detectors [22], [53], [62], [72]. We do that by additionally constraining $\boldsymbol{\Phi}_d$ to detect at most K keypoints per image (one per detected landmark). To that end, the modified K-means algorithm is executed *twice*: the first time, we cluster to K clusters to filter out duplicate occurrences of the same cluster in a single image (constraining out training set to at most K points per image). Note that the detection of less than K keypoints is allowed due to factors like occlusion. The second time we cluster the reduced set of features to M clusters to over segment the feature space and enable our method to recover multiple clusters per object landmark. An illustration of this in the form of a t-SNE [64] visualisation is shown in Fig. 2. Note that even though clustering is performed twice, using an accelerated similarity search method [24] this step can be executed very fast.



Feature pair are sampled on keypoints with different clustering assignments.

Feature pair are sampled on keypoints from the same image.

Fig. 3: Proposed negative pair mining strategy compared to [36]. In [36], negative pairs are sampled on keypoint locations with different clustering assignments. Since multiple clusters can track the same landmark, this can lead to inaccurate negative pairs *(red line)*. Sampling negatives from the same image, guarantees accurate pairs given that by definition, each landmark can only appear once per image.

3.4 Training Losses

After the correspondence recovery step described in Sec.3.3, the training set has now been augmented to include two different sets of pseudo-labels: the keypoint positions $\mathbf{p}_j$ and the corresponding cluster assignments $\mathbf{y}_j$. The next step consists then of training the network Φ, with both its backbone Φ_b and heads Φ_d and Φ_f, using the generated pseudo-labels. At the end of this step, the training set will be re-populated with the output's network: a new set of keypoints and descriptors will be generated, and new clustering assignments will be calculated.

The loss corresponding to the **detector head** is the standard MSE loss, defined as

$$\mathcal{L}_d(\mathbf{x}_j) = \|H(\mathbf{x}_j) - \boldsymbol{\Phi}_d(\boldsymbol{\Phi}_b(\mathbf{x}_j))\|^2, \qquad (2)$$

where the ground-truth heatmap H for a given image $\mathbf{x}_j$ is formed by placing 2D-Gaussian maps on each of the keypoint locations $\{\mathbf{p}_j^i\}_{i=1\ldots N_j}$. Our self-training approach confirms recent findings [2], [48] that show that over-parameterized neural networks tend to learn noiseless classes first, before overfitting to noisy labels in order to further reduce the training error. We observe such a pattern in learning object landmarks: a true landmark that commonly appears in the training set results in high detection confidence. Similarly, background locations that do not recurrently follow a specific pattern tend to be filtered out.

For the **feature extractor head** we propose the use of a contrastive loss. Note that this differs from [10] that uses a classifier to generate the pseudo-labels and a cross-entropy loss to update the network. Given the augmented training set at some training iteration t, $\mathcal{X}_t = \{\mathbf{x}_j, \{\mathbf{p}_i^j, \mathbf{y}_i^j\}_{i=1}^{N_j}\}$ our goal is to update Φ_f to produce features that, when extracted at some keypoints $\mathbf{p}_j^i$ and $\mathbf{p}_{j'}^{i'}$ for some locations j, j' on images i and i', respectively, are similar if and only if the corresponding pseudo clusters match, i.e. if $\mathbf{y}_i^j = \mathbf{y}_{i'}^{j'}$. To do so, we resort to a contrastive loss, where the goal is to bring pairs of features corresponding to the same cluster close whilst pulling features from different clusters apart. For a given pair of images $\mathbf{x}_j$ and $\mathbf{x}_{j'}$, and output locations i and i', the contrastive loss is formulated as:

$$\mathcal{L}_f(\mathbf{x}_i^j, \mathbf{x}_{i'}^{j'}) = \\ \mathbf{1}_{[y_i^j = y_{i'}^{j'}]} \|\mathbf{f}_i^j - \mathbf{f}_{i'}^{j'}\|^2 + \mathbf{1}_{[y_i^j \neq y_{i'}^j]} \max(0, m - \|\mathbf{f}_i^j - \mathbf{f}_{i'}^j\|^2),$$

where recall $\mathbf{f}_i^j = \Phi_f(\Phi_b(\mathbf{x}_j))^i$ is the d-dimensional feature vector extracted, for image j at the position $\mathbf{p}_i$, from the output of the feature head. A margin m is used to enforce features corresponding to negative pairs to be far apart.

As is common in unsupervised learning methods that build on contrastive learning, the choice of positive and negative pairs plays an important role in the learning process. Positive pairs can now be formed from different images where two keypoints are assigned the same cluster, as well as from two images where one is a synthetic deformation of the other. On the other side, negative pairs can be chosen in many different ways. While in [36] the negative feature pairs were selected randomly from the keypoint locations at different images (excluding those for which the pseudo-label was the same), in this work, we improve our negative mining by choosing all the negatives from the same image only. Given the over-segmentation of the underlying landmarks to M clusters, the same landmark in two different images could be assigned to different clusters, which would hinder the learning process. On the other hand, as noted above, each object landmark can only appear once per image. Thus, features extracted at any other location j' far from j, even when not corresponding to any proper keypoint $\mathbf{p}$, is a good, informative negative pair. An illustration of our negative pair mining strategy compared to that in [36] is show in Fig. 3.

Denoting by θ_b, θ_d and θ_f the parameters of $\mathbf{\Phi}_b$, $\mathbf{\Phi}_d$ and $\mathbf{\Phi}_f$, respectively, the full training procedure for Stage 1 is summarised in Algorithm 1.

Algorithm 1: Stage 1 training

```
Data: $\mathcal{X}_0 = \{\mathbf{x}_j, \{\mathbf{p}_i^j\}_{i=1}^{N_j}\}$
Compute $\mathbf{y}_i^j$ using Eqn. 1
Set $\mathcal{X}_0 = \{\mathbf{x}_j, \{\mathbf{p}_i^j, \mathbf{y}_i^j\}_{i=1}^{N_j}\}$
for $t = 1 : T$ do
    for $n = 1 : N_{iters}$ do
        Sample batch
        $(\theta_b, \theta_d) \leftarrow (\theta_b, \theta_d) - \nabla_{\theta_b, \theta_d} \mathcal{L}_d$
        $(\theta_b, \theta_f) \leftarrow (\theta_b, \theta_f) - \nabla_{\theta_b, \theta_f} \mathcal{L}_f$
    end
    Update $F$ and $p^j$ using frozen $\Phi$
    Compute $\mathbf{y}_i^j$ using Eqn. 1
    Update $\mathcal{X}_t = \{\mathbf{x}_j, \{\mathbf{p}_i^j, \mathbf{y}_i^j\}_{i=1}^{N_j}\}$

end
```

3.5 Bootstrapping

Initially, at round $t = 0$ the training set $\mathcal{X}_0$ only includes $\{\mathbf{x}_j, \{\mathbf{p}_i^j\}_{i=1}^{N_j}\}$ without point correspondences $\mathbf{f}_i^j$, needed for correspondence recovery as described in Sec. 3.3. In [36] the initial features were given by the generic keypoint descriptor, from where an initial clustering step could be performed. In this paper, we opt for a *warm up* pre-training stage where we train the feature extractor using only pairs of images in which one is a synthetic deformation of the other. We form known point correspondences through synthetic augmentations that can be used as initial positive pairs. This corresponds to initialising our backbone and feature extractor head using equivariance.

3.6 Learning an object landmark detector

At the end of Stage 1, the training set $\mathcal{X}$ is composed of a series of keypoints with landmark-aware descriptors. However, our goal is to train a network that can detect *a fixed number of K landmarks*.

Provided that the training set is now composed of $M \gg K$ clusters, training a landmark detector on K classes is not trivial because it is unknown which clusters correspond to the same landmark. In [36] this process was tackled by using a progressive merging step that was eventually reducing the number of clusters. However, thanks to the fact that the number of keypoints per image is now limited to K, as well as to the negative mining strategy, we observe that the learned features automatically form K well-separated clusters (as can be seen in Fig. 8 for $K = 30$). This observation thus eliminates the need for a progressive cluster merging step.

To finally populate our training set with K clusters only, we perform a last K-means clustering with K clusters only. Then, we can train $\mathbf{\Psi}$ using standard Heatmap Regression. For each image $\mathbf{x}_j$, we will produce a set of K heatmaps $H_i, k = 1, \ldots, K$ each of which is a Gaussian placed at the pseudo-ground truth landmark location for that image. An empty Heatmap is placed for the missing landmarks. The model is trained with an MSE loss over all output channels for which there is landmark-to-cluster assignment for that image: $L_d = \sum_k \|H(\mathbf{x}_k) - \Psi(\mathbf{x}_k)\|^2$.

3.7 Flipping augmentation

Flipping is a common augmentation strategy when training a landmark detector. In the supervised case, one can flip an image and mirror the ground-truth landmarks, given the naturally known correspondence between landmarks and their mirrored counterparts. In the unsupervised learning case, such correspondence is not known. In methods based on generative modelling or equivariance, one can only resort to flipping both the original and the synthetically generated image. This paper proposes to recover the symmetric landmark correspondences using clustering. At the correspondence recovery step (Sec. 3.3), pairs of features are sampled on both an image and its flipped version. We treat these features independently and produce 2 cluster assignments for each keypoint (one for the original and one for the flipped image). During the training of Stage 1, the cluster assignments of the flipped features are used when an image is randomly flipped. For Stage2, we find cluster symmetries by measuring maximal correspondence between clusters in the original and flipped images over the whole dataset. Note that in Stage2, flipping can be used both in training and test time as usually done with supervised landmark detectors.

4 Experimental Details

We first begin with describing the employed datasets (Sec. 4.1) as well as the general implementation details (Sec. 4.2). We then analyse the different parts of our method in Sec. 5, and compare the performance of our approach w.r.t. competing methods in Sec. 6.

4.1 Datasets

Facial datasets. We evaluate our method on the commonly used **CelebA-MAFL** [31], [76] and **AFLW** [26] datasets, as well as on the challenging **LS3D** [7]. The **CelebA** dataset contains $\sim$ 200K facial images manually annotated with 5 facial landmarks. We follow prior work and remove from the training the 1000 images corresponding to the MAFL partition [76] which is used for evaluation. The **AFLW** contains $10,112$ training images and $2,991$ test images annotated with 21 landmarks. Both CelebA and AFLW are annotated with a limited number of points which in practice limit the evaluation of unsupervised methods to capture proper geometric deformations. For this reason, we opt for re-annotating both datasets with 68 landmarks using the 2D detector of [7]. We evaluate both our and competing methods using the same set of detected points. The **LS3D** [7] dataset contains images of faces with large pose variations. It is constructed by re-annotating the images from 300W-LP [77], AFLW [26], 300VW [54], 300W [50] and FDDB [21] in a consistent manner with 68 points using the automatic method of [7]. Note that LS3D dataset is annotated with 3D points. Evaluation is performed on the LS3D-W Balanced test set, comprising 7200 images, including an equal number of images for each of the range of yaw angles $[0^o - 30^o]$, $[30^o - 60^o]$, $[60^o - 90^o]$.

Human Body datasets. We evaluate our method on **BBCPose** [12], **Human3.6M** [20] and **PennAction** [74]. **BBCPose** [12] is a dataset of 20 sign language videos (10 for training, 5 for validation and 5 for testing) annotated with 7 human pose landmarks (head, wrists, elbows, and shoulders). We form the training set by selecting 1 of every 10 frames leading to a set of 60885 images. Evaluation is performed on the standard test set (1000 images). **Human3.6M** [20] is an activity dataset with a constant background containing videos of actors in multiple poses under different viewpoints. We follow the evaluation protocol of [75] and use all 7 subjects of the training set (6 subjects were used for training and 1 for testing) on six activities (direction, discussion, posing, waiting, greeting, walking). We form our training set by extracting 1 every 50 (48240 training images) and 1 every 100 frames for testing (2760 images). Contrary to [75] we do not perform background subtraction to simplify landmark detection. **PennAction** [74] is a dataset of 2326 videos of humans participating in sports activities. For this experiments, we use the same 6 categories as in [33] (tennis serve, tennis forehand, baseball pitch, baseball swing, jumping jacks, golf swing). For this experiment, we do not use the provided $50\% - 50\%$ train-test split to ensure sufficient training data. We opt for using the 5 first videos for each category to form a separate test set. This results in 51661 training and 1776 testing images.

Other datasets: In addition to the above categories, we also evaluate our method on the **Cat Heads** [73] dataset, which consists of 9k images of cat heads annotated with 9 landmarks. We use the test-train split of [75] with 7747 training and 1257 testing images. Finally, we present a qualitative evaluation in the **CUB-200-2011** [65] dataset, which contains 11778 images of birds belonging to 200 species. We use the same setting as [33] and remove the seabird species.

4.2 Implementation Details

Network architecture: We use the Hourglass architecture of [38] with the residual block of [6] for both $\mathbf{\Psi}$ and $\mathbf{\Phi}$. The image resolution is set to 256×256. For network $\mathbf{\Phi}$, the localisation head produces a single heatmap with resolution 64×64, and the descriptor head produces a volume of $64 \times 64 \times 256$, i.e. a volume with the same spatial resolution containing the 256-d descriptors. The network $\mathbf{\Psi}$ produces a set of K heatmaps, each 64×64.

Training: Keypoints are initially populated by SuperPoint [16]. Before the training starts, we apply an automatic outlier removal step to filter out keypoints most likely to be of no use. We use the Faiss library [24] for this preliminary step, as well as for the K-means clustering. We perform warm-up for $30,000$ iterations as described in 3.5. Then, we apply clustering and update the pseudo-ground truth every $5,000$ iterations. The number of clusters M is set to 100 for all datasets. The algorithm takes around 200,000 iterations to converge in all datasets. For Stage2, we initialise the model Ψ from the weights of the model Φ resulting after Stage1, except for the weights of the last layer that are trained from scratch. To train the models, we used RMSprop [19], with learning rate equal to $2 \cdot 10^{-4}$, weight decay 10^{-5} and batchsize 16. All models were implemented in PyTorch [42]. Similarly to other recent methods, [22], [75] we also boost the training on video datasets by adding temporal supervision. To that end, image pairs for contrastive training are sampled both randomly (clustering correspondence) as well as from nearby frames (keypoint correspondence between frames is recovered through sparse optical flow calculation as in [75]). Note that our approach achieves good performance without temporal supervision, and optical flow is used only when explicitly stated.

Evaluation: Quantitative evaluation of unsupervised landmark detectors is often assessed by quantifying the degree of correlation between manually annotated landmarks and those detected by the proposed approach. This is accomplished by learning a simple regressor with no bias that maps the discovered landmarks to those manually annotated, using a variable number of images in the training set. Numerical evaluation is often measured by means of the Normalised Mean-squared Error (NME). In addition, we follow [53], and complement this measure (herein referred to as **Forward-NME**) by measuring the performance of a reverted regressor, i.e. one that maps the manual annotations into the discovered landmarks. As found by [53], this measure, known as **Backward-NME**, helps identify unstable landmarks. We also present Cumulative Error Distribution (CED) curves for these metrics, which permit a per-landmark comparison w.r.t. state-of-the-art methods. We use interocular distance to normalise errors in facial datasets (CelebA, AFLW and CatHeads), and shoulder distance for human pose datasets (BBCPose and Human3.6). Due to the

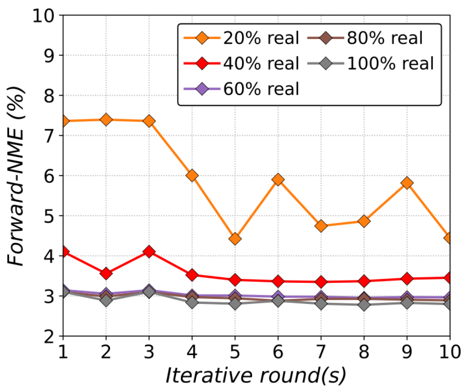
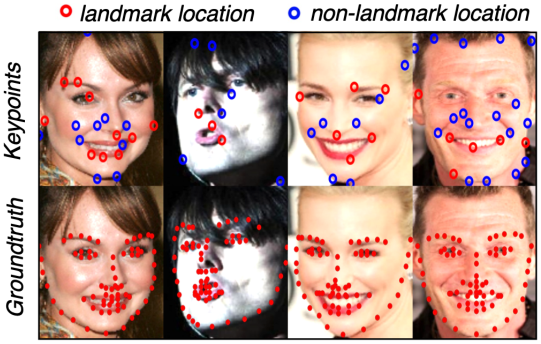


Fig. 4: Forward-NME (shown for the first 10 iterative rounds) of training the first stage of our method with varying ratios of real and random points. Experiment is performed on CelebA [31]. Real points are sampled from 15 facial landmarks and further perturbed spatially by a small offset sampled from $[-3px, +3px]$.

Keypoint Detector	Forward	Backward
SIFT [35] + *ANMS* [4]	4.07	7.79
ORB [49] + *ANMS* [4]	3.85	7.70
R2D2 [46]	3.71	7.97
SuperPoint [16]	**3.25**	**6.65**

TABLE 1: Evaluation of landmarks learned on the first stage of our framework on CelebA under different keypoint initialisation methods. Models are trained to for $K = 30$.



Precision (%) w.r.t 68 facial landamrks ($d = 10px$)

Keypoint Detector	Precision
SIFT [35] + *ANMS* [4]	35.2
ORB [49] + *ANMS* [4]	43.7
R2D2 [46]	50.7
SuperPoint [16]	**51.8**

Fig. 5: **(figure-top)** Examples of generic keypoints captured by SuperPoint on facial images along with the corresponding 68 ground-truth landmarks. Generic keypoints capture several object landmark locations (*red keypoints*) as well as non-corresponding background points (*blue keypoints*). **(table-bottom)** Precision of various generic keypoint detectors w.r.t 68-ground-truth landmark locations (on CelebA). As true positives we consider keypoints within $10px$ of a landmark location (image resolution 256×256).

large pose variation on LS3D and PennAction datasets, we opt for normalising the errors using the squared root of the bounding box area, where the bounding box is defined as the smallest rectangle that fits the ground-truth points.

We are also interested in assessing the quality of the discovered landmarks after Stage 1. Because not all the landmarks will be activated in each image after Stage 1, we need to complete the missing values before being able to compute the aforementioned metrics. To do so, we gather all discovered landmarks in a matrix $X \in \mathbb{R}^{K \times N}$, with N the number of training images and K the number of discovered landmarks, and use the Singular Value Thresholding method for Matrix Completion [9], leaving the detected points unchanged. At test time, we fill the missing landmarks with their corresponding mean positions, computed from the training set.

5 Ablation Studies

We perform a series of ablation studies to evaluate different aspects of our proposed method. In particular, we are interested in measuring how the initial conditions affect the training of our proposed approach, as well as the impact of the training components introduced in our method.

5.1 On the initial conditions

Robustness to noise We are firstly interested in measuring to which extend our method can recover semantic correspondence from noisy initialisations. A good initialisation is expected to have some consistent keypoints that overlap to some extend with proper landmarks; a huge number of random keypoints will hinder the learning of landmark correspondence. To evaluate such impact, we first conduct an experiment with synthetic initialisations, i.e. by initialising our training set with a mixture of ground-truth landmark locations and noisy points randomly sampled from the image domain. In particular, we populate each image with a set of 15 points that are either sampled from the ground-truth locations of 15 facial landmarks (eyes, eyebrows, nose, mouth, chin) or chosen at random, uniformely distributed over the image space. Our model is trained to detect 15 object landmarks, and we conduct experiments with varying mixture ratios to evaluate the effect of different noise levels. Fig. 4 shows the result of this experiment in terms of forward error. Interestingly we find that even with as much as only 20% of real object landmarks in the keypoint initialisation, our method can still perform reasonably well. Increases in the percentage of real points over 40% only result in slight performance gains.

Keypoint initialisation: We now evaluate the dependency of our method on initialisations as provided by real keypoint detectors. To this end, we compare the performance of our method, both by means of forward and backward errors, for the case where the initial keypoints are provided by SuperPoint [16], R2D2 [46], SIFT [35] and ORB [49]. Note that all these methods either are trained in an unsupervised manner (SuperPoint, R2D2), or do not even require training (SIFT, ORB), i.e. neither the initialisation nor our method require any manual supervision. Given that

# of clusters	Forward-NME	Backward-NME
$M = 30$	10.26	9.41
$M = 50$	7.99	6.99
$M = 100$	**7.95**	6.55
$M = 250$	8.58	6.26
$M = 500$	9.53	**6.19**

TABLE 2: Evaluation of landmarks learned from the first stage of our approach on LS3D [7] under various number of training clusters. M. All models are trained for $K = 30$. We see that $M >> K$ results in better performance as it allows appearance and viewpoint variantions of the same landamark to be captured by several clusters

#	Negative-Pairs	Correspondence	*NME*
1	*different clusters*	*Clustering*	11.15
2	*same image only*	*Equivariance*	10.02
3	*different clusters*	*Equivariance*	9.56
4	*same image only*	*Clustering*	**7.95**

TABLE 3: Ablation study of the proposed negative pair selection strategy (compared to the strategy of [36]), combined with either clustering or equivariance training. Experiment performed in the challenging LS3D [7] dataset. We report forward-NME error values.

Dataset	p.p.e		NME(%)	
	Stage1	Stage2	Stage1	Stage2
CelebA ($K = 30$)	*25.8*	*30*	3.3	**3.2**
AFLW ($K = 30$)	*23.4*	*30*	8.1	**7.4**
LS3D ($K = 30$)	*23.5*	*30*	7.9	**5.2**

TABLE 4: Comparison of the first and second stages of our framework in terms of Forward-NME. We also report average number of points detected per image (p.p.e) on each stage. The full landmark detector on the second stage detects one landmark per K channels so p.p.e is 30.

SIFT and ORB tend to detect large numbers of spatially clustered points (that is suboptimal for our purpose of detecting object landmarks), we combine them with Adaptive Non-Maximal Suppression (ANMS [4]) to ensure a homogeneous spatial distribution. The results shown in Table 1 show that all detectors allow our method to deliver competitive results, with SuperPoint proving to be the best choice.

Landmarks captured as keypoints: To further evaluate how the different initialisations affect the performance of our method, we measure to which extend each detector provides keypoints that are consistently close to a manually annotated landmark. To do so, we compute the precision of each of the detectors, measured as the percentage of keypoints that lie within a radius of 10 pixels around a ground-truth landmark. Fig.5 shows some visual examples of keypoints that overlap with manually annotated landmarks (red), as well as the computed precision. These results align with those in Table 1, showing that SuperPoint is a better choice to populate the training set.

Dataset	Flip*(Train)*	Flip*(Test)*[1]	**Stage1**	**Stage2**
CelebA	×	×	3.88	3.42
	✓	×	3.32	3.40
	✓	✓	3.32	3.25
LS3D	×	×	8.69	5.81
	✓	×	7.95	5.45
	✓	✓	7.95	5.26

TABLE 5: Experiments on the effect of flipping as a training augmentation and at test time. Results are given for both stages of our approach in terms of Forward-NME.

5.2 On the training design

Impact of number of clusters: We investigate the effect on the number of clusters in the training of our proposed approach. The results shown in Table 2 indicate that the best performance is attained for a larger number of training clusters. This over-segmentation of feature space is required for optimal clustering assignment as it allows for multiple clusters that capture different appearance variations of the same landmark, enabling the discovery of more stable landmarks (as demonstrated by smaller values of the backward error in Table 2). On the other extreme, for very big M values, the same underlying landmark is tracked by several clusters, each containing only very similar features. This hinders our method's ability to learn representations robust to viewpoint or appearance variations, and more diverse landmarks get filtered out (leading to an increase in Forward-NME). Note that our method essentially equates to equivariance training in extreme cases where M is equal to the number of detected keypoints (each cluster contains only one feature).

Negative-Pair Selection: We evaluate the proposed negative pair selection strategy (referred to as *same image only*), compared to that of [36] (referred as *different cluster*) where negative pairs were selected as keypoints with different clustering assignments. We also evaluate the effect of learning from unpaired images (enabled by correspondence recovery) compared to training on synthesised views of the same underlying image (equivariance training). Note that the experiments that use equivariance still utilise deep clustering (constraint the detector in detecting at most K landmarks and filtering out noisy keypoints). Results can be seen in Table 3.

We observe that our improved negative pair selection strategy is the best performing method when correspondence is recovered through clustering (*line 4*). The *different cluster* strategy separates features to M clusters (*line 1*) and results in poor performance when is not combined with an additional merging step (as in [36]). Also, our negative pair selection strategy is only beneficial when correspondence is recovered through clustering (not with equivariance). This is expected since, with equivariance training, point correspondences are known, and inaccurate negative pairs (similar to the ones shown in Fig. 3) do not emerge. As a result negative pairs from *different cluster* are more informative and result in better performance (*line 3* vs. *line 2*).

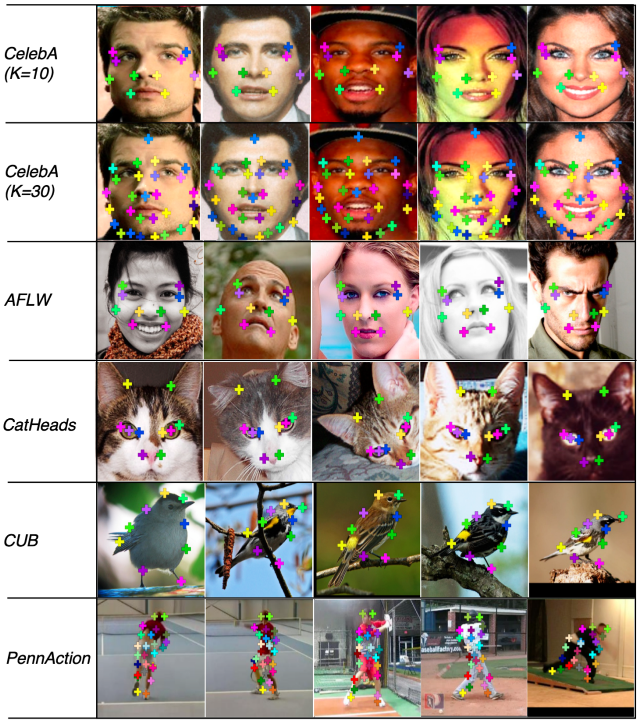


Fig. 6: Qualitative results of our proposed approach on various object categories

CatHeads Forward-NME (%)

Thewlis [62]	Zhang [75]	Lorenz [33]	Ours
26.94	14.84	9.30	9.31

TABLE 6: Performance on the CatHeads dataset [73]. All methods detect $K = 20$ unsupervised landmarks. Results for other methods are taken directly from the papers. Same as other methods, we regress 7 of the 9 annotated landmarks for this experiment (excluding landmarks on the ears).

Impact of Stage 2: For the first stage of our method, a set of points are detected per image for which correspondence is recovered through clustering. In the second stage, these points and correspondences are used to train a landmark detector with K output channels. The number of detected points per image on the first stage is $\leq K$ since there is no guarantee that each would appear in each image. On the contrary, our full landmark detector (output of the second stage) learns K unsupervised landmarks (one per output heatmap). In Table 4 we compare performance of the first vs second stage in terms of forward NME while also report the average number of points detected per image. We observe that the full landmark detector recovers the missing clusters in the second stage, resulting in lower error values. Performance increase is most notable on LS3D, where occlusion is extended due to large jaw angles.

Flipping: Finally, we conduct an ablation study on the proposed flipping augmentation strategy. Results for both CelebA and the more challenging LS3D database are given in Table 5. We observe that both flipping as a training augmentation and flipping at test time result in consistent performance improvement.

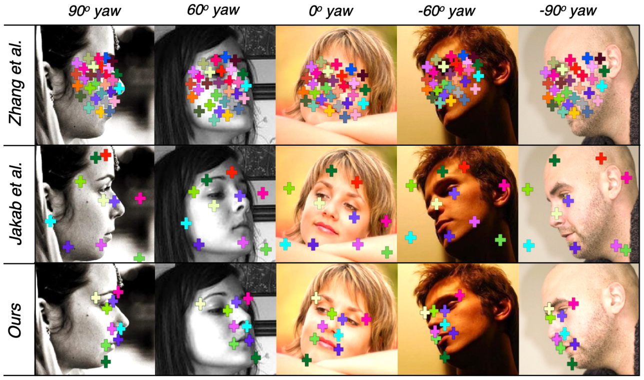


Fig. 7: Comparison between landmarks discovered by our approach and those of [22], [75] on LS3D facial images across the whole spectrum of facial pose.

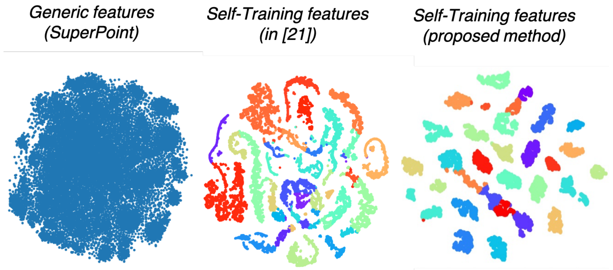


Fig. 8: T-SNE [64] visualisation of local features. Comparison with features produced by SuperPoint [16] and our previous work [36].

6 Overall evaluation

This Section presents the experiments carried out to validate the proposed approach against state of the art alternatives based on equivariance or image generation.

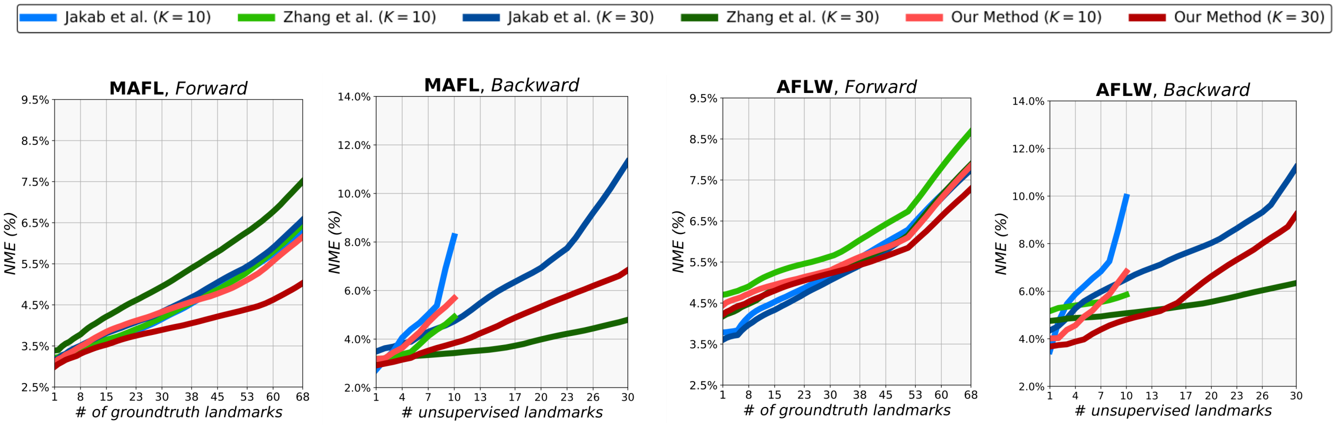
6.1 Qualitative results

We report qualitative results on various datasets in Fig. 6. We also present in Fig. 8 the t-SNE [64] representations of the features returned by SuperPoint (left), by [36](center), and by our method (right). Our method produces features that are clearly distinctive for each landmark, making the correspondence recovery effective.

6.2 Evaluation on facial datasets

Fig. 9 shows the results of our method on facial datasets. We report in the *Table* the commonly used forward error w.r.t. the 5 ground-truth facial landmarks. For the cumulative curves, the error is calculated w.r.t. 68-standard facial landmarks. As discussed in [53], for a method to work well, both forward and backward errors should be small. From our results on all datasets (*Figures*), we can see that overall our method provides the best results in terms of meeting both requirements. Notably, our method delivers state-of-the-art results for the challenging LS3D dataset, which contains large pose changes.

We also find that our approach surpasses other methods when evaluation is performed w.r.t all 68-facial landmarks (compared to standard 5 landmark evaluation on MAFL and



Method	MAFL	AFLW
Lorenz [33] *(K=10)*	3.24	-
Shu [56]	5.45	-
Jakab et al. [22] *(K=10)*	3.19	6.86
Zhang et al. [75] *(K=10)*	3.46	7.01
Sanchez [53] *(K=10)*	3.99	6.69
Sahasrabudhe [51]	6.01	-
Mallis [36]	4.12	7.37
Ours *(K=10)*	3.83	7.18

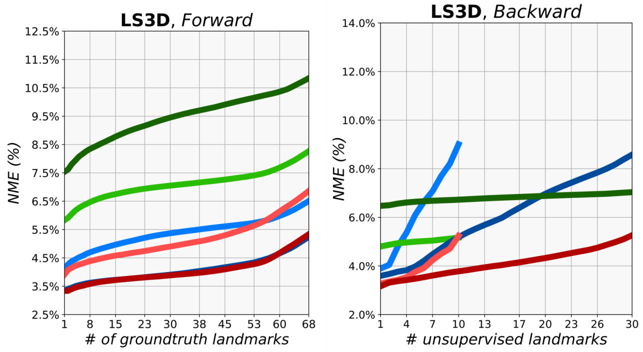


Fig. 9: Evaluation on facial datasets. ***(Table):*** Standard comparison on MAFL and AFLW, in terms of forward error. The results of other methods are taken directly from the papers (for the case where all MAFL training images are used to train the regressor and the error is measured w.r.t. to 5 annotated points). ***(Figures):*** CED curves for forward and backward errors. We compare our method with [22], [75] (for $K = 10, 30$). Where possible, we used pre-trained models, otherwise we re-trained these methods using the publicly available code. A set of 300 training images is used to train the regressors. Error is measured w.r.t. the 68-landmark configuration typically used in face alignment.

AFLW presented in Fig. 9 (*Table*) where we maintain competitive performance). One reason is that 5 facial landmarks include points in uniform areas and not repeatable edges or corners (centre of the eye, centre of the nose) that are not commonly tracked by generic keypoint detectors. On the contrary, our method is better suited to track the 68 commonly used facial landmarks. To further demonstrate that, we evaluate how accurately raw unsupervised landmarks track supervised landmark locations in Fig. 10. Each of the 68-facial landmarks is matched to the best corresponding unsupervised landmarks ($K = 30$ is used for all methods) through the Hungarian algorithm. We observe that most of our detected unsupervised landmarks track actual semantic object locations with high accuracy. In contrast, landmarks detected by [22], [75] are mostly uniformly spread over the objects' surface (to ensure stronger image generation/reconstruction) and do not tend to track manually annotated landmark locations.

Evaluation in terms of Forward-NME for the CatsHead dataset is shown in Table. 6. Our method reaches a similar error value as the best performing method of [33]. In addition, a set of qualitative examples is shown in Fig. 7 for the challenging LS3D data. We observe that landmarks produced by [22], [75] are not stable under 3D rotations and fail to capture large pose variations.

6.3 Evaluation on human pose datasets

Performance of our method on the BBCPose and Human3.6M datasets is shown in Fig. 11. Note that in this experiment, all methods are trained without temporal supervision. For both datasets, our approach demonstrates significantly better performance. As it can be seen from the forward error in Human3.6M, all three methods experience a sharp error increase when more than 22 landmarks are considered. We attribute this higher error to the fact that the hands are not captured by any method.

In Table 7 we measure the accuracy of regressed landmarks on the BBCPose database. For this experiment temporal supervision is available for all unsupervised methods. Even though this enables other approaches to achieve stronger performance, our model outperforms all other methods. Fig. 12 shows some examples of discovered landmarks that maximally correspond to ground-truth points.

We also note that due to the large degree of pose variation for human bodies, a simple linear layer does not suffice to learn a strong mapping between unsupervised and supervised landmarks. Hence, the forward errors are very high for all methods. To address this, we follow [22] and measure the accuracy of unsupervised landmarks that are found to maximally correspond to the provided ground-truth points (calculated through the Hungarian Algorithm) for Human3.6 and PennAction databases (Table 8). We ob-

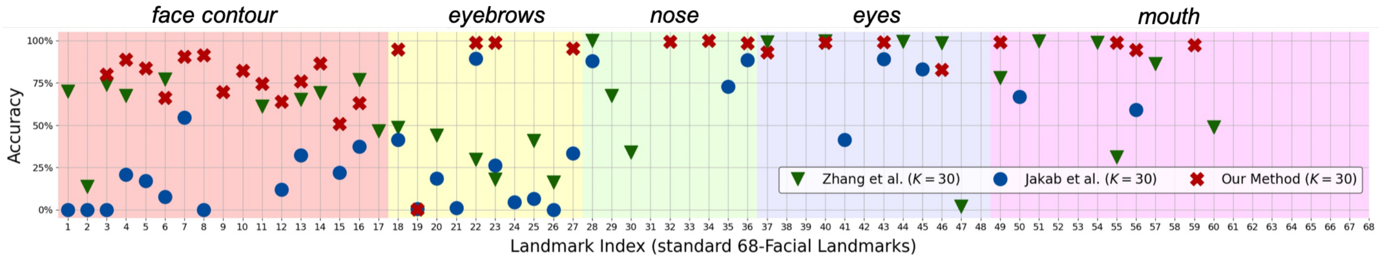


Fig. 10: Evaluation of the ability of raw unsupervised landmarks to capture supervised landmark locations on CelebA. Each unsupervised landmark is mapped to the best corresponding supervised landmark using the Hungarian Algorithm. Then accuracy is calculated for a distance threshold of $d = 10px$. Accuracy is shown for each of the 68-facial landmarks sorted by ascending order of index. Different landmark areas are highlighted with different colours (1-17 are facial contour landmarks, 18-27 are landmarks tracking the eyebrows, etc.)

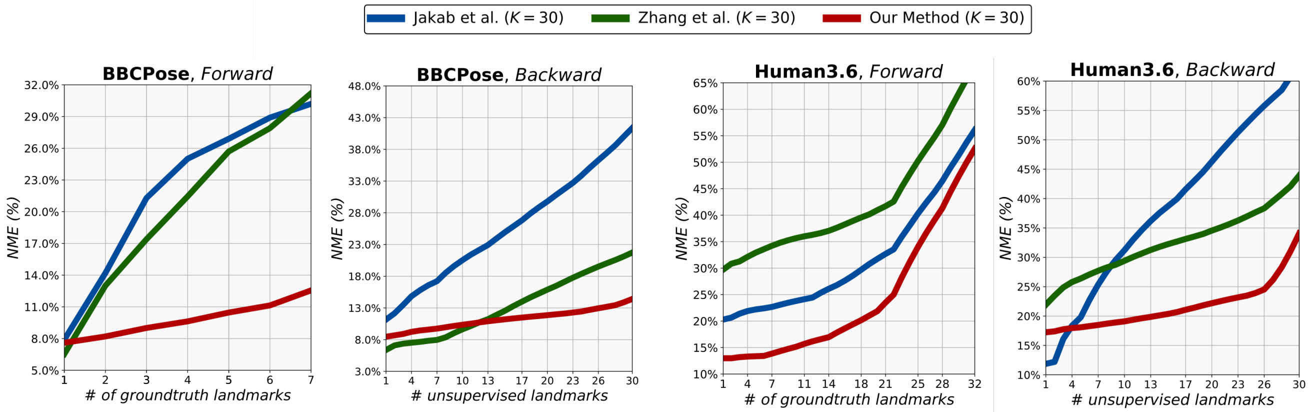


Fig. 11: Evaluation on BBCPose and Human3.6 datasets. CED curves for the forward and backward errors, computed for a regressor trained with 800 samples. We compare our method with [22], [75] (re-trained using the publicly available code). All methods are trained to discover 30 landmarks.

serve that our approach can discover unsupervised landmarks that robustly track several parts of the human body (except the hands for both Human3.6M and PennAction) and show much higher accuracy values compared to the other methods. Particularly for the challenging PennAction database that includes large pose variation and complicated backgrounds, we demonstrate strong performance, whereas [22] completely underperforms in this setting. Note that for both databases we do not utilise temporal supervision to train any examined method.

7 Conclusion

We presented a novel path for unsupervised discovery of object landmarks based on two ideas, namely self-training and recovering correspondence. The former helps our system improve by using its own predictions and constitutes a natural fit for training an object landmark detector starting from generic, noisy keypoints. The latter, although being a key property of object landmarks detectors, has not been previously used for unsupervised object landmark discovery. Compared to previous works, our approach can learn view-based landmarks that are more flexible in terms of changes in 3D viewpoint, providing superior results on a variety of challenging facial and human pose datasets.

Acknowledgments

Dimitrios Mallis' PhD studentship is funded by The Douglas Bomford Trust.

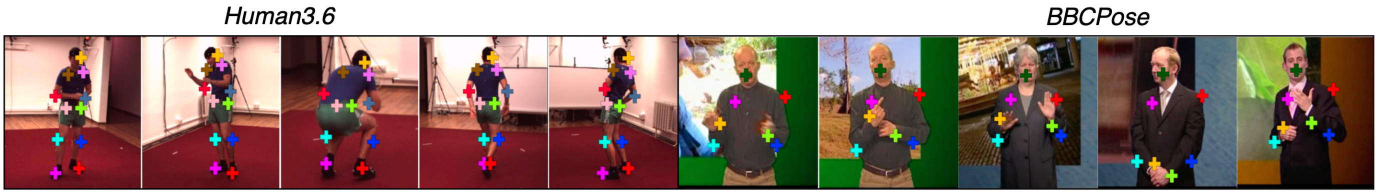


Fig. 12: Examples on Human3.6 and BBCPose databases. We show the unsupervised landmarks that maximally corresponding to the provided ground-truth (selected through the Hungarian Algorithm).

BBCPose Regressed Landmark Accuracy (%)

Method	Head	Shldrs	Elbws	Hands	Avg
		Supervised			
Yang [69]	63.40	53.70	49.20	46.10	51.63
Pfister [44]	74.90	53.05	46.00	71.40	59.40
Chen [13]	65.90	47.90	66.50	76.80	64.10
Charles [12]	95.40	72.95	68.70	90.30	79.90
Pfister [43]	98.00	88.45	77.10	93.50	88.01
		Unsupervised			
Jakab [22](selfsup)	81.01	49.05	53.05	70.10	60.79
Jakab [22]	76.10	56.50	70.70	74.30	68.44
Lorenz [33]	-	-	-	-	74.50
Ours	97.89	49.65	71.26	84.90	75.93

TABLE 7: Accuracy of regressed landmarks on BBCPose measured as %-age of points within $d = 6px$ from the ground-truth for a resolution of $128px$. Results for other methods taken directly from the papers. All unsupervised methods in this experiment utilise temporal information.

Human3.6 Raw Landmark Accuracy (%)

Method	Head	Shldrs	Elbws	Waist	Knees	Legs	Avg
Zhang [75]	20.9	53.1	51.0	43.7	85.6	2.0	42.7
Jakab [22]	0.5	52.2	32.4	26.1	3.7	24.6	23.2
Ours	81.1	89.8	39.7	94.2	93.6	64.4	77.1

PennAction Raw Landmark Accuracy (%)

Method	Head	Shldrs	Elbws	Hands	Waist	Knees	Legs
Jakab [22]	6.36	9.23	7.85	0.59	22.27	17.85	6.48
Ours	74.27	57.91	33.00	8.36	64.81	69.54	75.84

TABLE 8: Accuracy of raw discovered landmarks that correspond maximally (calculated through the Hungarian algorithm) to each ground-truth point measured as %-age of points within $d = 6px$ from the ground-truth [22] (image resolution of $128px$). For this experiment, examined methods do not utilise temporal information.

REFERENCES

[1] P. F. Alcantarilla, A. Bartoli, and A. J. Davison. Kaze features. In *ECCV*, 2012.

[2] D. Arpit, S. Jastrzebski, N. Ballas, D. Krueger, E. Bengio, M. S. Kanwal, T. Maharaj, A. Fischer, A. C. Courville, Y. Bengio, and S. Lacoste-Julien. A closer look at memorization in deep networks. In *ICML*, 2017.

[3] Y. M. Asano, C. Rupprecht, and A. Vedaldi. Self-labelling via simultaneous clustering and representation learning. In *ICLR*, 2020.

[4] O. Bailo, F. Rameau, K. Joo, J. Park, O. Bogdan, and I. S. Kweon. Efficient adaptive non-maximal suppression algorithms for homogeneous spatial keypoint distribution. *Pattern Recognition Letters*, 2018.

[5] H. Bay, T. Tuytelaars, and L. V. Gool. Surf: Speeded up robust features. In *ECCV*, 2006.

[6] A. Bulat and G. Tzimiropoulos. Binarized convolutional landmark localizers for human pose estimation and face alignment with limited resources. *ICCV*, 2017.

[7] A. Bulat and G. Tzimiropoulos. How far are we from solving the 2d and 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). *ICCV*, 2017.

[8] A. Bulat and G. Tzimiropoulos. Super-fan: Integrated facial landmark localization and super-resolution of real-world low resolution faces in arbitrary poses with gans. *CVPR*, 2018.

[9] J.-F. Cai, E. J. Candès, and Z. Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal on optimization*, 2010.

[10] M. Caron, P. Bojanowski, A. Joulin, and M. Douze. Deep clustering for unsupervised learning of visual features. In *ECCV*, 2018.

[11] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *NeurIPS*, 2020.

[12] J. Charles, T. Pfister, D. R. Magee, D. C. Hogg, and A. Zisserman. Domain adaptation for upper body pose tracking in signed tv broadcasts. In *BMVC*, 2013.

[13] X. Chen and A. Yuille. Articulated pose estimation by a graphical model with image dependent pairwise relations. In *NeurIPS*, 2014.

[14] Z. Cheng, J.-C. Su, and S. Maji. Unsupervised discovery of object landmarks via contrastive learning. *arXiv preprint arXiv:2006.14787*, 2020.

[15] J. Dai, K. He, and J. Sun. Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. *ICCV*, 2015.

[16] D. DeTone, T. Malisiewicz, and A. Rabinovich. Superpoint: Self-supervised interest point detection and description. *CVPR*, 2018.

[17] X. Dong, Y. Yan, W. Ouyang, and Y. Yang. Style aggregated network for facial landmark detection. In *CVPR*, June 2018.

[18] A. Faktor and M. Irani. Video segmentation by non-local consensus voting. In *BMVC*, 2014.

[19] A. Graves. Generating sequences with recurrent neural networks. *ArXiv*, 2013.

[20] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *TPAMI*, 2014.

[21] V. Jain and E. Learned-Miller. Fddb: A benchmark for face detection in unconstrained settings. Technical Report UM-CS-2010-009, University of Massachusetts, Amherst, 2010.

[22] T. Jakab, A. Gupta, H. Bilen, and A. Vedaldi. Unsupervised learning of object landmarks through conditional image generation. In *NeurIPS*, 2018.

[23] S. G. Jiabo Huang, Qi Dong and X. Zhu. Unsupervised deep learning by neighbourhood discovery. In *ICML*, 2019.

[24] J. Johnson, M. Douze, and H. Jégou. Billion-scale similarity search with gpus. *arXiv preprint arXiv:1702.08734*, 2017.

[25] A. Khoreva, R. Benenson, J. H. Hosang, M. Hein, and B. Schiele. Simple does it: Weakly supervised instance and semantic segmentation. *CVPR*, 2017.

[26] M. Köstinger, P. Wohlhart, P. M. Roth, and H. Bischof. Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. *ICCV Workshops*, 2011.

[27] H. W. Kuhn. The hungarian method for the assignment problem. 1955.

[28] T. D. Kulkarni, A. Gupta, C. Ionescu, S. Borgeaud, M. Reynolds, A. Zisserman, and V. Mnih. Unsupervised learning of object keypoints for perception and control. *NeurIPS*, 2020.

[29] K. Lenc and A. Vedaldi. Learning covariant feature detectors. In *ECCV Workshops*, 2016.

[30] D. Li, W.-C. Hung, J.-B. Huang, S. Wang, N. Ahuja, and M.-H. Yang. Unsupervised visual representation learning by graph-based consistent constraints. In *ECCV*, 2016.

[31] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. *ICCV*, 2015.

[32] D. Lorenz, L. Bereska, T. Milbich, and B. Ommer. Unsupervised part-based disentangling of object shape and appearance. In *CVPR*, 2019.

[33] D. Lorenz, L. Bereska, T. Milbich, and B. Ommer. Unsupervised part-based disentangling of object shape and appearance. *CVPR*, 2019.

[34] D. G. Lowe. Object recognition from local scale-invariant features. *ICCV*, 1999.

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

[36] D. Mallis, E. Sanchez, M. Bell, and G. Tzimiropoulos. Unsupervised learning of object landmarks via self-training correspondence. In *NeurIPS*, 2020.

[37] Z. Miao and X. Jiang. Interest point detection using rank order log filter. *Pattern Recognition*, 2013.

[38] A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. In *ECCV*, 2016.

[39] D. T. Nguyen, C. K. Mummadi, T. P. N. Ngo, T. H. P. Nguyen, L. Beggel, and T. Brox. Self: Learning to filter noisy labels with self-ensembling. In *ICLR*, 2020.

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

[41] M. Noroozi, A. Vinjimoor, P. Favaro, and H. Pirsiavash. Boosting self-supervised learning via knowledge transfer. *CVPR*, 2018.

[42] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer. Automatic differentiation in pytorch. 2017.

[43] T. Pfister, J. Charles, and A. Zisserman. Flowing convnets for human pose estimation in videos. *ICCV*, 2015.

[44] T. Pfister, K. Simonyan, J. Charles, and A. Zisserman. Deep convolutional neural networks for efficient pose estimation in gesture videos. In *ACCV*, 2014.

[45] M. Pietikäinen and G. Zhao. Two decades of local binary patterns: A survey. *ArXiv*, 2016.

[46] J. Revaud, C. R. de Souza, M. Humenberger, and P. Weinzaepfel. R2d2: Reliable and repeatable detector and descriptor. In *NeurIPS*, 2019.

[47] M. N. Rizve, K. Duarte, Y. S. Rawat, and M. Shah. In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning. *ICLR*, 2021.

[48] D. Rolnick, A. Veit, S. J. Belongie, and N. Shavit. Deep learning is robust to massive label noise. *ArXiv*, 2018.

[49] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski. Orb: An efficient alternative to sift or surf. *ICCV*, 2011.

[50] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. 300 faces in-the-wild challenge: The first facial landmark localization challenge. *ICCV Workshops*, 2013.

[51] M. Sahasrabudhe, Z. Shu, E. Bartrum, R. A. Güler, D. Samaras, and I. Kokkinos. Lifting autoencoders: Unsupervised learning of a fully-disentangled 3d morphable model using deep non-rigid structure from motion. *ICCVW*, 2019.

[52] S. Salti, A. Lanza, and L. di Stefano. Keypoints from symmetries by wave propagation. *CVPR*, 2013.

[53] E. Sanchez and G. Tzimiropoulos. Object landmark discovery through unsupervised adaptation. In *NeurIPS*. 2019.

[54] J. Shen, S. Zafeiriou, G. G. Chrysos, J. Kossaifi, G. Tzimiropoulos, and M. Pantic. The first facial landmark tracking in-the-wild challenge: Benchmark and results. *ICCVW*, 2015.

[55] K. J. Shih, A. Dundar, A. Garg, R. Pottorf, A. Tao, and B. Catanzaro. Video interpolation and prediction with unsupervised landmarks. *ArXiv*, 2019.

[56] Z. Shu, M. Sahasrabudhe, R. A. Güler, D. Samaras, N. Paragios, and I. Kokkinos. Deforming autoencoders: Unsupervised disentangling of shape and appearance. In *ECCV*, 2018.

[57] A. Siarohin, S. Lathuilière, S. Tulyakov, E. Ricci, and N. Sebe. Animating arbitrary objects via deep motion transfer. *CVPR*, 2019.

[58] K. Sohn, D. Berthelot, C.-L. Li, Z. Zhang, N. Carlini, E. D. Cubuk, A. Kurakin, H. Zhang, and C. Raffel. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *NeurIPS*, 2020.

[59] O. Stretcu and M. Leordeanu. Multiple frames matching for object discovery in video. In *BMVC*, 2015.

[60] J. Thewlis, S. Albanie, H. Bilen, and A. Vedaldi. Unsupervised learning of landmarks by descriptor vector exchange. *ICCV*, 2019.

[61] J. Thewlis, H. Bilen, and A. Vedaldi. Unsupervised learning of object frames by dense equivariant image labelling. In *NeurIPS*, 2017.

[62] J. Thewlis, H. Bilen, and A. Vedaldi. Unsupervised learning of object landmarks by factorized spatial embeddings. *ICCV*, 2017.

[63] J. Thewlis, H. Bilen, and A. Vedaldi. Modelling and unsupervised learning of symmetric deformable object categories. In *NeurIPS*, 2018.

[64] L. van der Maaten and G. E. Hinton. Visualizing data using t-sne. In *JMLR*, 2008.

[65] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona. Caltech-UCSD Birds 200. Technical Report CNS-TR-2010-001, California Institute of Technology, 2010.

[66] B. Xiao, H. Wu, and Y. Wei. Simple baselines for human pose estimation and tracking. In *ECCV*, 2018.

[67] Q. Xie, E. H. Hovy, M.-T. Luong, and Q. V. Le. Self-training with noisy student improves imagenet classification. *CVPR*, 2020.

[68] X. Yan, I. Misra, A. Gupta, D. Ghadiyaram, and D. K. Mahajan. Clusterfit: Improving generalization of visual representations. *CVPR*, 2020.

[69] Y. Yang and D. Ramanan. Articulated pose estimation with flexible mixtures-of-parts. *CVPR*, 2011.

[70] K. M. Yi, E. Trulls, V. Lepetit, and P. Fua. Lift: Learned invariant feature transform. In *ECCV*, 2016.

[71] D. Zhang, J. Han, and Y. Zhang. Supervision by fusion: Towards unsupervised learning of deep salient object detector. *ICCV*, 2017.

[72] J. Zhang, T. Zhang, Y. Dai, M. Harandi, and R. Hartley. Deep unsupervised saliency detection: A multiple noisy labeling perspective. In *CVPR*, 2018.

[73] W. Zhang, J. Sun, and X. Tang. Cat head detection - how to effectively exploit shape and texture features. In *ECCV*, 2008.

[74] W. Zhang, M. Zhu, and K. Derpanis. From actemes to action: A strongly-supervised representation for detailed action understanding. *ICCV*, 2013.

[75] Y. Zhang, Y. Guo, Y. Jin, Y. Luo, Z. He, and H. Lee. Unsupervised discovery of object landmarks as structural representations. *CVPR*, 2018.

[76] Z. Zhang, P. Luo, C. C. Loy, and X. Tang. Facial landmark detection by deep multi-task learning. In *ECCV*, 2014.

[77] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li. Face alignment across large poses: A 3d solution. In *CVPR*, 2016.

[78] C. Zhuang, A. Zhai, and D. Yamins. Local aggregation for unsupervised learning of visual embeddings. *ICCV*, 2019.